IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 17, Issue 3, Ver. 1 (May – Jun. 2015), PP 76-81
www.iosrjournals.org

Internet Worm Classification and Detection using Data Mining Techniques

Dipali Kharche[1], Anuradha Thakare[2]
1 (Department of Computer Engg., PCCOE, Savitribai Phule Pune University, Pune India)
2 (Department of Computer Engg., PCCOE, Savitribai Phule Pune University, Pune India)

Abstract: *Internet worm means separate malware computer programs that repeated itself and in order to spread one computer to another computer. Malware includes computer viruses, worms, root kits, key loggers, Trojan horse, and dialers, adware, malicious, spyware, rogue security software and other malicious programs. It is programmed by attackers to interrupt computer process, gatherDelicate Information, or gain entry to private computer systems. We need to detect a worm on the internet, because it may create network vulnerabilities and also it will reduce the system performance. We can detect the various types of Internet worm the worm like, Port scan worm, Udp worm, http worm, User to Root Worm and Remote to Local Worm. In existing process it is not easy to detect the worm, there is difficult to detect the worm process. In our proposed systems, internet worm is a critical threat in computer networks. Internet worm is fast spreading and self propagating. We need to detect the worm and classify the worm using data mining algorithms. For use data mining, machine learning algorithm like Random Forest, Decision Tree, Bayesian Network we can effectively classify the worm in internet.*

Keywords: *Bayesian Network, Classification,Data Mining, Decision Tree, Random Forest, Worm Detection.*

I. Introduction

Internet worm is a critical threat in computer networks. Internet worm is self propagating, and fast scattering. The internet worm [1] was released for the first time and more over hundred hosts were infected. After that the threat of internet worm has been increasing and causing more harm to network systems. Many research methods for internet worm detection have been projected. Most of internet worm detection is based on intrusion detection system (IDS) [2]. Automatic detection is challenging because it is tough to predict what form the next worm will take so, an automatic response and detection is becoming an imperative because a afresh released worm can infect lots of hosts in a substance of seconds. Internet worm based IDS can be divided into twocategories. That are network-based and host-based. The network-based internet worm detection reflects network packets before they spread to an end-host, whereas the host-based internet worm detection reflects network packets that already spread to the end-host. Moreover, the host-based detection studiesencoded network packets so that the stroke of the internet worm may be struck. When we focus on the network packet without encoding, we must studythe performances of traffic in the network. Numerous different types of machine learning techniques were used in the field of intrusion detection in general and worm detection. Data Mining has an important role and is essential in worm detection systems, which using different data mining techniques to build several models have been proposed to detect worms.

In this paper, we provide a new method for network-based internet worm detection. We preprocess the network packet data by mining a certain number of features of abnormal/normal traffic data and use three different data mining algorithms for data classification. Our model can detect internet worm with a detection ratenear to 99.6%, and false alarm is nearly zero.

The paper is structured as follows. In section II, provide related methods of internet worm detection. In section III, present details of irregularbehavior/patterns in the network traffic data. In section IV & V, present related study and our proposed model, respectively. In section VI & VII, experimental results and conclusion.

II. Related Work

Several recent researches in the few last years were proposed "Worms Detection" are based on data mining as an efficient ways to increase the security of networks. Classification techniques were the best for many recent researches.

Some data mining algorithms are operative to classify behaviors of internet worms. For example, internet worms by mining their features [6] from cleaned/infected platform. They made a data mining model and train it with

these performances and set up results of internet worm detection with greater overall accuracy and low false-positive rate.

Amethod [3] using association behavior to detect the internet worm. They considered the change of normal connections and worm connections. The worm connections were predictable to have a high number of failed connections. Moreover, the failure networks can be occurred when a source IP sends a request linking a packet to an unused IP address or some ports that no longer in service. After that, SYN/ACK packet, ICMP packet, and TCP RESET will be returned. So the amount of these packets will be high [4].

