Machine Learning Based Abnormal Flow Analysis of University Course Teaching Network

Shaobao Xu[1](✉) and Yongqi Jia[2]

[1] Philippine Christian University, 1004 Manila, Philippines
xxsb202@126.com
[2] Jose Rizal University, 1552 Mandaluyong, Philippines

Abstract. Due to the influence of abnormal traffic of university course teaching network, the accuracy of analysis results is low, this paper presents a method for analyzing the network traffic of university course teaching based on machine learning. According to the process of encapsulating user data into Ethernet data frames, a network traffic identification model based on machine learning is constructed. The screening strength is controlled by the screening coefficient, and the model deviation is reduced by measuring the task. After data collection and sorting, data cleaning and pretreatment, model evaluation and unknown traffic detection, the network traffic analysis steps are designed to analyze the abnormal situation of teaching network traffic. The experimental results show that the highest F1 score is 98%, and the accuracy of the analysis results is high, which provides sufficient network traffic for college course teaching.

Keywords: Machine learning · University courses · Teaching network · Abnormal traffic analysis

1 Introduction

As a vital part of the Internet, the campus network should control its security and performance effectively and timely. Traditional network management methods mainly use traffic statistical analysis tools to detect traffic and record related logs, but this lack of a more in-depth analysis of the traffic behavior of the entire campus network, but less of the inherent characteristics of the study, so when abnormal network traffic can not be handled in a timely manner, so we need to study the traffic behavior and characteristics of the campus network, to help the entire network better operation and maintenance. The main network security threats faced by campus network can be summarized as follows: network security problems existing in wired network, campus local area network also faces; information transmitted through campus local area network is generally not encrypted, or its encryption degree is poor, which will lead to information interception and tampering with by hackers; some lawbreakers can bypass the firewall set up by campus network, thus illegally stealing the internal network; in addition to the malicious attack of lawbreakers, campus local area network personnel can also set up the mode of wireless network card, so that they can communicate with external personnel, which

Y. Xu et al. (Eds.): ML4CS 2022, LNCS 13657, pp. 30–43, 2023.
https://doi.org/10.1007/978-3-031-20102-8_3

aggravates the security problem of campus network; in addition to the security problem of wireless local area network itself, there are few special security measures and products for wireless network in the market at present, and the technical level is difficult to meet the requirements. If the network traffic can be predicted in advance, the precautionary measures can be taken in advance to ensure the stability and security of the campus network. Therefore, the study of network traffic forecasting model is a necessary means to solve the security problem of campus network.

Network traffic analysis technology in the process of continuous development, but also according to changes in the network environment and constantly improve the identification algorithm and extraction of traffic characteristics. The existing methods of network traffic identification can be divided into two parts: one is based on port number mapping, the other is based on net load. Among them, the concrete realization process of traffic analysis method based on port number mapping is as follows: First, grab the packet and extract the port number information of the packet head. Then, according to the mapping table of port numbers formulated by the IP address assignment agency, the corresponding network applications are found. For example, FTP applications correspond to 20 port numbers, SMTP applications to 25 port numbers, and HTTP applications to 80 port numbers. The method of port number mapping is used to analyze the type of traffic application in P2P application, and the technology of port identification is used to develop the traffic billing system [1]. Then, the net payload of each packet in the network data stream is retrieved by deep packet detection technology. If the characteristic field of an application layer protocol is found, the specific application type of the network data stream can be identified [2].

However, due to the popularity of P2P and passive FTP as well as the widespread use of random port and network address conversion and proxy technologies, the flow identification method based on port number mapping can not meet the needs of practical application gradually. The drawback of the flow identification method based on net load feature is that although the method is simple, effective and easy to maintain, and the identification accuracy is much higher than the flow identification method based on port number mapping, this method may violate the privacy of both sides of communication, can not identify encrypted data flow, need to update the feature segment library in time for new applications, and the net load of analytical flow needs a large amount of calculation. Therefore, with the widespread use of data encryption technology and various network applications, this traffic identification method will become increasingly unable to meet the actual needs.

Aiming at the problems mentioned above, this paper puts forward a traffic analysis method of university course teaching network based on machine learning. Analyze the process of user data encapsulated into Ethernet data frames, and build a network traffic identification model based on machine learning. Use the screening coefficient to screen the data, improve the accuracy of the data, reduce the error of the network traffic model by measuring the task, collect and sort out the data, improve the practicability of the model, clean and preprocess the data to ensure the consistency of the data, evaluate the model to determine the model parameters, analyze the unknown traffic to determine the abnormal network traffic and normal network traffic. Machine learning has the ability of data mining, which can extract the implicit and regular information from big data.

