Research Article

An Encrypted Traffic Identification Scheme Based on the Multilevel Structure and Variational Automatic Encoder

Jiangtao Zhai ^{1,2}, Huaifeng Shi,³ Mingqian Wang,⁴ Zhongjun Sun,² and Junjun Xing²

¹School of Computer and Software, Nanjing University of Information Science & Technology, Nanjing, China

²School of Electronics & Information Engineering, Jiangsu University of Science & Technology, Zhenjiang, China

³School of Automation, Nanjing University of Science & Technology, Nanjing, China

⁴School of Information Engineering, Changzhou Vocational Institute of Mechatronic Technology, Changzhou 213164, China

Correspondence should be addressed to Jiangtao Zhai; jiangtaozhai@gmail.com

Received 17 March 2020; Revised 21 May 2020; Accepted 12 June 2020; Published 8 July 2020

Academic Editor: Yuan Tian

Copyright © 2020 Jiangtao Zhai et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

With the rapid growth of the encrypted network traffic, the identification to it becomes a hot topic in information security. Since the existing methods have difficulties in identifying the application which the encrypted traffic belongs to, a new encrypted traffic identification scheme is proposed in this paper. The proposed scheme has two levels. In the first level, the entropy and estimation of Monte Carlo π value as features are used to identify the encrypted traffic by C4.5 decision tree. In the second level, the application types are distinguished from the encrypted traffic selected above. First, the variational automatic encoder is used to extract the layer features, which is combined with the frequently-used stream features. Meanwhile, the mutual information is used to reduce the dimensionality of the combination features. Finally, the random forest classifier is used to obtain the optimal result. Compared with the existing methods, the experimental results show that the proposed scheme not only has faster convergence speed but also achieves better performance in the recognition accuracy, recall rate, and F1-Measure, which is higher than 97%.

1. Introduction

With the rapid development of network technology, more and more applications used encrypted algorithm to ensure their information security. As the encrypted traffic cannot be recognized well, it is becoming a good carrier of the network attacks. It is reported that the attacks represented by botnets [1], APTs (Advanced Persistent Threat) [2], and worms are becoming increasingly fierce. Meanwhile, new means of communication are constantly emerging (covert communication, virtual protocol network, and tunnels), some of which may constitute resource misuse of an organization's network system. The communication will bring risk to the network. The identification to the communication is important in network management and security. Thus, the identification to the encrypted traffic is favored by the researchers.

The identification methods to the encrypted traffic mainly can be divided into two types, feature-based and

machine learning-based. The feature-based method can also be divided into payload characteristics matching-based, host behavior-based, data packet distribution-based, and payload randomness-based [3]. A lot of achievements have been obtained by now [4, 5]. A researcher from Cambridge University [6] proposed a feature matching model which can identify various applications through matching the features of the protocol. The drawback of it is that the interaction phase of the encrypted traffic and the private protocols cannot be identified. Okada et al. [7] proposed a method which made decision through calculating the correlation between the encrypted and unencrypted traffic. In this method, the feature vector which has 29 dimensions is used in the machine learning algorithm. Although it has a better performance, it also has the shortcomings that much more features used may lead to a large calculation amount. Shen et al. [8] proposed a model called SOB based on the second-order Markov chain. It uses the SSL/TLS protocol certificate length and the size of the first application data as the feature.

Its experiments verify the effectiveness of this method. An encrypted traffic identification method based on weighted cumulation of time series has been proposed in [9]. Its experimental results show that it has a better performance to identify the encrypted traffic. However, as the private protocol has no common criteria, it will lead the method to be invalid.

Although the existing methods can achieve a better performance to the encrypted traffic, it may lead to many false alarms as the multimedia or compressed files transfer traffic existing. Meanwhile, as the private encrypted protocols have no common criteria, the current models may have difficulty to identify them accurately.