Anew method of internet worm detection[5] that categorizedalarm in source-destination ports that worms use for scattering themselves. They use K-L divergence to identify features of abnormal actions and use Support Vector Machine (SVM) to organize these actions. They obtain good results with a 90% detection rate for all endpoints and with false-alarm rate nearby zero.

We emphasis on a idea of network-based internet worm detection. We preprocess fresh network packets before it influences to an end user and consider association of source-destination IP addresses, association of source-destination ports and number of some abnormal packets that occur when some users produce internet worm traffic. Here, we use three different kinds of data mining algorithms that are Bayesian Network, Decision Tree and Random Forest to classify data into worm, normal data or network attack data (i.e., DOS and Port Scan).

III. Attack And Worm Characteristics

In this paper, we consider Blaster worm, which is one type of the public worms. Most worms have performances similar to those of the Port Scan and Denial of Service (DoS) attacks. Thus, our method is to classify and detect the Blaster worm, Port Scan and DoS attack performances.We consider UDP flood and HTTP flood in a DoS attack. Particulars about data type are presented below.

• Blaster worm activities a buffer overflow susceptibility of the DCOM RPC on Windows platforms by spreading to ports 135 and 4444 on TCP protocol and port 69 on UDP protocol. This worm can transfer and operate by itself. After that, the worm creates DoS attacks to escape patching update by makinga SYN flood to port 80.

• UDP flood is a sort of DoS attack. This attack will refer a lot of UDP packets to any target operators or a network system. This performance will consume more bandwidth.

• HTTP flood is a kind of DoS attack as well. This attack is as analogous as the UDP flood. The HTTP flood will send a lot of unusable packets to any target operators to consume high bandwidth on Web Server.

• Port Scan is a procedure to scan for accessible port or service that runs on any ports from any users.

IV. Classification Algorithms

4.1 C4.5 Decision Tree [8]:

It is famous data mining algorithm that classifies data set by using numerous nodes of the tree. It forms a tree by using a divide-and-conquer procedure. A Decision tree is approached with over-fitting on large datasets. The classification model of Decision tree is created by mining rules from the training set. These rules are used to calculate and classify a new or anonymous dataset called a testing set. The Decision tree will discover asolution class by starting at the root and crossing to a leafnode. The result of prediction and classification can be found in a leaf node. Moreover C4.5 Decision tree is an algorithm that is well-known and has an efficiency in classification.

4.2 Random Forest [9]:

It is an operational data mining algorithm since it can fix problem of over-fitting on large dataset and can train/test rapidly on large and complex data set. A tree is constructed using random data from a training dataset through replacement; major of these datasets is used for training, and the remaining of dataset is used for testing or result assessment. This model can calculate important features used in classification and un-pruned rules that are formed and estimated by the training dataset. There are many classification trees included in Random Forest model. Each classification tree is exclusive and is voted for a class. Finally, an solution class is assigned constructed on the maximum vote.

4.3 Bayesian network [10]:

It is a graphical model and a probabilistic model. A Bayesian network uses numerous nodes or positions that have probabilistic relation with each other. The Bayesian network studiesunexpected relation from the training dataset to classify or predict unknown cases. Moreover, it can avoid over-fitting with large data.

4.4 Information Gain:

Itis a proposition of feature selection. Information Gain computes for an entropy cost of each attribute. An entropy cost can be called as a rank. Rank of each feature represents its importance or association with an solution class that is used to recognize the data. So a feature with comparatively high rank will be one of the most important features for classification.

V. Proposed Model

5.1 Overview

Our worm detection model divides into preprocessing and classification part as shown in Figure 1.In the preprocessing, we insert the actual Blaster worm, obtained from a consistent online source, into a local area network (LAN). At the same time, we also produce UDP flood, HTTP flood and Port Scan attacks into a LAN (local area network).

