IOP Conf. Series: Materials Science and Engineering **263** (2017) 042103 doi:10.1088/1757-899X/263/4/042103

Comparison between genetic algorithm and self organizing map to detect botnet network traffic

Shinde Yugandhara Prabhakar[1], Pratishtha Parganiha[1], V MadhuViswanatham[1] and M Nirmala[2]
[1]School of Computer Science and Engineering, VIT University, Vellore-632014, Tamil Nadu, India
[2]School of Information Technology and Engineering, VIT University, Vellore-632014, Tamil Nadu, India

E-mail: vmadhuviswanatham@vit.ac.in

Abstract. In Cyber Security world the botnet attacks are increasing. To detect botnet is a challenging task. Botnet is a group of computers connected in a coordinated fashion to do malicious activities. Many techniques have been developed and used to detect and prevent botnet traffic and the attacks. In this paper, a comparative study is done on Genetic Algorithm (GA) and Self Organizing Map (SOM) to detect the botnet network traffic. Both are soft computing techniques and used in this paper as data analytics system. GA is based on natural evolution process and SOM is an Artificial Neural Network type, uses unsupervised learning techniques. SOM uses neurons and classifies the data according to the neurons. Sample of KDD99 dataset is used as input to GA and SOM.

1. Introduction

Robot and network makes up the term Botnet. Botnet is a network of infected computers to perform some activity. This botnets are used by an attacker to perform malicious activities. Mostly botnets are used to perform denial of service attack, in this, number of requests are sent to a single machine to break it down from a destination machine. Other attacks performed are distributed denial of service, cyber-espionage campaign, access to device, bombardment of spams.

Thousands of infected computers generates an artificial traffic known as Botnet traffic. The count of computers use depend on the attack, millions of computers are also used. Infected computers are nothing but a mafia practice, in this the computers are hacked with Trojan horse attack. Computer owner will be unaware about the malicious activities carried out from his/her computer. During the use of computer, the owner can not realize that a hidden browser is working through which attacker clicks and access whatever he/she wants.

Mimicking the process of biological evolution, the algorithm Genetic Algorithm [1] was developed, an evolutionary algorithm and heuristic search algorithm. GA is used to optimize the problems by a random search, representing an intelligent exploitation. They direct the search into the region which will give the optimized output thus increasing the performance, exploits the historical information. Like biological evolution, GA performs selection, crossover and mutation functions.

Self-Organizing Map (SOM) [6] an Artificial Neural Network (ANN) type produces low-dimensional map, using unsupervised learning. Map has the training sample input space in discretized representation. SOM reduces the dimensions. SOM apply cognitive learning which distinguish them

from other AN networks, which applies error-correction learning. Neighborhood function is used by SOM for topological property preservation of input data. Because of this it gives a low-dimensional visualization of high-dimensional data. It operates in two modes – training and map. Training mode creates a map for input example and map do classification by creating new input vector.

In this paper, literature survey is carried out by section 2, proposed method by 3,comparison by 4 and conclusion by section 5.

2. Literature survey

Yogita Danane and Thaksen Parvat[1] detection of intrusion traffic is presented by fuzzy-genetic approach. Memory allocation, execution time and accuracy shows the results. The memory usage is decreased by using genetic algorithm. Fuzzy rule is used to sort the network attack data. This proposed system gives better performance in mentioned aspects as compared to previous system.

Chen Yan [2] improved the efficiency of intrusion detection system by decreasing the false negative rate and false alert rate of existing system, using intelligent intrusion detecting model which uses genetic algorithm's global superiority and nerve's locality. Used genetic algorithm to optimize the weight of neural network. Dipika Narsingyani and Ompriya Kale [3] applied genetic algorithm on network intrusion detection, reduced false positive rate which increased the accuracy and performance of system. Hossein Ahmadzadegan and Mahdi GhalbiValian [4] reduced the false alarm rates and increased search speed by proposed intrusion detection model which uses Genetic Algorithm and the dataset is classified using K-Nearest Neighbor algorithm. Priya UttamKadam and Manjusha Deshmukh [5] proposed system which uses Genetic Algorithm, Fuzzy and Pattern matching algorithm. Network intrusion behavior is analyzed by Fuzzy. Best optimal solution is provided by Genetic Algorithm. Detection rules are obtained from pattern matching algorithm. Proposed model increased the accuracy and reduced false alarm rates. Duc C. Le, A. NurZincir-Heywood and Malcolm I. Heywood [6] tested the Self Organizing Map capability to detect known and unknown botnet traffic. Test resulted, SOM is capable of detecting unknown network traffic, possessing to be a data analytics tool. Angela Denise Landress [7] proposed a model which used simple K-means algorithm, self-organizing map and feature selection algorithm which employees J48 decision tree. Using this three the author reduced the false positive rate. To cluster data in less time K-means was used. The most prominent attack was found by using best subset of feature using J48 decision tree algorithm, it reduced size of data leading to reduced false positive rate. Accuracy was improved by using self-organizing map, it checked for the similarities between different attack types.

