IOSR Journal of Computer Engineering (IOSRJCE)
ISSN: 2278-0661 Volume 4, Issue 3 (Sep.-Oct. 2012), PP 01-04
www.iosrjournals.org

Utilization Data Mining to Detect Spyware

[1]Parisa Bahraminikoo, [2]Mehdi Samiei yeganeh, [3]G.Praveen Babu
[1, 2](M.Tech.(S/w. Eng.), [3](Associate Professor School of Information Technology, Jawaharlal Nehru Technological University, Iran)

Abstract: *Malicious software (malware) is any software that gives partial to full control of your computer to do whatever the malware creator wants. Malware can be a virus, worm, Trojan, adware, spyware, root kit, etc.Spyware is a type of malware (malicious software) installed on computers that collects information about users without their knowledge. In the year 1956, Artificial Intelligence (AI) was established at Dartmuth College during a conference. The technology developed so much that it started involving many other branches of engineering such as electronics, robotics etc. This eventually led to much more complex and smart machinery involving Artificial Intelligence. With the development of malware detection systems and Artificial Intelligence, as a new technology for them, Artificial Intelligence has been applied in anti-virus engines. There are several AI approaches that applied in spyware detection systems such as Artificial Neural Networks, Heuristic Technology and Data Mining (DM) Technique. Heuristic-based Detection performs well against known Spyware but has not been proven to be successful at detecting new spyware. In this paper we focus on DM-based malicious code detectors using Breadth-First Search (BFS) approach, which are known to work well for detecting viruses and similar software. BFS is a strategy for searching in a tree when search is limited to essentially two operations: (a) visit and inspect a node of a tree; (b) gain access to visit the nodes that are neighbor to currently visited node. The BFS begins at a root node and inspect all the neighboring nodes. Then for each of those neighbor nodes in turn, it inspects their neighbor nodes which were unvisited, and so on.*

Keywords: *Spyware, Artificial intelligence, Data mining, Breadth-First Search.*

I. Introduction

As the application of computer and Internet is more popular, it provides a convenient way to share the information among different people; however it also gives chances to malware activities, such as propagating malicious programs, including computer viruses [1]. Programs that have the potential to break the privacy and security of a system can be labeled as Privacy Invasive Software [2]. These programs include: spyware, adware, Trojans, greyware and backdoors [3]. Spyware is a type of malware (malicious software) installed on computers that collects information about users without their knowledge. The presence of spyware is typically hidden from the user and can be difficult to detect. Some spyware, such as keyloggers, may be installed by the owner of a shared, corporate, or public computer intentionally in order to monitor users. While the term spyware suggests software that monitors a user's computing, the functions of spyware can extend beyond simple monitoring. Spyware can collect almost any type of data, including personal information like Internet surfing habits, user logins, and bank or credit account information. Spyware can also interfere with user control of a computer by installing additional software or redirecting Web browsers. Some spyware can change computer settings, which can result in slow Internet connection speeds, un-authorized changes in browser settings, or changes to software settings.

The goal of spyware is generally not to cause damage or to spread to other systems. Instead, spyware programs monitor the behavior of users and steal private information, such as keystrokes and browsing patterns. This information is then sent back to the spyware distributors and used as a basis for targeted advertisement (e.g., pop-up ads) or marketing analysis [4]. AI is the science and engineering of making intelligent machines, especially intelligent computer programs [1]. AI is set to play an important role in our lives. Researchers produce new products which duplicate intelligence, understand speech, beat the opponent chess player, and acting in complex conditions. The major problems of Artificial Intelligence include qualities such as knowledge, planning, learning, reasoning, communication, perception and capability to move and control the objects [6].The aim of Artificial Intelligence is to develop the machines to perform the tasks in a better way than the humans [5]. As following we will describe the main application of artificial intelligence that is applied in spyware detection systems. The rest of this paper is organized as follows: section 2 briefly describes the Heuristic Technology, Section 3 explains Data mining Technique and section 4 briefly describes the Neural Network Technology.

II. Heuristic Technology

At the present, the first and main application for spam filtering technique based on artificial intelligence is Heuristic Technology. Current anti-spyware tools make use of signature-based methods by using specific features or unique strings extracted from binaries or heuristic-based methods by using on the basis of rules written by experts who define behavioral patterns as approaches against spyware. These approaches are often considered ineffective against new malicious code [7, 8].

