Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 78 (2016) 667 – 674

Procedia Computer Science

International Conference on Information Security & Privacy (ICISP2015), 11-12 December 2015, Nagpur, INDIA

Grouping the executables to detect malwares with high accuracy

Sanjay K. Sahay[a], Ashu Sharma[b,*]

[a]*Assistant Professor, Dept of Computer Science and Information System, BITS PILANI, K. K. Birla Goa Campus, India*
[b]*Research Scholar, Dept of Computer Science and Information System, BITS PILANI, K. K. Birla Goa Campus, India*

Abstract

The metamorphic malware variants with the same malicious behavior (family), can obfuscate themselves to look different from each other. This variation in structure lead to a huge signature database for traditional signature matching techniques to detect them. In order to effective and efficient detection of malwares in large amounts of executables, we need to partition these files into groups which can identify their respective families. In addition, the grouping criteria should be chosen such a way that, it can also be applied to unknown files encounter on computer for classification. This paper discusses the study of malwares and benign executables in groups to detect unknown malwares with high accuracy. We studied sizes of malwares generated by three popular second generation malwares (metamorphic malwares) creator kits viz. G2, PS-MPC and NGVCK, and observed that the size variation in any two generated malwares from same kit is not much. Hence we grouped the executables on the basis of malware sizes by using Optimal k-Means Clustering algorithm and used these obtained groups to select promising features for training (Random forest, J48, LMT, FT and NBT) classifiers to detect variants of malwares or unknown malwares. We find that detection of malwares on the basis of their respected file sizes gives accuracy up to 99.11% from the classifiers.

Peer-review under responsibility of organizing committee of the ICISP2015

Keywords: Anti-Malware; Static Malware Analysis; WEKA; Machine Learning

* Corresponding author. Tel.: +91-8975805861.
E-mail address: p2012011@goa.bits-pilani.ac.in

Peer-review under responsibility of organizing committee of the ICISP2015
doi:10.1016/j.procs.2016.02.115

1. Introduction

As new variants of malwares getting evolve every day, malwares defense becoming increasingly difficult task in detecting malware and protecting computers systems from them[1]. Recently 11 zero-day vulnerabilities reported during the month of August while 6 of these were reported in industrial control systems[2]. Even state sponsored highly skilled hackers are developing customized malwares to disrupt industries and for military espionage[3]. Many of countries continue to incur most costly data breaches. Among them two countries had the highest cost from data breach[4] (the U.S. at $5.4 million and Germany at $4.8 million).

Anti-malware industries are facing a major challenge of continuously increase of huge data, which need to be checked for potential malicious content. Microsoft reports that there real-time detection anti-malware products are present on over 160 Million computing devices throughout the globe and they daily analyze tens of millions of data files as potential malware[5]. Reason behind these high volumes of different files is that the malware authors introduce metamorphism to the malicious components. Metamorphic malware represent the next class of virus that can create an entirely new variant after reproduction[3]. The new variant produced is in no-way similar to the original variant which lead a huge increase in the malwares count.

In order to detect them with high accuracy, we need to group them to identify their respective families. In addition, such grouping criteria may be applied to new test executables to classify it to malware. In this paper we studied three popular second generation malwares creator kits viz. G2, PS-MPC and NGVCK and found that the size variation in any two generated malwares from same kit does not differ much. Hence in this work we grouped the executables on the basis of malware sizes by using optimal k-Means Clustering algorithm and promising features are selected separately from each groups. Further these obtained features are tested on random forest, J48, Logistic Model Trees (LMT), functional trees (FT) and naive bayes tree (NBT) classifier using machine learning technique.

The paper is organized as follow, in next section related work is discussed, In section 3 we present our approach, The section 4 discuss the experimental results and finally section 5 contains the conclusion and future directions.

2. Related work

The first virus was created in 1970[6] and since then there is a strong contest between the attackers and defenders. To combat the threats/attacks from the second generation malwares, Schultz et al. (2001) was the first to introduce the concept of data mining to classify the malwares[7]. In 2005 Karimet al.[8] addressed the tracking of malware evolution based on opcode sequences and permutations. O. Henchiri et al.(2006) proposed a hierarchical feature extraction algorithm and used ID3, j48, Naive Bayes and SMO classifier and obtained maximum of 92.56% accuracy[9]. In year 2005, Karimet al.[8] addressed the tracking of malware evolution based on opcode sequences and permutations. O. Henchiri et al.(2006) proposed a hierarchical feature extraction algorithm and used ID3, j48, Naive Bayes and SMO classifier and obtained maximum of 92.56% accuracy[9]. Bilar (2007) uses small dataset to examine the opcode frequency distribution difference between malicious and benign code[10]and observed that some opcodes seen to be a stronger predictor of frequency variation. In 2008, Yanfang Ye et. al.[11] applied association rules on API execution sequences for classifying the malwares. In 2008, Tian et al.[12] classified the Trojan using function length frequency and shown that the function length along with its frequency is significant in identifying malware family and can be combined with other features for fast and scalable malware classification. Moskovitchet al.[13] compared the different classifiers by byte-sequence n-grams (3, 4, 5 or 6). Among the classifiers they studied BDT, DT and ANN out-performed NB, BNB and SVM classifiers, exhibiting lower false positive rates. In year 2008, Siddiquiet al.[14] used variable length instruction sequence for detecting worms in the wild. They tested their method on a data set of 2774 (1444 worms and 1330 benign files) and got 95.6% detection accuracy. In 2009 S. Momina Tabish[15] used 13 different statical features computed on 1, 2, 3 and 4 gram by analyzing byte-level file content for classification of malwares. In year 2010, Bilal Mehdi et. al.[16] used hyper grams (generalized n-gram) and obtained 87.85% detection accuracy and claimed no false alarm. ChatchaiLiangboonprakong et al. (2013) proposed a classification of malware families based on N-grams sequential pattern features[17]. They used DT, ANN and SVM