Chet Langin and Mohammad Sayeh [8] proposed a model in which they combined knowledge discovery and intrusion detection techniques to classify and cluster the malignant network activity and botnet traffic of P2P by using SOM. SOM a self-trained is applied on Internet firewall log entries. New log entries of firewall is analyzed to discover unknown previous local P2P botnet traffic and similar network activity is classified. The model gave two advantages, firstly discovery of intrusions and malignant network activity and secondly the abstraction of data in the SOM and using non-local network privacy was maintained.

Shin-Ying Huang and Yennun Huang [9] used growing hierarchical self-organizing map (GHSOM) to analyze the network traces of victim also called as IP flow data. GHSOM generates a hierarchical architecture for input data, geometric distance between each attack pattern is calculated and from the topological space the signature of botnets are obtained. The attacks are grouped and distinguished by their sequential time stamp. Sequential time stamp can give different attack patterns. The results showed that GHSOM is efficient to identify several attack patterns.

3. Proposed method

In this method we are distinguishing data from input dataset based on source and destination traffic. Genetic Algorithm and Self Organizing Map are used to perform the clustering of data separately and then the accuracy of results are compared. MATLAB R2013a is used to cluster the data using GA and SOM.

Genetic Algorithm is an evolutionary based algorithm. It is used to solve optimization problems. The input to GA is called as population as in our case it is network traffic dataset. Fitness function is calculated for every member in population called as chromosomes. The fitness functions plays a major role in optimizing the solution. The chromosomes having the least fitness value are discarded. With the current chromosomes new chromosomes are produced by twisting, recombining or copying them, this stage is called as mutation. In crossover some features of two existing chromosomes are merged and a new chromosomes gets generated. Due to mutation we get distinct chromosomes. This process repeats till a termination criteria. The termination criteria can be anything, if no more distinct chromosomes are generated, when an absolute number of generations are reached. The flow chart depicts the working of GA.

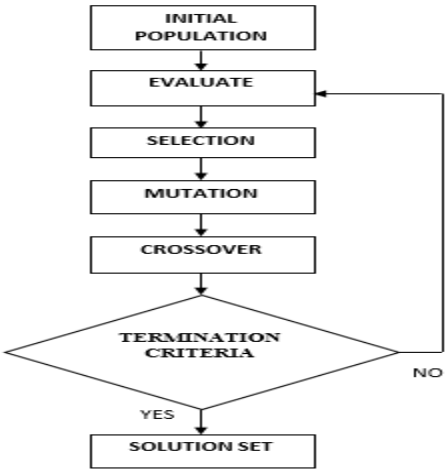

Figure 1. Genetic Algorithm

GA carried out clustering is shown in Fig 2, which shows the clustering of dataset based on source IP address. As we can see the cost for IP address between 0 and 3 is the highest. This means a lot of traffic have been sent from this IP addresses. We can say that this is used to put a load on some system or simply performing denial of service attack. Rest IP addresses have less cost which means this is a normal traffic.

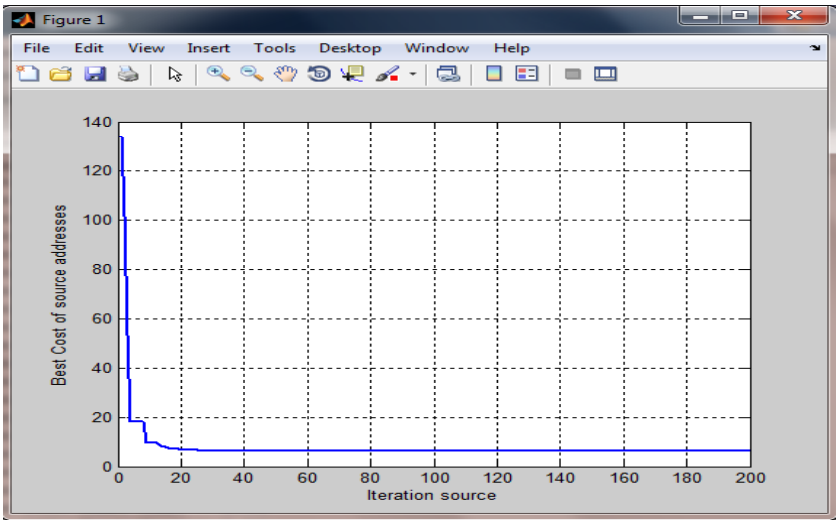

Figure 2. Dataset traffic classification based on Source IP addresses

Traffic classification based on destination IP addresses is shown in fig 3, shows the cost highest for IP

addresses between 0 and 5, by this we can say that this destination addresses have been targetted to perform some attack or even this IP's may be online shopping sites having sale so the traffic is high. Rest IP addresses have minimal cost so are normal.