Focusing on the abovementioned investigation, a new identification scheme based on a two-level structure is proposed in this paper. The first level divides the traffic into encrypted and unencrypted based on the entropy, the estimation of the Monte Carlo π value, and the C4.5 decision tree. The first level solves the problem that the existing method may produce false alarms to the encrypted and unencrypted traffic. The second level uses a finer-grained approach to identify the traffic. In this level, first, the feature selected automatically by VAE (variational automatic encoder) and the common features used in existing methods are combined. Then, the mutual information algorithm is used to get the feature set that has the greatest contribution to classification. It can avoid the efficiency and accuracy problem bring by the feature bias caused by feature redundancy. Finally, the experimental results show the effectiveness of the proposed scheme.

The main contributions of this paper can be concluded as follows: (I) a two-level encrypted traffic identification scheme is proposed. The first level judges whether the traffic is encrypted or not, and the second gives the detailed application which the encrypted traffic belongs to. (II) In the first level, to overcome the shortcomings of the entropy method, the Monte Carlo π value is introduced. (III) The features used by the proposed scheme are from the VAE and the existing common features. In addition, for reducing the calculation complexity, the mutual information algorithm is used to reduce some less contribution features.

2. Related Works

2.1. The Analysis of the Network Packet Payload. In the network traffic, the contents transferred are all represented by characters which are composed by ASCII codes. The characters' appearance in network conversations is generally found by statistical rules that the common characters appear more frequently than the uncommon [10]. The characters' appearance in the encrypted traffic will have more randomness than that in the unencrypted traffic. Thus, the entropy-based method is always used to identify the encrypted traffic.

2.2. The Brief Introduction to Entropy. The entropy is first proposed by C.E. Shannon. It is a measure of the number of possible arrangements [11]. The higher the entropy of an object, the more it is uncertain of its states. Assuming that

the number of possible events is N and the probability of their occurrence is described as $p_1, p_2, \dots, p_i$, the definition of its entropy can be given as follows:

$$H = -K \sum_{i=1}^N p_i \log(p_i). \quad (1)$$

From the abovementioned analysis of the payload randomness, the entropy may be different between the encrypted traffic and the unencrypted traffic. In most situations, the entropy of the encrypted traffic is larger than the unencrypted traffic. Thus, the researchers proposed many methods based on entropy to identify if the traffic is encrypted or not. However, there also exist some situations in which the unencrypted traffic has a larger entropy value. It will result in the entropy-based method being invalid. Figure 1 gives an example that the entropy-based method cannot identify whether it is encrypted in the abovementioned situations. From Figure 1, the entropy of the compressed traffic is like the encrypted traffic.

2.3. The Monte Carlo Simulation. Through the principle analysis of the compression technology, the short characters are used to replace the long ones in order to maximize space savings. The compressed traffic presents local randomness of characters [12]. The Monte Carlo simulation extracts every n character of the packet payload as a set of the simulation points. Its idea is that there is an inscribed circle in a square, the first $n/2$ characters are taken as the x -point of the coordinate axis, and the last $n/2$ characters are taken as the Y -point. The π value is calculated according to the number of coordinate points falling into the circle, and the error between it and the real π value is calculated. Figure 2 is a comparison diagram of the error π value between the unencrypted compressed traffic and the encrypted traffic. The error π value of the unencrypted compressed traffic is larger, while the encrypted traffic is smaller. The abovementioned results show that the error π value can be used to distinguish whether the traffic is encrypted or not.

2.4. The Variational Autoencoder. The variational autoencoder (VAE) is an unsupervised deep generative model proposed recently. The discriminant model is relative to the generation model. Generally speaking, given the observation variable X and the latent variable z , the discriminate model gets $P(z/X)$ and obtains the probability that the latent variable z appears according to the input observation variable X . However, the generation model is built on $P(X/z)$ and outputs the probability of the observed variable by inputting the latent variable. The core idea of VAE is to assume that the data is generated by some invisible continuous random variables. For complicated models and large-scale data, the calculation cost of $P(z/X)$ is very high. The distribution $Q(z/X)$ which is referred to as the encoder is used to infinitely approximate the decoder's $P(z/X)$. Then the Kullback–Leibler divergence is chosen as the similar measurement between $Q(z/X)$ and $P(z/X)$, which is given as follows:

FIGURE 1: The entropy value of different traffic.

$$D_{\text{KL}}\left[Q\left(\frac{z}{X}\right)\|P\left(\frac{z}{X}\right)\right] = \int Q\left(\frac{z}{X}\right) \log \frac{Q(z/X)}{P(z/X)} dz. \quad (2)$$

The following is obtained through some mathematical operations like Bayesian transformation and equation transformation:

$$\begin{aligned} \log P(X) - D_{\text{KL}}\left[Q\left(\frac{z}{X}\right)\|P\left(\frac{z}{X}\right)\right] &= E_{Q(z/X)}\left[\log P\left(\frac{X}{z}\right)\right] \\ &\quad - D_{\text{KL}}\left[Q\left(\frac{z}{X}\right)\|P(z)\right]. \end{aligned} \quad (3)$$

When X is given, $P(X)$ will be a fixed value. If $D_{\text{KL}}[Q(z/X)\|P(z/X)]$ is as small as possible, it would be equivalent to making the right-hand side as large as possible. The first item of the equation's right-hand side is based on the likelihood expectation of $Q(z)$, and the second is a negative KL divergence. In order to get an optimal $Q(z)$ and make it to $P(z/X)$ as close as possible, the log-likelihood expectation for the first item on the right should be maximized, while the KL divergence of the second item on the right should be minimized.

2.5. Feature Analysis. The features used in traffic identification mainly contain traffic features, host features, session features, and the behavior features. Among them, the traffic features are used mostly, and the most features are extracted from the transport or network layer. The traffic features are selected from a certain period traffic which has the same five-tuple information. Different application traffic has different characteristics, such as time and upload/download amount [13].

The Moore data set is a publicly available data set for the study of network traffic classification. More than 100 of the 249 network flow attributes in the Moore data set are obtained through the Fourier transform. Among them, redundant attributes are too few to represent the

FIGURE 2: The error π value between the encrypted and unencrypted traffic.

characteristics of the samples from the point of the machine learning view. However, too many features will also bring about redundancy, which will result in feature bias and reduce the performance of the classification efficiency [14]. Meanwhile, with the increase of the network traffic, if the Fourier transform is applied to each network flow, the computing load will be too heavy. Thus, in this paper, in the view of easy access to attributes, 23 network flow attributes commonly used for traffic identification are extracted, and the detailed introduction is shown in Table 1.

2.6. The Mutual Information. In probability and information theory, the mutual information of two random variables is a measure of their interdependence [15]. The mutual information is based on the concept of entropy. The entropy can be understood as the self-information of variables. The mutual information represents that one variable contains some information of the other. The larger the $I(X/Y)$, the higher the correlation between the output category and target category and the better the classification effect.

Formally, the mutual information of two discrete variables X and Y can be defined as follows:

$$I(X; Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}. \quad (4)$$

In equation (4), $p(x, y)$ is the joint probability distribution function of X and Y and $p(x), p(y)$, respectively, represent the edge probability distribution function of X and Y .

In the machine learning field, the feature bias caused by feature redundancy not only decreases the classification effect but also increases the calculation amount. Thus, the mutual information is used to simplify the feature set. The greater the weight of mutual information of one feature, the larger the contribution of the feature to classification.

TABLE 1: The traffic feature set.

No.	Abbreviation	Description in English
1	Total_fpackets	Total number of packets in the forward direction
2	Total_fvolume	Total number of bytes in the forward direction
3	Total_bpackets	Total number of packets in the backward direction
4	Total_bvolume	Total number of bytes in the backward direction
5	Mini_fpctl	Minimum forward packet length
6	Mean_fpctl	Mean forward packet length
7	Max_fpctl	Maximum forward packet length
8	Std_fpctl	Standard forward packet length
9	Mini_bpctl	Minimum backward packet length
10	Mean_bpctl	Mean backward packet length
11	Max_bpctl	Maximum backward packet length
12	Std_bpctl	Standard backward packet length
13	7th_fpctl	The first 7 packet sizes in the forward direction
14	7th_bpctl	The first 7 packet sizes in the backward direction
15	Min_fiat	Minimum forward interarrival time
16	Max_fiat	Maximum forward interarrival time
17	Mean_fiat	Mean forward interarrival time
18	Std_fiat	Standard forward interarrival time
19	Min_biat	Minimum backward interarrival time
20	Max_biat	Maximum backward interarrival time
21	Mean_biat	Mean backward interarrival time
22	Std_biat	Standard backward interarrival time
23	Flow_type	Type of flow