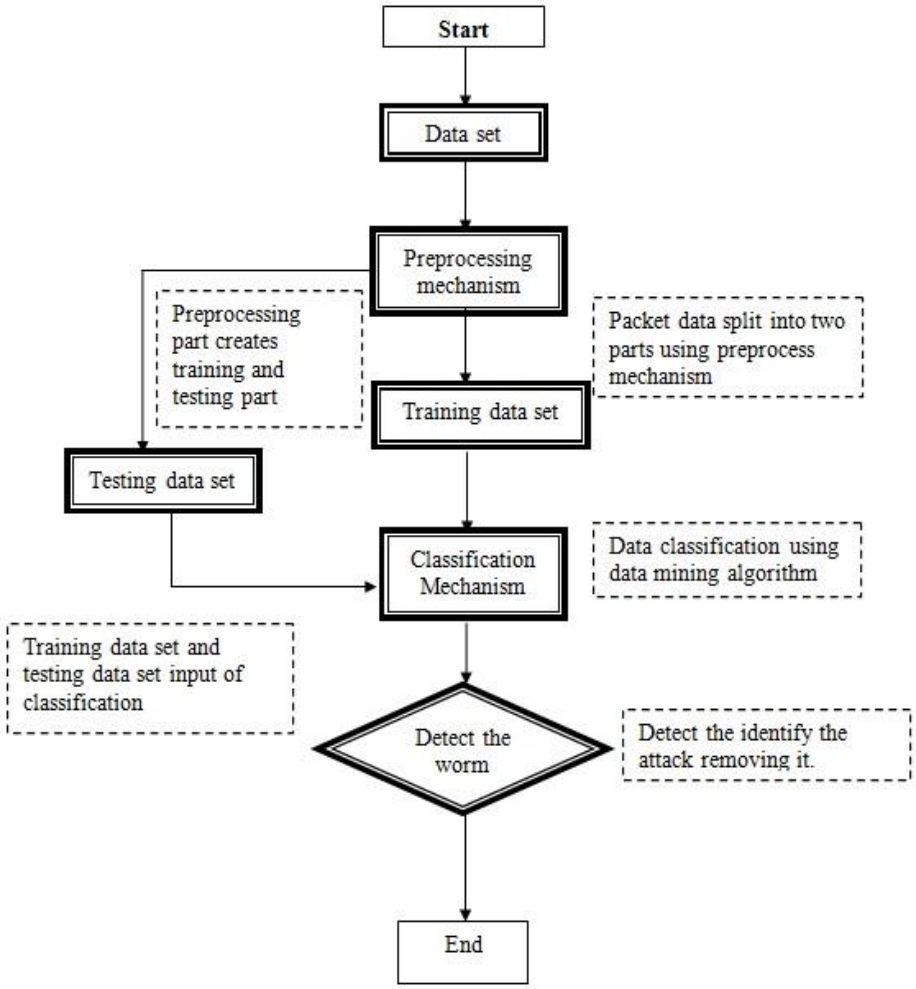

Fig. 1. Worm Detection Model

Here,snort raw network packets from the Local Area Network andchoose only some features from the packet header of all raw packet performances that is major and necessary to predict or classify the data. The preprocessing and feature selection technique will be shown in details in Section B. After the preprocessing part, separate the obtained datasets into two parts; one for training and the other one for testing. In the classification part,

using data mining algorithms to classify the features of Worm, Http flood, UDP flood, Port Scan and Normal network behavior. These will be discussed in more detail in Section C.

5.2 Preprocessing Part

Each source IP address togetherat one second is one record. Moreover, each record has 13 features that mine from entire packets in 1 second. Detail of the features is shown below.

• Number of individually source IP address in 1 second
• Numeral of destination IP address
• Number of TCP header packet
• Number of ICMP header packet
• Number of UDP header packet
• Number of SYN (Synchronization) flag (bit 1)
• Number of ACK (Acknowledgement) flag (bit 1)
• Number of RST flag (bit 1)
• Total of source port
• Total of destination port
• Number of difference packet size
• Port ratio is the number of source port separated by number of destination port
• SYN ratio is the number of SYN flag bit 1 shared by number of destination IP

In Preprocessing is the major task in data mining. After preprocessing the data we can split the data into two set one is training set and another one is testing set. We can perform the preprocessing in the worm detection dataset. And the importing the dataset, then perform preprocessing.In preprocessing part, we can extract the training test based on the source IP address collected at 1 second is one record Moreover, each record has 13 features that extract from all packets in 1 second.Finally, the preprocessing part creates a training dataset and testing dataset. The testing dataset has half size of the training set.