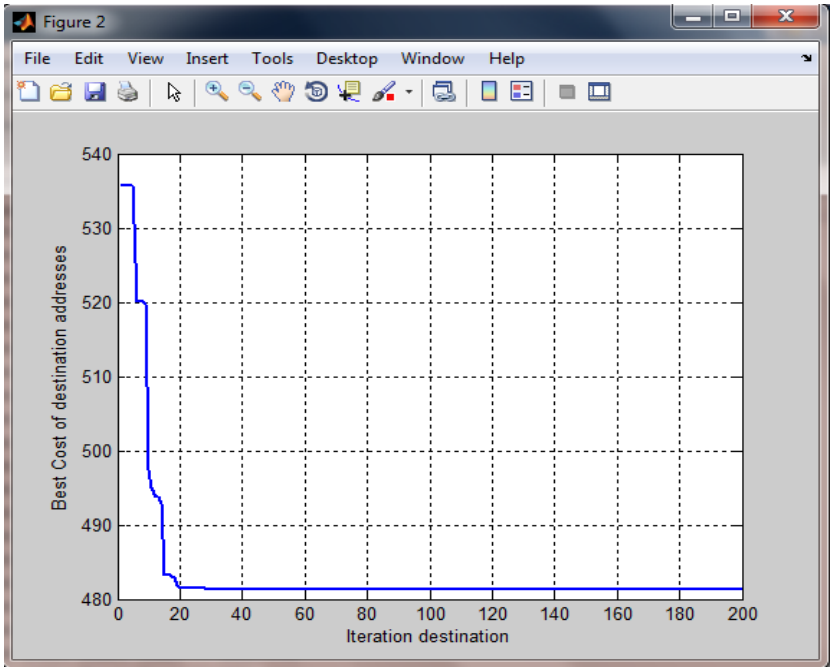

Figure 3. Dataset traffic classification based on Destination IP addresses

Self Organizing Map fig 4 is also called as Kohonen's Self Organizing Feature Maps. SOM do the mapping of n dimensional data to 2 dimensional data. SOM have a feature which learns to classify the data without any supervision that is the unsupervised data. SOM requires no target vector to compare its output. Classification of data is done by following the neurons. Data called as nodes will get clustered with neuron nearest to it. To calculate the distance Euclidean formula is used. Each neuron shifts its position and the nodes closest to it get into this neurons cluster. This process continues till all the nodes gets clustered in some neurons.

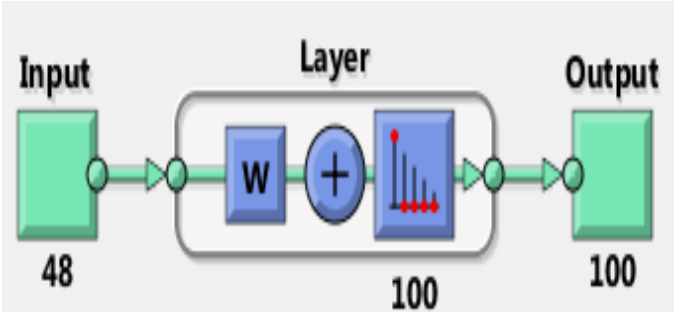

Figure 4. Self Organizing Map

Correlating with the network traffic data, the column names can be considered as neurons. Several steps involved in SOM algorithm are:

1. Weight initialization of each node's– the weight is initialized between 0 and 1.
2. Choosing of random vector from the set of training data. Then it is presented to lattice
3. The weight of the nodes are compared with input vector, the nodes with close value as of input vector are called as Best Matching Unit (BMU) – This is done by calculating the

distance between nodes weight and the input vectors weight, the one with closest weight is tagged as BMU.

4. The neighbourhood radius of BMU is now calculated. This starts by considering the whole lattice and decreses with each step. The nodes in this radius, calculated by using Pythagoras, are considered to be inside the neighbourhood of BMU.
5. The nodes found in step 4, there wieghts are altered to match with input vector. More the node is close to BMU, more its weight is altered.
6. Step 2 is repeated for n iterations.