3. The Proposed Scheme

The whole scheme of the proposed method is shown in Figure 3. First, it divides the traffic into encrypted and unencrypted. Then, the VAE is used to identify the detailed application it belongs to.

3.1. Encrypted Traffic Filtered Based on Entropy and the Monte Carlo Method. The current state-of-the-art algorithms have the shortcomings that cannot accurately and efficiently differentiate among the encrypted and compressed packets (such as .zip, .rar, and so on), image packets, or video packets. They may have the drawback that they identify the unencrypted compressed or multimedia traffic as encrypted. Also, they have poor performance to the private protocols. In this paper, an improved encrypted traffic identification model based on the payload randomness is proposed. The proposed method uses entropy and the Monte Carlo estimation value as the input feature vector of the classifier. The classifier used in this paper is the C4.5 decision tree. The flow chart of the abovementioned method is shown in Figure 3.

The detailed steps are as follows:

Step I: the network traffic is captured according to the five-tuple. For the TCP traffic, a link is between three successive SYN packets and the final FIN or RST packet. Also, for the UDP traffic, a link is determined by the time between the first packet received and no packet received during 60 seconds.

Step II: the first packet of one link is extracted, and it is determined whether the length of the first packet is larger than 1024 B. If it is, its payload will be extracted,

and otherwise, it will be discarded and go to the next packet until N packets are all extracted. Using equal spacing algorithm, the Shannon entropy formula is used to calculate the entropy value of each character in the whole data package.

Step III: each N characters of the extracted packet's payload are used as a set of Monte Carlo simulation points. The first N/2 characters are taken as the x-point of the coordinate axis, and the last N/2 characters are taken as the Y-point. The π estimation value is calculated according to the number of coordinate points falling into the circle, and the error with a real π value is also calculated.

Step IV: the entropy value H and Monte Carlo estimation π value error P are standardized, and then the two features are input into the C4.5 decision tree classifier to get the classification results.

3.2. Encrypted Traffic Identification Method. In order to further distinguish which application the filtered encrypted traffic belongs to, an encrypted traffic identification model based on the variational automatic encoder has been proposed in this paper. The identified encrypted traffic data set should be preprocessed. The first n bytes of the data stream are truncated, and the number of n bytes is not enough to fill 0. In order to prevent the impact of physical hardware on classification, it is necessary to drop the link layer data of the packets. Meanwhile, as the UDP header is 12 bytes less than the TCP, in order to eliminate the influence of experimental error, 12 zero need to be filled to the UDP header. In order to get the best classification effect, it is also necessary to normalize the extracted packet bytes.

FIGURE 3: The framework of the proposed method.

The detailed preprocessing of the encrypted traffic is shown in Figure 3.

Then, the model automatically extracts features through VAE algorithm, and the feature vector with the largest contribution to the classification through mutual information algorithm with the flow feature set is obtained. Also, finally, the feature set is input into the random forest classifier.

Let $X = \{X_1, X_2, X_3, \dots, X_n\}$ presents the network traffic set, and X_i presents one flow. $Y = \{Y_1, Y_2, Y_3, \dots, Y_m\}$ presents all the m sets of network traffic types. The function of the identification model designed in this paper is to realize the mapping from set X to Y , so as to realize the accurate identification of encrypted traffic.

According to Figure 3, the detailed identification steps can be concluded as follows:

Step I: the preprocessed data is input into the VAE model. Then, the n -dimensional hidden layer variables Z of the VAE model are extracted.