Using machine learning technology, the overall situation of network traffic is analyzed by extracting the statistical and behavioral characteristics of network traffic.

2 Construction of Network Traffic Identification Model Based on Machine Learning

The encapsulated data packet of university course teaching network consists of a packet head and a payload. The payload part is the data to be transmitted. Taking the TCP transport protocol as an example, the user data is encapsulated through the TCP/IP stack as shown in Fig. 1.

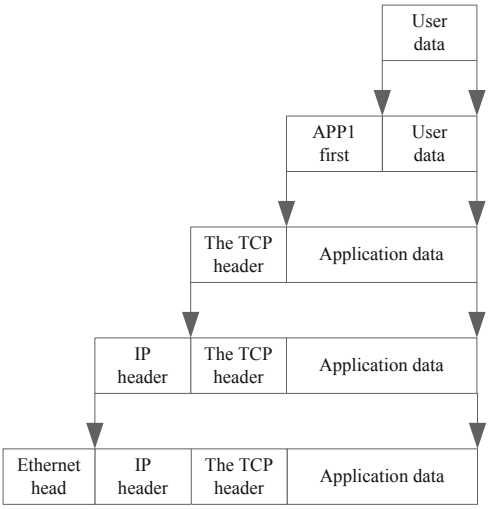

Fig. 1. Encapsulation of user data into Ethernet data frames

As can be seen from Fig. 1, when the source host sends data, the user data first passes through the application layer, adding the Appl header of the application layer, and obtains the application data transmitted in the network, that is, the load part of the packet; then, the application data is encapsulated through the transport layer, adding the TCP header, and obtains the TCP data segment transmitted in the transport layer; then, the TCP data segment is encapsulated through the network layer, adding the IP header, and obtains the IP datagram transmitted in the network layer; finally, the IP datagram is encapsulated through the link layer, adding the Ethernet header and tail, and obtains the Ethernet data frame transmitted in the link layer [3]. Ethernet data frame is the data transmitted directly on the physical link. When the destination host computer receives the data, it peels off the first part from the link layer to the application layer. In the encapsulated data packets described above, the applied data part is the payload, or the deep packet detection part, to be examined in depth [4]; the rest is the head part.

Network traffic identification model based on machine learning, as shown in Fig. 2.

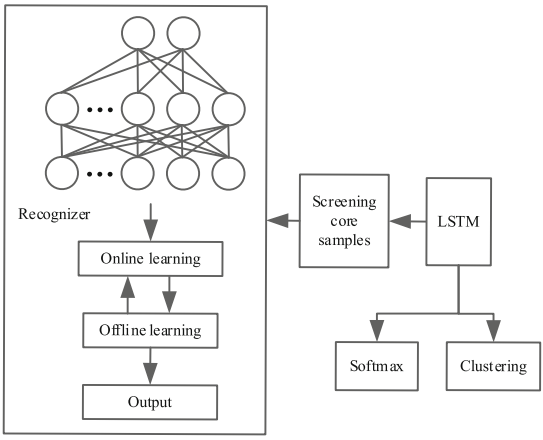

Fig. 2. Network traffic identification model based on machine learning

As can be seen from Fig. 2, the method combines the CNN -based incremental learning network traffic identification model and the LSTM -based semi-supervised network traffic clustering model. Based on the flow level semi-supervised network traffic clustering model, further filtering is added and the filtering results are fed back to the online learning model for updating training [5]. The model not only can give the result of traffic identification in time, but also can adapt to the change of network environment automatically. In addition, semi-supervision also can reduce the cost of traffic data annotation.

3 Traffic Analysis of University Course Teaching Network Based on Machine Learning

3.1 Data Screening

Because the filtered data is used for online learning model training, it is necessary to trade-off between the amount of data and the accuracy of the results. When the amount of data filtered is large, the accuracy of the data is relatively low. Although the amount of data filtered for updating the training online learning model is large and the number of training executions is large, the accuracy of the samples used for training is relatively low, which may lead to the false characteristics of the model [6]; conversely, when the amount of data filtered is small, the accuracy of the data is relatively high, although the accuracy of the data filtered for updating the training online learning model is very high, the number of samples available for training is small and the number of training executions is small, which may lead to the failure of the model to learn the new environment in a timely manner [7].

