Abnormal Traffic Flow Detection Based on Dynamic Hybrid Strategy

Yang Liu[✉], Hongping Xu, Hang Yi, Xiaotao Yan, Jian Kang, Weiqiang Xia, Qingping Shi, and Chaopeng Shen

Beijing Institute of Astronautical System Engineering, Beijing, China
yangliu_npu@163.com

Abstract. Efficient and accurate analysis of the traffic data contained in the network is the key measure to detect the abnormal behavior, resist the invasion and protect the information security. In this paper, we make a comprehensive utilization of the characteristics of port mapping identification, payload identification, statistical analysis and SVM machine learning, and propose the dynamic hybrid strategy. Firstly, the machine learning training samples are obtained through port mapping and load feature recognition. Then, on the basis of information gain feature selection, the SVM machine learning model is built and trained. Finally, through the voting mechanism, we achieve comprehensive analysis of the traffic data. The experimental results show that the accuracy of the proposed algorithm is as high as 99.1%, and the number of manual decision analysis is greatly reduced at the same time.

Keywords: Port mapping · Payload feature matching
Dynamic hybrid strategy · Machine learning

1 Introduction

The traffic of the Measurement and Control Network carrying the key information of the system, the majority of abnormal or aggressive behavior will make the system network traffic presents specific differences, through in-depth analysis of the system flow, we can quickly identify non-compliance flow, timely find the information redundancy or abnormal behavior, and ensure the reliability of the data communication network. In 2004, Lang uses port based protocol identification method to obtain pure network traffic, and verifies the effectiveness of the method [1]. In 2006, Liang prove that the port based protocol identification method is not suitable for dynamic port applications, but it still has high accuracy in traditional network applications [2]. In 2012, Lin et al. proposed a method based on packet length distribution and port to identify network traffic. In 2013, Moore reduced the time and space complexity by reducing the length and number of load identification [4]. In 2013, Zhang et al. proposed support vector machine (Support Vector Machine, SVM) and statistical feature classification method [5]. Xiao in 2015 proposed a hierarchical support vector machine method

G. Sun and S. Liu (Eds.): ADHIP 2017, LNICST 219, pp. 481–488, 2018.
https://doi.org/10.1007/978-3-319-73317-3_55

to solve the classification problem of network flow, and achieve the recognition accuracy of 94%, [6] has achieved good application effect in large networks.

The traffic identification based on port mapping is efficient, however, the error rate is high. Recognition method based on feature matching of load has a high accuracy, but can't identify encrypted traffic. Statistical identification can identify the encryption protocol based on statistical features, but the statistical feature is difficult to select, prone to false positives, and is relatively poor for real-time traffic analysis. The recognition method based on machine learning traffic protocol is intelligent and has high recognition accuracy, but it depends on the correct training data and the appropriate network flow characteristics.

In this paper, we proposed a new method, comprehensively using the port identification, load and precise feature matching, and the accurate statistical identification method to construct the hybrid identification strategy. After obtaining the sample data of more known labels, the support vector mechanism is used to build the self-learning mechanism, constantly update and replace the statistic optimization method. And finally form a self-iteration update network traffic comprehensive recognition mechanism.

2 Traffic Flow Anomaly Detection Based on Dynamic Hybrid Strategy

2.1 Traffic Data Acquisition and Preprocessing

Before flow analysis, the general flow capture tools such as Sniffer, Wireshark, NetFlow, flow-tools and fprobe [7,8], can quickly collect the traffic data. In order to meet the needs of real-time processing, we carry out the flow separation pretreatment according to the five tuple (source IP, destination IP, source port, destination port, transport layer protocol number) before the flow analysis. Flow table is built to store the separated network data. Messages belonging to the same specific data stream have many similar attributes, By calculating the five tuple information of the network packet, we can get the *hash value* as Eq. 1. Packets will be divided into different flow according to the hash value.

$$hash\_index = HASH(I) \tag{1}$$

The information of each stream is saved in the flow table, which provides data support for the flow protocol identification.

2.2 Port Based Identification

In the complex Internet environment, due to the use of dynamic port technology, many applications no longer use the standard port, the accuracy of traditional traffic identification method based on port mapping is reduced greatly. However, most of the protocols or applications still use standard ports for communication in the network environment of the launch vehicle. Traffic data can achieve efficiently identification through the port mapping table (port_table) fast mapping.

$F[*]$ express the port application mapping function.

$$Protocol = F[port\_table] \tag{2}$$

The specific application in the Measurement and Control Network, through the planning and design of port construction in advance, constructing the port and application mapping table, and using the port mapping to analysis the network data.