Step II: the stream level features related to time and packet length from the identified encrypted traffic data set are extracted to obtain the stream feature set.

Step III: the n -dimensional hidden layer variable Z obtained in step I and the flow feature collection obtained in step II are input to mutual information algorithm to obtain the feature vector with the largest contribution to classification. This step can help to reduce the feature dimension.

Step IV: the feature set obtained in Step III is input to the random forest classifier as the feature vector, the classifier parameters are debugged through cross validation, the optimal classifier model is obtained, and the decision is made.

4. Experiments and Analysis

In this section, the experimental environment, experimental data set, and the performance evaluation index are given.

4.1. Experimental Environment and Data Set

4.1.1. Experimental Environment. The configuration of the experimental computer used in this paper is as follows: Windows 7 Professional, Intel (R) Core (TM) i5-3230M CPU @2.60 GHz, 8G RAM. The third-party software and API used are as follows: VMware Workstation 12, Ubuntu 16.04, Wireshark 2.2.1, LibPcap, Scapy, Sklearn, and Tensorflow.

4.1.2. Experimental Data Set. The following experimental data used in this paper are all captured in our Lab. Our Lab is in Nanjing, Jiangsu province, whose ISP is China Education and Research Network (CERNET). The detailed information of the data set is shown in Table 2. A total of 15,000 encrypted and 9,000 unencrypted traffic streams are collected. The encrypted traffic data set includes Skype, Gmail, SFTP down, Tor Twitter, YouTube, ICQ, and Facebook. Meanwhile, the unencrypted traffic data set includes HTTP, FTP, and Socket file transfer (the file types include.txt, .zip, .doc, and .pdf).

4.1.3. Performance Evaluation Index. In order to evaluate the performance of the algorithm objectively, in this paper, the accuracy P , recall R , and F1-measure are selected as the three scoring references. The recall rate is the proportion of correct prediction to the total actual positive. The F1-measure is a comprehensive evaluation index, which is defined as the harmonic mean of the accuracy rate and recall rate. The calculation formula of the abovementioned three indices is shown as follows:

$$\begin{aligned} P &= \frac{T_p}{(T_p + F_p)}, \\ R &= \frac{T_p}{(T_p + F_N)}, \\ F_m &= \frac{2PR}{(P + R)}. \end{aligned} \quad (5)$$

In the abovementioned formulas, T_p represents the number of correctly identified samples representing the encrypted traffic. F_p indicates the number of encrypted traffic with a wrong identification. F_N represents the number of correctly identified samples representing the unencrypted traffic.

4.2. Experimental Results

4.2.1. Encrypted Traffic Identification Results Based on Load Randomness

(1) Experiments on the Relationship between the Detection Window Size and the Accuracy. The number of packets in the observation window has a great influence on the recognition rate of the model. If the length of packets is too small, it cannot reflect the randomness of the load, so it is necessary to extract packets with a load greater than 1024 bytes. The experimental results are shown in Figure 4.

TABLE 2: The experimental data set.

Traffic type	Data stream number	Encrypted situation
HTTP	3000	Unencrypted
TCP (socket)	3000	Unencrypted
FTP	3000	Unencrypted
SFTP down	2000	Encrypted
Skype	2000	Encrypted
Gmail	1000	Encrypted
Tor Twitter	3000	Encrypted
YouTube	2000	Encrypted
ICQ	2000	Encrypted
Facebook	3000	Encrypted

The average accuracy of the recognition model at the beginning is proportional to the number of data packets. When the number of data packets is small, the accuracy of the model is low. From the statistical point of view, as the amount of data is not enough to fully reflect the characteristics of network traffic, the limitations are too large. When the number of packets reaches 10, the average accuracy reaches 94.98%, and then the two fluctuate up and down in an oscillating relationship.

(2) Experiment between the Number of Characters and the Accuracy of Coordinate Points. The number of characters in the coordinate points of the Monte Carlo simulation point also affects the accuracy of the recognition model. The experimental results are shown in Figure 5. When the number of coordinate point characters is 2, the accuracy of the model is 89.02%, which is not different from that of using only information entropy. As the number of characters in the coordinate point increases to 6, the accuracy of the model is the highest and then decreases with the number of characters in the coordinate point. When the observation window of the recognition model is set to 6, the pseudorandom characteristics of unencrypted compressed traffic can be distinguished mostly.

