International Journal Online of Sports Technology & Human Engineering

ISSN: 2349-0772| Volume III Issue V October 2016

www.ijosthe.com

A Review on Malware Analysis by using an Approach of Machine Learning Techniques

Aayushi Priya[1], Kajol Singh[2] , Rajeev Tiwari[3]

Department of Computer Science and Engineering, Bhopal, India[1,2,3] aayu.rec@gmail.com[1], kajolsingh01415@gmail.com[2] , rajeevrnt@yahoo.in[3]

Abstract : In the Internet age, malware (such as viruses, trojans, ransomware, and bots) has posed serious and evolving security threats to Internet users. To protect legitimate users from these threats, anti-malware software products from different companies, including Comodo, Kaspersky, Kingsoft, and Symantec, provide the major defense against malware. Unfortunately, driven by the economic benefits, the number of new malware samples has explosively increased: anti-malware vendors are now confronted with millions of potential malware samples per year. In order to keep on combating the increase in malware samples, there is an urgent need to develop intelligent methods for effective and efficient malware detection from the real and large daily sample collection. One of the most common approaches in literature is using machine learning techniques, to automatically learn models and patterns behind such complexity, and to develop technologies to keep pace with malware evolution. This survey aims at providing an overview on the way machine learning has been used so far in the context of malware analysis in Windows environments. This paper gives an survey on the features related to malware files or documents and what machine learning techniques they employ (i.e., what algorithm is used to process the input and produce the output). Different issues and challenges are also discussed.

Keywords: Windows log file, Executable files, Malware analysis, Machine learning.

1. Introduction

As computers and Internet are increasingly ubiquitous, the Internet has been essential in everyday life. It has been reported by the ITU (International Telecommunication Union) that the number of Internet users worldwide, who always use Internet services such as e-banking, e-commerce, instant communication, education, and entertainment, has reached 2.92 billion as of 2014 [1]. Just like the physical world, there are people with malicious intentions (i.e., cyber-criminals) on the Internet. They try to take advantage of legitimate users and benefit themselves financially. Malware (short for malicious software), is a generic term widely used to denote all different types of unwanted software programs. These programs include viruses, worms, trojans, spyware, bots, rootkits, ransomware, and so on. Malware has been used by cybercriminals as weapons in accomplishing their goals. In particular, malware has been used to launch a broad range of security attacks, such as compromising computers, stealing confidential information, sending out spam emails, bringing down servers, penetrating networks, and crippling critical infrastructures. These attacks often lead to severe damage and significant financial loss. To put this into perspective, according to a recent report from Kaspersky Lab, up to $1 billion was stolen in roughly 2 years from financial institutions worldwide due to malware attacks [2]. In addition, Kingsoft reported that the average number of infected computers per day was between 2-5 million [3].

Numerous malware attacks have posed serious and evolving security threats to Internet users. To protect legitimate users from these threats, anti-malware software products from different companies provide the major defense against malware, such as Comodo, Kaspersky, Kingsoft, and Symantec. Typically, the signature-based method is employed in these widely-used malware detection tools to recognize various threats. A signature is a short sequence of bytes, which is often unique to each known malware, allowing newly encountered files to be correctly identified with a small error rate [2].

However, due to the economic benefits, malware authors quickly developed automated malware development toolkits. These toolkits use techniques, such as instruction virtualization, packing, polymorphism, emulation, and metamorphism to write and change malicious codes that can evade the detection. These malware creation toolkits greatly lower the novice attackers' barriers to enter the cyber-crime world (allowing inexperienced attackers to write and customize their own malware samples) and lead to a massive proliferation of new malware samples due to their wide availability. As a result, malware samples have been rapidly gaining prevalence and have spread and infected computers at an unprecedented rate around the world. In 2008, Symantec reported that the release rate of malicious programs and other unwanted codes might exceed that of benign software applications.

