International Journal of Security and Its Applications
Vol.9, No.11 (2015), pp.97-104
http://dx.doi.org/10.14257/ijsia.2015.9.11.10

Botnet Detection Based on Genetic Neural Network

Chunyong Yin[1], Ardalan Husin Awlla[1], Zhichao Yin[2] and Jin Wang[1]

[1] *Jiangsu Key Laboratory of Meteorological Observation and Information Processing, School of Computer and Software, Jiangsu Engineering Center of Network Monitoring, Nanjing University of Information Science & Technology, Nanjing 210044, China*[2], *Nanjing No.1 Middle School, Nanjing 210001, China*

Abstract

Botnet have turned into the most serious security dangers on the present Internet framework. A botnet is most extensive and regularly happens in today's cyber-attacks, bringing about the serious risk of our system resources and association's properties. Botnets are accumulations of compromised computers (Bots) which are remotely regulated by its creator (BotMaster) under a typical Command-and-Control (C&C) framework. Botnets cannot just be implemented utilizing existing well-known applications and additionally developed by unknown or inventive applications. This makes the botnet detection a challenging issue. In this paper proposed an anomaly detection model based on genetic neural network system, which joined the significant global searching capability of genetic algorithm with the precise local searching element of back propagation feed forward neural networks to improve the initial weights of neural network.

Keywords*: ANN, GA, GNN, Botnet, Bot, BotMaster*

1. Introduction

These days, the most dangerous exhibition malware is Botnet. For a superior comprehension of Botnet, we need to know two terms to begin with Bot and BotMaster, and after that, we can legitimately characterize Botnet.

Bot – Bot is another kind of malware introduced into a compromised computer that can be controlled remotely by BotMaster for executing a few requests through the received commands. After the Bot code has been introduced into the compromised computers, the computer turns into a Bot. Bots can get commands from BotMaster and are used as a part of attack platform. BotMaster –BotMaster is a man or a gathering of persons that remote control Bots. Botnets-Botnets are systems comprising of a vast number of Bots. [1] Bots typically circulate themselves over the Internet by searching for helpless and unprotected computer to taint. The Bot stay covered up until they are educated by their BotMaster to perform an attack or assignment. [2-3] Botnet detection bear on the classification and recognition issue with a substantial number of non-linear conditions, which make it crucial to contemplate non-linear integrated approaches to solving the issue [4-8]. Artificial neural system (ANN), regularly simply called "neural network" (NN), is a numerical model or computational model given natural neural systems. It comprises of an interconnected gathering of artificial neurons and procedures information utilizing a connectionist approach to computation. Much of the time an ANN is an adjusting system that progressions its structure based on external or internal information that moves through the system during the learning stage. In more useful terms, neural network systems are non-linear statistical data modeling tools. They can be utilized to model

ISSN: 1738-9976 IJSIA

complex connections between of inputs and output or to discover patterns in data. The capability to learn and adjust to vulnerabilities of ANN is only suitable to solve the botnet detection issue. In any case, an ANN effortlessly drops into a local minimum, so it may not search the global optimum [9-11]. For this surrender, the paper will propose an anomaly identification model based on Genetic Neural Network (GNN), which consolidates the significant global searching function of genetic algorithm with the exact local searching feature of back propagation networks to enhance the initial weights of neural network systems.