A screening coefficient is used to control the screening intensity, thereby controlling the trade-off between the amount of data and the accuracy. The formula can be as follows:

$$s_i = \lambda \cdot s_c \tag{1}$$

In formula 1, $\lambda > 0$ is the screening coefficient and s_c is the truncated distance. By adjusting the coefficient, we can control the range of density boundary, and then control the intensity of selecting sample data [8]. On the contrary, the smaller the screening coefficient is, the smaller the threshold value is, the larger the number of core samples is, and the lower the precision of core samples is. The setting of the filter coefficient is a trade-off between the amount and precision of the training data updated from the online learning model [9]. Setting appropriate coefficients can effectively improve the performance of online learning model.

3.2 Model Training

The analysis of normal network traffic and the analysis of abnormal network traffic are two closely related problems. From the point of view of machine learning, they both make some predictions based on the current network traffic (From this point of view, the detection of abnormal traffic can also be regarded as a special predictive problem). In learning, both of these problems rely on experience derived from historical traffic data over time; from the type of learning, the prediction of network traffic is a supervised learning problem, while the detection of abnormal traffic is an unsupervised learning problem [10].

Normal Network Traffic Analysis
In the process of analyzing normal network traffic, the task of learning is to predict the traffic at a certain time in the future according to the current observations of traffic. The improvement of task performance is measured by the deviation between the predicted result and the true value of traffic flow. The learning experience that relies on improving performance metrics comes from historical traffic data over time.

For normal network traffic analysis problems, offline training shall be adopted. The detailed steps are as follows:

Step 1: Collect network traffic data needed for training;
Step 2: Pre-training CNN model and LSTM-based encoder -decoder model:;
Step 3: Test CNN model and encoder -decoder model, if the test results do not meet the performance indicators;
Step 4: Output pre-trained CNN model and LSTM encoder model.

Analysis of Abnormal Network Traffic
In the process of analyzing abnormal network traffic, the task of learning is to judge whether the network traffic is abnormal according to the current observed value of traffic. The measure of performance improvement of a task is the deviation between the result of detection and the real traffic situation (normal or abnormal), and the learning process relies on the experience of improving the performance measure from the historical traffic data over time.

For the abnormal network traffic analysis problem, the online training mode is adopted, and the detailed steps are as follows:

Step 1: Receive a network packet;
Step 2: Extract the five-tuple information of the packet, i.e. Source IP, Destination IP, Source Port, Destination Port, Protocol >, as an identifier for a stream, if it exists in the database, go to Step 4;
Step 3: Increase the storage space for this identifier in the database;
Step 4: Store the packet header data of the packet in the corresponding position of the identifier, if the packet corresponding to the identifier has been stored P, go to step 5, perform packet level network traffic identification; otherwise go to step 6;
Step 5: Input the data of P packets into the packet level recognizer, get the recognition results, and output the recognition results, to step 6;
Step 6: If the flow for this identifier is complete, go to Step 7, otherwise go to Step 1;
Step 7: Input the data of the whole stream into the flow level recognizer, and get the recognition result. If the recognition result is the edge of the cluster, it indicates that the recognition is not very sure.
Step 8: If the sample is new, go to Step 9 and perform a structural update, otherwise go to Step 10;
Step 9: Increase the output node of the package level recognition model;
Step 10: Update the internal parameters of the package level identification model, including weights, offsets, and proficiency;
Step 11: Free up the storage space for this identifier and go to Step 1.

3.3 Design of Network Traffic Analysis Steps

The process of network traffic analysis based on machine learning can be divided into four steps: data collection and collation, data cleaning and preprocessing, model evaluation and unknown traffic detection.

In the network traffic analysis process, firstly, network traffic data will be collected and sorted, and the training set and the verification set will be separated after data cleaning and preprocessing; then the training set will be input into the machine learning model for parameter training, and after training, the verification set will be used to evaluate the performance of the model; if the model performs well on the verification set, the model will be saved; when there is any data with unknown traffic that needs to be detected, it will be cleaned and preprocessed as the input of the saved model, and finally the detection results of the traffic type will be output. The specific process is as follows:

Data Collection and Consolidation
There are two ways to obtain network traffic data, one is to use open traffic dataset, the quality of the data obtained in this way is high, do not need a lot of data cleaning operations, and the traffic type label is accurate. However, the difficulty of traffic detection using public data sets is that the characteristics of the current network environment are often different from those of the data sets. This method is based on the history of network traffic log or data frame to extract the available features manually, and then tag the data according to the real category of traffic. The data obtained by this way are noisy and

need more careful cleaning and preprocessing, but if we can ensure the accuracy of the training set and the training label, the model will be more practical.