2.3 Identification Based on Load Feature Matching

According to the characteristics of the traffic data load, we can judge whether the load has special characteristics or not, and realize the analysis and identification of traffic data. Based on the precise feature matching method, each network packet is split, and the application data is extracted for feature matching. According to the actual characteristics of the launch vehicle network information, firstly, we extract the features and identify the protocol according to the special domain. As shown in formula (2), $G[*]$ express the traffic feature extraction function.

$$Protocol = G[flows] \xrightarrow{matching} protocol\_feature \tag{3}$$

For example, Protocol_1/0:0xEB_1:0x90_2:0x00_3:0x20, Protocol_2/0: 0x70_2: 0x10_3:0x80 respectively express the feature information of Protocol_1 and Protocol_2, by extracting the protocol fingerprint of traffic data, and using the AC/SRS multi pattern matching algorithm, we can achieve efficient and rapid identification analysis. Process flow is shown in Fig. 1.

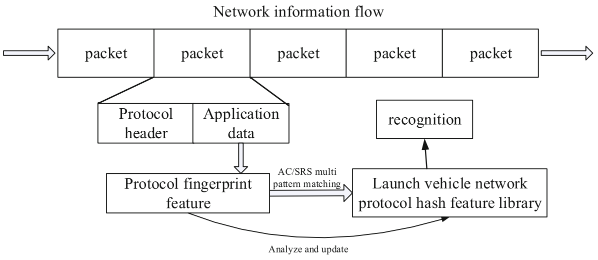

Fig. 1. Traffic identification based on accurate feature matching.

The precision of method based on the feature matching is relatively high, however, the speed is slow because of the need for each packet analysis, besides, the encryption protocol or some special protocol cannot be resolved, it is difficult to achieve analysis.

2.4 Detection and Recognition Based on Statistical Features and Machine Learning Models

In the launch vehicle network, the flow of different applications usually presents some unique statistical characteristics at the network level, such as idle time, the average length of the stream flow density, average packet length, packet interval and so on. For some specific application, application layer features represented by the ratio of source to destination communication data is unique, discrimination can be achieved by joint analysis of multiple dimensions. A large number of studies have shown that the identification based on the statistical characteristics of network traffic data is effective [9–11]. Especially in some specific application scenarios, it can realize the fast identification of encrypted protocols. In 2005, Moore gave a 249 dimensional feature set for summarizing and analyzing network traffic characteristics. Network traffic feature selection is important for the identification, in the actual analysis, some network features contain few information, correlation and redundancy, which has no contribution to the improvement of detection accuracy, but increases the time complexity and space complexity. The feature dimension reduction can be achieved by data feature selection, which can reduce the computational complexity and improve the efficiency of system detection.

The information gain of feature A is defined as the difference between the original information demand and the new demand [12], represented by: $gain(A)$

$$\begin{aligned} gain(A) &= \inf(D) - \inf_A(D) \\ &= -\sum_{1}^{m} p_i \log_2(p_i) - \sum_{j=1}^{n} \frac{|D_j|}{|D|} \times \inf(D_j), \end{aligned} \tag{4}$$

where $p_i = |C_{i,D}| \,/\, |D|$ indicates the nonzero probability that any sample in data set D belongs to C_i, $|D|$ indicates the total sample size, $|C_{i,D}|$ represents the number of samples belonging to class C_i, m is the number of sample classes, $\inf(D)$ represents the average amount of information required to identify the category of tuples D. $\inf_A(D)$ represents the desired information for the classification of tuples in set D based on feature A, and n indicates the number of D_j subsets.

The key problem of network traffic identification is to determine the mapping relationship between network flow and application categories. For the statistical characteristics of a large number of different dimensions, it is difficult to achieve the mapping through the intuitive rules such as threshold. The support vector machine (SVM) method based on the statistical learning theory has strong cognitive ability, especially for small sample learning problems, we can grasp the potential rules of irregular description by statistical learning, and realize the multi feature joint mapping. The basic characteristics and statistical characteristics of network flow are obtained in the unit time after feature selection.