SOM is clustering the data is shown by fig 5. Highest traffic from and to the respective IP addresses are shown in Fig 5 the sample hits by 3 cells. The 3 cells shows the traffic for IP addresses with smaller IP address as compared to other IP addresses present in dataset. The difference from GA result is, SOM is clustering the traffic taking into consideration both the source and destination traffic.

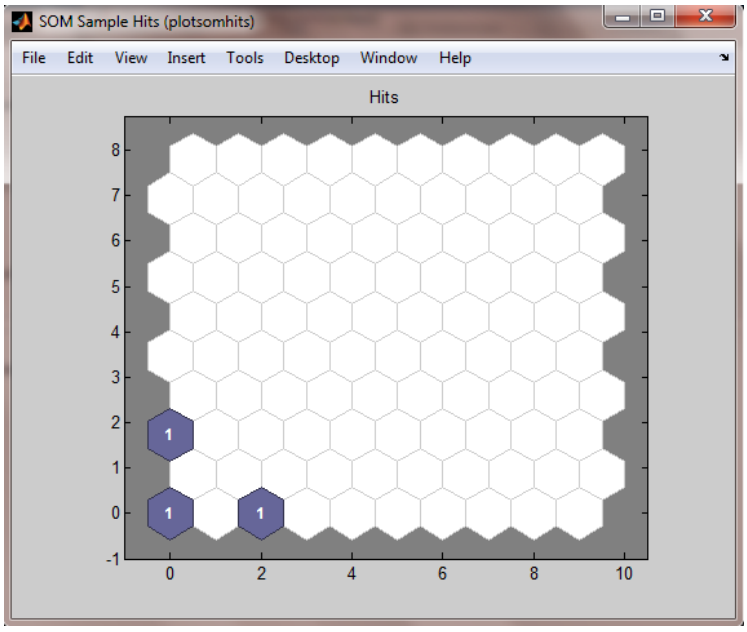

Figure 5. Dataset classification by SOM hits

4. Comparison

In terms of accuracy GA is more accurate. Both algorithms were able to classify the traffic based on the source address, destination address. But SOM gives an unclear result.
In terms of speed to perform the clustering, SOM took less time than GA. GA repeats all the functions till distinct chromosomes are obtained. Due to this it takes comparatively more time than SOM.

5. Conclusion

Genetic algorithm and Self Organizing Map were compared to perform clustering of network traffic from input dataset as normal and malicious. The SOM performs the clustering faster than GA. Due to the operations performed by GA it takes time to complete the clustering process. GA gave more accurate result, we can easily identify the greatest used IP address. The results showed the distinction of traffic by source address and destination address. Our future work will be to generate rule set from both algorithms and compare.

6. Reference

[1] Danane Y and Parvat T 2015 Intrusion detection system using fuzzy genetic algorithm. In *Pervasive Computing (ICPC), International Conference on* 1-5
[2] Yan C 2015 Intelligent Intrusion Detection Based on Soft Computing. In *Measuring Technology and Mechatronics Automation (ICMTMA), 2015 Seventh International Conference on* 577-580
[3] Narsingyani D and Kale O 2015 Optimizing false positive in anomaly based intrusion detection using Genetic algorithm. In *MOOCs, Innovation and Technology in Education (MITE), 2015 IEEE 3rd International Conference on* 72-77
[4] Ahmadzadegan M H, Khorshidvand A A and Valian M G 2015 Low-rate false alarm intrusion detection system with genetic algorithm approach. In *Knowledge-Based Engineering and Innovation (KBEI), 2015 2nd International Conference on* 1045-1048
[5] Kadam P U and Deshmukh M 2016 Real-time intrusion detection with Genetic, Fuzzy, Pattern matching algorithm. In *Computing for Sustainable Global Development (INDIACom), 2016 3rd International Conference on* 753-758
[6] Le D C, Zincir-Heywood A N and Heywood M I 2016 Data analytics on network traffic flows for botnet behaviour detection. In *Computational Intelligence (SSCI),IEEE Symposium Series* 1-7
[7] Landress A D 2016 A hybrid approach to reducing the false positive rate in unsupervised machine learning intrusion detection. In *SoutheastCon,* 1-6
[8] Langin C, Zhou H, Rahimi, S, Gupta B, Zargham M, and Sayeh M R 2009 A self-organizing map and its modeling for discovering malignant network traffic. In *Computational Intelligence in Cyber Security, 2009. CICS'09. IEEE Symposium* 122-129
[9] Huang S Y and Huang Y 2013 Network forensic analysis using growing hierarchical SOM. In *Data Mining Workshops (ICDMW), 2013 IEEE 13th International Conference* 536-543