This suggests that traditional signature-based malware detection solutions may face great challenges since they can be outpaced by the malware writers. For example, according to Symantec's report, about 1.8 million malware signatures were released in 2008, which resulted in 200 million detections per month. In 2013, the suspicious files collected by the anti-malware lab of Kingsoft reached 120 million, 41.26 million (34%) of which were detected as malware. While many malware samples have been detected and blocked, a large number of malware samples (e.g., the so-called "zero-day" malware [4]) have been generated or mutated and they tend to evade traditional signature-based anti-virus scanning tools. This has prompted the anti-malware industry to rethink their malware detection methods, as these approaches are mainly based on variants of existing signature-based models.

2. Overview of Malware and Anti-Malware Industry

Malware is the software program that deliberately meets the harmful intent of malicious attackers [5]. It has been designed to achieve the goals of attackers. These goals include disturbing system operations, gaining access to computing system and network resources, and gathering personal sensitive information without user's permission. As a result, malware often creates a menace to the integrity of the hosts, availability of the Internet, and the privacy of the users. Malware can reach the systems in different ways and through multiple channels. These different ways are summarized below:

- The vulnerable services over the network allow malware to infect accessible systems automatically.
- The downloading process from the Internet: It has been shown that 70–80% of the malware come from popular websites. By exploiting the web browser's vulnerability, a drive-by download is capable to fetch malicious codes from the Internet first and then execute the codes on the victims' machines.
- The attackers can also lure the victims into deliberately executing malicious codes on their machines. Typical examples include asking the users to install a provided "codec" to watch the movies which are hosted on the website, or clicking/opening images attached to spam emails.

In some cases, malware may only affect the system performance and create overload processes. In case of spying, malware hides itself in the system, steals critical information about the computer, and sends information to the attackers. To protect legitimate users from the malware attacks, the major defense is the software products from anti-malware companies.

However, the more successful the anti-malware industry becomes in detecting and preventing the attacks, the more sophisticated malware samples may appear in the wild. As a result, the arms race between malware defenders and malware authors is continuing to escalate. In the following sections, we introduce the taxonomy of malware, elaborate the development of malware industry, and then describe the progress of malware detection.

3. Types of Malware

Based on the different purposes and proliferation ways, malware can be categorized into various types. This section provides a brief overview of most common types of malware, such as viruses, worms, trojans, spyware, ransomware, scareware, bots, and rootkits. Viruses: A virus is a piece of code that can append itself to other system programs, and when executed, the affected areas are "infected" [6]. Viruses cannot run independently since they need to be activated by their "host" programs [7]. The Creeper virus written by Bob Thoma was an experimental self-replicating program, which was first detected in the early 1970s [8].
Worms: Unlike a virus which requires its "host" program be run to activate it, a worm is a program that is able to run independently. Note that a worm can propagate a fully working copy of itself to other machines. The Morris worm was the first publicly known program instance that exhibited worm-like behavior. During the Morris appeal process, based on the estimate of the U.S. Court of Appeals, the cost of removing the Morris worms was around

$100 million. The infamous worms, such as Love Gate, CodeRed, SQL Slammer, MyDoom, and Storm Worm, have successfully attacked tens of millions of Windows computers and caused great damages.

Trojans: Compared with a worm, which is apt to propagate a fully working version of itself to other machines, Trojan is a software program that pretends to be useful but performs malicious actions in the backend [8]. One of the recent notable trojans, Zeus (also called Zbot) is capable of carrying out many malicious and criminal tasks. Zeus has often been used to steal banking-related information by keystroke logging and form grabbing [9]. In June 2009, security company Prevx discovered that over 74,000 FTP accounts had been compromised by Zeus on the websites of many companies (including ABC, Amazon, BusinessWeek, Cisco, NASA, Monster.com, Oracle, Play.com, and the Bank of America).