Data Cleaning and Pretreatment

The effect of data cleaning and preprocessing has a great influence on the model. In order to shorten the system response time, the flow detection model with few parameters and simple structure is adopted in the current network environment. Network traffic logs or data frames contain a large amount of redundant information. Data cleaning mainly refers to extracting the original traffic features or statistical features needed from logs or available files; and preprocessing refers to the operation of removing and filling the extracted data to improve the data quality and transform them into structured data that meets the model input standards. Before model training, we need to divide the data set into training set and verification set. The training set is used for model training, and the verification set is used for model evaluation. In addition, the analysis of unknown flow also requires the same cleaning and preprocessing operations to ensure the consistency of training data.

Model Evaluation

In the model evaluation phase, a small size selector is used for feature selection and sample construction. Then these new samples are input into a stochastic forest and a limit tree model for evaluation. In addition, the size and number of selectors are dynamically determined. Larger feature sizes are used to combine features and build samples until the training loss function of the multigranularity traversal module is no longer reduced. Multi-granularity traversal is different from multi-granularity scan in deep forest model, the former is to search for different combinations of all features in structured data, and the latter is to construct new samples in image data by scanning pixels of different regions with a certain step size through a sliding window.

The traversal flow of the two combinations is shown in Fig. 3.

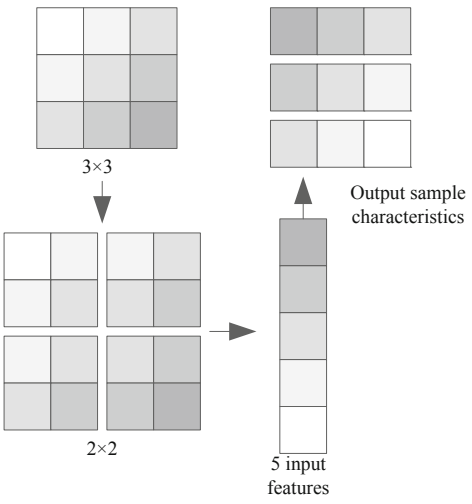

Fig. 3. Traversal flow of two combinations

As can be seen from Fig. 3, the input image of 3 × 3 size is taken as an example to illustrate the process of multi-granularity scanning. When the size of the sliding window is 2 × 2, when the step size is 1, the sliding window can slide down to the right until the right edge and the lower edge of the sliding window coincide with the original input image, resulting in four samples. With the structured data of 5 original features as an example, if the size of the selector is 3, then all possible combinations of 3 features can be traversed, resulting in ten samples according to the combination formula.

After the model is trained, the performance of the model needs to be evaluated with the help of the verification set. The common evaluation indexes are precision, recall, precision and F1 score. If the evaluation index of the model is lower than the expected value, it is necessary to analyze the reasons from feature validity, model validity, experimental environment setting and other aspects to find out the problem and make pertinent correction to the model; if the model performance meets the expected value, the model can be directly saved as the final network traffic analysis model.

Analysis of Unknown Flows

After the parameters of the model are determined, the detection of the unknown flow also needs to be washed and pretreated in the same way as the training set, and then input it to the trained model for detection. For the multi-classification problem of network traffic detection, the output of the model is the probabilistic vectors of each traffic type, and the traffic type corresponding to the highest one is the final analysis type.

The internal product matrix is constructed from the training data set, and the formula is:

$$E=\left(e\left(x_i, y_j\right)\right) \tag{2}$$

In formula (2), $e\left(x_i, y_j\right)$ represents some inner product of sample x_i and y_j through kernel function e. The eigenvalue decomposition of the inner product matrix is used to obtain the eigenspectrum. Calculate the large principal component vector of the sample to be tested, which is:

$$f_d = \sum_{j=1}^{x} \frac{\left(y_{telt}^1\right)^2}{\phi_j} \tag{3}$$

In formula (3), ϕ represents the characteristic spectrum, and y_{telt}^1 represents the principal component vector.

