A Review of Machine Learning Application in Botnet Detection System

Z. ISMAIL, A. JANTAN

School of Computer Science, University Sains Malaysia,

Received 28th August 2016 and Revised 17th October 2016

Abstract: The purpose of this study was to review various machine learning techniques for Botnet detection system by looking at their advantage and limitation, and propose our Botnet detection system. In this paper, we summarized different machine learning techniques used in previous research. Recently, machine learning has become prominent in developing Botnet detection system especially peer to peer Botnet, and most of them are capable to detect decentralized Botnet. Further study has been made on Fuzzy Expert System (FES) and Self Organizing Map (SOM) techniques because we believe both techniques have the capability to fulfill the features required in our Botnet detection system which are autonomous, high accuracy and real time detection. Then, there will be method proposed for future work. The method is divided into six phases and in the future, we will conduct the experiments to assess and prove the effectiveness of these techniques in Botnet detection system.

Keyword: Botnet detection, machine learning, fuzzy expert system, self organizing map

1. INTRODUCTION

Botnet has become a major concern in computer industry. With user engaging in daily life surfing to the Internet, there was high risk becoming a victim. Botnet is becoming an increasingly widely-used method by hacker to gain recognition from their peer hacker and at a certain degree for financial gain. Botnet vigorously effect many computers and build large network of infected machine through command and control (CandC) action, with the aim to spread malicious code, launch DDoS attack, sending spam, phishing, adding spyware and many nefarious activity throughout the network.

The major problem with Botnet occurs when they are used for attack purpose Botnet is capable to intimidate national security, caused terrible disturbance and high usage of network resource therefore government have to spend a lot of money to prevent and treat Botnet attack. Botnet attack is launch using thousand and even millions of compromised computers thus the impact of the attack is huge. Several authors like Solomon and Evron highlighted Botnet spamming as a major concern because of large amount distribution of spam which will use a lot of network resource. Their concern was supported by McAfee Avert Labs report stated that more than 70 percent of spam email is caused by Botnet. Therefore Botnet issuing a great concern not just in the area of detection, but email forensic research area. There are Botnets responsible for spam activity like Grum, Bobax, Rustock, Bagle, Maazben and so on. Detection is one of the components in email forensic. The Botnet data collected from detection will be use for the analysis phase in email forensic.

Furthermore, Botnet are not easily to be detected because it can conceal it presence in infected system. Schiller et al. (2007) reported that many bots they saw usually detected until the botmaster had abandoned the computers. Moreover, cleaning on detected system can be difficult because volume of network traffic created by bots is massive making it impossible to perform update on infected machine. To make the problem worse, new kind of Botnet has been developed to support the capability of making antivirus tool ineffective and bots able to modify registry entries so they remain active even when the infected machine is booted in a safe mode. Some of the Botnet even respond vigorously if they notice there are efforts made trying to detect their presence. In addition, the emergence of peer to peer Botnet added a challenge in Botnet detection because peer to peer Botnet has decentralized architecture which is exceptionally hard to detect [6]. Botnet attack is serious and it is a tough challenge to develop effective Botnet detection. Therefore, the improvement of Botnet detection system is crucial especially to prevent new variant of Botnet which can threaten the national infrastructure of most countries and resulted in country-wide outages.

Starting (2009) onward, many researches on Botnet have been conducted and there was growing number of publication in accordant to the increasing number of Botnet attack. These have shown that the Internet community developed immense interest in Botnet attack and mitigation. However research in Botnet is still in its infancy as Silva et al. (2012) stated that existing studies remain somewhat limited in scope and do not generally include recent research and development. Bailey et al.
Recent research requires improvement in detection because it is a system defense mechanism and technology. Basically our work will focus on Botnet studies as Botnet evolves and will create new technology. Basically our work will focus on Botnet detection because it is a system defense mechanism and recent research requires improvement in detection technique since Botnet is a dynamic threat.

