Machine learning approach detects DDoS attacks

Nguyen Thi Khanh Tram, Doan Trung Son, Nguyen Thi Thu Huong, Tran Thi Thu

***Abstract*— Denial of Service attacks have been around since the dawn of the internet age. Along with the development and explosion of the Internet, denial of service attacks are also increasingly powerful and become a serious threat in cyberspace. The article aims to evaluate machine learning algorithms: K-nearest neighbor (KNN) algorithm, Decision Tree, Random Forest algorithm and Support Vector Machine (SVM) on various metrics in detecting DDoS attacks. The main objective of the paper is to analyze the algorithms, collect data and evaluate the effectiveness of the algorithms in DDoS attack detection.**

***Tóm tắt*— Tấn công từ chối dịch vụ đã xuất hiện từ những năm khởi nguyên của thời đại internet. Song hành cùng sự phát triển và bùng nổ của mạng Internet, tấn công từ chối dịch vụ cũng ngày càng mạnh mẽ và trở thành mối đe dọa nghiêm trọng trên không gian mạng. Bài báo hướng tới đánh giá các thuật toán học máy: Thuật toán K láng giềng gần nhất (K-nearest neighbor - KNN), cây quyết định (Decision Tree), thuật toán rừng ngẫu nhiên (Random Forest) và máy vector hỗ trợ (Support Vector Machine - SVM) trên các chỉ số đánh giá khác nhau trong việc phát hiện các cuộc tấn công DDoS. Mục tiêu chính của bài báo nhằm phân tích các thuật toán, thu thập đánh giá dữ liệu và tiến hành so sánh hiệu quả các thuật toán vào phát hiện tấn công DDoS.**

Keywords—DDoSt;KNN; Decision Tree; Random Forest; SVM.

Từ khóa— DDoS; KNN; Cây quyết định; Rừng ngẫu nhiên; SVM.

I. INTRODUCTION

Distributed Denial of Service (DDoS) attack is accomplished by increasing online traffic from multiple sources to the server. This causes the server to run out of resources and bandwidth. DDoS first appeared in 1999.

Vietnam is facing a great risk of being attacked and distributed by DDoS attacks with the 6th position globally after China, the US, France, Russia and Brazil, the 2nd position in the region. Asia Pacific region and leading in Southeast Asia [1].

DDoS involves making requests from a network of computers made up of millions of computers with different IP addresses over which control has been previously established (Botnet). Computers and other networked resources such as IoT devices together create "Tsunamis" of traffic. A DDoS attack can be understood as a sudden traffic jam that blocks a highway, preventing normal traffic from reaching its destination. Because it is dispersed into many access points with different IP ranges, DDoS is much stronger than DoS, and it is often difficult to recognize or prevent DDoS attacks.

Different types of DDoS attacks target different components of a network connection. Based on the target and behavior, DDoS attacks can be classified into three types traffic/fragmentation attacks, bandwidth/volume attacks, and application layer attacks.

In late 1999, CERT first published its report on the threat of DDoS attacks and outlined specific prevention actions to mitigate this threat [2]. A few months later, the Internet suffered its first large-scale DDoS attack [3], and successive attacks of increasingly large scale in the following years. Since then, researchers have analyzed a number of tools used to launch DDoS attacks [4, 5, 6], measured their impact on the Internet, and come up with a number of defense methods [7]. Accordingly, these research efforts have resulted in a number of effective and reliable anti-DDoS products offered as stand-alone devices or cloud-based services.

In recent years, along with the strong development of Artificial Intelligence (AI), machine learning (ML) and deep learning methods are being used more and more in detecting DDoS attacks. Sambadi and Gondi propose an approach that uses multiple linear regression to detect DDoS attacks [8].

P. Sangkatsanee et al. [9] built a real-time detection mechanism applying machine learning techniques. In it, 12 essential network traffic characteristics are proposed, which distinguish between normal data and DDoS.

Sofi et al. [10] upgraded a new dataset consisting of 27 features and five different traffic classes. Four machine learning algorithms namely Naive Bayes, SVM, decision tree and MLP have been applied to identify DDoS attacks. In which, the MLP algorithm gives the best results.

Mahadev et al [11] used the Naive Bayes classifier in the weka tool to analyze the network traffic flow and found it to provide 99% accuracy in detecting DDoS attacks.

S Duque et al. [12] show that the K-means clustering algorithm gives increased efficiency with the correct usage of the number of clusters. Furthermore, note that with an increase in the number of clusters over the number of data types, the false-negative, detection rate decreases, but the false-positive rate increases.