When calculating the small principal component vector of the sample to be tested, the formula may be:

$$f_x = \sum_{j=1}^{x} \frac{\left(y_{telt}^2\right)^2}{\phi_j} \tag{4}$$

In formula (4), y_{telt}^2 represents a small principal component vector.

Suppose there is a pair of false alarm parameters α_1 and α_2, if the calculation result of formula (3) and (4) is larger than the false alarm parameter, the network traffic shall be determined as abnormal traffic, otherwise it shall be deemed as normal traffic.

4 Experimental Analysis

4.1 Demand Analysis of University Course Teaching Network

A certain university course teaching network is a relatively large campus network, four campus is located in a relatively distant geographical location, the building to 100 trillion Ethernet network access, between the building by gigabit optical fiber connection, with more than 20 core and convergent switches, access to more than 600 switches, more than 60 network servers to undertake core services, distributed in the whole school more than 13000 information points, thousands of people at the same time. The network topology is shown in Fig. 4.

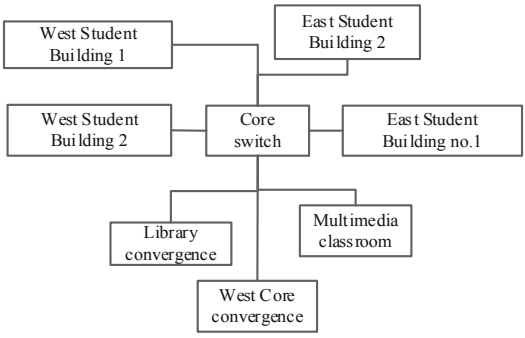

Fig. 4. University course teaching network topology

As a platform of teaching and scientific research, University Course Teaching Network must be able to provide a larger bandwidth, and it needs a long time network service. The main features of the university course teaching network are:

Multi-user, Wide Range
With the development and popularization of the network, campus network has been not only limited to the original computer room and classroom, but all over the teaching buildings, dormitories and so on. Today's campus network has covered the entire campus, the user base is also gradually expanding.

Frequent Traffic and Various Network Applications
In addition to the normal network applications of scientific research and teaching, there are quite a number of services also rely on the network, such as online games, campus network BBS, E-mails, etc. And some need high bandwidth, such as network game and FTP server.

Export Uniformity
Although the campus network covers a wide range of traffic types, but are basically

limited to a few exports, namely, telecommunications, Unicom and education network exports, mostly one or several routers connected to the external network.

Low Safety
Because the university curriculum teaching network's scope is big, the flow many characteristics, causes the university curriculum teaching network the security question to be day by day serious. Upgrading the operating system or installing firewalls and antivirus software for all the hosts of the University Course Teaching Network is a clear waste of manpower and resources. If you use the enterprise-class firewall policy, you may be part of the research nature of the application to shut out.

4.2 Experimental Indicators

The evaluation performance indicators include accuracy rate, precision, recall rate, F1 score, etc. F1 score is an index used to measure the accuracy of binary classification model in statistics. The accuracy rate is the most simple and commonly used performance indicator in classification problems, which is used to measure the percentage of correct samples identified against the total samples. The calculation formula is as follows:

$$N = \frac{M}{m} \tag{5}$$

In formula (5), M represents the total number of samples; m represents the number of samples correctly identified. Precision, recall rate and F1 score are commonly used performance indicators in traffic analysis problems. For traffic analysis problems, the performance indicators of each category should be calculated. When calculating the corresponding performance indicators of Z_0, Z_0 is regarded as a positive example, and the other categories are regarded as counterexamples. Based on the combination of the real results and the model identification results, the sample is divided into four cases: real case, false positive case, true negative case and false negative case. If the T_T, F_T, T_J, F_J represents the corresponding number of samples respectively, the following are:

$$T_T + F_T + T_J + F_J = m \tag{6}$$

$$T_T + F_T = M \tag{7}$$

According to the confusion matrix, the precision, recall and Fl fraction can be calculated. The precision represents the percentage of positive examples that are identified, also known as the precision ratio. The formula is:

$$P = \frac{T_T}{T_T + F_T} \tag{8}$$

The recall rate indicates how many of the samples that should be positive are correctly identified, also known as recall rate. The formula is:

$$R = \frac{T_T}{T_T + F_J} \tag{9}$$

F1 scores shall be defined as the harmonic average of precision ratio and recall ratio, and the calculation formula shall be:

$$\frac{1}{F1} = \frac{1}{2} \cdot \left(\frac{1}{P} + \frac{1}{R} \right) \tag{10}$$

In this experiment, the accuracy rate and F1 score were used as evaluation indexes.

