Application of Ensemble Learning in DDoS Detection

Yun Feng[1, *], Fu Kun[2]

[1]Jiangnan University Department of Internet of Things Engineering, Wuxi Jiangsu, 214122, China

[2]Baotou Jiuyuan Comprehensive Law Enforcement Bureau of Urban Management, Baotou, Inner Mongolia Autonomous Region, 014000, China

*Corresponding author

Keywords: Internet of Things; Anomaly Detection; DDoS; Machine Learning; Feature Engineering; Ensemble Learning

Abstract: A growing number of Internet of Things (IoT) devices are appearing on the Internet. Yet these devices are facing more and more security risks, such as Distributed denial of service (DDoS) attack. In recent years, the scale of DDoS attacks is getting larger and larger, and each DDoS attack will cause huge losses. This motivates the development of new techniques to automatically detect attack traffic and improve detection accuracy as much as possible, so as to reduce losses. This paper demonstrates several ensemble learning methods which show a higher accuracy DDoS detection in IoT network traffic than using a single machine learning method. Experimental results show that ensemble learning can complement the limitations of using a single machine learning method to improve the DDoS detection accuracy.

1. Introduction

As the progress of science and technology, more and more IoT devices appear in our daily life. IoT devices which can connect with the Internet greatly facilitate our lives. However, these IoT devices are at risk of being victims of network attacks. Distributed denial of service (DDoS) attack is one of the most common and challenging problems on the Internet. Since DDoS traffic is very similar to normal traffic, the DDoS attack is really difficult to detect.

There are several DDoS attack events in 2016.

On April, Squad launched DDoS attacks on Blizzard's Battle Net servers, including StarCraft II, World of Warcraft, Diablo 3, and other important games that were an offline outage, which prevented players from landing[1].

On September, Krebson Security, a security research institute, was attacked by Mirai, which was considered one of the biggest cyber attacks ever. Soon, however, OVH, the French host service provider, was attacked twice, mainly by Mirai. When Krebson Security was attacked, the traffic reached 665 GB, while when OVH was attacked, the total traffic exceeded 1 TB[1].

On October, Dyn DNS, which provides dynamic DNS services, was attacked by a large-scale DDoS attack, which mainly affected its services in the Eastern United States. The attack resulted in access problems for many websites using Dyn DNS services, including GitHub, Twitter, Airbnb, Reddit, Freshbooks, Heroku, SoundCloud, Spotify and Shopify. Attacks have paralyzed these sites for a time, and Twitter has even had nearly 24 hours of access[1].

Faced with the growing threat of DDoS attack, some scholars have given some solutions to recognize the anomaly traffic. For example, machine learning is used to identify attack traffic. Although some of these methods have shown some good results (Like k-Nearest Neighbor[2], Support Vector Machine[3], Naïve Bayes[4], Decision Tree[4]). Identifying anomaly traffic in a separate method still has a lower accuracy because of distribution of data. With the development of ensemble learning technology, which combine multiple machine learning algorithms, ensemble learning technology can often achieve significantly better generalization performance than a single machine learning algorithm.

So, my goal is to apply the ensemble learning into recognition of DDoS attack. Using multiple machine learning algorithms can compensate for the limitations of a single machine learning algorithm to improve accuracy. If the accuracy of identifying DDoS can be improved, it will be easier to defend against DDoS, thereby reducing losses and making devices safer. For example, for the classification problem (linear separable problem), a simple linear classifier can be used to achieve good classification results. But if the distribution of data is not linearly separable, using simple linear classifier can not achieve good classification results. To solve this problem, the more commonly used method is ensemble learning. Improving accuracy can reduce losses. I will build a network, simulate DDoS attacks, collect traffic, preprocess data, use some ensemble learning methods, and compare the advantages and disadvantages.

2. Materials and Methods

IoT devices used to implement DDoS attacks are becoming more and more intense. Because there are a very large number of security vulnerabilities in IoT devices in the network. So attackers can distribute malicious programs over the Internet to other computers. In this way, attackers build an army of infected computers which is called Botnet to perform the DDoS attack. Then the computers of this Botnet can attack a web server simultaneously. When this happens, it depletes the server's system resources (like CPU, memory and network bandwidth).

In order to prevent generating a real Botnet. I will set up a small network with only a few IoT devices, which can carry out a normal simulation experiment. There are a DDoS attack source, a DDOS victim, and three IOT devices in this small network. And I connect them through a router. The experiment is to let the attack source infect three IoT devices into botnets via routers. Then, they implement DDoS attack on victims together. My goal is to collect normal traffic and attack traffic through routers within a specified period of time.

