Journal of Computers Vol. 32 No. 5, 2021, pp. 31-43
doi:10.53106/199115992021103205003

DDoS Attack Detection System Based on RF-SVM-IL Model Under SDN

Jingyuan Fan, Guiqin Yang*, Jiyang Gai

School of Electronic information and Engineering, Lanzhou Jiaotong University, Lanzhou, China
yangguiqin@mail.lzjtu.cn

Received 6 February 2021; Revised 5 July 2021; Accepted 1 August 2021

Abstract. The data forwarding plane in the Software Defined Network (SDN) is decoupled from the network control plane, which can realize the unified control of the whole network by the controller. Although centralize control provide great convenient in many cases, it is vulnerable to malicious attacks, especially for one of the most threatening attack - Distributed Denial of Service (DDoS). We innovatively propose a machine learning hybrid DDoS attack detection model which provide high precision within short-term. We named our model as RF-SVM-IL, which represents an integration of integrates Random Forest (RF), Support Vector Machine (SVM) and Incremental Learning (IL). The combination of RF and SVM can detect detect attacks in two layers and filter out the easily misclassified samples. Then IL is added to filter new samples to avoid repeated iteration training, and improve the adaptability of the model to dynamic data. Compared with other methods, RF-SVM-IL can detect DDoS attacks in SDN with higher accuracy and shorter time. The experimental results show that the average detection accuracy of RF-SVM-IL model is as high as 98.54%, and the detection time is as low as 2.386s.

Keywords: SDN, DDoS, SVM, random forest, incremental learning, RF-SVM-IL

1 Introduction

Different from Denial of Service (DoS) attack, Distributed Denial of Service attack (DDoS) send a lot of malicious information to the victims by controlling the distributed zombie hosts, which eventually makes the victims consume a lot of resources and provide ineffective services. The schematic diagram of DDoS attack is shown in Fig. 1. According to the white paper of Cisco's Annual Internet Report (2018-2023), the total number of DDoS attacks will increase from 7.9 million in 2018 to 15.4 million in 2023 [1]. As an architecture of network, Software Defined Network (SDN) has been widely used in the Cloud Computing, 5G and other fields, and has been recognized as the trend of future network development [2]. In the traditional network, the system is closed and integrated, the construction of multi-functional network can only be achieved by the protocols with stacking specific functions [3]. This kind of network has low expansibility and is difficult to fulfill the requirements of diversified modern network development. The core characteristic of SDN is to decouple the data plane from the control plane, and the controller can manage the whole network centrally. The structural comparison between SDN and traditional network is shown in Fig. 2. In order to launch a more lethal attack on the whole SDN network, some attackers always treat the controller as an important attack object. When a switch receives a large number of malicious flow tables, it usually does not have forwarding rules for these malicious flows. and needs to send pack in to request operation instructions from the controller. At this time, the switch needs to send a large number of Pack-in to request operation instructions from the controller. The controller sends forwarding rules to the switch through Pack-out. If a large number of malicious packets with unknown rules enter SDN, the resource consumption of the controller will be great, leading to the failure of the SDN network to run normally.

[*] Corresponding Author

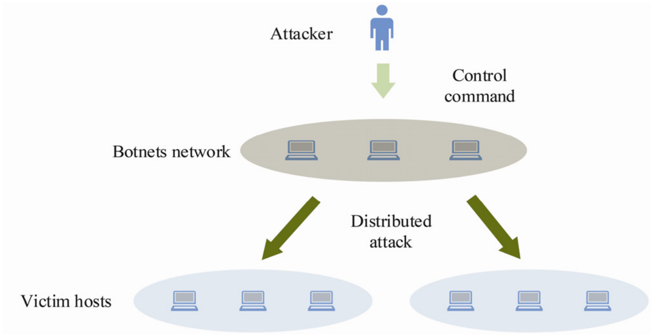


Fig. 1. The schematic diagram of DDoS attack

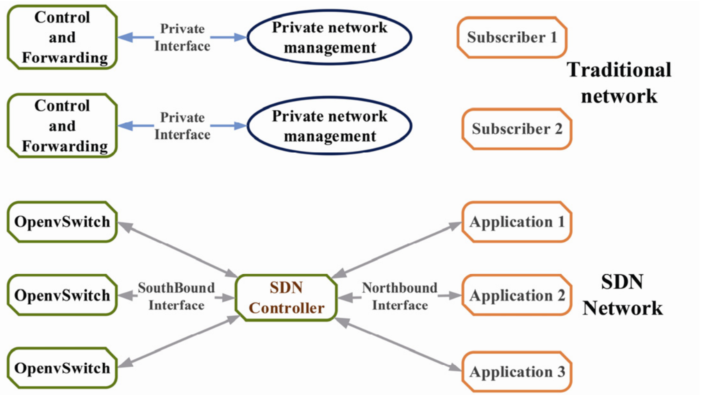


Fig. 2. SDN and traditional network structure comparison

To slove above issues, we proposed a machine learning Fusion Model which constructed by combining, cooperate, and improve Random Forest (RF), Support Vector Machine (SVM) and Incremental Learning (IL) simultaneously. At the same time, the advantage of the centralized control network of the controller is brought into full play, so that it can detect DDoS attacks quickly and accurately. DDoS attack has the characteristics of large traffic and long attack time. Aiming at this characteristic, this paper combines RF and SVM for two-layer detection to improve the detection accuracy. In the first layer, RF takes advantage of highly parallel training to detect and classify large-scale data. At the same time, RF eliminates the data samples which are easy to be misclassified in SVM model. After the above preprocessing, the results are input into SVM for the second level detection, which can give full play to the advantages of SVM in processing small-scale, nonlinear and high-dimensional data [10]. The combination of RF and SVM overcomes the problem of poor classification effect of SVM for big data, and improves the detection accuracy through double-layer detection. Incremental learning algorithm in the face of new data does not need iterative training, through the positive and negative support vector similarity to screen valuable training samples. The addition of IL shortens the detection time of the whole model and improves the dynamic adaptability of the detection model.

The rest of the paper is organized as followed. In Section 2, we describe the related work of DDoS detection. In Section 3, we details the main contributions of this paper, Section 4 shows some experimental results of our method, and Section 5 presents the conclusions and future directions.