Heuristic Technology means "the ability of self-discovery" or "the knowledge and skills that use some methods to determine", and intelligently analyze codes to detect the unknown virus by some rules while scanning [9].Heuristics are used quite often in Artificial Intelligence based research. They are new and constantly being refined by most antivirus companies over the last five years or so. Computational they are much faster than signature based techniques. Heuristics look for a set of characteristics within a file in order to determine whether or not it may be a potential threat. In a sense, heuristic anti-malware attempts to apply the processes of human analysis to an object. In the same way that a human malware analyst would try to determine the process of a given program and its actions, heuristic analysis performs the same intelligent decision-making process, effectively acting as a virtual malware researcher. As the human malware analyst learns more from and about emerging threats he or she can apply that knowledge to the heuristic analyzer through programming, and improve future detection rates. Antivirus software may use one or several techniques to proactively detect malware. The main essence of each method is to analyze the suspicious file's characteristics and behavior to determine if it is indeed malware.

The main concern with heuristic detection is that it often increases false positives. False positives are when the antivirus software determines a file is malicious (and quarantines or deletes it) when in reality it is perfectly fine and/or desired. Because some files may look like viruses but really aren't, they are restricted and stopped from working on your computer. In Heuristics based detection we can use Generic Signature, This technique is particularly designed to locate variations of viruses. Several viruses are re-created and make themselves known by a variety of names, but essentially come from the same family (or classification). Genetic detection uses previous antivirus definitions to locate these similar "cousins" even if they use a slightly different name or include some unusual characters. The best way to illustrate this idea is with identical twins. They may have slightly different fingerprints, but their DNA is identical.

III. Neural Network Technology

An Artificial Neural Network (ANN) (Bishop, 1995) is an information processing paradigm that is inspired by the way biological nervous systems (i.e., the brain) are modeled with regard to information processing [13]. A neural network is designed to simulate a set of neurons, usually connected by synapses. In the nervous system, a synapse is a structure that permits a neuron to pass an electrical or chemical signal to another cell. Just as in biological systems, learning involves adjustments to the synaptic connections that exist between the neurons. Neural networks can differ on: the way their neurons are connected; the specific kinds of computations their neurons do; the way they transmit patterns of activity throughout the network; and the way they learn including their learning rate. Neural networks are being applied to an increasing large number of real world problems [14].

The neural network is configured for a specific application, such as data classification or pattern recognition, through a learning process called training [15]. In [16] has introduced how to use Single layer neural classifier to detection boot viruses, and the generic virus detector was incorporated into IBM Antivirus in May, 1994.Its structure has been shown in Fig 1.

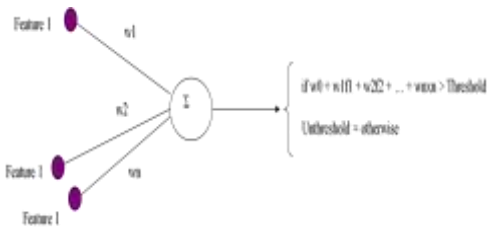

Figure 1: Single layer neural classifier

William Arnold and Gerald Tesauro constructed multiple neural network classifiers which can detect unknown Win32 viruses by combining the individual classifier outputs using a voting procedure, following a technique described in previous work (Kephart et al, 1995) on boot virus heuristics [17]. And the system has 508 achieved effectively.Authors of [18] presented a new rule generation method from neural networks formed using a genetic algorithm (GA) with virus infection and deterministic mutation. This method can extract rules (regularities) for a pattern classification and chaotic system identification by using the same system [1].

IV. Data mining Technique

With the rapid development of Information Technology, the rapid growth of data has exceeded the ability of the manual processing of data. So how to help people to extract the general knowledge from the mass of data has become more and more important. In order to implement it, data mining technique is put forward and soon becomes an active research direction. Data mining analyzes the observed sets to discover the unknown relation and sum up the results of data analysis to make the owner of data to understand. Data Mining Algorithm that is from Statistics, Pattern Identification, Machine Learning [5, 11], and Database and so on, has developed comprehensive [1].

In [10] presented a data-mining framework that detects new, previously unseen malicious executables accurately and automatically. The 2001 data mining study of malicious code [8] used. Three types of features, i.e., Dynamic-link Library resource information, consecutive printable characters (strings) and byte sequences.In [12] presented a network virus precaution system based on data mining shown in Fig 2.It can detect the abnormal connecting behavior of network in real-time to discover the trace of worm virus, especially the precaution action to the new worm virus to make administrator to adopt corresponding measure to avoid tremendous loss.