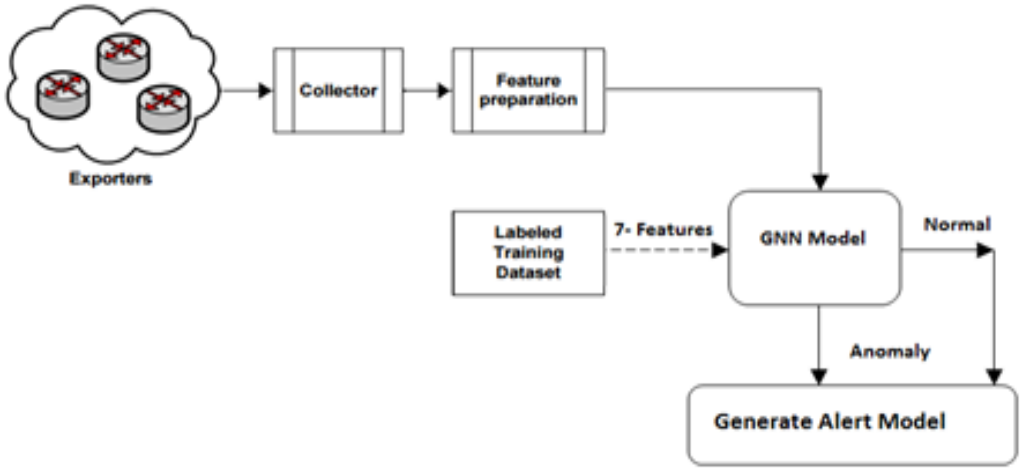


Figure 1. System Diagram of GNN Botnet Detection

2. Introduce BP Feed Forward Neural Network and Genetic Algorithm

The feed forward neural network system is an interconnected neural network system that can execute a few functions that are based on the comprehension of the human brain neural system. It is a data processing framework that is constructed by impersonating the brain neural network system structure and function.

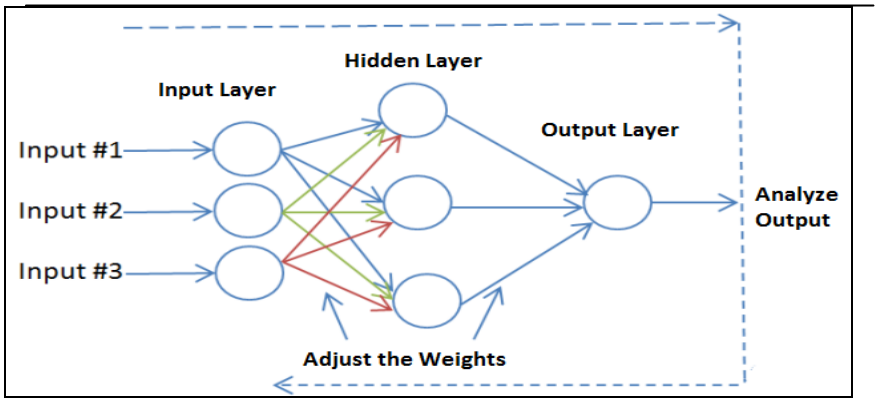


Figure 2. BP Feed Forward Neural Network Structure

In this model [5], from the input layer neurons receive the information from the outside and pass the information without any changing to the next layer that is a hidden layer. Hidden layer that is an inward layer to preparing data and based on the need can be separated into a single layer and multiple layers. The hidden layer is capable of changing the received information and exchanging the information to the output layer. At the point when the information reaches output layer, the output layer examines the information and

outputs results. When the actual output does not agree with the wanted worth, it will start to enter the error back propagation, which is orderly a reverse adjustment until the last result to reach acceptable error or the maximum number of learning as indicated by the gradient descent method.

The BP neural network can be used in many fields but mainly used in pattern recognition, classification, and compression data, but also using it has some disadvantages. It includes first the amount of the learning rate is fixed which means it has an influence on a speed of the network. The second decision to choose the numbers of hidden layers for a network has no any hypothetical basis for guiding. Third the minimum vale we can achieve in BP is a local minimum value.

Genetic algorithms mimic nature when attempting to find an optimal solution for a particular problem, the operation that is derived from biological evolution. The data model that is utilized as a part of the Genetic algorithms is a representation of an arrangement of unknown variables as genes in a chromosome. As portrayed in [6] each genetic algorithm has four sections in common:

Selection: is the process choosing of parents from the population based on fitness function for later breeding.

Crossover: is mating between parents to produce one offspring. If the fitness of the new offspring is desired, the algorithm ends here moving to step 4.

Mutation: is the changing randomly a single gene or more in the current offspring to produce offspring.

Replace: is the process of replacing the weakness individuals in the population with new offspring.