Certain researchers adopt conventional approach to detect Botnet, but their focus mainly on centralized Botnet (2010). In reality, we are facing with sophisticated types of Botnet which has decentralized and unstructured architecture like peer to peer Botnet. Peer to peer Botnet is harder to locate, detect, shutdown, monitor and hijack, thus extra effort is needed to encounter them. Each bot in peer to peer Botnet behaves as a client and server, without centralized CandC server; consequently they do not suffer from single point of failure. The shifted of the trends from centralized to decentralized Botnet have been stated by Ritu and Kaushal. Thus conventional approach did not suitable to countermeasure peer to peer and other decentralized Botnet. We believe machine learning technique is suitable and could be the best solution to detect different types of Botnet. Machine learning has been proven by previous research as being able to solve many issues like accuracy and real time in Botnet detection.

Machine learning provides reliable solution to detect the presence of Botnet because it is knowledge based approach focusing on pattern recognition, and Botnet produces distinguishable traffic patterns and behaviors. In order to understand the role of machine learning in Botnet detection, we have looked into similar works which has been done previously. Based on the literature study conducted, we have come out with proposal on developing Botnet analysis and detection system using FES and Self Organizing Map (SOM). The components of the system will involve analysis and detection system. Neural network has been widely used for intrusion detection system. SOM is an unsupervised neural network. Even though SOM has the pattern recognition capability, SOM alone will not provide the best solution in detecting Botnet attack. Therefore, we are proposing FES to be incorporated with SOM. Expert system technology provides better decision support capability, situation awareness and knowledge. The expert system will use fuzzy logic rather than Boolean logic in the rule generation because fuzzy set rule will provide vagueness and covers the wide range of rules definition. The combination of FES and SOM will provide superiority in the areas of development flexibility, fast response for Botnet attack and offer autonomous learning of inference rules in its decision support system. Plus, the knowledge base part of expert system which contains both factual knowledge and heuristic knowledge provide the capability of autonomous learning to the system. Basically, the work will consider the usability of the learning process (fuzzy inference rules and knowledge base component of expert system, and SOM), development of intrusion detection system algorithms and, the reports generation of the events to the system administrator. The rest of this paper is organized as follows: Section 2 generally summarized various machine learning approaches in Botnet detection. This section focuses on the application of machine learning algorithm for Botnet detection with the advantage, objective, challenge and limitation of several machine learning techniques. The discussion on research direction and future work is presented in section 3. Finally there will be conclusion in section 4.

2. MACHINE LEARNING IN BOTNET DETECTION SYSTEM

Hyslip [10] stated that traditional Botnet detection techniques rely on passive techniques, primarily Honey net, and that Honey net are not effective at detecting peer to peer and other decentralized Botnet. Recent research has shown machine learning are effective techniques for detecting various forms of Botnet, including peer to peerBotnet. Machine learning approaches have been used extensively in analyzing various forms of network traffic data and some of the researcher believed machine learning has more capability to handle new variant of Botnet compared to conventional approach. There are several papers discussing Botnet detection using machine learning approach. We have summarized Botnet detection approach by as a reference.

Ritu and Kaushal (2015) performed an experiment to compare different supervised machine learning techniques for determining peer to peer Botnet detection accuracy. The techniques used for the comparison are: Nearest Neighbour Classifier, Linear Support Vector Machine, Artificial Neural Network, Gaussian Based Classifier and Naive Bayes classifier. All the techniques gave accuracy between 99.10% to 100% except for Nearest Neighbour Classifier which varies from 33.3% to 97.10% for different classes of Botnet traffic. However the result might be inaccurate because these results are obtained using only fraction of dataset for preliminary study, so most probably a near perfect accuracy of 100% may not hold true when learning algorithms are applied on entire dataset.
Bilge et al. (2014) developed EXPOSURE system, the Botnet and other malicious activities detection, using genetic algorithms technique for feature selection. They conducted a controlled experiment with a large, real-world data set consisting of billions of DNS requests to prove their system works well and produce reliable detection rate of 98%. Hammadi presented a host-based behavioral approach for detecting Botnet based on correlating different activities generated by bots by monitoring function calls within a specified time window. Hammadi used dendric cell algorithm inspired by the immune system. The evaluation shows that correlating different activities generated by IRC/P2P bots within a specified time period achieves high detection accuracy. In addition, using an intelligent correlation algorithm not only states if an anomaly is present, but it also exposes the source of anomaly.