2 Related Work

Oo et al. [4] used the improved SVM model to detect DDoS, and deployed the detection module to the application layer. They have improved the traditional SVM algorithm by minimizing $\|\omega\|^2/2$, which is equivalent to maximizing the distance between two classification planes of SVM, so that the detection accuracy of SVM model is higher. However, the computational complexity of SVM model in dealing with large-scale data will be greatly improved, resulting in the increase of detection time and the decrease of detection efficiency. Li et al. [5] defined a Combination Correlation Degree (CCD) of network flow through the analysis of attack flow, and improved the key parameters of RF based on the genetic algorithm of CCD feature sequence. The key parameters include the maximum number and maximum depth of decision tree. Although the detection accuracy of the improved RF has been improved to a certain extent, due to the use of a single machine learning algorithm, the detection accuracy still does not reach a high level.

The above related work is using a single machine learning model for detection, the following describes the work related to machine learning hybrid model. Kim et al. [6] uses SVM and DT for fusion detection. SVM classifies the traffic into normal and attack categories, the DT classifies the flow table information at the edge for the second time to ensure the high accuracy of detection. The detection accuracy of this method has been improved to a certain extent, but the amount of data is not reduced, so the total detection time is not shortened. Li et al. [7] uses Dynamic Convolution Neural Network - Deep Stack Autoder (DCNN-DSAE) hybrid model. In the early detection stage, DCNN was used to detect and classify the traffic. In order to ensure that the attack traffic could be further accurately identified, DSAE was used to detect again. In addition, DCNN and DSAE adopted two methods of supervised learning and unsupervised learning respectively, which effectively avoided the local optimization problem of the model. However, using deep learning algorithm for detection requires higher requirements on the graphics card of the detection device, higher utilization of CPU and GPU, and longer detection time. Aamir et al. [8] combines machine learning algorithm and clustering algorithm for DDoS detection. Principal Component Analysis (PCA) algorithm mainly extracts the feature of information, clustering algorithm classifies and labels the extracted feature information, and finally uses SVM, RF and other machine learning algorithms to train the model. Gong et al. [9] fused KNN and SVM in machine learning algorithm, and introduced relief algorithm into KNN-SVM model by weighting. In SDN, the detection of DDoS is based on the traffic characteristics. This method needs repeated iterative training of data set when new data is added, which has low dynamic adaptability and slow response speed.

3 RF-SVM-IL Detection Model

3.1 Design Idea of RF-SVM-IL Hybrid Model

In this section, the proposed model is described. RF-SVM-IL mainly improves the effection of detection from aspects of both time and precision. As shown in Fig. 3, the dual module collaborative working mode is used. The first module (RF-SVM) mainly implements two-layer detection. In the first layer, it uses the advantage of RF high parallel training to detect and classify large-scale data, then eliminates the data samples that are easy to be misclassified in SVM according to $M_{\text{margin}}(x)$. After that, the filtered samples are inputted into SVM for the second layer detection. The second module is an incremental learning screening sample module. In the process of SDN being attacked, the amount of data to be detected is continuously increasing. The module selects the training samples by judging the positive and negative support vector similarity of new data samples, which shortens the detection time and improves the dynamic adaptability of the detection model.

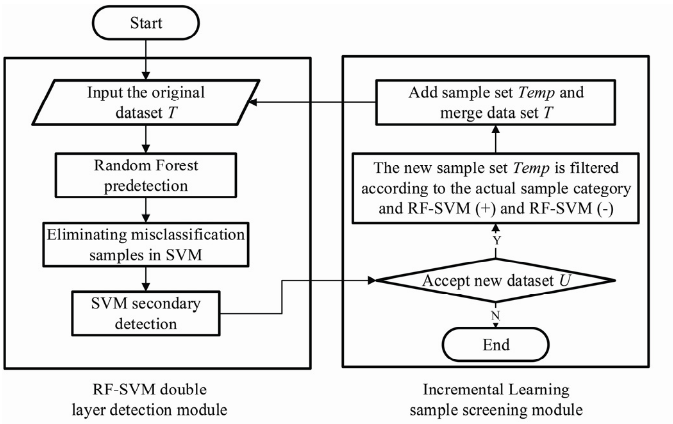

Start
Input the original dataset *T*
Random Forest predetection
Eliminating misclassification samples in SVM
SVM secondary detection
Add sample set *Temp* and merge data set *T*
The new sample set *Temp* is filtered according to the actual sample category and RF-SVM (+) and RF-SVM (-)
Y
Accept new dataset *U*
N
End
RF-SVM double layer detection module
Incremental Learning sample screening module

Fig. 3. The workflow of RF-SVM-IL model

3.2 RF-SVM Double Layer Detection Module

SVM mainly solves the problem of binary classification. In addition, it also overcomes many other machine learning problems, including but not limited to over learning, dimension disaster and so on [11]. SVM divides the data into two categories according to each feature quantity, and maximizes the distance between the classified data samples and the hyper-plane. If the data set *T* needs to detect *G* features, that is to say, it needs to use *G-1* hyper-planes to classify the data set, $\boldsymbol{w}$ represents the normal vector of the hyper-plane, $\mathbf{b}$ represents the intercept of the hyper-plane, and any hyper-plane can be written as

$$\mathbf{w}^{\mathbf{T}} x + \mathbf{b} = 0. \quad \textbf{(1)}$$

The basic form and the optimization goal of SVM is to find the most suitable $\boldsymbol{w}$ and $\mathbf{b}$ according to the training samples, so as to obtain the optimal interval classifier. The optimal solution must satisfy both the original problem and the dual problem, so the $(\mathbf{w}, \mathbf{b}, \alpha)$ must satisfy the Karush Kuhn Tucker (KKT) condition, where α_i is Lagrange Multiplier. The optimal solution $(\mathbf{w}, \mathbf{b}, \alpha)$ and KKT can be expressed as equation (2) and (3) respectively:

$$\min_{\mathbf{w},\mathbf{b}} \frac{1}{2}\mathbf{w}\mathbf{w}^{\mathbf{T}} \quad s.t \quad y^{(i)}(\mathbf{w}^{\mathbf{T}}\mathbf{x}^{(i)} + \mathbf{b}) \geq 1 \quad i = 1, 2, ..., N, \quad \textbf{(2)}$$

$$\left.\begin{array}{l} y^{(i)}\left(\mathbf{w}^{\mathbf{T}}\mathbf{z}^{(i)} + b\right) \geq 0 \\ \alpha_i \geq 0 \\ \sum_{i=1}^{N} \alpha_i y^{(i)} = 0; \mathbf{w} = \sum_{i=1}^{N} \alpha_i y^{(i)}\mathbf{z}^{(i)} \\ \alpha_i\left(1 - y^{(i)}\left(\mathbf{w}^{\mathbf{T}}\mathbf{z}^{(i)} + b\right)\right) = 0 \end{array}\right\} . \quad \textbf{(3)}$$