(3) Comparison Experiments. The algorithms proposed in [10, 16] are used to compare with the proposed one. The results are shown in Figures 6–8. Compared with the two typical algorithms, the average accuracy, recall, and F1-measure of our method are over 94.98%, 90.05%, and 92.45%. The experimental results show that our proposed method achieves the best performance among the existing. As the encrypted traffic and unencrypted compressed traffic show similar characteristics in information entropy (especially, the file types are .zip and .flv), only using the entropy value will lead to misjudgments between them. The average accuracy, recall, and F1-measure of the model in [11] are 85.45%, 83.43%, and 84.42%. Meanwhile, the proposed method is better than the experimental results of flow feature-based recognition model proposed in [16], and the average accuracy, recall, and F1-measure of the recognition model in [16] are only 92.34%, 88.50%, and 90.38%. It is because the recognition model based on flow characteristics cannot accurately identify the situation of byte filling for

FIGURE 4: Experiment on the effect of the packet amount.

FIGURE 5: Experiment on the effect of the coordinate point character number.

data packets and the traffic of too short network flow, which leads to the effect difference between models.

4.2.2. Encrypted Protocol Identification Based on VAE.

(1) *The Experiment about the Traffic Length.* The length of the data stream has great influence on the recognition rate. The experimental results about it are shown in Figure 9. From the results, with the increase in the data length, the average accuracy also increases. When the length is larger than 1,000, the detector can achieve a better performance, whose average accuracy is about 97.86%.

The dimension of hidden layer Z also affects the average accuracy of the proposed method. The experimental results of the relationship between the dimension of hidden layer and the average accuracy are shown in Figure 10.

From the results shown above, when the dimension of hidden layer Z is larger than 2, the average accuracy can

—●— Our scheme
—▲— The scheme of Ref. [20]
—★— The scheme of Ref. [11]

FIGURE 6: The accuracy rate among different models.

—●— Our scheme
—▲— The scheme of Ref. [20]
—★— The scheme of Ref. [11]

FIGURE 7: The recall rate among different models.

achieve 94.5%. When the dimension is 6, the average accuracy can achieve the best performance.

The convergence rate of the model is also an important index. There is also an experiment which tests the trend of accuracy and loss rate in the training process of the recognition model that has been performed. The results are shown in Figure 11.

From the results shown above, the loss rate of the proposed method in the first 10 rounds of training decreases rapidly. Then, the loss rate decreases continuously and finally tends to be stable. It presents that the model proposed in this paper has a faster convergence speed.

FIGURE 8: The F1-measure among different models.

FIGURE 9: Experiment on the effect of the traffic length.

FIGURE 10: Experiment on the effect of the dimension of Z.

FIGURE 11: Experiment on the effect of the Epoch number.

(2) *Comparison Experiments.* An encrypted identification model based on VAE has been proposed by this paper. The VAE model is often used as malicious traffic monitoring [17]. The model parameters of VAE are shown in Table 3. The input of the VAE model is a 1000-dimension original bytes vector. The encoder has two full-link layers. The input of the first full-link layer is a 256-dimension vector, and the second connects two output networks in a parallel structure. The final output of the encoder is a 46-dimension hidden layer variable Z . The decoder has two steps. The first step converts the abovementioned 46-dimension vector to a 256-dimension output vector, and the second converts the 256-dimension vector to the 1000-dimension output vector. Then, the vector Z and flow characteristics are used to calculate their mutual information. Finally, 10 features with the largest weight are input to the random forest classifier.

In order to test and compare the performance of the proposed method, the most basic deep learning model MLP (a recognition model based on flow characteristics proposed in [18]), the recognition model based on CNN proposed in [19] and the multiple classifiers fusion-based method proposed in [20] are selected for comparison. The selected MLP model has one input layer which has 784 neurons and two hidden layers. The two hidden layers, respectively, have 256 and 64 neurons. The activation function is ReLu. The MLP model has one output layer, which has 16 neurons, whose activation function is SoftMax.

