Journal of Computer Virology and Hacking Techniques
https://doi.org/10.1007/s11416-021-00404-z

ORIGINAL PAPER

Metamorphic malware detection using structural features and nonnegative matrix factorization with hidden markov model

Yeong Tyng Ling[1] · Nor Fazlida Mohd Sani[2] · Mohd Taufik Abdullah[2] · Nor Asilah Wati Abdul Hamid[2]

Received: 23 April 2021 / Accepted: 13 September 2021

Abstract
Metamorphic malware modifies its code structure using a morphing engine to evade traditional signature-based detection. Previous research has shown the use of opcode instructions as feature representation with Hidden Markov Model in the context of metamorphic malware detection. However, it would be more feasible to extract a file feature at fine-grained level. In this paper, we propose a novel detection approach by generating structural features through computing a stream of byte chunks using compression ratio, entropy, Jaccard similarity coefficient and Chi-square statistic test. Nonnegative Matrix Factorization is also considered to reduce the feature dimensions. We then use the coefficient vectors from the reduced space to train Hidden Markov Model. Experimental results show there is different performance between malware detection and classification among the proposed structural features.

Keywords Hidden Markov model · Metamorphic malware · Nonnegative matrix factorization · Structural feature

1 Introduction

Malware writers use various obfuscation techniques to evade signature-based detection that is typically employed by traditional antivirus scanners. The most commonly adopted obfuscation techniques include code packing, polymorphism, and metamorphism [8]. Among these, it is well known that metamorphic malware is extremely difficult to detect [20]. It poses a special threat with its unique way of constantly mutating its internal structure while staying in an infected system. By obfuscating repeatedly, two copies from the same malware will be syntactically different while keeping the same malicious behavior [50]. It is hence imperative to find an effective strategy for metamorphic malware detection.

Hidden Markov Model (HMM) has been extensively applied to the metamorphic malware detection problem [1,11,18,36,47]. Previous research using HMM approaches employed opcode instruction sets as the feature representation of an executable file. However, these approaches require disassemble tool to translate an executable file into instruction sets format in assembly language that is specific to the underlying machine architecture [7,37]. Disassemble tool relies on heuristics-based method to find the reachable code, which can lead to inaccurate recovery of assembly language [45]. Hence, the opcodes feature generated may not represent an executable file correctly.

File structural entropy analysis approach [6,29,46] has shown effective at detecting malware files. It is based on the assumption that files with high entropy are relatively likely to have encrypted or compressed sections inside them and thus make adequate distinctive features when compared with normal files [27]. Thus, a file can be represented as stream of entropy where its content is divided into a series of byte chunks, and the entropy of each chunk is computed. This creates the structural entropy [40] of a file in the form of *time series* which reveals different data complexity throughout an executable file.

Motivated by this, in this paper, we expand the study in [25] with HMM. Specifically, we employ compression ratio, entropy, Jaccard similarity coefficient on hexadecimal sets, Jaccard similarity coefficient on integer sets, and Chi-square statistic test to derive file features and exploit Nonnegative Matrix Factorization (NMF) [22] to reduce the dimension of proposed structural features in HMM. NMF has been suc-

✉ Nor Fazlida Mohd Sani
fazlida@upm.edu.my

[1] Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak, Sarawak, Malaysia

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

Published online: 31 October 2021