Journal of Computer Virology and Hacking Techniques (2019) 15:195–208
https://doi.org/10.1007/s11416-019-00331-0

ORIGINAL PAPER

Nonnegative matrix factorization and metamorphic malware detection

Yeong Tyng Ling[1] · Nor Fazlida Mohd Sani[1] · Mohd Taufik Abdullah[1] · Nor Asilah Wati Abdul Hamid[1]

Received: 5 August 2018 / Accepted: 20 March 2019 / Published online: 26 April 2019

Abstract
Metamorphic malware change their internal code structure by adopting code obfuscation technique while maintaining their malicious functionality during each infection. This causes change of their signature pattern across each infection and makes signature based detection particularly difficult. In this paper, through static analysis, we use similarity score from matrix factorization technique called Nonnegative Matrix Factorization for detecting challenging metamorphic malware. We apply this technique using structural compression ratio and entropy features and compare our results with previous eigenvector-based techniques. Experimental results from three malware datasets show this is a promising technique as the accuracy detection is more than 95%.

Keywords Metamorphic malware · Nonnegative matrix factorization · Dimension reduction · Structural analysis

1 Introduction

Malware are threatening today's computing world by its rapid evolving. According to latest report in [1], malware implant to legitimate software were up to 200% in year 2017. Therefore, this leads to urge for new detection technique.

It is generally believed that the detection of metamorphic malware remains a difficult task in the area of research [2,35]. Metamorphic malware utilize code obfuscation techniques to disguise themselves from traditional pattern matching signature based detection. During each propagation, they change their internal syntax or structure while maintaining underlying malicious functionality. A recent study has shown that a well-designed metamorphic malware was capable of evading statistical analysis using Hidden Markov Model (HMM) [29] detection. Due to this reason, several file structural analysis approaches such as in [10,14,19,33], have shown to be very effective at classifying this type of metamorphic malware.

Motivated by this, we adopt structural analysis on two types of file feature, namely, the compression ratio and entropy from previous works in this paper. We also modified the implementation of [14] through inspiration of facial recognition technique based on matrix factorization known as Nonnegative Matrix Factorization (NMF) [17]. NMF learns a parts-based representation of facial images by representing faces with a set of basis images correspond to parts of faces. This recognition technique bears a resemblance to malware detection where detecting a set of malware can actually be represented by detecting a smaller subset of their own. Hence, our contribution is to adapted NMF technique using structural analysis of file compression ratio and entropy features on three types of challenging metamorphic malware families, in particularly, on real malware sample. Our approach use the similarity scores derived from NMF technique to find the similarity of structural pattern that exist in both malware and benign files. As expected, we show that NMF technique yields comparable result when test on small scale datasets with both features. We observed that using structural entropy can produce better detection than using structural compression ratio feature.

The rest of this paper is organized as follows. In Sect. 2 we briefly reviewed metamorphic techniques and related work in detection, the difference between SVD and NMF techniques. Then, in Sect. 3 the application of NMF on our malware detection will be described in details. Section 4 presents the experimental and result analysis. Finally, we conclude and suggest future work in Sect. 5.

✉ Yeong Tyng Ling
ling.yt@student.upm.edu.my

Nor Fazlida Mohd Sani
fazlida@upm.edu.my

[1] Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Serdang, Selangor, Malaysia