In many anomaly detection algorithms, the complexity of the algorithm increases exponentially with the increase of the dimension of the data to be classified [12]. Different from these algorithms, the way of SVM classification is to find the hyper-plane, and there is no limit to the dimension of the classified data, so SVM overcomes the problem of dimension disaster in the face of DDoS attack data. When the total number of samples in the dataset is *D*, the training space and complexity of SVM are D^2 and D^3, which indicates that when the data set is too large, the training space and complexity of SVM increase on a large scale [13], so the processing capacity of SVM on large-scale data sets is poor. To overcome the weakness of SVM, RF is introduced in out model due its ability to parallelly deal with large-scale sample data. In addition, it has good fault tolerance for noise and other factors in data [14]. RF also has the characteristics of high detection accuracy and less trend to overfit. In the RF-SVM model, Classification and Regression Trees (CART) algorithm is used as the node classification method. The algorithm has good classification effect on discrete data. According to Gini coefficient, CART algorithm divides data set *T* into left and right decision trees (T_1, T_2). When P_i is the probability of category C_{j} in the sample set T, the calculation formula of $Gini(T)$ can be expressed as equation (4):

$$Gini(T) = 1 - \sum_{i=1}^{m} P_i^2 . \quad \textbf{(4)}$$

In the solution of discrete attributes, through calculation and comparison, an attribute node is selected for classification, and then the binary recursive method is used to construct the decision tree and generate the decision rules. $Gini_{spirt}(T)$ coefficient of each node can be expressed as equation (5):

$$Gini_{spirt}(T) = \frac{T_1}{T} Gini_{spirt}(T_1) + \frac{T_2}{T} Gini_{spirt}(T_2) . \quad \textbf{(5)}$$

In the process of SVM classification, firstly finds out the hyperplane according to the distribution of the samples, divides the samples into two categories outside the region, in which we regard the hyperplane as the center and Margin as the distance. However, some samples in that region are prone to be misclassified [15], as shown in Fig. 4.

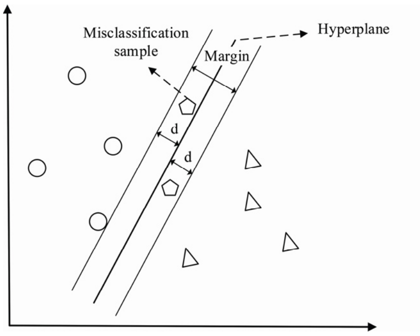


Fig. 4. The classification of sample in SVM

RF used to remove the data which can be misclassified after predetection of large scale samples. The specific process is organized as followed. After the original data set *T* is input into the training model, the Bootstrap sampling technology is used to extract data samples from *T*, and then send them to the basic classifier. The main function of the basic classifier is to vote the data to determine the final classification result. The RF-SVM module calculates the $M_{\text{margin}}(x)$ for each data sample and selects the data sample with a higher $M_{\text{margin}}(x)$ to input the SVM model. Samples with smaller $M_{\text{margin}}(x)$ are excluded, because these data samples are closer to the hyperplane of SVM and are more likely to be misclassified [16]. The integration interval $M_{\text{margin}}(x)$ for the data samples can be expressed as equation (6):

$$M_{\text{m}argin}(x) = \frac{v_{c_1} - v_{c_1}}{\text{X}} = \frac{\max\limits_{\text{c}=1,\ldots,\text{M}}(v_c) - \max\limits_{\text{c}=1,\ldots,\text{M}\cap\text{c}\neq\text{c}_1}(v_c)}{\text{X}}. \tag{6}$$

c_1 represents the class with the highest number of data *x* scored votes, v_{c_1} is the number of votes obtained for the class of c_1, c_2 represents the second ranked class with the number of data *x* scored votes, v_{c_2} is the number of votes obtained for the c_2 class, X represents the number of base classifiers, M represents the lowest $M_{\text{margin}}(x)$ ranked times that were fed into the SVM model, and also represents the total number of samples that entered the SVM. From the analysis above, it follows that when *x* is a fixed constant, then the smaller the difference of v_{c_1} from v_{c_2}, the smaller the $M_{\text{margin}}(x)$ value of data *x*.

In conclusion, combining with the advantages of SVM and RF can complete double-layer high-precision detection for DDoS attacks in SDN. The workflow of RF-SVM model is shown in Fig. 5.

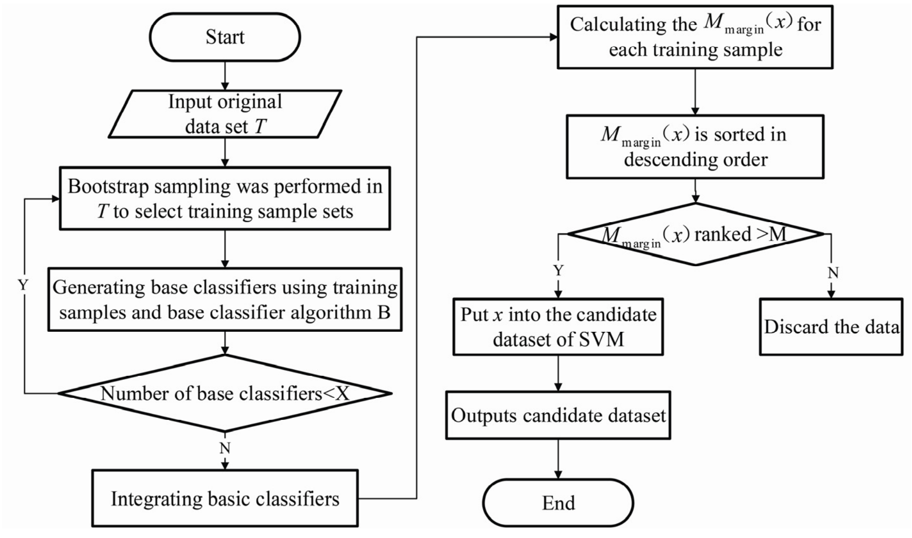

Start
Input original data set *T*
Bootstrap sampling was performed in *T* to select training sample sets
Generating base classifiers using training samples and base classifier algorithm B
Number of base classifiers<X
Y
N
Integrating basic classifiers
Calculating the $M_{\text{margin}}(x)$ for each training sample
$M_{\text{margin}}(x)$ is sorted in descending order
$M_{\text{margin}}(x)$ ranked >M
Y
N
Put *x* into the candidate dataset of SVM
Discard the data
Outputs candidate dataset
End

Fig. 5. The workflow of RF-SVM model

3.3 Incremental Learning Sample Screening Module

As a dynamic machine learning algorithm, the IL combines supervised learning and unsupervised learning. With the continuous expansion of data, new data is learned according to the known supervised learning training results. Compared with other algorithms, IL does not need iterative training of large-scale data when it receives new data, but slices and detects the new data. The sliced data will be trained by IL in turn until all data are learned.