The experimental results are shown in Table 4. The average accuracy, recall rate, and F1-measure of the proposed model are, respectively, 97.68%, 97.30%, and 97.49%. It has the best performance among all the comparison models. MLP is the basic deep learning method whose training process is a little simple. However, its average accuracy, recall rate, and F1-measure are only 94.50%, 94.32%, and 94.41%. The average accuracy, recall rate, and F1-measure of the identification model based on the convolutional neural network and stack automatic encoder proposed in [19] are, respectively, 95.60%, 95.34%, and 95.44%. Compared with the method proposed in [19], on the basis of using the deep learning algorithm to automatically

TABLE 3: The model parameter of our experiments.

Model parameter	Value
vae_input	1000
vae_hidden_1	256
vae_hidden_2	6
vae_learning_rate	0.001
vae_training_epochs	400
vae_batch_size	256
vae_loss	MSE
rf_n_estimators	250
rf_criterion	Entropy
rf_max_depth	4

TABLE 4: The comparison of the experimental results.

Methods	Accuracy (%)	Recall (%)	F1-measure (%)
Our model	97.68	97.30	97.49
MLP model	94.50	94.32	94.41
Ref. [18]	94.85	93.74	94.30
Ref. [19]-CNN	95.60	95.34	95.44
Ref. [19]-SAE	96.57	94.20	95.37
Ref. [20]	97.37	95.80	96.58

extract features, our proposed method innovatively combines the idea of VAE algorithm to automatically extract features with the idea of using the knowledge in the field of network traffic; thus, our proposed method can get the best sample features in the load sample feature vector. Meanwhile, the comparison experiments between the proposed method and [18] have also been performed. The average accuracy, recall rate, and F1-measure of the model in [18] are 94.85%, 97.74%, and 94.30%. It is obvious that the performance of the proposed method is better than [18]. That is because the model in [18] only uses the length of the former n packets. There is a similar work proposed in [20]. Its method first drops the packets that has no payload and sets the burst threshold to 1 s. Then, it extracts several features and uses multiple classifiers fusion to give the results. Its average accuracy, recall rate, and F1-measure are, respectively, 97.37%, 95.80%, and 96.58%. It is a little poorer than the proposed method. As it uses the time feature of the traffic, it is vulnerable to network conditions.

5. Conclusions and Future Work

An encrypted traffic identification scheme based on the multilevel structure is proposed in this paper. In the first level, the traffic is divided into encrypted or not by using entropy and the Monte Carlo π value as classification features. The experimental results show that the proposed method has a better performance than the existing methods. For identifying the application within the encrypted traffic more finely, the idea that VAE algorithm can automatically extract features and the network traffic domain knowledge can be used to extract features are combined and used in this paper. Also, the feature set with the largest contribution to classification is obtained through mutual information

algorithm, which avoids the feature bias problem. The comparison of the experimental results shows that the proposed method has achieved a better performance than the existing ones.

In the future, more network situations and applications should be considered to identify in the second level of the proposed scheme.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the National Science Foundation of China (Grant no. 61702235 and 61801073), the Natural Science Foundation of the Higher Education Institutions of Jiangsu Province (Grant no.19KJB510019), and the Startup Foundation for Introducing Talent of NUIST.