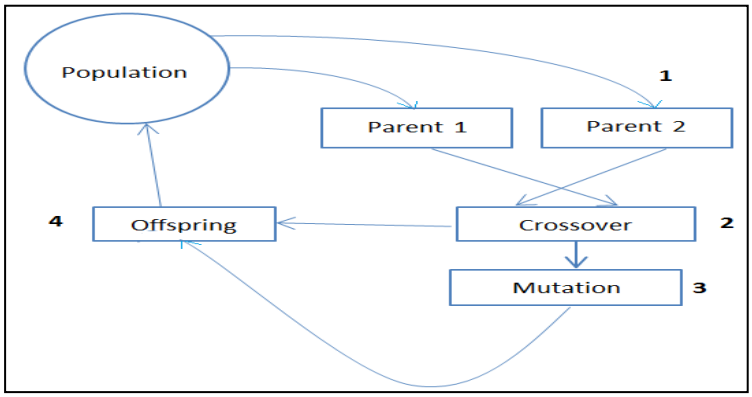


Figure 3. The Basic Properties of Genetic Algorithm

These steps or cycle continue until the user gets the desired output. The desired output can be produced after finding the proper average fitness in the population or until the end of the fixed iterations.

3. Briefs about Combination of ANN-GA for Detection of a Botnet

According to the above examination, it can be seen that BP feed forward neural system is exceptionally suitable for botnet detection. By using chose illustrations to train the neural network we get a network structure of botnet detection, and afterward can use it to recognize any network information to judge whether there is a botnet or not. When the inputs and outputs have been given, it is a critical issue how to get the best neural network structure and association weights that will have a significant effect on last detection results. Genetic Algorithm is regularly used to solve optimized issues, so we can use it to find the best neural network structure and association weights. The present paper

demonstrates the use of GA for initializing and improving the connection weights of BP. and uses it in botnet detection

Summarized steps of the system structure based on figure 1 are steps. The first step is responsibility of this module to gather flow data which is exported from one or few exporters. The received data should be perceived by protocol and-and transfer into an internal format. The collected data is always being sent to the feature readiness module. The second step is organizing training and testing dataset. The third step is setting the genetic algorithm to the back propagation neural network. The fourth step is applying the consequences of step 2 and step 3 to learn, train and approve lastly generates a botnet detection based on GNN Figure 4. The fifth step is inputting another and random network packet data into the botnet detector to make discoveries. And in the sixth step, if bots are detected, the system informs the system security administrator to take suitable efforts to establish safety to guarantee network system.