The workflow of Incremental Learning screening is shown in Fig. 6. After inputting the original data set T to the RF-SVM model for training, the detection results of RF-SVM(+) and RF-SVM(-) can be obtained. When DDoS attacks continually send malicious traffic to SDN, the data samples that need to be detected by the model will continue to increase. In order to reduce the repeated iterative training of old data, and the amount of data involved in training in the new data, we take several data processing precoders. Firstly, by comparing the actual category and the support vector similarity of the sample, the incremental data set U is filtered. The specific process is organized as followed, each sample in data set U needs the calculation of both positive and negative support vector similarity, then sorted them in descending order. $S_+(x_i)$, $S_-(x_i)$ are the positive and negative support vector similarity of incremental sample x respectively, $S_+(x_i)$, $S_-(x_i)$ can be expressed as equation (7) and (8) respectively.

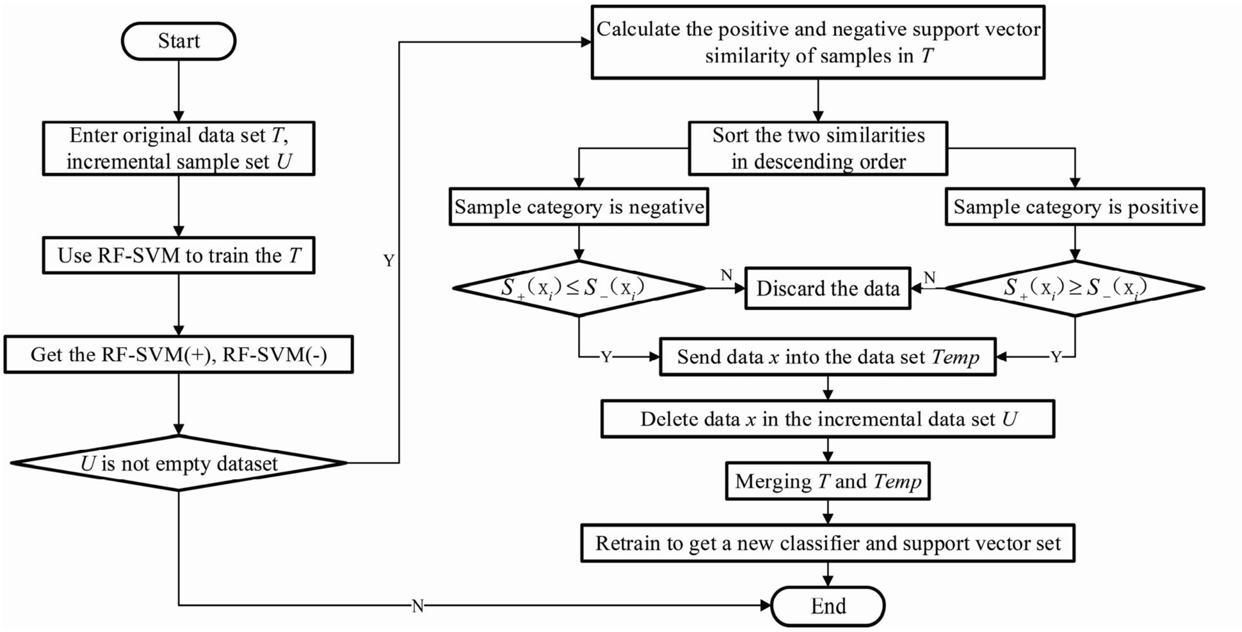

Start
Enter original data set T, incremental sample set U
Use RF-SVM to train the T
Get the RF-SVM(+), RF-SVM(-)
U is not empty dataset
Y
N
Calculate the positive and negative support vector similarity of samples in T
Sort the two similarities in descending order
Sample category is negative
Sample category is positive
$S_+(x_i) \le S_-(x_i)$
$S_+(x_i) \ge S_-(x_i)$
N
Discard the data
N
Y
Send data x into the data set $Temp$
Y
Delete data x in the incremental data set U
Merging T and $Temp$
Retrain to get a new classifier and support vector set
End

Fig. 6. The workflow of IL sample screening module

$$S_+(x_i) = \frac{l_+}{\sum_{j=1}^{l_+} d(x_i, y_j)} . \tag{7}$$

$$S_-(x_i) = \frac{l_-}{\sum_{j=1}^{l_-} d(x_i, y_j)} . \tag{8}$$

Where l_+ and l_- represent the number of positive and negative support vectors respectively, and $d(x,y)$ represents the distance between sample x and y in kernel space, $d(x,y)$ can be expressed as equation (9).

$$d(x,y) = \sqrt{K(x,x) + K(y,y) - 2K(x,y)} . \tag{9}$$

Considering the large number of eigenvectors, the Gaussian kernel function (RBF) is chosen to map the data to an infinite dimension. The RBF kernel function $K(x,y)$ can be expressed as equation (10).

$$K(x,y)=\exp(-g\|x-y\|^2)\,. \tag{10}$$

Take the actual attack data as an example. Assume actual category is negative. Also, positive similarity $S_+(x_i)$ is smaller than negative similarity $S_-(x_i)$. Then, it indicts that the sample has strong correlation with the outcome after the classification of the initial test [17]. That means, it will be selected into the new sample set *Temp* that participates in the training. Finally, the Temp is trained after merging with the original sample set *T*.

4 Experiments and Result Analysis

4.1 RF-SVM-IL Attack Detection System and Evaluation Indicators

This experiment is based on Ubuntu 16.04 operating system, using Mininet to simulate the real SDN environment. The Ryu controller is selected to control the whole network, the Hping3 is used to launch DDoS attack to the target host. The data packets are analyzed by Netsniff-ng to generate the corresponding data statistics report.

The framework of RF-SVM-IL attack detection system in SDN is shown in Fig. 7. Openvswitch (OVS) looks up the local flow table entry after receiving the information [18]. If there is no corresponding forwarding rule for malicious information in the flow table entry, the malicious information is sent to Ryu controller through Pack-in. The Ryu controller extracts the characteristics of the data and conducts attack detection. After then Ryu sends the detection results and corresponding forwarding rules to OVS through Packet-out [19]. At the same time, Ryu sends the detection result information to the attacked host, so that the attacked user can see the real-time detection result more clearly and intuitively.