II. MACHINE LEARNING ALGORITHM

The four algorithms for performing DDoS attack detection in this paper refer to KNN, Decision Tree, Random Forest and SVM. These are all commonly used classical machine learning algorithms.

A. KNN

The K-nearest neighbor (KNN) algorithm is one of the simplest supervised learning algorithms (which is effective in some cases) in machine learning. When training, this algorithm does not learn anything from the training data, all calculations are performed when it needs to predict the outcome of the new data. With KNN, in the classification problem, the label of a new data point is directly inferred from the K nearest data points in the training set using distance measures such as Euclidean distance, Manhattan distance and Minkowski distance.

Implementation steps:

Step 1. Calculate the distance

Step 2. Find nearest neighbors

Step 3: Predict labels.

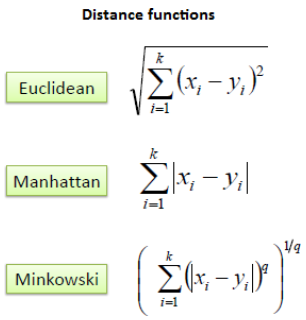

Figure 1. Distance formula in KNN

B. DECISION TREE

Decision Tree - is a supervised and non-parametric learning algorithm used for classification and regression. The methods create a highly accurate, stable, and easy-to-follow tree model, eliminating unnecessary attributes. Each inner node is equivalent to a variable, each arc goes to a child node corresponding to the possible value of that variable. The leaves correspond to the predicted target values for the variables.

Decision tree learning is also a very popular method in data mining. Where a decision tree describes a tree structure in which leaves represent classes and branches represent combinations of features that lead to classification. A tree can be learned by dividing the source set into subsets based on the values of the test attributes. This process is repeated on each obtained subset. The recursion ends when it cannot be divided any further or when each element of the subset has been labeled. Decision trees are described by calculating conditional probabilities. Decision trees can be described as a combination of techniques learning and computational algorithms that support the description, classification, and generalization of a given data set.

C. RANDOM FOREST

Random Forest builds many decision trees using the Decision Tree algorithm, but each decision tree will be different (with a random element). The prediction results are then aggregated from the decision trees. Random forest is a supervised family algorithm that can solve both regression and classification problems. Random Forest works in 4 steps:

Step 1. Select random samples from the given data set.

Step 2. Set up a decision tree for each sample and get prediction result from each.

Step 3. Vote for each prediction results.

Step 4. Select the most predicted result as the final prediction.

In addition, Random Forest has the following notable characteristics:

- A collection of unrelated trees performing the same task is better than having each tree count one by one;
- Assuming the trees are independent of each other in error rate or have little correlation with each other to ensure independence;
- Feature selection must be good enough for the tree to classify better than random selection;
- The predictability and error of each tree have little correlation with each other.

D. SVM

Support vector machine (SVM) is a supervised machine learning algorithm that is very commonly used today in classification or regression problems.

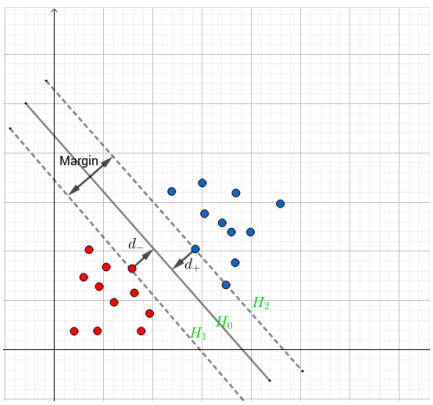

Figure 2. Hyperplane selection model in SVM

SVM was proposed by Vladimir N. Vapnik and his colleagues in 1963 in Russia and then became popular in the 90s thanks to its application to solving non-linear problems (Nonlinear).

The idea of SVM is to find a hyperplane (to separate the data points. This hyperplane will divide the space into different domains, and each domain will contain a type of data.

The optimal hyperplane we need to choose is the split hyperplane with the largest margin. Machine learning theory has shown that such a hyperplane minimizes the error limit.

III. DATA PROCESSING AND PARAMETER IMPLEMENTATION

A. DATA SET

The authors have collected the data set based on the available documents under the link []. The dataset has been pre-processed and labeled, the authors download and use it according to the model's requirements. The new dataset collected by the authors contains four types of DDoS attacks as follows: (HTTP Flood, SIDDOS, UDP Flood) and no redundant or duplicate records. Table 1 lists the log counts for these types of attacks. Table 2 shows the processed features of the data set.