The high dimensional sample feature vector data was constructed as $X = \{x_1, x_2, \ldots, x_l\}$, each network traffic sample can be marked as $D(X, y_i)$, y_i represents the class label for this type of traffic sample, and $y_i \in \{+1, -1\}$. Optimal classification surface used for distinguishing different categories can be expressed by $\overrightarrow{w} \cdot \overrightarrow{x} + b = 0$ which can make the biggest difference between different categories. Maximizing the interface is equivalent to solving the following optimization problem.

$$\min \quad \frac{1}{2}\sum_{i=1}^{n} w_i^2 \tag{5}$$

$$\text{Subject to} \quad y_i(\overrightarrow{w} \cdot \overrightarrow{x} + b) - 1 \geq 0, i = 1, \ldots, n, \tag{6}$$

where, n indicates the number of sample, w is not only related to the location of sample points, but also related to the category of samples. Under the constraint of formula (6), the formula (5) can be solved by convex quadratic optimization. For the two classification problem, the SVM discriminant function can be expressed as

$$f(x) = \operatorname{sgn}\{(w, x) + b\} = \operatorname{sgn}\{\sum_{i=1}^{l} \alpha_i \cdot y_i \cdot (x_i \cdot x) + b\}, \tag{7}$$

where α_i indicates the optimized Lagrange operator, (w, b) determining the equation of the classification surface $< w \cdot x > +b = 0$. For the multi classification problem, we design the SVM discrimination model between any two categories. As for k categories it needs C_k^2 categories. For the sample to be classified, the class with the most votes is the category of the sample.

2.5 Network Traffic Anomaly Detection and Analysis Based on Hybrid Strategy

In this paper, we design a hybrid optimization strategy, make a comprehensive utilization of all kinds of detection methods with their advantages to realize the accurate use of network traffic data, and effectively detect the abnormal traffic data. Network traffic anomaly detection and analysis algorithm based on hybrid strategy is shown as follows.

First of all, based on network traffic data distribution on the pretreatment, obtain the preliminary classification results by port mapping identification. At the same time, use the load feature matching to analysis and get the results. Compare the two results and analysis the inconsistent results manually to determine the protocol type. Then we can obtain the label training data for machine learning classification. Next, extract the feature of network traffic and construct the feature vector for machine learning classification. Support vector machine is used for training and learning based on the training data, and the knowledge classification model is obtained. Carry out the training process and test the accuracy constantly until the error rate is lower than the set threshold. Then change the recognition strategy, the identification results of port mapping, load feature matching and machine learning recognition are used to decide the final result by Voting Mechanism. The results are constantly used to train the SVM learning model, and update adaptively. The flow chart of the algorithm is shown in Fig. 2.

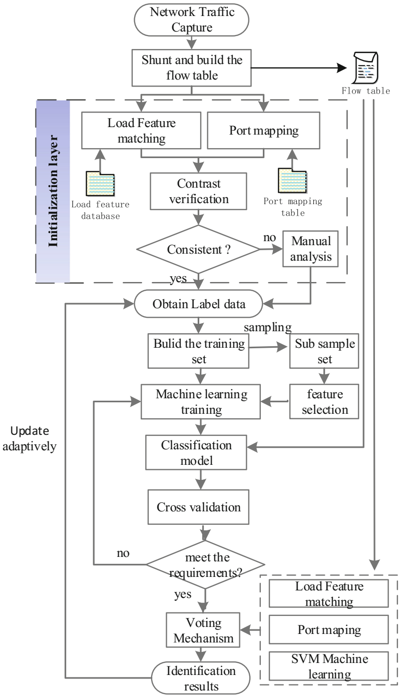

Fig. 2. Flow chart of anomaly detection algorithm.

3 Experiments and Results

In order to verify the applicability of the algorithm based on dynamic hybrid strategy, in this paper, we choose the actual data in the Measurement network of launch vehicle. The network topology of the Measurement network is shown as Fig. 3. Different system devices exchange the instruction and the data with each other through switches. For the test analysis, the whole data of the network can be obtained by the port of the core switch. Main configuration of the computer for experiments is as follows, Intel i5-3450 processor packaged by LGA1155, 4G 1333 DDR3 memory, and 1T SATA 7200r/s mechanical hard drive.

Select a subset of samples and do the feature selection by means of the information gain method described in formula (4), and the threshold parameter used to measure SVM machine learning performance is set to 90%. The experimental data is about 13.2G, and the actual processing time is 2 min 50 s, which can realize real-time data processing. The experimental data were processed with 9742 streams, of which the private network traffic of the network was about 49.18%. In order to verify the effectiveness of the proposed algorithm, the port mapping method and the deep packet detection (DPI) method are compared with the method in this paper. The experimental results are shown in Table 1.