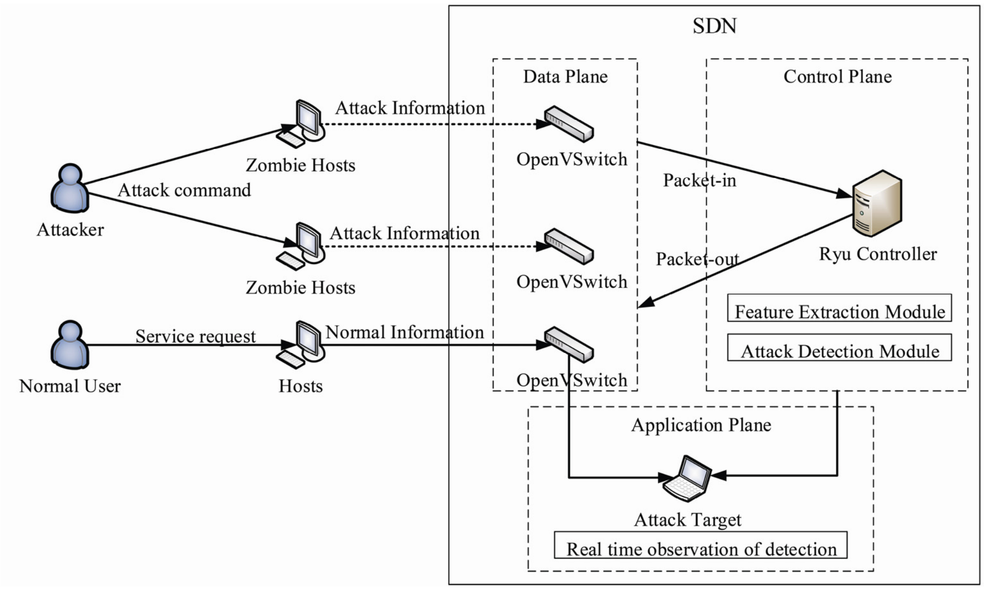
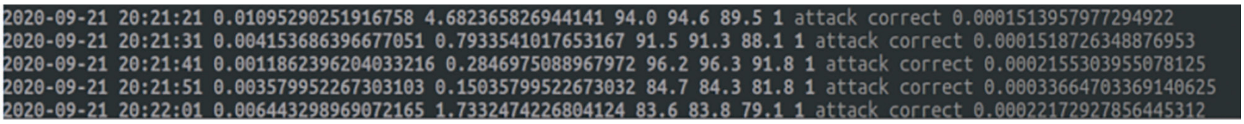


Fig. 7. The frame of attack detection system under SDN

In this system, two functional modules are put into the controller, which are feature extraction module and attack detection module. The feature extraction module extracts the following five features: the average number of packets, the average bits of packets, the port growth rate, the flow growth rate, and the source IP growth rate. The type of flow table is marked according to the feature extraction, the normal flow is marked “1” and abnormal flow marked “0”. In order to better observe the detection effect in real time, the following three indicators are visually displayed, including the type of flow table, whether the detection is correct and the detection time, as shown in the Fig. 8. The first real-time monitoring information is shown in Table 1.



```
2020-09-21 20:21:21 0.01095290251916758 4.682365826944141 94.0 94.6 89.5 1 attack correct 0.0001513957977294922
2020-09-21 20:21:31 0.004153686396677051 0.7933541017653167 91.5 91.3 88.1 1 attack correct 0.0001518726348876953
2020-09-21 20:21:41 0.0011862396204033216 0.2846975088967972 96.2 96.3 91.8 1 attack correct 0.0002155303955078125
2020-09-21 20:21:51 0.003579952267303103 0.15035799522673032 84.7 84.3 81.8 1 attack correct 0.00033664703369140625
2020-09-21 20:22:01 0.006443298969072165 1.7332474226804124 83.6 83.8 79.1 1 attack correct 0.00022172927856445312
```

Fig. 8. Result of real time attack detection

Table 1. Data characteristic index of real-time detection

Feature Indicators	Data Stream
Real time stamp	2020-09-21 20:21:31
Average packet number of the flow	0.0109529025191675
Average bit of the flow packet	4.682365826944141
Growth rate of port (bit/s)	94.0
Growth rate of the flow (bit/s)	94.6
Growth rate of the source IP (bit/s)	89.5
The type of flow	1
Traffic type of model checking	Attack
Detect correctness	Correct
Detection time (s)	0.0001513957977294

To better illustrate the performance of RF-SVM-IL model, detection accuracy Acc_{detect} and detection time T_{detect} are used as evaluation indexes.

Acc_{detect} represents the percentage of correct detection of the input data by the model and can be expressed as equation (11):

$$Acc_{\text{detect}} = \frac{T_p + T_N}{T_p + T_N + F_p + F_N} . \tag{11}$$

Where T_p, F_p represent the number of correctly classified, incorrectly classified samples for the DDoS attack data. T_N, F_N represent the number of correctly classified, incorrectly classified samples for the normal data, respectively.

T_{detect} represents the total time used by the model to detect an attack and can be expressed as equation (12):

$$T_{\text{detect}} = t_1 - t_0 . \tag{12}$$

Where t_0 represents the moment at which detection began and t_1 represents the moment at which detection ended.

In order to compare the performance of KNN-SVM-IL and RF-SVM-IL model. The precision rate P, the recall rate R and the false alarm rate F are be used. The P represents the percentage of normal data among all data that are correctly classified, the R represents the proportion of normal data among all samples judged to be normal, the F represents the proportion of the miscalled normal data among all the miscalled data samples. P, R and F can be expressed as equation (13), (14) and (15) respectively.

$$P = \frac{T_N}{T_P + T_N} , \tag{13}$$

$$R = \frac{T_N}{T_N + F_P} , \tag{14}$$

$$F = \frac{F_N}{F_N + F_P} . \tag{15}$$

4.2 System Performance Evaluation

This experiment uses a real collection data set which contains normal traffic and DDoS attacks. Among them, Hping3 simulates three types of DDoS attacks, namely the UDP, ICMP and SNP attacks. To enable the data set to reflect the authenticity performance of RF-SVM-IL model, normal traffic and attack traffic for the same time were collected at a rate of 10s/slip. To better demonstrate the better performance of the RF-SVM-IL proposed in this paper, the experimental steps in this section are as follows. The first step compares the detection effect of a single model (RF, KNN, SVM and CNN). The second step integrate RF and KNN with algorithms with SVM, respectively. Analyze and compare the detection effect of KNN-SVM to RF-SVM. On the basis of the above above, the incremental learning algorithm is added. The detection effects of KNN-SVM-IL and the two models were analyzed and compared with RF-SVM-IL.