TABLE I. NUMBER RECORD OF DATA SET BY ATTACK TYPE

Attack Type	**Number of records**
SIDDOS	6550
UDP Flood	201344
HTTP Flood	4110

TABLE II. PROCESSED CHARACTERISTICS OF THE DATASET

STT	***Description***	***Type***
1	SRC ADD	Continuous unit
2	DES ADD	Continuous unit
3	PKT ID	Continuous unit
4	FROM NODE	Continuous unit
5	TO NODE	Continuous unit
6	PKT TYPE	Continuous unit
7	PKT SIZE	Continuous unit
8	FLAGS	Symbolic unit
9	FID	Continuous unit
10	SEQ NUMBER	Continuous unit
11	NUMBER OF PKT	Continuous unit
12	NUMBER OF BYTE	Continuous unit
13	NODE NAME FROM	Symbolic unit
14	NODE NAME TO	Symbolic unit

15	PKT IN	Continuous unit
16	PKTOUT	Continuous unit
17	PKTR	Continuous unit
18	PKT DELAY NODE	Continuous unit
19	PKTRATE	Continuous unit
20	BYTE RATE	Continuous unit
21	PKT AVG SIZE	Continuous unit
22	UTILIZATION	Continuous unit
23	PKT DELAY	Continuous unit
24	PKT SEND TIME	Continuous unit
25	PKT RESEVED TIME	Continuous unit
26	FIRST PKT SENT	Continuous unit
27	LAST PKT RESEVED	Continuous unit

The proposed data collection system follows these steps:

- Collect and control: all network traffic from NIDS is collected and examined;
- Preprocessing data format: remove redundant and duplicate records;
- Feature extraction: extract feature parameters from the collected network traffic and assign each feature to each data column; they will be used as a vector in the new dataset;
- Statistical measurements: in this step, the features are additionally calculated using statistical equations.

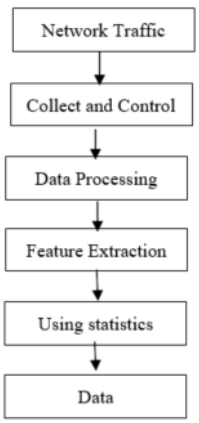

Figure 3. The process of building a new dataset

The authors use a data collection system inherited from the topic "Building an application to collect transmission data for network investigation" of author Nguyen The Hoang.

After collecting the data set, it is fed into the system to identify denial of service attacks. The steps are as follows training model's weight and the model's accuracy evaluation parameters.

- Receiving input dataset: the system receives user-provided network attack datasets;
- Machine learning model training: the system stores machine learning algorithms commonly used in network attack detection, then trains those algorithms with the input dataset;
- Changing model parameters: the system makes adjustments to change some parameters with each certain algorithm to increase the accuracy of the algorithm;
- Display training and model evaluation results: the system will output the results of the training model's weight and the model's accuracy evaluation parameters;
- Make a conclusion whether the network behavior is a denial of service attack or not.

B. Data Processing

Figures with the above dataset, process the data before putting it into the experiment. The input information must be processed at the same cost. Therefore, data cleaning is always the first step in designing a machine learning model. Remove the symbolic features (Symbolic) such as PKT_TYPE, FLAGS, NODE_NAME_FROM, NODE_NAME_TO, PKT_CLASS and unimportant features like SRC_ADD, DES_ADD.

Because the data set has a relatively high number of records belonging to normal behavior, to balance the machine learning model, take 10000 records for 2 labels Normal and UDP Flood. The input data set is divided into training and testing sets in the ratio of 7:3.

C. Hyperparameter selection

Hyperparameter Tuning is an important step in machine learning techniques. Hyperparameters are user-defined parameters that control the training process of the model and play an important role in determining the performance of the model. Such

parameter tuning is usually done by traversing a predefined grid of parameters. This parameter grid can be defined values, or it can also be random following a definite distribution or condition. In this paper, the parameter grid with defined values is used as shown in the following table:

TABLE III. HYPERPARAMETRIC GRID

Algorithm	Parameter name	Value
KNN	Neighbors number	[10, 100, 1000]
DT	Evaluation function	Gini impurity or Information gain (Entropy)
RF	Number of Tree	[10, 100, 1000]
SVM	C	[-1, 1, 3]
	γ	[-1, 1, 3]

```
HYPER_GRID = {
    'K-Nearest Neighbors': {"n_neighbors": [10, 100, 1000]},
    'Decision Tree': {"criterion": ["gini", "entropy"]},
    'Random Forest': {"n_estimators": [10, 100, 1000]},
    'SVM': {"C": np.logspace(-1, 1, 3), "gamma": np.logspace(-1, 1, 3)},
```

Figure 4. Hyperparameter selection

D. RESULT EVALUATION INDEX

The indicators used to evaluate the results include:

Accuracy: this is the ratio of correctly predicted points to the total number of points in the test dataset.

