SCIS 2022 2022 Symposium on
Cryptography and Information Security
Osaka, Japan & Online, Jan. 18 – 21, 2022
The Institute of Electronics,
Information and Communication Engineers

ドキュメント化されていないヘッダを活用した機械学習によるマルウェア分類
Malware Classification by Machine Learning Using Undocumented Header

小久保 博崇 * 大山 恵弘 †
Kokubo Hirotaka Oyama Yoshihiro

キーワード マルウェア, 機械学習, Rich ヘッダ

あらまし

近年，マルウェアは増加の一途を辿っており，危険なマルウェアに素早く対処するためには高速かつ効果的なトリアージ技術が必要である．トリアージに必要な情報は，マルウェアを表層解析，静的解析，動的解析することで得ることができ，その中でも表層解析は機械的かつ高速に実行できる解析手法である．

表層解析とは，マルウェアのファイル名，ハッシュ値，文字列等リソースの情報，セクションの情報，PE ヘッダの情報など，マルウェアを実行したり逆アセンブルしたりすることなく得ることのできる表層的な情報を収集する解析手段である．このようなマルウェアの表層解析結果を特徴量とした機械学習によるマルウェア分類は，高速かつ効果的なトリアージ技術の一例である．

しかし，マルウェア作者は表層解析で得られる情報が解析に利用されていることを把握しており，パッカー等でこれらの改竄を行うことがしばしばある [1]．このような妨害を受けた場合，よりコストのかかる静的解析や動的解析によって解析を進めざるを得ない．

この問題に対して，本研究ではこれまで表層解析の対象となっていることが少なく，マルウェア作者の注目も集めていない，ドキュメント化されていないヘッダ（Rich ヘッダと呼称する）に着目することで解決を図る．Rich ヘッダは Microsoft 社のリンカを使用して作られた PE ファイルに付与されるヘッダであり，その PE ファイルのビルド時の環境情報を含んでいる．また，George ら [1] によれば使用するパッカーによってはパック後も Rich ヘッダは影響を受けずに残存するため，攻撃者の妨害を受けにくいと考えられる．

我々は，この Rich ヘッダのみを特徴量として機械学習によるマルウェア分類を試みた．使用したデータセットは Richard ら [2] が公開しているマルウェアバイナリを含むデータセット SoReL-20M から抽出した，Rich ヘッダを持つユニークなハッシュ値のマルウェア 15,847 検体 129 ファミリから成る部分データセットである．Triplet Network を採用した深層学習によりマルウェア分類を行い，層化 5 分割交差検証を実施したところ，約 82.9%の精度で分類が成功した．また，Rich ヘッダを含む実行ファイルをパックしたときに Rich ヘッダが本当に残存するかどうかについて，複数のパッカーを用いて検証したところ，著名なパッカーの中でも Rich ヘッダを消去せず残存させるパッカーがあることがわかった．

参考文献

[1] George D. Webster and Bojan Kolosnjaji and Christian von Pentz and Julian Kirsch and Zachary D. Hanif and Apostolis Zarras and Claudia Eckert,“Finding the Needle: A Study of the PE32 Rich Header and Respective Malware Triage,” DIMVA, 2017.

[2] Richard Harang and Ethan M. Rudd,“SOREL-20M: A Large Scale Benchmark Dataset for Malicious PE Detection,” arXiv, 2020.

* 富士通株式会社, 〒 211-8588 神奈川県川崎市中原区上小田中 4-1-1, Fujitsu Limited, 4-1-1 Kamikodanaka Kawasaki, Kanagawa 211–8588, Japan
† 筑波大学, 〒 305-8577 茨城県つくば市天王台 1-1-1, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki, 305–8577, Japan