Combined with the research content of this paper, single machine learning model selects the KNN, RF and CNN. Table 2 represents the average of the above three models in the Acc_{detect}, T_{detect} and F. As k increases from 0.1 to 0.9, the comparison of detection accuracy of the above three models is shown in Fig. 9. Under the same conditions, the comparison of F of the above three models is shown in the Fig. 10. The k is the proportion of test data in the total data, k can be expressed as equation (16):

$$k = \frac{D_1}{D_0 + D_1}. \tag{16}$$

Where D_0 represents the amount of training data, D_1 represents the amount of test data. *Incre* represents the proportion of the incremental data amount among the total data amount and can be expressed as equation (17):

$$Incre = \frac{I_1}{D_0 + D_1}. \tag{17}$$

Where I_1 represents the number of test samples that are subsequently added to the model.

Table 2. The average of the singel machine learning model in Acc_{detect}, T_{detect} and F

Name of Model	RF	KNN	CNN
Average of Acc_{detect}	0.9540	0.9203	0.9844
Average of T_{detect}	0.4093	0.5170	48.3167
Average of F	0.0444	0.0554	0.0135

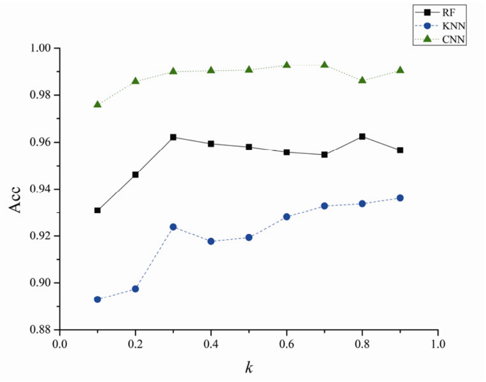


Fig. 9. Acc_{detect} comparison of the single model

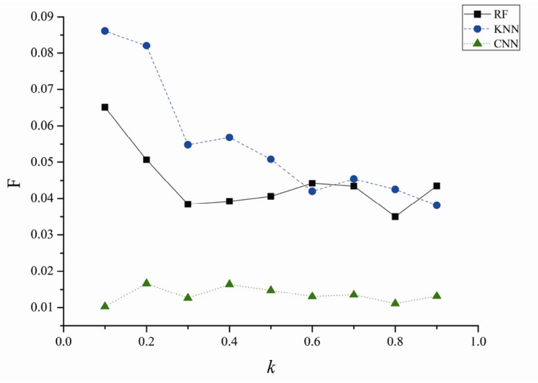


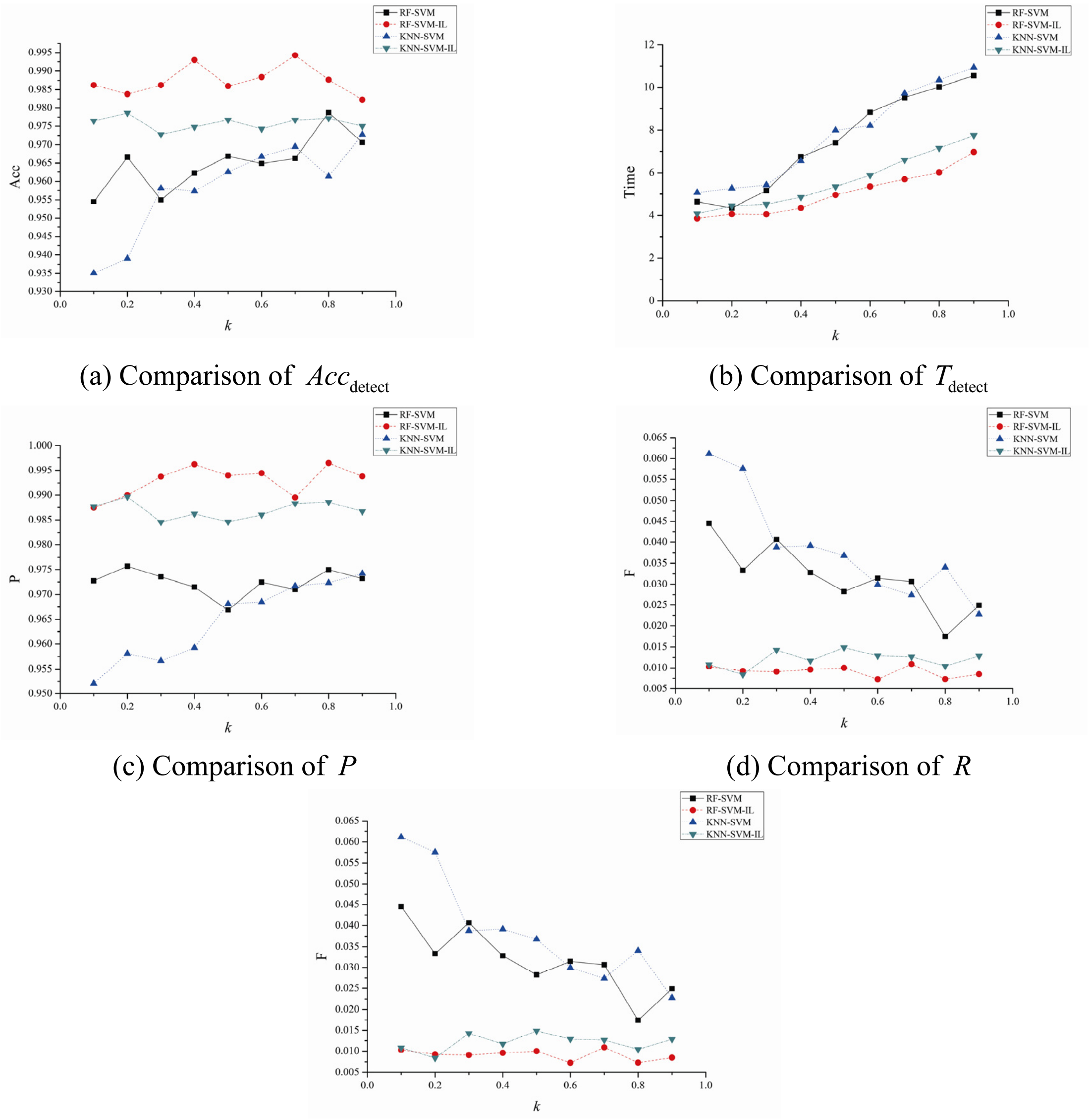
Fig. 10. F comparison of the single model

It can be seen from Table 2 that the detection time of CNN is several tens of times higher than that of other single models, so the curve image is not used to express the contrast effect of the detection time. It can be analyzed from Table 2 that CNN has the highest detection accuracy, which is 0.0451 higher than RF and 0.0698 higher than KNN. However, the average detection time of CNN is 118 times and 93 times higher than that of RF and KNN, respectively.

