International Journal of Engineering Research in Computer Science and Engineering (IJERCSE) Vol 2, Issue 9, September 2015

Botnet Identification System Using Clustering And Machine Learning C5.0

[1]Ankita Bhaiyya, [2]Miss. Sonali Bodkhe
[1]Student,[2] HOD
Department of Computer Science Engineering
G.H.Raisoni College Academy of Engineering and technology Nagpur, (INDIA)
[1]ankita.bhaiyya88@gmail.com[2] sonali.mahure@gmail.com

Abstract: **One of the most significant current issues in computer network security is BOTNET. It is an active focus of the research community and industry due to a sharp rise of attacks on individual and organizational computers. BOTNET is a large network of compromised computers used to attack other computer systems for malicious intent. Botnets are one of the most destructive threats to the cyber security. A botnet is a collection of compromised machines (bots) receiving and responding to commands from a server (the C&C server) that serves as a rendezvous mechanism for commands from a human controller.**

Recently, HTTP protocol is frequently utilized by botnets as the Command and Communication (C&C) protocol. In this work, we aim to detect HTTP-based botnet activity based on machine learning approach. To achieve this, botnet analysis system is implemented by employing two different machine learning algorithms, C5.0 and k means-bisecting algorithm. This Bisecting K-means algorithm is a clustering algorithm that give trained data by taking the desired iteration. The data obtained by the k-means algorithm is processed by a machine learning C5.0 algorithm. Then the probable botnets are identified using this algorithm. Thus botnet can be blocked from the system by using these two effective algorithms.

I. INTRODUCTION

Malicious software for short has acquired an important position in high-tech modern life. Starting from the earliest use of programmable systems, approaches to infecting them with software containing malicious functionality have existed but, in the past, malware often had just limited or local impact. The success of the Internet also became a starting point for reports about widespread malware infections affecting several million systems around the globe. Almost concurrently, remotely controlled networks of hijacked computers, so-called botnets, became popular. One critical observation is that the motivation for creating malware has changed dramatically over the last decade. No longer is it the primary aim to establish a reputation within an almost mystical community of technically highly-skilled individuals. With the Internet easily accessible to everyone and the use of financially-oriented services such as electronic shopping and banking now widespread, casual users with minimal technical knowledge have become promising targets for criminals. Financial gain is now the leading motivation for online criminal operations and malware creation.

A.Botnet:

Botnet is malicious software running in a group of a computer without knowing the owners knowledge. Their versatility proved to be an enabler for a wide range of criminal business models ranging from spam delivery to phishing, blackmail, and espionage, triggering counteractions by those operating networks or responsible for securing information infrastructures. Since infected machines participating in botnets are often owned by a large and diverse group of individuals and organisations, scattered over a large number of jurisdictions, their efforts were often focused on denying the operator control over its botnet [1].

B. HTTP filtering:

There is a wide range of the HTTP usage on the Internet; most recent botnets employ HTTP protocol to hide their malicious activities among the normal web traffic. Their C&C channels utilize HTTP protocol to communicate with their bots. Therefore, to investigate the effect of protocol filtering on botnet detection, specifically on false alarm rates, we employed an HTTP filter to select the only HTTP related traffic. Citadel and Zeus are the two most powerful botnets that have affected the legitimate Internet realm the most in the past few years.

C. Machine learning classifier

Machine learning deals with the construction and study of the systems that can learn from data rather than follow only explicitly programmed instructions. It has strong artificial intelligence and optimization. It employs in a range of a computing task.

II. BOTNETS MALICIOUS ACTIVITIES:

A.Identity Theft:

Major use of botnets, with the intention of gaining financial benefits, is for the automated extraction of user data and credentials from infected hosts. Key targets include passwords for various services like e-mail accounts, web shops, banking platforms or social networking platforms. This technique is often called identity theft because it enables botmasters to impersonate the victim, making further actions, like fraud, possible.

B. Click Fraud And Pay Per Install:

Allows for the provision of services based on current demand requirements. This is done automatically using software automation, enabling the expansion and contraction of service capability, as needed. This dynamic scaling needs to be done while maintaining high levels of reliability and security.Another way of monetising botnets is what is called click fraud. First, the attacker sets up an account with an online advertiser, who pays for page visits or additional advertising links by, for example, clicking on a banner. Second, the attacker uses the controlled bots to visit those pages and to generate clicks on the target banners. This is possible because he has full control of the victim‘s machine and may use it to simulate surfing behavior. In this case, the attacker gains money directly from the advertising company, which in turn does not benefit from the traffic generated. In the context of botnets, clicks can be sold to third-party advertisers to optimize their click popularity and ranking in search engines. Alternatively, generating such traffic can be used to influence online polls.