Fig. 1: Botnet Detection Approach

Apparently, the most prominent technique founds for Botnet detection is neural network and its variables. Self Organizing Map (SOM) is an unsupervised neural network and has been widely use in intrusion detection. Unfortunately, there are not so many works discuss SOM for Botnet detection, but generally in intrusion detection. However, SOM is a promising approach especially for developing autonomous Botnet detection system. Langin et al. (2009) used SOM to cluster and classify peer to peer Botnet traffic and other malignant network activity by analyzing firewall log entries. Hammadi used hexagonal Self Organizing Map for clustering, and then used for classification of new firewall log data to look for additional bots in the network. Guntuku et al. (2013) proposed and implemented a hybrid framework for detecting peer to peer Botnet in live network traffic by integrating Neural Networks with Bayesian Regularization for detection of newer and unseen Botnet in live traffic of a network. It was conclusively shown through the statistical tests that the trained Bayesian Regularization - Neural Network model is able to generalize very well and is able to predict the activity of unknown bots’ malicious activity. Thus Botnet detection activities successfully achieved with an accuracy of 99.2%.

Salvador et al. (2010) presented a new approach using neural networks, which is able to detect Botnet based on the historical traffic profiles presented by licit and illicit network applications. The results obtained show that the proposed methodology is able to achieve good identification results, being at the same time computationally efficient and easy to deploy in real network scenarios. Nogueira et al. (2010) has extended the framework proposed by Salvador and developed the Botnet Security System called BoNeSSy. Nogueira developed a Botnet detection system that is based on the collection of flow statistics, such as packet sizes and number of packets passing through a network segment, in order to identify known and unknown traffic profiles corresponding to licit and illicit applications using neural network technique. The results obtained from the performance tests showed that the system is a feasible and efficient, since it provides high detection rates with low computational overhead. (Table-1) summarized previous work on Botnet detection using machine learning techniques.
Table 1: Botnet Detection Techniques