The hybrid models analyzed in this paper include RF-SVM, KNN-SVM, RF-SVM-IL and KNN-SVM-IL. As *k* increases from 0.1 to 0.9, the incremental sample *Incre* remains to 0.1, the average of the five detection metrics is shown in Table 3, the comparison of the above four hybrid models in the five detection indexes are shown in Fig. 11. The comparison results of detection accuracy can be analyzed from Table 3. RF-SVM is 0.007 higher than KNN-SVM. After adding incremental learning algorithm, RF-SVM-IL is 0.023 higher than RF-SVM, KNN-SVM-IL is 0.018 higher than KNN-SVM, and RF-SVM-IL is 0.012 higher than KNN-SVM-IL. The comparison results of detection time can be analyzed from Table 3. RF-SVM is 0.115s higher than KNN-SVM. After adding incremental learning algorithm, RF-SVM-IL is 0.172s less than RF-SVM, KNN-SVM-IL is 0.077s higher than KNN-SVM, and RF-SVM-IL is 0.133s lower than KNN-SVM-IL.

Table 3. The average of five detection metrics in hybrid models

Name of Model	RF-SVM	KNN-SVM	RF-SVM-IL	KNN-SVM-IL
Average of ACC_{detect}	0.9651	0.9580	0.9875	0.9758
Average of T_{detect}	7.4706	7.7232	5.0365	5.6297
Average of P	0.9724	0.9646	0.9928	0.9869
Average of R	0.9683	0.9634	0.9907	0.9851
Average of F	0.0316	0.0386	0.0091	0.0121



(a) Comparison of Acc_{detect}

(b) Comparison of T_{detect}

(c) Comparison of P

(d) Comparison of R

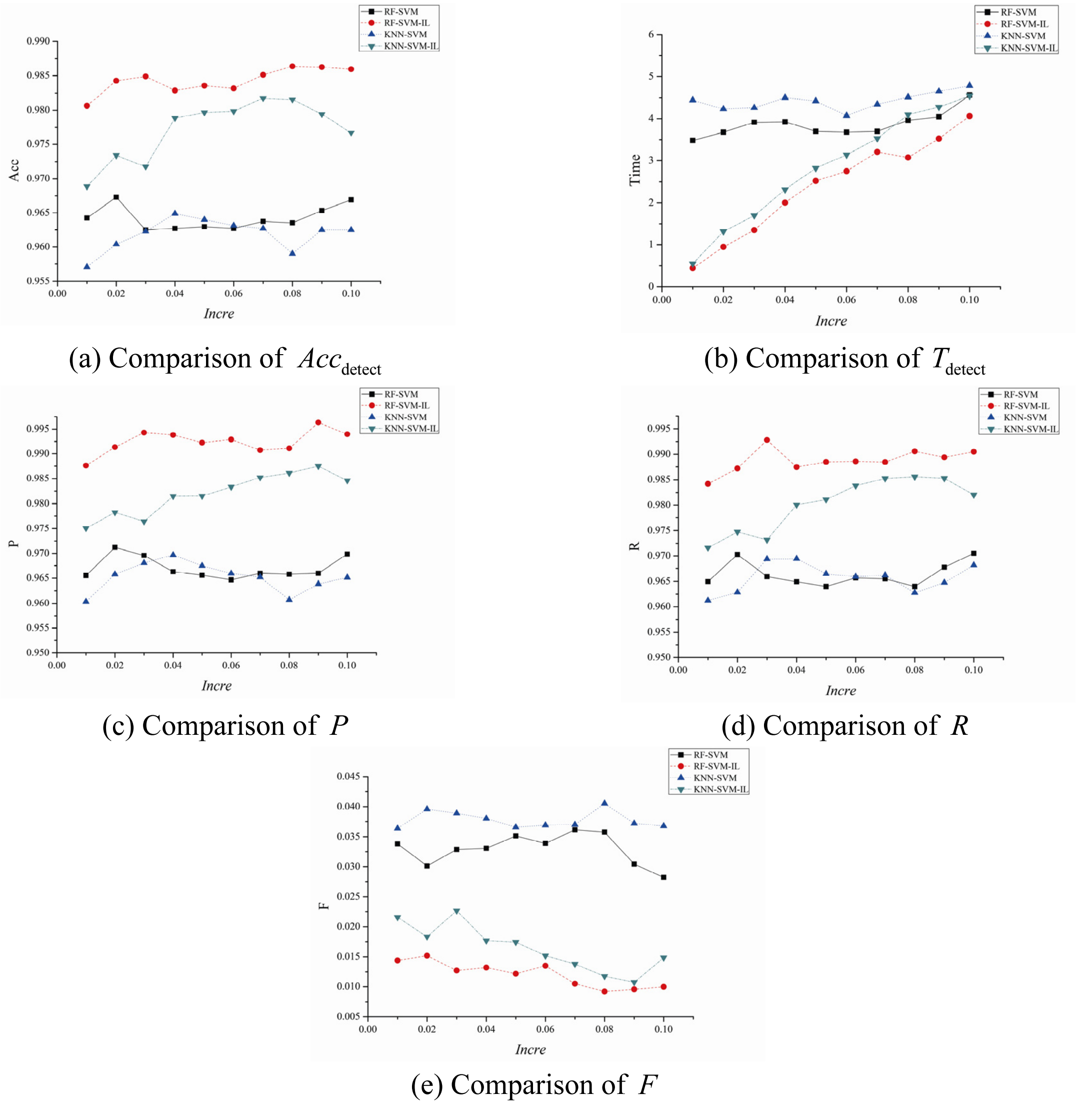
(e) Comparison of F

Fig. 11. The comparison of the above four hybrid models in the five detection indexes (*Incre* = 0.1)

In the above experiments, *Incre* remained unchanged, and the detection performance of the four models was compared by changing the *k*. In fact, once the attack is launched, the amount of data entering the system is increasing, so it is necessary to deal with the dynamic increase of data better. The following experiment keeps $k = 0.5$ unchanged, *Incre* increases from 0.1 to 1.0. Fig. 12 shows the comparison of the four hybrid models in five detection indexes, and the average values of above detection indexes are listed in Table 4. The conclusions drawn from the analysis in Table 4 are as follows. In terms of detection accuracy, RF-SVM is 0.007 higher compared to KNN-SVM, after adding incremental learning algorithm, RF-SVM-IL is 0.023 higher than RF-SVM, KNN-SVM-IL is 0.018 higher compared to KNN-SVM, and RF-SVM-IL is 0.012 higher than KNN-SVM-IL. In terms of detection time, RF-SVM is 0.115s higher than KNN-SVM. After adding incremental learning algorithm, RF-SVM-IL is 0.172s less than RF-SVM, KNN-SVM-IL is 0.077s higher than KNN-SVM, and RF-SVM-IL is 0.133s lower than KNN-SVM-IL. In terms of other detection indexes, RF-SVM-IL models have the highest *P* and *R* and the lowest *F*. It can be seen that the RF-SVM-IL model proposed in this paper has good detection results.