Precision or Positive predictive value (PPV): Is the ratio of the number of points in the attack behavior that the model correctly predicts to the total number of points the model predicts in the attack behavior. The higher the Precision metric, the higher the number of points the model predicts that an attack is an attack. Precision = 1, i.e. all scores that the model predicts as an attack are correct, or none of the scores labeled as normal behavior that the model mistakenly predicts is an attack.

Recall: The ratio of the number of points that are correctly predicted by the model attack to the total number of points that are actually the attack (or the total number of points labeled as the original attack). The higher the recall, the lower the score is that the attack is missed. Recall = 1, i.e. all points labeled as attack behavior are recognized by the model. Recall is also known as True Positive rate (TPR), Sensitivity, Hit rate.

F1-score: Is the harmonic mean between Precision and Recall when these two quantities are non-zero. Calculated by the formula:

$$F_1 = 2\frac{precision * recall}{precision + recall}$$

False positive rate is (FPR) also known as False Alarm Rate is false detection rate, a behavior is normal but the model considers it as attack behavior.

IV. RESULTS AND DISCUSSION

The results of running the 4 mentioned algorithms are presented in the following table:

TABLE IV. RESULTS WHEN RUNNING 4 ALGORITHMS

Algor ithms	Accuracy	Precision	Recall	F1-score	FPR
KNN	0.9475	0.9541	0.9495	0.9494	0.0003
DT	0.9093	0.9093	0.9093	0.9093	0.0902
RF	0.9508	0.9440	0.9412	0.9411	0.0194
SVM	0.9489	0.95 43	0.9497	0.9496	0.0000

According to the results from Table 4, the decision tree algorithm gives the lowest probability of correct detection (90.93%) as well as the highest false detection rate, the Random Forest algorithm gives the highest probability (95.08%), the algorithm gives the highest probability (95.08%). SVM with longest running time, lowest false detection rate. In general, the 4 algorithms using scikit-learn library provide

relatively good results and are optimized for better performance.

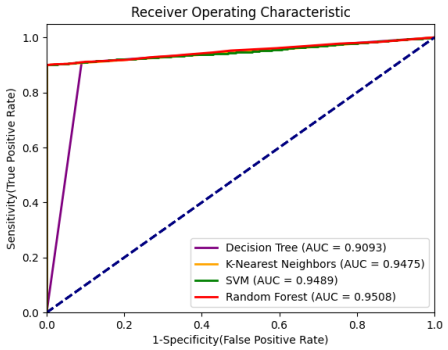

Figure 5. ROC curves of 4 algorithms

Besides, the experimental results are also evaluated based on the ROC (Receiver Operating Characteristic) curve, which is a graphical chart illustrating the performance of the binary classification system. Each point on the ROC curve is the coordinate corresponding to the true positive frequency (sensitivity) on the vertical axis and the false positive frequency (1- specificity) on the horizontal axis. Performance line the more you deviate to the top and to the left, the clearer the distinction between the two states. The ROC curve when running 4 algorithms is recorded in Figure 5. The AUC (Area under the ROC Curve) values of the decision tree algorithms, Random Forest, KNN, SVM are 0.9093, 0.9508, 0.9475, 0.9489, respectively. These are all values in the excellent threshold, where the decision tree algorithm gives the lowest result and the Random Forest algorithm gives the best prediction.

VI. CONCLUSION

Based on the newly collected dataset containing four types of DDoS attacks as follows: (HTTP Flood, SIDDOS, UDP Flood) and no redundant or duplicate records, the author conducted experiments with 4 machine learning algorithms. for DDoS attack detection. As a result, all 4 algorithms are capable of detecting DDoS attacks with high accuracy, fast speed and efficiency.