<table>
<thead>
<tr>
<th>Author</th>
<th>Technique</th>
<th>Advantage</th>
<th>Research Objective</th>
<th>Type of Botnet</th>
<th>Detection Approach</th>
<th>High Accuracy</th>
<th>Real Time Detection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Langin et al. (2009)</td>
<td>SOM (clustering and classification)</td>
<td>SOM has self-trained capability</td>
<td>- analyzed new firewall log entries in a case study - classify similar network activity, and discovered previously unknown local P2P bot traffic and other security issues</td>
<td>P2P Botnet</td>
<td>Anomaly</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Salvador et al. (2009)</td>
<td>Artificial Neural Network</td>
<td>ANN is good in identifying Botnet traffic pattern</td>
<td>detect Botnet based on the historical traffic profiles presented by &quot;licit&quot; and &quot;illicit&quot; network applications</td>
<td>Botnet</td>
<td>Anomaly</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Nogueira et al. (2010)</td>
<td>Artificial Neural Network</td>
<td>ANN is good in identifying Botnet traffic pattern</td>
<td>develop stable, fast and efficient Botnet detection system – BoNeSSy</td>
<td>Botnet</td>
<td>Anomaly</td>
<td>✓</td>
<td>-</td>
</tr>
<tr>
<td>Basheer Al-Duwairi and Lina Al-Ebbini (2010)</td>
<td>Fuzzy logic</td>
<td>FL is a powerful technique for dealing with human reasoning and decision making process</td>
<td>develop fuzzy inference system for Botnet detection (BotDigger)</td>
<td>Botnet</td>
<td>Anomaly</td>
<td>✓</td>
<td>-</td>
</tr>
<tr>
<td>Langin et al. (2010)</td>
<td>Hexagonal SOM</td>
<td>A Vulture Fest Model of intrusion detection was combined with a hexagonal SOM to visually compare and contrast traditional, malicious, and wireless network traffic.</td>
<td>compare wireless network traffic with malicious traffic, and wireless traffic with traditional traffic, and visualize network landscape presenting Botnet in wireless and wired network</td>
<td>P2P Botnet</td>
<td>Anomaly</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Yousof Ali Abdulla Al-Hammadi (2010)</td>
<td>Dendritic Cell Algorithm</td>
<td>DCA classifies the data into normal or anomalous based on the categorization of signals. It performs multi-sensor data fusion on a set of input 'signals' which reflect some activities in the host. Applicable to work in high dimensional datasets. DCA has many novel ideas.</td>
<td>detect a bot running on the system by correlating different activities (represented by signals) and trace the suspect causing these activities from multiple sources and produce the state of the entity, either normal or anomalous.</td>
<td>IRC and P2P Botnet</td>
<td>-</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Vural and Venter (2010)</td>
<td>Fuzzy system</td>
<td>FS able to define boundaries between normal and abnormal traffic</td>
<td>determine normal and abnormal traffic -&gt; save data storage costs for spam emails</td>
<td>Spamming Botnet</td>
<td>Anomaly</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Wang et al. (2011)</td>
<td>Fuzzy Pattern-based Filtering</td>
<td>resource-efficient and can identify inactive Botnet to indicate potential vulnerable hosts</td>
<td>to identify bot-relevant domain names and IP addresses by inspecting network traces.</td>
<td>IRC, HTTP and P2P Botnet</td>
<td>Anomaly</td>
<td>✓</td>
<td>-</td>
</tr>
</tbody>
</table>
A Review of Machine Learning…

<table>
<thead>
<tr>
<th>Tarng et al. (2011)</th>
<th>J48 decision tree (classifying), K-Means divide normal and abnormal traffic flow using groups distance calculation</th>
<th>K-Means divide normal and abnormal traffic flow</th>
<th>detect abnormal P2P traffic flows</th>
<th>P2P Botnet</th>
<th>Anomaly (P2P)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Guntuku et al. (2013)</td>
<td>Neural Networks with Bayesian Regularization</td>
<td>BR-achieving better generalization of the dataset</td>
<td>detection of newer and unseen Botnet in live traffic of a network</td>
<td>P2P Botnet</td>
<td>Anomaly</td>
</tr>
<tr>
<td>Roshna and Ewards (2013)</td>
<td>Fuzzy logic and neural network</td>
<td>Time taken to learn ANFIS is very short compared to neural network. Both techniques overcome each other limitation.</td>
<td>less time to learn ANFIS reaches to the target faster than neural network. ANFIS is more preferable than neural network for handling complex problem</td>
<td>Botnet</td>
<td>Anomaly</td>
</tr>
<tr>
<td>Pragati Chandankhede (2013)</td>
<td>K-Means Algorithm (Clustering)</td>
<td>K-Means separate normal and abnormal traffic in unsupervised way</td>
<td>detect bot from the given traffic in unsupervised way</td>
<td>Bot</td>
<td>Anomaly</td>
</tr>
<tr>
<td>Bilge et al. (2014)</td>
<td>J48 decision tree algorithm (classifier), Genetic Algorithm (feature selection)</td>
<td>GA - lower false positive rate, high accuracy</td>
<td>identify and block potentially dangerous domain names</td>
<td>DNS bot</td>
<td>Anomaly (DNS)</td>
</tr>
</tbody>
</table>