References

- [1] L. D. Carli, R. Torres, G. Modelo-Howard et al., "Botnet protocol inference in the presence of encrypted traffic," in *Proceedings of the IEEE Infocom Conference on Computer Communications*, IEEE, Atlanta, GA, USA, 2017.
- [2] Yu Zhang, X. Pan, L. I. U. Q. Zhong et al., "APT attacks and defenses," *Journal of Tsing Hua University (Natural Science Edition)*, vol. 57, no. 11, pp. 1127–1133, 2017.
- [3] W. Pan, G. Cheng, X. Guo et al., "Review and perspective on encrypted traffic identification research," *Journal of Software*, vol. 37, no. 9, pp. 154–167, 2016.
- [4] A. Pektaş and T. Acarman, "Identification of application in encrypted traffic by using machine learning," in *Proceedings of the International Conference on Man-Machine Interactions*, Springer, Berlin, Germany, pp. 545–554, 2017.
- [5] N. Namdev, S. Agrawal, and S. Silkari, "Recent advancement in machine learning based internet traffic classification," *Procedia Computer Science*, vol. 60, pp. 784–791, 2015.
- [6] A. W. Moore and K. Papagiannaki, "Toward the accurate identification of network applications," *Lecture Notes in Computer Science*, pp. 41–54, 2005.
- [7] Y. Okada, S. Ata, N. Nakamura, and Y. Nakahira, I. Oka, "Comparisons of machine learning algorithms for application identification of encrypted traffic," in *Proceedings of the 10th IEEE International Conference on Machine Learning and Applications and Workshops*, vol. 2, pp. 358–361, Honolulu, HI, USA, 2011.
- [8] M. Shen, M. Wei, L. Zhu, and M. Wang, "Classification of encrypted traffic with second-order Markov chains and application attribute bigrams," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 8, pp. 1830–1843, 2017.
- [9] B. Zhao, H. Guo, Q. Liu, and J. X. Wu, "Protocol independent identification of encrypted traffic based on weighted cumulative sum test," *Journal of Software*, vol. 24, no. 6, pp. 1334–1345, 2014.

- [10] L. Chen, L. Zhang, X. Ban et al., "Encrypted session detection approach based on information entropy," *Computer Science*, vol. 42, no. 2, pp. 142–174, 2015.
- [11] Y. Zhang, S. Wang, and L. Yang, "Detection of encrypted data-flow based on entropy," *Computer & Digital Engineering*, vol. 4, no. 4, pp. 555–558, 2014.
- [12] G. Cheng and Y. Chnen, "Identification method of encrypted traffic based on support vector machine," *Journal of Southeast University*, vol. 47, no. 4, pp. 655–659, 2017.
- [13] P. Xu and S. Lin, "Internet traffic classification using C4.5 decision tree," *Journal of Software*, vol. 20, no. 10, pp. 2692–2704, 2009.
- [14] G. Lu, X. Z. Yu, H. L. Zhang et al., "Traffic classification framework based on ensemble clustering," *Journal of Software*, vol. 27, no. 11, pp. 2870–3288, 2016.
- [15] C. S. Miao, C. Q. Yuan, X. W. Wang et al., "A hybrid feature selection algorithm based on mutual information and memetic framework to optimize traffic classification," *Journal of Northeastern University*, vol. 35, no. 11, pp. 1530–1534, 2014.
- [16] Y. Huang, L. I. Yun, and B. Qiang, "Internet traffic classification based on min-max ensemble feature selection," in *Proceedings of the International Joint Conference on Neural Networks*, pp. 3485–3492, IEEE, Vancouver, BC, Canada, 2016.
- [17] J. Sun, X. Wang, N. Xiong, and J. Shao, "Learning sparse representation with variational auto-encoder for anomaly detection," *IEEE Access*, vol. 6, no. 6, pp. 33353–33361, 2018.
- [18] M. Zhang, H. Zhang, B. Zhang, and G. Lu, "Encrypted traffic classification based on an improved clustering algorithm," in *Trustworthy Computing and Services*, pp. 124–131, Springer Berlin Heidelberg, Heidelberg, Germany, 2013.
- [19] M. Lotfollahi, M. Jafari Siavoshani, R. Shirali Hossein Zade, and M. Saberian, "Deep packet: a novel approach for encrypted traffic classification using deep learning," *Soft Computing*, vol. 24, no. 3, pp. 1999–2012, 2020.
- [20] G. Aceto, D. Ciuonzo, A. Montieri, and A. Pescapé, "Multi-classification approaches for classifying mobile app traffic," *Journal of Network and Computer Applications*, vol. 103, pp. 131–145, 2018.