C. Spam E-Mail:

One of the most popular uses of botnets is for unsolicited mass mailing, also known as spamming. While classic spamming has been achieved with single computers or comparatively small networks owned directly by the spammers, botnets have enabled them to perform a cost shift, in terms of computation power, bandwidth and reputation, towards the owners of compromised computers. Also, the ability of botnets to use bots‘ IP addresses to hide the true originator of the spam email complicates countermeasures such as the blacklisting of suspicious IP addresses.

D. Distributed Denial Of services:

Botnets usually consist of such large numbers of remote machines that their cumulative bandwidth can reach multiple gigabytes of upstream traffic per second. This enables botmasters to start targeted sabotage attacks against websites. By commanding bots to contact a website frequently, the servers are rendered unreachable because they cannot handle the incoming traffic. This attack is called Distributed Denial of Service (DDoS) attack; distributed, because a large number of geographically-distributed bots are involved in the attack. These attacks happen regularly, and the profit scheme connected with this use of botnets is extortion. Many companies depend on web-based services, e.g. web shops, and downtime causes a loss of business volume.

III. PROPOSED MODEL:

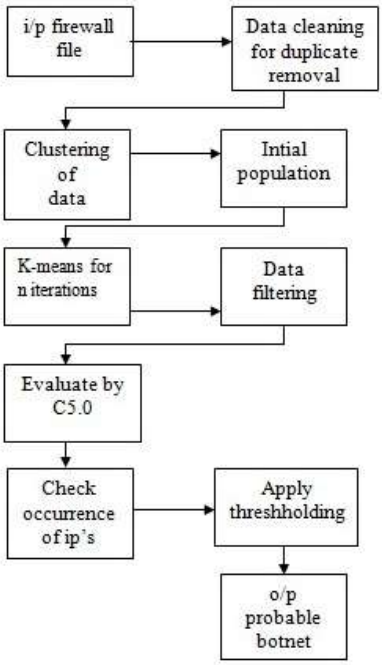

Fig. 4.1: System Architecture

The architecture Shows the complete working for identification of botnets from the system. Initially, the data packets are collected from the machine itself from the c drive of a computer in a "p firewall log" folder. Then the Data cleaning process is done for removal of the duplicate data that is nothing but the pre-processing applied on the raw data. After that clustering of data is done by the k-means algorithm And the probable botnets are found through the machine learning C5.0 algorithm.

***A. C5.0 over C4.5 and Naive Bayes*:**

Previously used algorithms are C4.5 and Naive Bayes algorithm. Thus, these algorithms have many drawbacks.

Improvement in C5 from C4.5 Algorithm are:

- C5 is faster than C4.5and naive Bayes.
- Memory usage is more efficient in C5.0 than C4.5 and naive Bayes.
- C5.0 gets smaller decision trees in comparison with C4.5 And Naive Bayes.

- The C5.0 rule sets have lower error rates on unseen cases.

So comparing with C4.5 and Naive Bayes the accuracy of a result is good with C5.0 algorithm.C5.0 automatically allows removing unhelpful attributes.

IV. PROPOSED ALGORITHM

A. Bisecting k-means algorithm:

Bisecting k-Means is like *a* combination of k-Means and hierarchical clustering. It starts with all objects in a single cluster. The pseudocode of the algorithm is displayed below:

- Pick a cluster to split.
- Find two sub-clusters using the basic k-Means algorithm (Bisecting step)
- Repeat step 2, the bisecting step, for ITER times and take the split that produces the clustering with the highest overall similarity.
- Repeat steps 1, 2 and 3 until the desired number of clusters is reached.
- The critical part is which cluster to choose for splitting. And there are different ways to proceed, for example, you can choose the biggest cluster or the cluster with the worst quality or a combination of both.