Recently, with the continuous development of 5G, a large number of insecure Internet of Things (IoT) devices are connected to the Internet, which presents great challenges to protect against attacks. DDoS attacks, especially when attackers are trying to "recruit" more devices to the Botnet (Example Mirai Botnet) to increase the frequency, size and throughput of DDoS attacks worldwide. In the future, attackers will most likely take advantage of artificial intelligence and machine knowledge that allows automatic alteration of attacks so that they evolve to more optimal attack techniques. In that case, it is necessary to improve the DDoS attack detection algorithms towards real-time processing of the raw data of the attacks obtained.

ACKNOWNLEDGMENT

The author would like to thank VINIF for their financial support to Nguyen Thi Khanh Tram as master student in VNU University of Engineering and Technology.

REFERENCES

[1]. Hội thảo "Bảo vệ mạng và dữ liệu khỏi các cuộc tấn công từ chối dịch vụ (DDoS) nhằm vào các tổ chức, doanh nghiệp" - ngày 3-5-2019, Cục An toàn Thông tin, Báo VietnamNet, tổ chức Nexusguard Limited tổ chức.

[2]. CERT Coordination Center, "Results of the Distributed-systems Intruder Tools Workshop", năm 1999. Software Engineering Institute.

[3]. L. Garber, Denial-of-Service Attacks Rip the Internet", IEEE Computer, 33(4):12–17, 2000.

[4]. D. Dittrich, "The DoS Project's "trinoo" Distributed Denial of Service Attack Tool", 21 tháng 10 năm 1999.

[5]. D. Dittrich, "The "stacheldraht" distributed denial of service attack tool", https://staff.washington.edu/dittrich/misc/stacheldraht.analysis/, 31 tháng 12 năm 1999.

[6]. D. Dittrich, "The Tribe Flood Network" Distributed Denial of Service Attack Tool"- https://staff.washington.edu/dittrich/misc/tfn.analy sis/, 1999.

[7]. D. Kumar, G. Rao, M. K. Singh, and G. Satyanarayana, "A Survey of Defense Mechanisms countering DDoS Attacks in the Network", Intl. Journal of Advanced Research in Computer and Communication Engineering, 2:2599–2606, tháng 7 năm 2013.

[8]. Swathi Sambangi và Lakshmeeswari Gondi, "A Machine Learning Approach for DDoS (Distributed Denial of Service) Attack Detection Using Multiple Linear Regression" trong hội thảo quốc tế INTER- ENG 2020 Interdisciplinarity in Engineering lần thứ 14 tại Mures, Romania, 08/9/2020.

[9]. P Sangkatsanee, N Wattanapongsakorn and C Charnsripinyo, "Practical real-time intrusion detection using machine learning approaches", ELSEVIER Computer Communications 34(2011) 2227-2235.

[10]. I Sofi, A Mahajan and V Mansotra, "Machine Leaming Techniques used for the Detection and Analysis of Modem Types of DDoS Attacks", International Research

Journal of Engineering and Technology (IRJET), Tập:04, tháng 06/2007.

[11] Mahadev, V Kumar and H Sharma, "Detection and Analysis of DDoS Attack at Application Layer Using Naive Bayes Classifier", Intemational Journal of Computer Engineering & Technology (IJCET), tập 9, 2018, pp. 208-217, Article IICET_09_03_025.

[12]. S Duque, M Nizam bin Omar, "Using Data Mining Algorithms for developing a Model for Intrusion Detection System (IDS)", ELSEVIER Procedia Computer Science 61 (2015) 46-51.

ABOUT THE AUTHOR

Doan Trung Son

Workplace: Faculty of Information Security, People's Security Academy.

Email: *son.doantrung@gmail.com*

Education:

University: Faculty of Information Technology - People's Security Academy

Master: Faculty of Information Technology, Hanoi University of Science and Technology

Doctorate: Hagen University, Germany

Recent research direction: Cybersecurity, High-tech Crime Prevention, Artificial Intelligence, Data Science, Trust and Distributed Systems, Modern issues in information technology.

Nguyen Thi Khanh Tram

Workplace: Phenikaa School

Email: khanhtramt2k23@gmail.com

Education:

University: Faculty of Information Technology - People's Security Academy

Master student: University of Engineering and Technology, Hanoi, Vietnam

Tran Thi Thu

Workplace: Hanoi Law University

Email: thutran@hlu.edu.vn

Education:

University: Information Technology, University of Natural Sciences, Vietnam National University, Ho Chi Minh City

Master: Business Economics, University of Toulouse.