Spyware: Spyware is a type of malicious program that spies on user activities without the users' knowledge or consent [9]. The attackers can use spyware to monitor user activities, collect keystrokes, and harvest sensitive data (e.g., user logins, account information). Ransomware: Ransomware is one of the most popular malware in recent years, which installs covertly on a victim's computer and executes a cryptovirology attack that adversely affects it [10]. If the computer is infected by this malware, the victim is demanded to pay a ransom to the attackers to decrypt it. Scareware: Scareware is a recent type of malicious file that is designed to trick a user into buying and downloading unnecessary and potentially dangerous software, such as fake antivirus protection [11], which has posed severe financial and privacy-related threats to the victims.

Bots: A bot is a malicious application that allows the bot master to remotely control the infected system. Typical spread methods of bots are exploiting software vulnerabilities and employing social engineering techniques. Once a system has been infected, the bot master can install worms, spyware, and trojans, and transform the individual victimized systems into a botnet. Botnets are widely used in launching Distributed Denial of Service (DDoS) attacks [11], sending spam emails, and hosting phishing fraud. Agobot and Sdbot are two of the most notorious bots.

Rootkits: A rootkit, a stealthy type of software, is designed to hide certain processes or programs and enable continued privileged access to computers [12]. Rootkit techniques can be used at different system levels: they can instrument Application Programming Interface (API) calls in user-mode or tamper with operating system structures as a device driver or a kernel module.

Hybrid Malware: Hybrid malware combines two or more other forms of malicious codes into a new type to achieve more powerful attack functionalities. Some other categories of commonly encountered Internet pests can also be a nuisance to computer users, such as "Spamware," "Adware," and the like. Actually, these typical types of malware are not mutually exclusive. In other words, a particular malware sample may belong to multiple malware types simultaneously.

4. Literature Review

Despite the significant improvement of cyber security mechanisms and their continuous evolution, malware are still among the most effective threats in the cyber space. Malware analysis applies techniques from several different fields, such as program analysis and network analysis, for the study of malicious samples to develop a deeper understanding on several aspects, including their behavior and how they evolve over time. Within the unceasing arms race between malware developers and analysts, each advance in security technology is usually promptly followed by a corresponding evasion. Part of the effectiveness of novel defensive measures depends on what properties they leverage on. For example, a detection rule based on the MD5 hash of a known malware can be easily eluded by applying standard techniques like obfuscation, or more advanced approaches such as polymorphism or metamorphism [13].

For a comprehensive review of these techniques, refer to Ye et al. [14]. These methods change the binary of the malware, and thus its hash, but leave its behavior unmodified. On the other side, developing detection rules that capture the semantics of a malicious sample is much more difficult to circumvent, because malware developers should apply more complex modifications. A major goal of malware analysis is to capture additional properties to be used to improve security measures and make evasion as hard as possible. Machine learning is a natural choice to support such a process of knowledge extraction. Indeed, many works in literature have taken this direction, with a variety of approaches, objectives and results. Some of important contribution in the field of malware detection are discussed in table I.

Table I: Existing Contributions in Malware Analysis using Machine Learning Approach

Author Name	Description	Result and Conclusion
Anderson et al. [5]	SVM	By combining both static and dynamic analysis, it was tested on a dataset of 780 malware and 776 benign instances giving an accuracy of 98.07%.
Santos et al. [6]	DT, kNN, BN, and SVM	It has been found that the hybrid approach enhanced the performance of both approaches when run separately, based on the static and dynamic analysis.
Islam et al. [7]	DT, Random Forest, SVM, and Instance-based Classifier	By combining both static and dynamic analysis, the obtained results showed that meta-Random Forest performed best.
Karampatziakis et al. [8]	Regression Classifier	Based on the graphs induced by file relationships, the system's detection accuracy could be significantly improved using the proposed method, particularly with low false positive rates
Tamersoy et al. [9]	Back Propagation Neural network	Based on the file-to-file relation graphs, the developed system attained early labeling of 99% of benign files and 79% of malicious files.
Saxe and Berlin [10]	Deep Neural Network and Bayesian Calibration Model	Using the statically extracted features, their system achieves a 95% detection rate at 0.1% false positive rate, based on more than 400,000 software binaries
Hardy et al. [11]	DL Architecture using SAEs	Based on the extracted Windows API calls, the developed deep learning framework outperformed ANN,

