Structural features with nonnegative matrix factorization for metamorphic malware detection

ABSTRACT

Metamorphic malware is well known for evading signature-based detection by exploiting various code obfuscation techniques. Current metamorphic malware detection approaches require some prior knowledge during feature engineering stage to extract patterns and behaviors from malware. In this paper, we attempt to complement and extend previous techniques by proposing a metamorphic malware detection approach based on structure analysis by using information theoretic measures and statistical metrics with machine learning model. In particular, compression ratio, entropy, Jaccard coefficient and Chi-square tests are used as feature representations to reveal the byte information existing in malware binary file. Furthermore, by using Nonnegative Matrix Factorization, feature dimension can be reduced. The experimental results show the Jaccard coefficient on hexadecimal byte as feature representation is effective for Windows metamorphic malware detection with an accuracy rate and F-score as high as 0.9972 and 0.9958, respectively. Whereas for Linux morphed malware detection, the Chi-square statistic test shows as effective feature representation with an accuracy rate and F-score as high as 0.9878 and 0.9901, respectively. Overall, the proposed feature representations and the technique of dimension reduction can be useful for detecting metamorphic malware.

Keyword: Metamorphic malware; Compression ratio; Entropy; Jaccard coefficient; Chi-square; Nonnegative matrix factorization