2.1 Traffic Collection

For normal (non-DoS) traffic collection, I connected three IoT devices to each other through a router to form a small LAN network. The three IoT devices interacted with each other for 10 minutes. Then, I recorded pcap files and all packets sent during that time period.

For DoS traffic collection, a Kali Linux virtual machine is used as the DoS source and an Apache Web Server as the DoS victim. Both devices are connected via Wi-Fi to the router. In this small LAN network, I simulated the three most common classes of DoS attack (a TCP SYN flood, a UDP flood, and an HTTP GET flood). Each DoS attack targeted the victim’s IP for almost 1.5 minutes once.

To make it appear as if the IoT devices simultaneously produced normal traffic and conducted DoS attacks, I combined the DoS traffic with the normal traffic, spoofing source IP addresses, MAC addresses, and packet send times. Each of the three DoS attack classes appeared to be executed once within a 10-minute interval on Each of the three IoT devices. The attacks occurred in a random order for a random duration ranging uniformly from 90 to 110 seconds each.

2.2 Feature Engineering

In this section, I will demonstrate the features which are different between normal traffic and DoS attack traffic.

1) Packet Size

The packet sizes are different between normal traffic and DoS attack traffic. A vast majority of attack packets are under 100 bytes. Because attackers want to use smallest size of the packets to exhaust as much the victim server’s resources as possible. On the contrary, normal packets are usually more than 100 bytes.

2) Inter-packet Interval

Most packets are sent at regular interval with a certain time interval between packets. This means IoT network pings or other automated network activities. In comparison, inter-packet intervals (ΔT) and high first and second derivatives of inter-packet intervals of DoS attack traffic are closed to zero.

3) Protocol

The protocol compositions of normal and DoS attack traffic are different. In attack traffic, TCP packets are more than three times as much as UDP packets. In contrast, in normal traffic, UDP is nearly four times as much as TCP in attack traffic. Moreover, normal traffic contains more protocol totally.

4) Bandwidth

Network traffic was divided by source device and the average bandwidth (over a 10-second time intervals) was calculated to measure the instantaneous bandwidth associated with each device.

5) IP Destination Address

Another key characteristic of IoT device traffic is that the set of destination IP addresses rarely changes over time. According to the literature written by Doshi, Apthorpe and Feamster[5], two features were crafted to reflect this behavior. First, a count of distinct destination IP addresses within a 10-second window; more endpoints may indicate attack traffic. Second, we calculate the change in the number of distinct destination IP addresses between time windows; new endpoints might suggest that the device is conducting an attack.

2.3 Ensemble learning

It is limited to simply use a single machine learning method to identify DDoS attack. Because a single machine learning method is difficult to fit different kinds of dataset. Thus, my idea is to combine a variety of machine learning methods to reduce the limitation of a single classifier. Thereby the generalization ability of the overall classifier is improved. In this section, I mainly use 5 types of ensemble learning methods.

1) Forests of Randomized Trees
Random Forests
Extremely Randomized Trees
2) Bagging
Basic Classifier: Decision Tree
3) Boosting
AdaBoost
Basic Classifier: Decision Tree
4) Voting Classifier
Majority Class Labels (Majority/Hard Voting)
Basic Classifier: Random Forest Classifier, Extra Trees Classifier
Weight: 1:1
Weighted Average Probabilities (Soft Voting)
Basic Classifier:
K-Neighbors, Random Forest, Extra Trees
Weight: 1:5:5
5) Stacking
Basic Classifier:
Extra Trees, Random Forest, K-Neighbors
Meta_classifier:
Logistic Regression

I selected several single machine learning algorithms to recognize traffic separately. Then I used the above machine learning algorithms as the basic classifier of ensemble learning methods and compare the accuracy of individual machine learning algorithms and ensemble learning algorithms.

3. Results and Discussions

I tested several ensemble learning algorithms to recognize normal traffic and attack traffic. Sklearn Python library, xgboost Python library and mlxtend Python library are used to implement these ensemble learning models. I used a training set with 85% of the combined normal and DoS traffic to train each classifier and calculated separately the classification accuracy of single machine learning methods and the classification accuracy of ensemble learning methods.