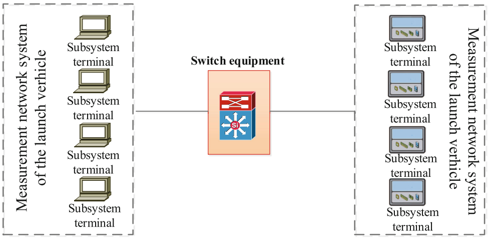

Fig. 3. Network topological of the launch vehicle network.

Table 1. Experimental results.

Methods	Accuracy rate	Proportion of unrecognized flow
Port mapping	84.9%	4.1%
Deep packet detection	92.2%	3.9%
Port mapping with DPI	81.9%	17.5%
Proposed method	99.1%	0.57%

Port mapping with DPI means that the port mapping method is compared with the output result of the DPI. If the results are consistent, then output the results, otherwise, identified as unknown traffic, and need for manual analysis. Inconsistent flows of Port mapping with DPI processed by manual analysis, the accuracy rate can be improved to 99.4 From Table 1, the accurate rates of proposed dynamic traffic identification based on hybrid strategy analysis was 99.1%, far higher than other methods, and the artificial processing required is less than the other three methods. The proposed method not only reduces the manual analysis, but also improves the recognition accuracy rate.

4 Conclusion

In this paper, we propose an algorithm for network anomaly detection based on dynamic hybrid strategy, comprehensively use the port mapping, load feature matching, statistical analysis and machine learning, design a dynamic hybrid strategy, and achieve the identification by using the voting mechanism. Not only ensure the accuracy rate of identification, greatly reduce the manual analysis, but also adaptive update. Greatly improve the intelligence and automation level of the anomaly detection and analysis. In the future, by long term analysis and iterative update of the actual data in the launch vehicle network, the traffic data can be gradually transparent and credible, effectively guarantee the safe and reliable operation of the network system.

References

1. Lang, T., Branch, P., Armitage, G.: A synthetic traffic model for Quake3. In: ACM Sigchi International Conference on Advances in Computer Entertainment Technology, Singapore, June 2004, pp. 233–238 (2004)
2. Liang, C., Jian, G., Xuan, X.: Identification of application- level protocols using characteristic. J. Comput. Eng. Appl. **42**, 16–19 (2006)
3. Lin, Y.-D., Lu, C.N., Lai, Y.-C., Peng, W.-H., Lin, P.-C.: Application classification using packet size distribution and port association. J. Netw. Comput. Appl. **32**, 1024–1030 (2009)
4. Moore, A.W., Papagiannaki, K.: Toward the accurate identification of network applications. In: Dovrolis, C. (ed.) PAM 2005. LNCS, vol. 3431, pp. 41–54. Springer, Heidelberg (2005). https://doi.org/10.1007/978-3-540-31966-5_4
5. Jun, Z., Yang, X., Yu, W., Wanlei, Z., Yong, X., Yong, G.: Network traffic classification using correlation information. J. IEEE Trans. Parallel Distrib. Syst. **24**, 104–117 (2013)
6. Xiao, L., Cheng, L.: State classification algorithm for bus based on hierarchical support vector machine. In: 2015 8th International Symposium on Computational Intelligence and Design (ISCID), Hangzhou, China, vol. 2, pp. 649–652 (2015)
7. Renuka, D.S., Yogesh, P.: A hybrid approach to counter application layer DDoS attacks. Int. J. Crypt. Inf. Secur. **2**, 649–652 (2012)
8. Gao, Y., Zhou, W., Han, J., Meng, D.: An online log anomaly detection method based on grammar compression. Chin. J. Comput. **37**, 73–86 (2014)
9. Wang, C., Zhang, H., Ye, Z., Du, Y.: A peer to peer traffic identification method based on support vector machine and artificial bee colony algorithm. In: 2015 IEEE 8th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), Warsaw, vol. 2, pp. 982–986 (2015)
10. Yu, W., Chao, C., Yang, X.: Unknown pattern extraction for statistical network protocol identification. In: IEEE Conference on Local Computer Networks, Clearwater Beach, FL, pp. 506–509 (2015)
11. Chen, T., Liao, X.: An optimized solution of application layer protocol identification based on regular expressions. In: 2016 18th Asia-Pacific Network Operations and Management Symposium (APNOMS), Kanazawa, pp. 1–4 (2016)
12. He, H., Tiwari, A., Mehnen, J., Watson, T., Maple, C., Jin, Y., Gabrys, B.: Incremental information gain analysis of input attribute impact on RBF-kernel SVM spam detection. In: 2016 IEEE Congress on Evolutionary Computation (CEC), Vancouver, BC, pp. 1022–1029 (2016)