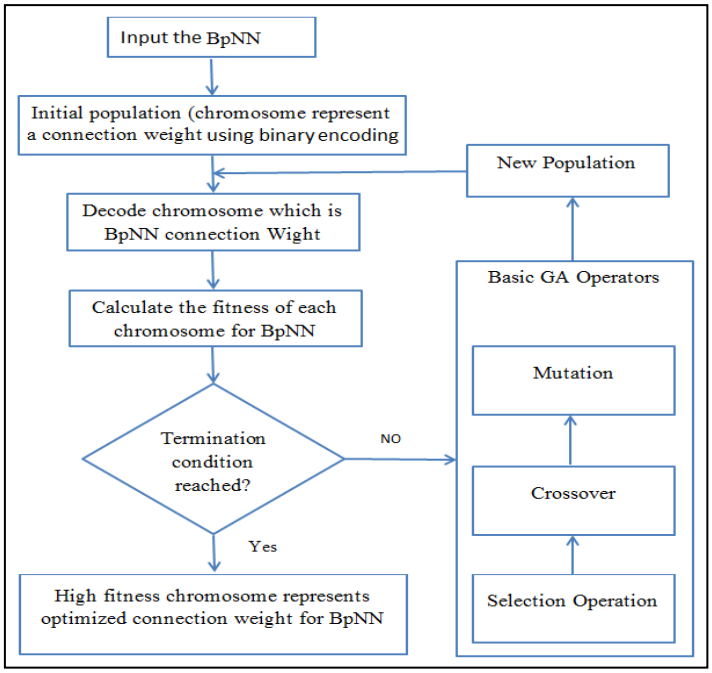


Figure 4. Combination of the Neural Network and Genetic Algorithm

4. The Entire GNN is Explained as the Following

1. First of all, the structure of the neural network should be determined which is mean the number of hidden layers, number of nodes in each layer, the initial population size ,range of the weights, the quantity of system structures in every generation, the amount of crossover and mutation rate.
2. Randomly initial the range number of nodes in each hidden layer, the population size, and the weight range produce (generate) the network structure as a parent.
3. For each available network structure in population network structure, the system uses present connection weights and thresholds to implement a calculation process from the input layer to output layer and get output:
4. Test the samples and figure expectation outputs:
5. After getting the output, we should calculate the prediction of error of the model.

$$E = \frac{1}{2} \sum_p \sum_j (tpj - opj)\ 2$$

Where the summation is calculated, complete output nodes pj and tj the desired cost of output oj for a given input vector.

6. Whenever the evolutionary system generation meets the prerequisite, or the best system structure is discovered, the procedure is over, and the last generation is the best system structure; otherwise, go to step 7.
7. From the population, pick paired binary chromosome as parent's i1 and i2 whose fitness are no under average fitness for predictable crossover. For the selection process, we use fitness function that the best chromosome can be chosen.

$f = \frac{1}{E}$ Where E is the root mean square error, the fitness of each chromosome is between 0 and 1 where the best chromosome fitness is close to 1.Crossover operator is replacing part bits of each parent individuals to generate new individuals; crossover operator is the most significant operation to produce the high fitness, individual. For crossover operator, we represent the chromosome as binary and use the uniform operator for crossover between two parents to produce a new individual. For example:

Figure 5. Uniform Crossover Operation

Parent A		Parent B	=	Offspring
11001011	+	11011111	=	11011111

Mutation is the flip or converting one or two bit in each chromosome after crossover because we use binary operation. We select one bit randomly and flip it after selection. For example:

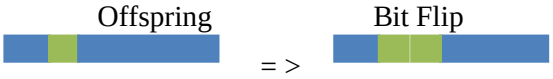


Figure 6. Bit Flip Mutation Operator

Offspring		Bit Flip
11011111	=>	11001111

8. Step 7 should be repeated until the proper genetic neural network structure is gotten.

5. Feature Selection for Botnet Detection

The most vital step in building botnet detection is how to choose the essential feature. They will affect in identification rate and enhanced false alarms. By enhancing feature, the data space will also be improved, so the training and time for dataset will be more proficient for classification that works under continuous environment.

The module of preparation feature receives and processes network flow data, a sequence of packets from a source node to a destination node in the network is a network flow, which is sent from the flow accumulation module. The fundamental of this module is to extract the essential features that are necessary for botnet detection; the processing of feature preparation that involves normalization all features by mapping value each of them into range of (0, 1) must be done before sending them to the identification module. The most important features are the following seven features:

Average Flow Size: it gives a valuable clue to anomalous events such as, port scan, and it is ordinarily small so as to increase the proficiency of attacks.

Packet Average Size: another variable is the measure of every packet in the flow; if the average size is little; it maybe an indication of an anomaly. For instance, in TCP flooding attacks, commonly packets of 120 bytes are sent.

Average Number of Packets: one of the primary features of DoS attack is the source IP spoofing that makes the following attacker's actual source extremely troublesome. A reaction is the generation of flows with a little number of packets (*i.e.*) around three packets in the flow. This varies from ordinary traffic that is not a higher number of packets per flow.

Different numbers of flow to the same destination IP: This feature checks the quantity of flow to the same destination IP address. If the number of flows is high, cloud mean a port scan flood attack.

Number of flows to distinctive Destination Ports: also, it has an impact on distinguishing attack. A strangely enormous number of unique destination ports imply that the system is under port scan attack.

SYN - SYN/ACK: many researchers used this feature [7], comparing the quantities of SYN and SYN/ACK packets a host gets and returns individually, to detect Dos Attack. Under ordinary conditions, the two numbers should be adjusted since each SYN packet is replied by an SYN/ACK packet. Therefore, a high number of unanswered SYN packets are a sign of SYN flood.