The following are the experimental results:

3.1 Single machine learning methods (Table I):

Gaussian Naïve Bayes (GNB)
Bernoulli Naïve Bayes (BNB)
Logistic Regression (LR)
Linear Support Vector Machine (LSVM)
Decision Tree (DT)
K-Neighbors (KN)

Table I: machine learning methods

	GNB	BNB	LR	LSVM	DT	KN
Normal Precision	0.76495	0.99513	0.99289	0.99232	0.99652	0.99897
Attack Precision	0.96698	0.98859	0.98976	0.99083	0.99952	0.99941
Normal Recall	0.52375	0.83722	0.85466	0.86997	0.99326	0.99163
Attack Recall	0.98859	0.99971	0.99956	0.99952	0.99975	0.99993
Normal F1	0.62178	0.90937	0.91860	0.92713	0.99489	0.99529
Attack F1	0.97766	0.99412	0.99464	0.99516	0.99964	0.99967
Accuracy	0.95782	0.98895	0.98994	0.99092	0.99932	0.99938

3.2 Ensemble learning methods (Table II):

Extra Trees (ET)
Random Forest (RF)
Bagging (BG)
AdaBoost (AB)
Majority Class Labels (Majority/Hard Voting) (MCL)
Weighted Average Probabilities (Soft Voting) (WAP)
Stacking (STK)

Table II: Ensemble learning methods

	ET	RF	BG	AB
Normal Precision	0.99959	0.99876	0.99958	0.99917
Attack Precision	0.99983	0.99983	0.99980	0.99980
Normal Recall	0.99752	0.99752	0.99710	0.99710
Attack Recall	0.99997	0.99991	0.99997	0.99994
Normal F1	0.99855	0.99814	0.99834	0.99813
Attack F1	0.99990	0.99987	0.99988	0.99987
Accuracy	0.99981	0.99976	0.99978	0.99976
	MCL	WAP	STK	
Normal Precision	0.99855	0.99938	0.99958	
Attack Precision	0.99992	0.99983	0.99981	
Normal Recall	0.99896	0.99752	0.99731	
Attack Recall	0.99990	0.99996	0.99997	
Normal F1	0.99876	0.99845	0.99845	
Attack F1	0.99991	0.99989	0.99989	
Accuracy	0.99984	0.99980	0.99980	

From the above experimental results, we can see Extra Trees, Random Forest, AdaBoost, Bagging, Majority Class Labels, Weighted Average Probabilities, Stacking have better results than the best machine learning method.

The experiment also shows that the accuracy of combining multiple machine learning methods is higher than that of using a single machine learning method.

For further work, it is very meaningful to use more suiTable machine learning or ensemble learning algorithms as basic classifiers to improve accuracy. In addition, preventing over-fitting is also very important. Another direction is to adjust the parameters of the ensemble learning algorithms, such as the weight of each basic classifier. Make the ratio of each basic classifier is the best to improve accuracy.

About this experiment, only three IoT devices were used, so the type of device is limited. Further research can try to apply the ensemble learning method to many different types of IoT devices on the Internet. Meanwhile, a larger dataset also can be tried to improve the accuracy of recognition.

4. Conclusion

In this work, I applied some ensemble learning algorithms to improve DDoS detection accuracy. The limited features can train the machine learning models and ensemble learning models with the dataset. Then I compared the detection accuracy of the two methods. These ensemble learning methods can distinguish more accurately than using a single machine learning algorithm, which proves it is more accurate to combine machine learning algorithms than to use them alone. Improving accuracy by the ensemble learning method also facilitates further researches, such as DDoS defense. In this way, higher detection accuracy can reduce the losses caused by DDoS Attacks as much as possible.

Acknowledgment

Thanks to the professor Nick Feamster for his academic support on this project.

References

[1] Available: http://m.sohu.com/a/120983171_416377

[2] L. Ertoz, E. Eilertson, A. Lazarevic, P.-N. Tan, V. Kumar, J. Srivastava, and P. Dokas, “Minds minnesota intrusion detection system,“ In Data Mining: Next Generation Challenges and Future Directions, 2004.

[3] E. Eskin, W. Lee, and W. Stolfo, “Modeling system call for intrusion detection using dynamic window sizes,“ 2001.

[4] Singh, K. J. , & De, T. . (2015). An Approach of DDOS Attack Detection Using Classifiers. Emerging Research in Computing, Information, Communicationnd Applications. Springer India.

[5] Doshi, R. , Apthorpe, N. , & Feamster, N. . (2018). [ieee 2018 ieee security and privacy workshops (spw) - san francisco, ca, usa (2018.5.24-2018.5.24)] 2018 ieee security and privacy workshops (spw) - machine learning ddos detection for consumer internet of things devices. 29-35.