5.3 Classification Part

In this part, first we train the data mining techniques which are Random Forest, C4.5 Decision tree and Bayesian Network using the WEKA tool [7] with training dataset and then testing these techniques with a different testing data set. Here, test our models by classifying normal data, UDP flood, HTTP flood ,Blaster worm and Port Scan, using 13-features of preprocessed dataset.

VI. Experimental Evaluation

6.1 Parameter Evaluation

The performance of each classification model is compared and measured by using the detection rates, which are True Positive and False Alarm defined as follows:

- True Positive: a process classifies the input data correctly.
- False Alarm: a process misclassifies normal input data, and reports it as having anomalousperformance.

6.2 Experimental Results

For our experiment, our classification outcomes in terms of detection rate and false-alarm rate. Three different data mining techniques are considered and estimated one by one. From Table I, with our 13-feature input data, each of the techniques can classify normal internet data, UDP flood, internet worm, HTTP flood and Port Scan attacks with a detection rate over 97.8% data. In particular, the Decision tree ,Random Forest and Bayesian Network techniques give 99.4% , 99.6%and 97.8 detection rates, respectively. Additionally, Bayesian Network offers the lowest true-positive rate in worm detectionthat is 91.6%, while the UDP flood detection is perfect with

100% true-positive detection rate. From Table II, with our 13-feature input data, the Random Forest, Decision tree and Bayesian Network models can detect and classify internet worm giving false-alarm rates equal to 0.3% ,0.2% and 1.9%, respectively. Essentially, each of the techniques can classify network attacks which are HTTP flood, UDP flood and Port Scan attacks, giving the false-alarm rate equal to zero.

Table I. Detection Rate And True Positive

Model	Detection Rate (%)	True Positive				
		Normal (%)	Worm(%)	UDP Flood(%)	HTTP Flood(%)	Port Scan (%)
Bayesian Network	97.8	98.2	91.6	100.0	98.0	99.8
C4.5 Decision Tree	99.4	99.6	99.0	100.0	98.2	99.8
Random Forest	99.6	99.7	99.2	100.0	98.8	99.8

Table II. False Alaram Rate

Model	False Alarm			
	Worm(%)	UDP Flood (%)	HTTP Flood(%)	Port Scan (%)
Bayesian Network	1.9	0.0	0.0	0.0
C4.5 Decision Tree	0.2	0.0	0.0	0.0
Random Forest	0.3	0.0	0.0	0.0

From Table I and Table II the resulte of the classification techniques in term of worm detectionmodel true positive and false rate is shown in the Figure 2.