Land: this feature is in charge of checking whether there is a land attack in the system or not. (*i.e.*SrcIP=DestIP,SrcPort=DestPort)

6. Performance of the Proposed Model

The network packets that are collected are divided into two parts. The first part about eight hundred records is used to train genetic neural network module. The second part is about two hundred records used to test the Botnet detection. The accuracy of the neural network depends on the number, type and amount of features used to train the neural network.

We use the Java programming language to build our system Botnet detection. Input the prepared dataset into the system that is trained genetic neural network figure the identification rate, So as to assess the execution of a botnet identification strategy, we have to present a quantitative estimation. In our Botnet detection system, we essentially classify the network traffic into two groups that are normal and anomalous. So we need to represent the true positive (TP), true negative (TN), false positive (FP) and finally false negative (FN) to define true positive rate (TPR) and false negative rate (FNR).

True positive rate (TPR) and false negative rate (FNR) can be computed using the following mathematical equations.

$$TPR = \frac{TP}{(TP+FN)}, \quad FPR = \frac{FP}{(FP+TN)}$$

The true positive rate (TPR) assesses execution of botnet detection technique regarding the probability of a suspicious data reported correctly as anomalous data. Then, again the false positive rate (FPR) estimates the execution of botnet detection technique as far as the probability of a regular traffic reported as abnormal data or anomalous.

For training artificial neural network, the learning rate is 0.7, momentum is 0.9, maximum error to reaches is 0.01, weight and threshold values are randomly initialized before training the range of values between [-1, 1]. For training Genetic algorithm, we initialize that the population is equal to 5000, the selection probability 0.05, the crossover percentage is 0.1, and the mate percentage is 0.25.

We can see from the table 1 and table 2 that the identification rate of mixed both genetic algorithm and neural network GNN algorithm is 95.7%, the rate of false report is 4.3%, the identification rate of BP feed forward neural network is 90.3%, false report is 9.7%, the identification rate of Genetic algorithm is 93.4% and false report is 6.6% with regardless the amount of training and test dataset. Whenever GNN algorithm needs 24517 iterations for training, but BP feed forward neural network algorithm needs 600 thousand iterations and may drop into a few local minimum.

Table 1. Comparing Bp, GA, and GNN

Input Dataset	BPNN	GA	GNN
600	0.034045	0.009422	0.002776
2600	0.002309	0.001129	0.000786
8000	0.008119	0.000791	0.000750
10000	0.002013	0.000760	0.000568
14000	0.001279	0.000660	0.000430

Table 2. Comparing of Detection Rate

Input Dataset	BPNN	GA	GNN
Detection Rate	90.3%	93.4%	95.7%
False Negative	9.7%	6.6%	4.3%

6. Conclusions and Future Work

The bot is another kind of malware introduced into a compromised computer, and it can be controlled remotely by BotMaster for executing a few requests through the received commands (C&C). Botnet detection is one of the new sorts of network security technology that is used to analyze network packets online or offline to detect a bot and protect the system from attack. While detecting the attack, it can inform the network administrator for existing attack by an alarm. Botnet detection belongs to the classification and identification issues. The paper applied genetic algorithm to BP feed forward neural network for botnet detection where the genetic neural network can learn knowledge from an enormous number of dataset for training and testing the result for detection. The experiment result demonstrated that the using feed forward back propagation is a real local minimum algorithm, and genetic algorithm is a good global search algorithm or optimization algorithm based on the practical. The experimental results imply that the accuracy both of them is lower than applied GA to BP feed forward algorithm.

Acknowledgements

This paper is a revised and expanded version of a paper entitled “A Novel framework towards Botnet Detection” presented at COMCOMS 2015, Hanoi, Vietnam, October 22-24, 2015. This work was funded by the National Natural Science Foundation of China (61373134, 61402234), and by the Industrial Strategic Technology Development Program (10041740) funded by the Ministry of Trade, Industry and Energy (MOTIE) Korea. It was also supported by the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD), Jiangsu Key Laboratory of Meteorological Observation and Information Processing (No.KDXS1105) and Jiangsu Collaborative Innovation Center on Atmospheric Environment and Equipment Technology (CICAEET). Prof. Jin Wang is the corresponding author.

