

Detecting Unknown Malicious Executables Using Portable Executable Headers

Tzu-Yen Wang
Dept. of Computer Sci.
National Chiao Tung Univ.
Hsinchu, Taiwan
tzu-yen.wang@hotmail.com

Chin-Hsiung Wu*
Dept. of Info. Tech. & Comm.
Shih Chien Univ. Kaohsiung Campus,
Kaohsiung, Taiwan
chwu@mail.kh.usc.edu.tw

Chu-Cheng Hsieh
Dept. of Computer Sci.
Univ. of California,
Los Angeles, USA
chucheng@ucla.edu

Abstract—Even though numerous kinds of anti-virus software packages have been used for many years, previously unseen malware is still a serious threat to computer and information system. By analyzing portable executable header entries of executables, a malware detection model which consists of four stages: attribute extraction, attribute binarization, attribute elimination, and feature selection and classifier training was carried out in this study. First, we collected header entries from all executables in our dataset and viewed each entry as a potential attribute. Second, information gain and gain ratio were used to binarize numerical and nominal attributes. Next, useless and redundant attributes were eliminated in the third stage. Finally, by using support vector machine which is a classification algorithm of conspicuous generalization ability, feature selection was simultaneously performed with classifier training to reduce the number of attributes and retain the performance of classifier in a cost-effective. We evaluated our model by 1,908 benign programs and 7,863 malicious files (virus, email worm, trojan and backdoor) and estimated its generalization ability by cross validation. The experiment results showed that our model had promising performance for detecting virus and email worm.

Keywords- *previously unseen malware; information gain; gain ratio; feature selection; support vector machine*

I. INTRODUCTION

Malware, which refers to codes performing malicious activities such as crashing computer systems, jamming network flows and stealing personal information, etc., has been a notorious threat to information security for nearly 30 years. Commercial anti-virus packages, which have been widely used for malware detection, seemed to be reliable to fight against malicious codes. In general, these commercial programs are based on signatures generated by both static analysis and dynamic analysis. Static analysis usually constructs signatures by segmenting the binary code sequences from known malware while dynamic analysis executes the malware in a safe environment and observes its own behaviors. By matching malware to those predefined patterns, anti-virus software could accurately detect known malicious codes without any false alarm. However, according to the report from AV-Comparatives.org¹ in May

2008, the proactive detection abilities of sixteen anti-virus software packages were examined by detecting new malicious code without updating their virus definitions. The examination results were utterly unsatisfying that none of them had proactive detection rate higher than 60%. Therefore, the main problem of current signature-based method is that it is ineffective against zero-day attacks launched by previously unseen malicious files or even polymorphic ones due to the lack of virus definitions of them. Every minute people wait for an update of virus definitions is another minute that their computer is vulnerable to any vicious damage. Hence it is no doubt that malware is remaining a noticeable security challenge and the needs of continuing improvements in unknown malware detection methods are crucial.

The goal of this research is to propose a novel model to detect unknown malware in Portable Executable (PE) [13] format since it is the most widely distributed type of malicious codes. To fulfill the purpose of preventing malware from crashing users' computers, we adopted static analysis method to exploit the information of PE headers entries. With the assistance of Support Vector Machine (SVM) [20], which is a classification algorithm of conspicuous generalization ability, these attributes could be used to construct an effective and reliable model to distinguish malicious executables from benign programs.

The rest of this paper is divided into five sections. The first section is devoted to some related studies in the area of unknown malware detection. After which the second section sketches the content of PE header entries. This is followed by the methodology and procedures of our malware detection model. The experiment results for detecting unknown malware are illustrated in the next section. Finally, some concluding remarks and future works are made in the final section.

II. RELATIVE WORKS

Owing to the lack of proactive detection ability of nowadays commercial virus scanners, considerable concern has arisen over the field of unknown malware detection. To detect previously unseen malware in advance, we only focused on the use of static analysis methods and data mining techniques in this study and gave some literature review which is related to our work in this section.

Schultz and Zadok [15] presented some kinds of features such as Dynamic Linked Libraries (DLLs), Application

*Corresponding author. This work was partially supported by the National Science Council under the contract no. NSC-97-2221-E-158-003.

¹ www.av-comparatives.org/