4.3 Dataset Description

The correctness rate of traffic analysis of university course teaching network is verified from two aspects: one is the correctness rate of identifying normal network traffic and abnormal network traffic;

The specific experimental procedures are as follows:

①In the local area network, use the NetMate tool to collect network data packets and calculate the network traffic characteristics, according to the data flow grouping and marking is normal network traffic applications. The collected traffic is divided into 3 datasets named Al, A2 and A3 respectively.
②Extracting stream related information from data source files, vectorizing 3D feature vectors to form training and classification data sources.
③The traffic data collected and sorted from the core router exit of the University Course Teaching Network Center is shown in Table 1.

Table 1. Dataset/KB

Time/min	A1	A2	A3
2	20	35	42
4	20	30	44
6	18	32	46
8	18	35	44
10	−10	−38	−40
12	−15	−40	−42
14	−22	−42	−45

In Table 1, "-" represents a traffic exception. In each group, select 1000 data and divide them into 800 groups and 200 groups in 8: 2 ratio. Add 10% samples to the test set to verify the correctness of the machine learning-based traffic analysis method used in university course teaching network.

4.4 Experimental Results and Analysis

Accuracy Rate

In order to verify the effectiveness of the network traffic analysis method based on machine learning for university course teaching, it is compared with the traffic analysis method based on port number mapping and the traffic analysis method based on payload characteristics, as shown in Fig. 5.

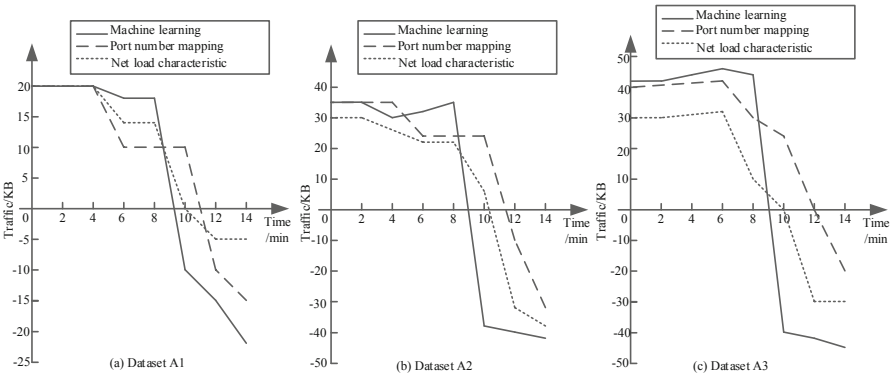

Fig. 5. Comparison of network traffic analysis results using three methods

As can be seen from Fig. 5 (a), the maximum value of network traffic analyzed using the port number-based traffic analysis method is 20 KB, and the minimum value of network traffic is −15 KB; the maximum value of network traffic analyzed using the traffic analysis method based on net load characteristics is 20 KB, and the minimum value of network traffic is −5 KB; the maximum value of network traffic analyzed using the network traffic analysis method based on machine learning -based teaching courses is 20 KB, and the minimum value of network traffic is −22 KB.

As can be seen in Fig. 5 (b), the maximum value of network traffic analyzed using the port number-based traffic analysis method is 35 KB and the minimum value of network traffic is −32 KB; the maximum value of network traffic analyzed using the traffic analysis method based on net load characteristics is 30 KB and the minimum value of network traffic is −38 KB; and the maximum value of network traffic analyzed using the network traffic analysis method based on machine learning -based teaching courses is 35 KB and the minimum value of network traffic is −42 KB.

As can be seen from Fig. 5 (c), the maximum value of network traffic analyzed using the port number-based traffic analysis method is 42 KB and the minimum value of network traffic is −20 KB; the maximum value of network traffic analyzed using the traffic analysis method based on net load characteristics is 32 KB and the minimum value of network traffic is −30 KB; and the maximum value of network traffic analyzed using the network traffic analysis method based on machine learning -based teaching courses is 46 KB and the minimum value of network traffic is −45 KB.

From the above analysis results, it can be seen that the result of the traffic analysis based on machine learning is consistent with the data in Table 1, and the accuracy is high.

F1 Score

Three methods of F1 score contrast analysis, the results are shown in Fig. 6.