Training neural networks is important and may require considerable effort. Even though neural network was the prominent technique used in intrusion detection, training neural network required extensive amount of computing time, cause to higher computing specifications requirements. FES has the potential to overcome the limitations. Most current approaches to the process of detecting intrusions utilize some forms of rule-based analysis. Expert system is the most common form of rule-based intrusion detection approaches. Most existing behavior based techniques are not able to detect and predict the Botnet as they change their structure and pattern. Adaptive Neuro Fuzzy Inference System (ANFIS) proposed new technique which trains the system for future prediction. However, the limitation of this work is the restriction of fuzzy rules and fuzzy sets for the comparison purpose. Therefore, the proposed work should be able overcome the limitations by increasing the number of rules generated using the Botnet features and information gain from Honey net. Fuzzy pattern recognition proposed by Wang et al. (2011) intend to identify bot-relevant domain names and IP addresses by inspecting network traces. The algorithm developed involves traffic reduction, feature selection and pattern recognition. Fuzziness in pattern recognition helps to detect bots which are hidden or camouflage. Performance evaluation results based on real traces show that the proposed system can reduce more than 70% input raw packet traces and achieve a high detection rate (about 95%) and a low false positive rates (0–3.08%). Furthermore, the proposed FPRF algorithm is resource-efficient and can identify inactive Botnet to indicate potential vulnerable hosts. Bot Digger proposed by Duwairi and Ebbini (2010) utilizes fuzzy logic in order to define logical rules that are mainly based on some statistical facts and important features that identify Botnet activities. The key advantage of the architecture designed in this research is that it allows the integration of wide range of traffic specifications.

Most of the machine learning techniques discussed above stated the advantages of the system by highlighting real time capability in their proposed work. Real time detection involves discovering malicious activities through live traffic of the network. Researchers focus on developing real time Botnet detection system for example Guntuku et al. (2012), Wang et al. (2011) and Salvador et al. (2010). Chandankhede proposed the new autonomous model for Botnet detection using K-means algorithm, one of the simplest unsupervised learning algorithms that solve the well-known clustering problem.

3. **RESEARCH DIRECTIONS**

We are proposing Botnet Analysis and Detection System (BADS) to analyze and detect the presence of bots. BADS will be developed using the component of...
FES and SOM. Basically, the algorithm developed for BADS comprises of usability of the learning process (generation of fuzzy inference rules and defining knowledge base content of expert system) and self-train capability of SOM which suitable for unsupervised learning to generate autonomous, real time Botnet detection system. SOM is unsupervised machine learning algorithm. In the context of Botnet detection, unsupervised machine learning algorithm commonly used for the clustering of bot-related observation. The Botnet features obtained from Honey net is the source for knowledge base components and generation of rules in FES. BADS analyzes data before connection is allowed and drop connection if bots detected. BADS provides fast response for Botnet attack and autonomous learning in its decision support system which contains both factual knowledge and heuristic knowledge. Autonomous learning enables BADS to handle and detect different Botnet variants because Botnet evolves every now and then. Thus, there are high chances Botnet might come with new variant. BADS has the capability of self-manage which requires less intervention of network administrator. The report of Botnet attack will be generated for system administrator. The system is just not aims at warning the system administrator for expected Botnet attack, but at the same time protects the system. Generally, BADS offers Botnet detection system which is autonomous, provide real time detection and reduce false alarm rate. BADS will be employed in the network to detect and countermeasure.

It is very essential to start the research by reviewing and studying about Botnet and detection methods. The clear definition of these topics will help to get better understanding in the later work. Then, the research work will focus on the exploration of FES and SOM for Botnet detection system. Finally, the research will deal with discovering the new algorithm for Botnet detection system. To achieve all the targets, the research activities will be divided into six phases as follows:

**Phase 1: Defining Botnet attack, Botnet detection system, FES and SOM**

This phase will require deep excavation on each component. The study starts with Botnet characteristic, lifecycle, type, and attack mechanism. Then, we will study Botnet detection system and its components. Finally, the focus will be on defining FES and SOM, and reviewing previous research that use both techniques in Botnet detection system.