B. *Machine Learning C5.0 algorithm:*

The formulae give the simplest form of exponential smoothing; The C5.0 algorithm is a new generation of Machine Learning Algorithms (MLAs) based on decision trees [16]. It means that the decision trees are built from the list of possible attributes and set of training cases, and then the trees can be used to classify subsequent sets of test cases. C5.0 was developed as an improved version of well-known and widely used C4.5 classifier, and it has several important advantages over its ancestor [17]. The generated rules are more accurate and the time used to generate them is lower (even around 360 times on some data sets). In C5.0 several new techniques were introduced:

- **Boosting:** several decision trees are generated and combined to improve the predictions.
- **Variable misclassification costs:** it makes it possible to avoid errors that can result in harm.
- **New attributes**: Dates, times, timestamps, ordered discrete attributes.values can be marked as missing or not applicable for particular cases. Supports sampling and cross-validation.

1. Create a root node for the tree
2. Check for the base case
3. Apply Feature Selection using Genetic Search
4. bestTree = Construct a DT using training data
5. Perform Cross Validation

a. Divide all examples into N disjoint subsets, E = E1, E2, ..., EN

b. For each i = 1, ..., N do

- Test set = Ei
- Training set = E – Ei
- Compute decision tree using Training set
- Determine performance accuracy Pi using Test set

c. Compute N-fold cross-validation estimate of performance = (P1 + P2 + ... + PN)/N

6. Perform Reduced Error Pruning technique
7. Perform Model complexity
8. Find the attribute with the highest info gain (A_Best)
9. Partition S into S1,S2,S3... according to the value of A_Best
10. Repeat the steps for S1, S2, S3
11. Classification: For each ti ϵ D, apply the DT to determine its class

V. METHODOLOGY

Firstly the data is generated from the machine itself. If the machine is connected from the internet. Initially, the data packets are collected from the C drive of a computer in a “p firewall log” folder. Generated data is preprocessed for removing redundancy. Thus, the data that have source and destination IP address is same are counted as one IP with the sum of the packets exchanged.

Preprocessed data is are processed by the K-mean Bisecting algorithm. Which take iterations up to the desirable output is generated. This iterative data is processed by the C5.0 machine learning algorithm that gives the probable botnets. Now this Probable botnet is checked in URL to fine. whether it's a botnet or not. Thresholding is 0.1 which is used to check the efficiency.

VI. COMPARATIVE ANALYSIS WITH DIFFERENT AVAILABLE ALGORITHM

The Accuracy of the botnet detection is obtained by the different available algorithm. It all depends upon the No. Of occurrence of the botnet and the actual botnets by which opening each IP's in URL. We are taking Ten percent of data at a time. Now by the available algorithm C4.5 and Naive Bayes are compared with the C5.0 algorithm.Where C4.5 machine algorithm is the predecessor of C5.0 algorithm.

In this work, we used the following metric in our evaluations:

1) Detection Rate (DR): DR is the fraction of all the correctly labeled instances.

2) False Positive (FP) and True Positive (TP) Rates: In general, positive means "identified" and negative means "rejected". Therefore, FP means incorrectly identified, and TP means correctly identified. Thus, FP Rate (FPR) means the ratio of incorrectly identified samples and TP Rate (TPR) means the ratio of correctly classified samples of each class.

3) Complexity: The definition of complexity often depends on the concept of the "system". Speaking of classifiers, complexity can be measured on different criteria.

Such as memory consumption, time or solution. In this work, three complexity criteria are utilized. Firstly, computation time, which is a typical scale for learning algorithms during training procedure denoted as training time.

	Data Set	Detection Rate	Botnet		Legimite		Time Complexity
			TPR	FPR	TPR	FPR	
C4.5	HTTP	91.5	91.5	8	91	7.5	0.13
Naive Bayes	Non HTTP Filter	66	8.5	1	99	91.5	0.07
C5.0	HTTP	94	97.6	1	99	1	0.10

Table 6.1 Comparative AnalysisWith Different Available Algorithm

VII. CONCLUSION

With the steep rise in computer network attacks mostly due to Botnets has significantly highlighted the issue to work on an effective and efficient remedy for Botnet. That is why one has to work in advance then the hackers not only on it's after effect but before the attacks are done. In this system, we generate the logs of packets then applied the clustering schemes for filtering the data and then trained data is taken and applied machine learning algorithm C5.0 on it for the identification of the botnet.

Moreover, C5.0 can choose the most appropriate features from all the features given to it, C4.5 and naive Bayes does not have this ability. The threshold on data generated by C5.0 machine learning classifier is applied for finding efficiency.