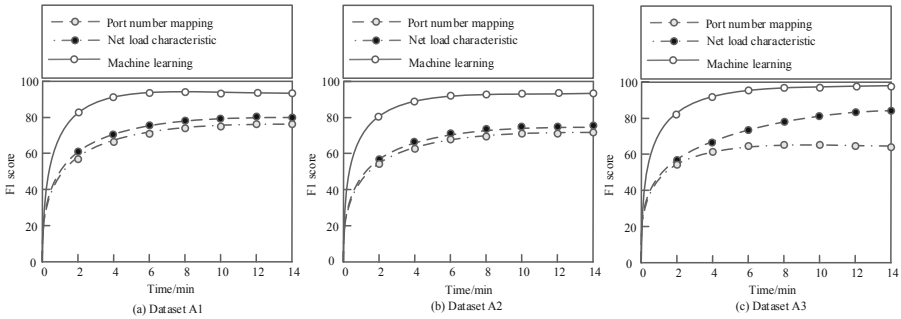

Fig. 6. Comparison of F1 score analysis results of three methods

As shown in Fig. 6 (a), using the port-number-based mapping approach to traffic analysis, the highest F1 score was 77 per cent; using the net-load character-based approach to traffic analysis, the highest F1 score was 80 per cent; and using the machine-learning -based approach to network traffic analysis for university courses, the highest F1 score was 96 per cent.

As shown in Fig. 6 (b), the maximum score of F1 is 70 per cent using the port-number-based mapping approach to traffic analysis; the maximum score of F1 is 73 per cent using the net-oad character-based approach to traffic analysis; and the maximum score of F1 is 94 per cent using the machine-learning -based approach to traffic analysis for university course teaching networks.

As shown in Fig. 6 (c), the maximum score of F1 was 64 per cent using the port-number-based mapping approach to traffic analysis; the maximum score of F1 was 82 per cent using the net-load character-based approach to traffic analysis; and the maximum score of F1 was 98 per cent using the machine-learning -based approach to traffic analysis for university course teaching networks.

Through the above analysis, it can be seen that the F1 score is the highest by using machine learning -based traffic analysis method.

5 Conclusion

Aiming at the problem of poor real-time performance of current traffic detection methods, this paper puts forward a traffic analysis method for university course teaching network based on machine learning. By constructing the traffic analysis model of university course teaching network, the original traffic features are filtered according to the importance scores of the features, and then a new sample is constructed into the model.

References

1. Yanan, W., Xue, Y., Haotao, Z., et al.: Algorithm of logical topology mapping for resource optimization based on reinforcement learning. Optical Communication Technol. **44**(06), 46–50 (2020)
2. Xiaowei, L., Yao, M., Yongle, C., et al.: Shodan traffic identification based on load characteristics and statistical characteristics. Comput. Eng. **47**(01), 117–122 (2021)
3. Liu, S., Liu, D., Muhammad, K., Ding, W.: Effective template update mechanism in visual tracking with background clutter. Neurocomputing **458**, 615–625 (2021)
4. Qilong, Z.: Short-term traffic prediction simulation of multi-scale network based on discrete variables. Computer Simulation **38**(05), 423–426+476 (2021)
5. Liu, S., et al.: Human memory update strategy: a multi-layer template update mechanism for remote visual monitoring. IEEE Trans. Multimedia **23**, 2188–2198 (2021)
6. Xue, B., Erbuli, N.: Research on visual analysis method based of multi-view collaboration on network traffic data. J. Chinese Computer Systems **41**(09), 1893–1897 (2020)
7. Meng, Y., Qin, T., Zhao, L., et al.: Network anomaly detection method based on residual analysis. J. Xi'an Jiaotong University **54**(01), 42–48+84 (2020)
8. Shuai, L., Shuai, W., Xinyu, L., et al.: Fuzzy detection aided real-time and robust visual tracking under complex environments. IEEE Trans. Fuzzy Syst. **29**(1), 90–102 (2021)
9. Xiaohui, L., Chaoyang, C., Huawei, Y., et al.: Large scale network traffic prediction based on cloud computing and big data analysis. J. Jilin University (Engineering edition) **51**(03), 1034–1039 (2021)
10. Huixiang, X., Min, C., Yingying, M.: Combined prediction model for nonlinear network flow based on big data analysis. J. Shenyang Univ. Technol. **42**(06), 670–675 (2020)