**Phase 2: Data collection**

The advanced study of Botnet attack and the characteristic will require Honeynet to be setup. Low interaction Honeypot will be used which work by emulating certain services and operating system, and have limited interaction, meaning attacker’s activities are limited to the level of emulating provided by the Honeypot. Honeypot tool that will be use to inspect Botnet’s activities is Nepenthes. Multiple Honeypots will be run virtually in single physical host.

**Phase 3: Algorithm development**

SOM provides speed of operation in term of pattern recognition and classification, and at the same time offers self-learning and self-manage capabilities. It is capable in detecting Botnet as a threat to the system. While the knowledge-based part of expert system contains both factual knowledge and heuristic knowledge also provide the capability of autonomous learning to the system. Basically, the work will consider the usability of the learning process (SOM, and fuzzy inference rules and knowledge-based component of expert system), development of Botnet detection system algorithms and the reports generation of the events to the system administrator. The combination of the FES and SOM provides superiority in the areas of development flexibility, fast response for Botnet attack, high accuracy and offer autonomous learning of inference rules in its decision support system.

**Phase 4: Algorithm implementation**

The algorithm develop will be implemented in Snort. Then, program or plug in will be developed to evaluate the training data using determined attributes. For the testing part, module will be write that will process the equivalent libpcap format data and determine how well the rules created from the training compared to the test set. PREDICT Botnet dataset will be used for training and testing.

**Phase 5: System testing**

The open stack testbed will be setup and multiple virtual machines will be created. Several virtual machines will be used to carry out intrusion attempts and one virtual machine to run Snort.

**Phase 6: Report generation**

Log will be generated from the test. This log can be used for further investigation, for example email forensic to study spam email caused by Botnet.
4. CONCLUSION

Base on the previous work, most of the researchers focused on peer to peer Botnet detection or other decentralized Botnet rather than centralized Botnet. Basically, some of the researchers believed, the issues in centralized Botnet have been successfully solved by previous researchers. The shifted of the trends have been stated by Ritu and Kaushal (2015) . Furthermore, peer to peer Botnet as the bigger challenge since it was harder to locate, detect, shutdown, monitor and hijack, thus extra effort is needed to encounter them. The previous works also showed that machine learning techniques are capable to handle high amount of traffic and provide real time detection. The detection approach by previous researchers mainly focuses on anomaly based detection. Anomaly based detection capable to detect new attack because the system already developed the pattern (predefined behavior) of normal traffic. Any anomalies are categorized as abnormal traffic. They also claimed their work has increased the detection accuracy based on experiments conducted. A top of previous works reviewed, we have proposed Botnet analysis and detection technique using FES and SOM. The data obtained from this work can be used for further research in email forensic to study spam email because spam email is one of the major Botnet contribution. This system is an online detection system that analyzes data before connection is allowed. The connection will be dropped if it detects intrusion. The system is just not aims at warning the system administrator for expected Botnet attack, but at the same time protects the system.

ACKNOWLEDGEMENT

This work is supported by MOSTI FRGS grant number 203/PKOMP/6711426, School of Computer Science, University Sains Malaysia (USM).

REFERENCES:


Guntuku S. C., P. Narang and C. Hota (2013), Real-time Peer-to-Peer Botnets Detection Framework based on Bayesian Regularized Neural Network, Networking and Internet Architecture Section, Cornell University Library


Vural I. and H. S. Venter (2010), Using Network Forensics and Artificial Intelligence Techniques to Detect Botnets on an Organizational Network, 2010 Seventh International Conference on Information Technology, 12-14 Las Vegas, Nevada USA.