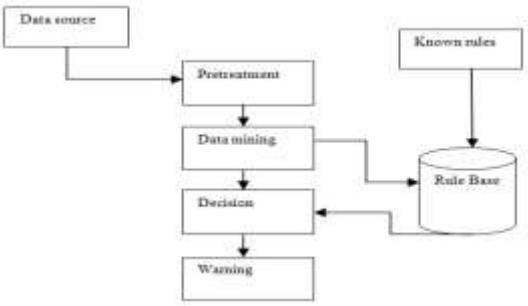

Figure 2: The structure of warning system

Data mining techniques perform better than traditional techniques such as signature-base detection and Heuristic-based detection. The focus of our analysis is executable files for the Windows platform. We use the Waikato Environment for Knowledge Analysis (Weka) to perform the experiments. Weka is a suite of machine learning algorithms and analysis tools, which is used in practice for solving data mining problems, first, we extract features from the binary files We extract the features by using the Common Feature-based Extraction (CFBE). The purpose of employing this approach is to evaluate two different techniques that use different types of data representation, i.e., the occurrence of a feature and the frequency of a feature. CFBE method are used to obtain Reduced Feature Sets (RFSs) which are then used to generate the ARFF files and includes instances from the frequency range of 50-80. And we then apply a feature reduction method in order to reduce data set complexity. In experiments for the detection of malware, sequences of bytes extracted from the hexadecimal dump of the binary files have been represented by n-grams [3]. Finally, we convert the reduced feature set into the Attribute-Relation File Format (ARFF). ARFF files are ASCII text files that include a set of data instances, each described by a set of features [3].

Data mining base malicious approach have proven to be successful in detecting viruses and worms. Overall accuracy of 90.5% is achieved with the BF-tree algorithms.

We evaluate each learning algorithm by performing cross-validation tests to ensure that the generated classifiers are not tested on the training data. From the response of the classifiers the relevant confusion matrices were created. Four metrics define the elements of the matrix: True Positives (TP), False Positives (FP), True Negatives (TN) and False Negatives (FN).

Metric	Abbreviation	Meaning
True Positives	TP	Number of correctly identified benign programs.
False Positives	FP	Number of wrongly identified Spyware programs.
True Negatives	TN	Number of correctly identified Spyware programs.
False Negatives	FN	Number of wrongly identified benign programs.

Table 1 Evaluation metrics

We shall now demonstrate how multi-criteria metrics can be used as an approach to trade-off some of the important aspects of EULA (End User License Agreement) classification [19]. The performance of each classifier was evaluated using the true positive rate, false positive rate and overall accuracy which are defined as follows:

True Positive Rate (TPR): Percentage of correctly identified benign programs (TP / TP+FN)
False Positive Rate (FPR): Percentage of wrongly identified Spyware programs (FP / TN+FP)
Overall Accuracy (ACC): Percentage of correctly identified programs (TP+TN / TP+TN+FP+FN)

Table 2 show that Using the feature set produced by the CFBE feature selection method for n = 4, the BFT decision tree classifier achieves the highest accuracy results in Frequency Range 50-80.

Algorithm	Type	TPR	FPR	ACC
BFT Frequency Range 50-80	Trees	0.992 0.977	0.731 0.730	89.896 (5.104) 88.5222 (5.899)
Random Forest Frequency Range 50-80	Trees	0.979 0.960	0.665 0.720	89.489 (5.520) 87.077 (6.477)
Naive Bayes Frequency Range 50-80	Bayes	0.973 0.916	0.730 0.705	88.209 (6.174) 83.703 (8.202)
SMO Frequency Range 50-80	Function	0.935 0.946	0.665 0.515	86.566 (8.130) 88.5222 (7.530)

Table 2 Comparison of Algorithms for N-gram size = 4

V. Conclusion

Spyware technique has become the most important Prevention technique. With this technologies, the system can detect virus invasion in real-time, and enlarge security management capacity of system administrators to enhance the integrity of the infrastructure of information security.