Table 4. Comparison of detection accuracy and average detection time of four models

Name of Model	RF-SVM	KNN-SVM	RF-SVM-IL	KNN-SVM-IL
Average of ACC_{detect}	0.9642	0.9649	0.9854	0.9772
Average of T_{detect}	3.8605	4.4239	2.3866	2.8276
Average of P	0.9670	0.9652	0.9924	0.9819
Average of R	0.9664	0.9657	0.9888	0.9803
Average of F	0.0330	0.0378	0.01205	0.0164



(a) Comparison of Acc_{detect}

(b) Comparison of T_{detect}

(c) Comparison of P

(d) Comparison of R

(e) Comparison of F

Fig. 12. The comparison of the above four hybrid models in the five detection indexes ($k = 0.5$)

5 Conclusion

Aiming at the problem of DDoS attack detection in SDN network, this paper proposes a hybrid machine learning model of RF-SVM-IL, which combines RF, SVM and IL. The main purpose is to improve the detection accuracy, shorten the detection time and realize accurate and efficient dynamic detection. In order to improve the detection accuracy, this paper combines RF and SVM for double-layer detection, and uses RF to remove the samples that are easy to be misclassified. On the basis of the above, the IL algorithm is added to filter the increasing input samples to reduce the amount of data processed by the model, so that the model can quickly detect in the face of massive attack traffic. This paper simulates the real SDN environment and puts the RF-SVM-IL module into the controller for testing. The experimental results show that, compared with single machine learning model (RF, KNN, CNN) and four hybrid models, RF-SVM-IL model has lower detection time and higher detection accuracy, and has better effect in P, R and F.

Future research can start from the following aspects: after accurately detecting attacks, load balancing technology can be used to quickly alleviate the damage caused by attack traffic to SDN, and DDoS attacks can be eliminated by attack cleaning technology. The defense module is imported into the controller to achieve the overall defense of SDN network.

Acknowledgements

We would like to express our appreciation to the support by Gansu plateau transportation information engineering and control key laboratory open project (No. 20161106), and Gansu radio monitoring and positioning innovation team (2017C-09).

Reference

[1] Cisco Annual Internet Report (2018-2023) White Paper. <https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html>, 2020 (accessed 20.05.15).

[2] X. Peng, Z.-P. Liu, W. Li, Software defined network distributed control channel construction protocol, Journal of Chinese Computer Systems 39(04)(2018) 763-768.

[3] Q. Yan, F. Yu, Distributed denial of service attacks in software-defined networking with cloud computing, IEEE Communications Magazine 53(04)(2015) 52-59.

[4] M.M. Oo, S. Kamolphiwong, T. Kamolphiwong, V. Sangsuree, Advanced Support Vector Machine-(ASVM-) based detection for Distributed Denial of Service (DDoS) attack on Software Defined Networking (SDN), Journal of Com-puter Networks and Communications 3(4)(2019) 01-12.

[5] M.-Y. Li, The DDoS attack detection method based on Random Forest, [dissertation] Haikou: University Of Hainan, 2019.

[6] Y.-P. Kim, D.-H. Choi, T.-P. Van, M.-T. Long, DDoS Detection System Based on Multiple Machine Learning Combination for Software Defined Networking, Journal of Korean Institute of Communications and Information Sciences 42(8)(2017) 1581-1590.

[7] C.-H. Li, Y. Wu, Z.-Z. Qian, W.-J. Wang, DDoS attack detection and defense based on a deep learning hybrid model under SDN, Journal of Communications 39(7)(2018) 176-187.

[8] M. Aamir, S.M.A. Zaidi, Clustering based Semi-Supervised Machine Learning for DDoS Attack Classification, Journal of King Saud University Computer & Information Sciences 33(4)(2021) 436-446.

[9] R. Gong, Research on Load Balancing and DDoS Attack Detection Technology Based on SDN, [dissertation] Hefei: University Of Anhui, 2016.

[10] L. Guo, S. Boukir, Fast data selection for SVM training using ensemble margin, Pattern Recognition Letters 51(2015) 112-119.

[11] W. Wei, Y.-B. Dong, D.-M. Lu, DDoS Defense based on Support Vector Machine and Multi-resource Maximum and Minimum Equity, Journal of Zhejiang University (Engineering Science) 44(2)(2010) 265-270.

[12] A. Papana, E. Siggiridou, Shortcomings of Transfer Entropy and Partial Transfer Entropy: Extending Them to Escape the Curse of Dimensionality, International Journal of Bifurcation and Chaos 30(16)(2020) 205-250.

[13] M. Shi, Research on adaptive defense technology of DDoS attack based on SDN, [dissertation] Qingdao: Qingdao University of science and technology, 2017.

[14] Q. Xiao, K.-Y. Su, Detection of botnet traffic based on Random Forest, Microelectronics & computers 36(3)(2019) 43-47.

[15] L. Demidova, Y. Sokolova, A novel SVM-kNN technique for data classification, in: Proc. 2017 6th Mediterranean Conference on Embedded Computing (MECO), 2017.

[16] Y. Hou, Research and application of feature extraction and ensemble learning algorithm, [dissertation] Beijing: Beijing University of science and technology, 2015.

[17] C.P. Diehl, G. Cauwenberghs, SVM incremental learning, adaptation and optimization, in: Proceedings of the International Joint Conference on Neural Networks, 2003.

[18] OpenvSwitch. <http://openvswitch.org>, 2011 (accessed 19.06.03).

[19] C.-H. Li, W. Yan, X.-Y. Yuan, Z.-J. Sun, Detection and defense of DDoS attack based on deep learning in OpenFlow based SDN, International Journal of Communication Systems 31(5)(2018) e3497.1.