5. Conclusion

In recent years, a few research efforts have been conducted on surveys of data-mining based malware detection methods. The authors reviewed the malware propagation, analysis and detection and surveyed the feature representation and classification methods for malware detection. Many researchers surveyed automated dynamic malware analysis techniques and tools. In this paper, we not only overview the development of malware and antimalware industry and present the industrial needs on malware detection, but also provide a comprehensive study on data-mining-based methods for malware detection based on both static and dynamic representations as well as other novel features. Furthermore, we also discuss the additional issues and challenges of malware detection using data mining techniques and finally forecast the trends of malware development. In these methods, the process of malware detection is generally divided into two steps: feature extraction and classification/clustering. In order to achieve the best detection performance in real applications, it is often better to have enough training samples with balanced distributions for both classes (malware and benign files).

References

[1] ITU. 2014. ITU releases 2014 ICT figures. Retrieved from https://www.itu.int/net/pressoffice/press_releases/ 2014/23.aspx.

[2] Kaspersky. 2015. The Great Bank Robbery. Retrieved from http://www.kaspersky.com/about/news/virus/2015/ Carbanakcybergangsteals-1-bn-USDfrom-100-financial-institutions-worldwide.

[3] Kingsoft. 2016. 2015-2016 Internet Security Research Report in China. Retrieved from http://cn.cmcm.com/ news/media/20160114/60.html.

[4] Wikipedia. 2017, Zero-day (computing). Retrieved from https://en.wikipedia.org/wiki/Zero-day_(computing).

[5] Blake Anderson, Curtis Storlie, and Terran Lane. 2012. Improving malware classification: Bridging the static/dynamic gap. In Proceedings of 5th ACM Workshop on Security and Artificial Intelligence (AISec).

[6] Igor Santos, Jaime Devesa, Felix Brezo, Javier Nieves, and Pablo Garcia Bringas. 2013. OPEM: A staticdynamic approach for machine learning based malware detection. In Proceedings of International Conference CISIS-ICEUTE, Special Sessions Advances in Intelligent Systems and Computing.

[7] Rafiqul Islam, Ronghua Tian, Lynn M. Batten, and Steve Versteeg. 2013. Classification of malware based on integrated static and dynamic features. Journal of Network and Computer Application 36, 2 (2013), 646–656.

[8] Nikos Karampatziakis, Jack W. Stokes, Anil Thomas, and Mady Marinescu. 2013. Using file relationships in malware classification. In Proceedings of the Conference on Detection of Intrusions and Malware and Vulnerability Assessment.

[9] Acar Tamersoy, Kevin Roundy, and Duen Horng Chau. 2014. Guilt by association: Large scale malware detection by mining filerelation graphs. In Proccedings of ACM International Conference on Knowledge Discovery and Data Mining (ACM SIGKDD). [10] Joshua Saxe and Konstantin Berlin. 2015. Deep neural network based malware detection using two dimensional binary program features. In Proceedings of the 10th International Conference on Malicious and Unwanted Software (MALWARE).

[11] William Hardy, Lingwei Chen, Shifu Hou, Yanfang Ye, and Xin Li. 2016. DL4MD: A deep learning framework for intelligent malware detection. In Proceedings of the International Conference on Data Mining (DMIN).

[12] J. Gardiner, S. Nagaraja, On the security of machine learning in malware c&c detection: A survey, ACM Comput. Surv. 49 (3) (2016) 59:1–59:39.

[13] A. Souri, R. Hosseini, A state-of-the-art survey of malware detection approaches using data mining techniques, Human-centric Computing and Information Sciences 8 (1) (2018) 3.

[14] Y. Ye, T. Li, D. Adjeroh, S. S. Iyengar, “A survey on malware detection using data mining techniques”, ACM Computing Surveys (CSUR), vol 50, issue 3, 2017, pp. 41.