With development of Artificial Intelligence technology has provided new methods and ideas for spyware detection system. Intergraded spyware detection with AI will greatly improve the performance of the existing spyware detection system, promote more effective artificial intelligence algorithms to be proposed, and be applied in the popular detection field .The main objective of the present work is to establish a method in Spyware detection research using data mining techniques. These techniques are used for information retrieval and classification. Data mining-based malicious code detectors have been proven to be successful in detecting clearly malicious code, e.g., like viruses and worms. Results from different studies have indicated that data mining techniques perform better than traditional techniques against malicious code. However, spyware has not received the same attention from researchers but it is spreading rapidly on both home and business computers. Overall accuracy of 90.5% is achieved with the BF-tree algorithms.

References

[1]. Review on the application of Artificial Intelligence in Antivirus Detection System", Xiao-bin Wang Guang-yuan Yang Yi-chao Li Dan Liu.

[2]. M. Boldt and B. Carlsson, "Privacy-invasive software and preventive mechanisms," 2nd International Conference on Systems and Networks Communications, (ICSNC 2006), Oct. 28- Nov.2, IEEE Computer Society.

[3]. Detection of Spyware by Mining Executable Files" Raja Khurram Shazhad, Syed Imran Haider, Niklas Lavesson, 2010International Conference onAvailability,Reliability and Security.

[4]. Behavior-based Spyware Detection" Engin Kirda and Christopher Kruegel, Greg Banks, Giovanni Vigna, and Richard A. Kemmerer.

[5]. Review on use of Reinforcement Learning in Artificial Intelligence" Mehdi Samieiyeganeh , Parisa Bahraminikoo ,G.Praveen Babu ,12th International Conference of Science and Technology Impact on Development and Justice held at Maulana Azad National Urdu University, Hyderabad, India, on 7th & 8th February, 2012.

[6]. Chuck Williams. "A BRIEF INTRODUCTION TO ARTIFICIAL INTELLIGENCE", 10.0 109 83 IEEE.

[7]. C. D. Bozagac, "Application of Data Mining based Malicious Code Detection Techniques for Detecting new Spyware", White paper, Bilkent University, 2005.

[8]. M. Schultz, E. Eskin, F. Zadok, and S. Stolfo, "Data mining methods for detection of new malicious executables," Proceedings of IEEE Symposium on Security and Privacy, 14-16 May 2001, Los Alamitos,

[9]. Xianwei Zeng, Zhijun Zhang, and Zhi Zhang, "Heuristic skill of computer virus analysis based on virtual machine," Computer Applications and Software, Vol. 22(9), 2005, pp. 125-126.

[10]. Matthew G. Schultz, Eleazar Eskin, Erez Zadok, and Salvatore I. Stolfo, "Data Mining Methods for Detection of New Malicious Executables," The 2001 IEEE Symposium on Security and Privacy, Oakland. CA, 2001, pp.38-49.

[11]. I.H. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, 2nd ed. Morgan.

[12]. Yufeng Yang, "The Network Virus Precaution System Based on Data Mining," Journal of Shaoguan University, Vol. 26(12), 2005, pp. 31-33.

[13]. Detection of Unknown Computer Worms based on Behavioral Classification of the Host", Robert Moskovitch, Yuval Elovici, Lior Rokach.

[14]. Review on Artificial Intelligence and Artificial neural networks" Mehdi Samieiyeganeh , ParisaBahraminikoo,G.PraveenBabu, International Conference on Computing,Communications, Systems & Aeronautics (ICCCSA-12) , Organized by Malla Reddy College of Engineering & Technology From March 30-31,2012. Hyderabad, India.

[15]. Artificial Intelligence: Neural Networks Simplified "Indranarain Ramlall1 University of Technology, Mauritius.

[16]. Kephart, J.O., "Biologically inspired defenses against computer viruses," Proceedings of International Joint Conference on Artificial Intelligence, 1995, pp. 985-96.

[17]. R. Moskovitch, C. Feher, N. Tzachar, E. Berger, M. Gitelman, S. Dolev, and Y. Elovici, "Unknown malcode detection using OPCODE representation," 1st European Conference on Intelligence and Security Informatics, (EuroISI 2008), 3-5 Dec., Berlin, Germany: Springer-Verlag, pp. 204-215.

[18]. Y. Elovici, A. Shabtai, R. Moskovitch, G. Tahan, and C. Glezer., "Applying Machine Learning Techniques for Detection of Malicious Code in Network Traffic", Proceedings of the 30th annual German conference on Advances in Artificial Intelligence, KI 2007, 10-13.

[19]. Niklas Lavesson , Martin Boldt , Paul Davidsson ,Andreas Jacobsson, "Learning to detect spyware using end user license agreements", Springer-Verlag London Limited 2009.