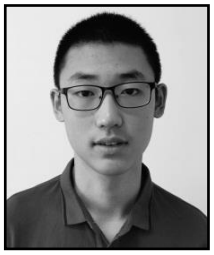
References

[1] H.R; Manaf, Zeidanloo, "Botnet Command and Control Mechanisms", "Second International Conference on Computer and Electrical Engineering", ICCEE., (**2009**), Dec, pp. 564-568, 28-30,
[2] R G Bace, "Intrusion Detection", "Macmillan Technical Publishing", Indianapolis, IN 46290 USA, (**2000**).
[3] P. Ramstedt , B McAnderson, "Intrusion Detection Technology: Today and Tomorrow", "12th Annual First Conference", (**2000**), June, 25-30,.
[4] Yu Zhao, "Novel approach of P2P Botnet Node-based detection and applications", "Journal of Chemical and Pharmaceutical Research", 2014, vol. 6, no. 7,: (**2014**) , pp 1055-1063,.
[5] S hi Y, Gu Y, Wang J, Efficient intrusion detection based on multiple neural network classifiers with improved genetic algorithm journal of Software, vol 7, no 7 **(2012),** Jul. (**2012**). pp. 1641-1648,
[6] G.W. Flake. "The computational beauty of nature: Computer explorations of fractals, chaos, complex systems, and adaptation", The MIT Press, (**2000**) 31, January,
[7] C. Yin, "Towards accurate node-based detection of P2P botnets", "Scientific World Journa"l, 2014: , 24, June, (**2014**). pp .425491-425491.
[8] C. Yin, M. Zou, D. Iko and J. Wang, "Botnet Detection Based on Correlation of Malicious Behaviors", "International Journal of Hybrid Information Technology", 2014, vol. 6, no. 6, Nov. 06, (**2013**). pp. 291-300.
[9] B. Gu, V. S Sheng, K. Yeow Tay, W. Romano, and S. Li, "Incremental Support Vector Learning for Ordinal Regression", "IEEE Transactions on Neural Networks and Learning Systems", 2015, vol. 26, no.7, Aug. 12, (**2014),** pp. 1403-1416,
[10] B. Gu, V. S. Sheng, Z. Wang, D. Ho, S. Osman, S. Li, "Incremental learning for ν-Support Vector Regression", Neural Networks, 2015, 67:140-150, 13, 03 (**2015**).
[11] D. Zhang, H. Wang, and K. G. Shin, .SYN-dog: Sniffing SYN Flooding Sources, In Proc. Of 22nd International. Conference on Distributed Computing Systems, 2002, , 2-4, July, (**2002**), pp. 421-428.
[12]] Jin Wang, Jeong-Uk Kim, Lei Shu, Yu Niu and Sungyoung Lee, A distance-based energy aware routing algorithm for wireless sensor networks, Sensors, 10, 10, (**2000**).
[13] J. Wang, Y. Yin, J. Zhang, S. Lee, and R. Simon Sherratt, "Mobility based energy efficient and multi-sink algorithms for consumer home networks", IEEE Transactions on Consumer Electronics, 59, 1, (**2013**).

Authors

Chunyong Yin, he is currently an associate Professor and Dean with the Nanjing University of Information Science & Technology, China. He received his Bachelor (SDUT, China, 1998), Master (GZU, China, 2005), PhD (GZU, 2008) and was Post-doctoral associate (University of New Brunswick, 2010).He has authored or coauthored more than twenty journal and conference papers. His current research interests include privacy preserving and network security.

Ardalan Husin Awlla, he received his bachelor degree in 2010 from University of Sulaimaniyah. He is studying for his master's degree in Nanjing University of Information Science & Technology. His main research interests include data mining, network security and privacy protection.



Zhichao Yin, he is studying in Nanjing No.1 Middle School. His current research interests include network security and mathematical modeling.