Thus, the no. of occurrences of botnets are checked, by putting it into URL if it opened then that particular packet is not a botnet. If It does not open, then that particular packet would be a botnet.

FUTURE WORK:

The future work may include:

A prediction related to the growing availability of mobile Internet access concerns the increasing probability that smartphones will be compromised on a large scale. Smartphones are becoming attractive to criminals because of the devices increasing computing capacity and ability to connect to the Internet. The spreading of the worm was supported by access to personal contact data stored on infected phones. However, user interaction was still needed to enable this worm to spread.

REFERENCES:

[1] P. Wurzinger, L. Bilge, Th. Holz, J. Goebel, Ch. Kruegel, and E. Kirda, "Automatically generating models for botnet detection," in 14th European conference on research in computer security (ESORICS), pp.232-249, 2009

[2] Paul Royal. Maliciousness in Top-ranked Alexa Domains. [Online].

[3] G. Ollmann, "Botnet communication topologies," Jun 2009.

[4] Wen-Hwa Liao, Chia-Ching Chang "Peer to Peer Botnet Detection Using Data Mining Scheme" 978-1-4244-5143-2/10/2010 IEEE

[5] Vrizlynn L. L. Thing, Morris Sloman, and Naranker Dulay "A Survey of Bots Used for Distributed Denial of Service Attacks", 22nd IFIP International Information Security Conference (SEC), Sandton.

[6] K. P. Clark, M. Warner, and F. M. T. Brazier, "BOTCLOUDS -TheFuture of Cloud-based Botnets?," in Proceedings of the 1st InternationalConference on Cloud Computing and Services Science (CLOSER), 2011, pp. 597-603

[7] J. Calahorrano and D. Chow V. Buitron. (2007) northwestern..http://www.cs.northwestern.edu/~ychen/classes/msit458- s09/Botnets_defense.ppt

[8] Dae-il, Minsoo Kim, Hyun-Chul Jung, Bong-Nam Noh"Analysis of HTTPP2P Botnet: case study wale DAC", 978-1-4244-5532-4/09 /2009 IEEE.

[9] E. Yuce, "A Literature Survey About Recent Botnet Trends," GÉANT Network, ULAKBI M,Turkey, Rep. JRA2 T4, 2012

[10]J. Dae-il, C. Kang-yu, K. Minsoo, J. Hyun-chul, and N. Bong-Nam,"Evasion Technique and Detection of Malicious Botnet," in Proceedingsof the International Conference for Internet Technology and SecuredTransactions (ICITST), 2010, pp. 1-5

[11]A. Berger and M. Hefeeda, "Exploiting sip for botnet communication," 2009 5th IEEE Workshop on Secure Network Protocols, pp. 31–36,2009. [Online]. Available: http://tinyurl.com/8n9rkex

[12]I. Burke, "Who needs botnets if you have google?" Presented at ZaCon2, Johannesburg, South Africa, 2010.

[13]Micorsoft, "Flame malware collision attack explained," http://tinyurl.com/dxxlb5j, June 2012, [Online; accessed 1-Aug-2012].

[14]K. Bong and J. Brozyck, "Managing large botnets," http://tinyurl.com/blcuxbo, 2007, [Online; accessed 15-December-2011].

[15]N. R. Ramay, S. Khattak, A. A. Syed, and S. A. Khayam, "Bottleneck:A generalized, flexible and extensible framework for botnet defense," in IEEE Symposium on Security and Privacy, 2012, May 2012, poster Paper.

[16]BotHunter, "Bothunter: A network-based botnet diagnostic system," http://www.bothunter.net/, [Online; accessed 12-December-2011].

[17]M. Feily, A. Shahrestani, and S. Ramadass, "A survey of botnet and botnet detection," 2009 Third International Conference on Emerging Security Information Systems and Technologies, pp. 268–273, 2009. [Online]. Available: http://tinyurl.com/9njpehq

[18]Seculert, "Mahdi - the cyberwar savior?" http://tinyurl.com/brp64k4, July 2012, [Online; accessed 12-Dec-2012].

[19]InternetWorldStats, "Internet growth statistics," http://www. internetworldstats.com/emarketing.htm, [Online; accessed 12- December-2011].

[20]K. Bong and J. Brozyck, "Managing large botnets," http://tinyurl.com/ blcuxbo, 2007, [Online; accessed 15-December-2011].