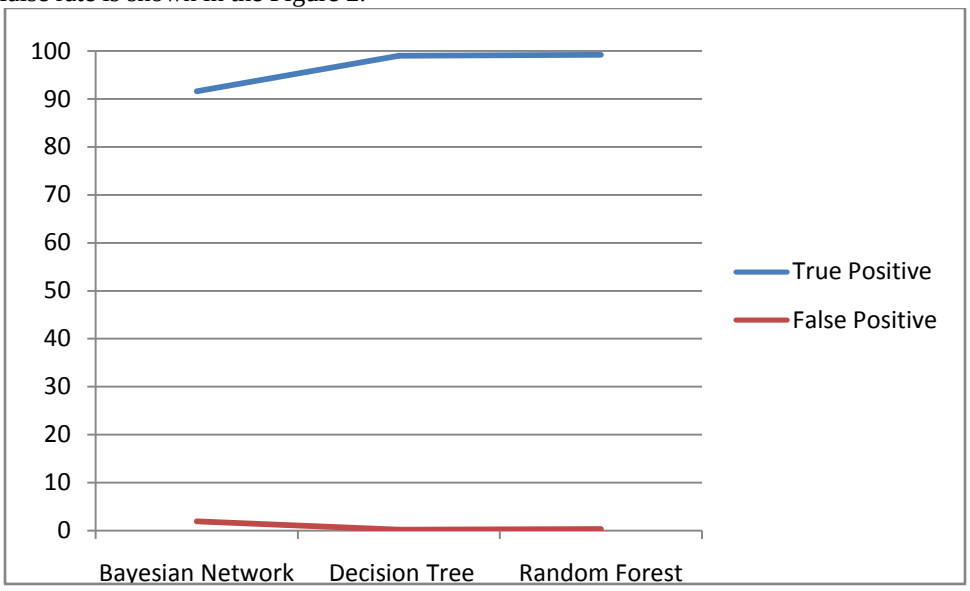

Fig.2. Performance of Worm Detection Model

VII. Conclusion

In this paper, our worm detection model consists of preprocessing and classification techniques. The propose model consist of a preprocessing method with 13 features mined from the network packets.

Three data mining algorithms which are Random Forest, Bayesian Network and Decision tree are measured to classify performances of Normal network data, UDP flood, Http flood, Blaster Worm and Port Scan. Most internet worms have performances similar to Port scan and DoS attack. So proposed model not only has efficiency to detect internet worms, but also can classify attack types such as HTTP flood, UDP flood and Port Scan with low false-alarm rate and high detection rate. Especially, Bayesian Network gives the percentage of internet worm classification less than 99% as 91.6% and percentage of false-alarm as 1.9% so that in practice, 1.9% of false-alarm rate is very high. However, we found that the Random Forest and the Decision Tree algorithms can detect internet worm and classify DOS and Port Scan attacks with a detection rate over 99% and false-alarm rate close to zero.

References

[1]. N. Weaver, V. Paxson, S. Staniford and R. Cunningham, "Taxonomy of computer worms," Proc of the ACM workshop on Rapid malcode, WORM03, 2003, pp. 11-18.

[2]. C. Smith, A. Matrawy, S. Chow and B. Abdelaziz, "Computer Worms: Architecture, Evasion Strategies, and Detection Mechanisms," J. of Information Assurance and Security, 2009, pp. 69-83.

[3]. M. M. Rasheed, N. M. Norwawi, O. Ghazali, M. M. Kadhum, "Intelligent Failure Connection Algorithm for Detecting Internet Worms", International Journal of Computer Science and Network Security, Vol. 9, No. 5, 2009, pp. 280-285.

[4]. D. R. Ellis, J. G. Aiken, K. S. Attwood, S. D.Tenaglia, "A Behavioral Approach to Worm Detection," Proceedings of the 2004 ACM workshop on Rapid malcode, 2004, pp. 43-53.
[5]. S. A. Khayam, H. Radha and D. Loguinov, "Worm Detection at Network Endpoints Using Information-Theoretic Traffic Perturbations", IEEE Inter Conf on Communications (ICC), 2008, pp. 1561-1565.
[6]. M. Siddiqui, M. C. Wang and J, Lee, "Detecting Internet Worms Using Data Mining Techniques", Cybernetics and Information Technologies, Systems and Applications: CITSA, 2008.
[7]. Weka 3.7.0 tools [Online], Available: www.cs.waikato.ac.nz/ml/weka/ [2009, July 2]
[8]. N. Pater, "Enhancing Random Forest Implementation in Weka", Machine Learning Conference Paper for ECE591Q, 2005
[9]. D. Heckerman, "A Tutorial on Learning with Bayesian Networks" Microsoft Research Advanced Technology Division Microsoft Corporation, 1996.
[10]. Classification via Trees in WEKA [online], Avilable : http:/maya.cs.depaul.edu/~classses/ect584/WEKA/classify.html
[11]. Tawfeeq S. Barhoom, Hanaa A. Qeshta," Adaptive Worm Detection Model Based on Multi classifiers" 978-0-7695-4984-2/13, 2013 Palestinian International Conference on Information and Communication Technology