Scikit-Learn 活用時におけるマルウェア検出精度の検討

Study of Detection accuracy of malware by making use of Scikit-Learn

中村 祐輝† 笠間 貴弘†† 宮保憲治†
Yuuki Nakamura† Takahiro Kasama†† Noriharu Miyaho†
† 東京電機大学 情報システム工学部 ††国立研究開発法人情報通信研究機構
†Department of information System Engineering, Tokyo Denki University †† National Institute of Information and Communications Technology

1. はじめに

近年，サイバー攻撃の被害件数が増加している．これは，Exploit Kit 等の使用頻度が急速に増加しているためと考えられる．Exploit Kit はサイバー攻撃を容易に可能とする攻撃ツールの一種であり，本検討では，多くの機械学習アルゴリズムが実装されている Python の Scikit-Learn を活用し，Exploit Kit 使用時の悪性通信の分類を行った結果を報告する．

2. 提案方式

図 1 に Scikit-Learn を活用した分類手法を示す．最初に，パケットデータから特徴量の抽出を行う．次に，Scikit-Learn を活用して機械学習モデルを実装し，Exploit Kit を使用した悪性通信の分類を行った．

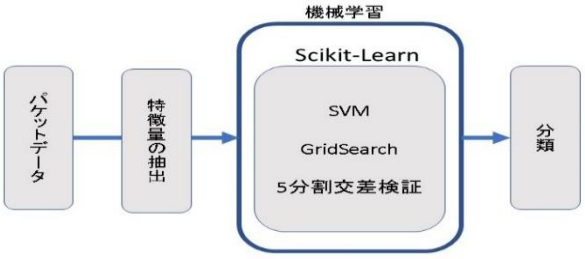

図 1Scikit-Learn を活用した分類手法

3. 使用したデータセット

悪性データとして，nao_sec により Web クライアントハニーポットで収集された augmaDataSet2020[1]を使用した．augmaDataSet2020 はサイバー攻撃で頻繁に使用される Exploit Kit を用いた通信を観測した pcap ファイルのデータセットである．augmaDataSet2020 内の 90%の pcap ファイルが Underminer，Fallout，GrandSoft であり，各 100 件(合計 300 件)使用した．良性データとして，人気 Web サイトランキングである Amazon Alexa の The top 500 sites on the Web[2]を使用した．ランキング上位の Web サイトから WireShark を用いて pcap ファイルを 100 件収集した．

4. 機械学習モデル

Python のライブラリの一種である Scikit-Learn を用い，GridSearch と交差検証を組み込んだ SVM(Support vector machine)を実装した．実装した SVM モデルと表 1 に示す受信パケットに関する特徴量を用いて分類精度実験を行った．表 1 に示す特徴量は規則性が見出された 9 種類である．規則性の例として大きさが 54 バイトのパケットが明らかに多く Exploit Kit の通信に存在した．また，GridSearch では， コストパラメータである C(1≦C≦100)と RBF のカーネルパラメータである γ(1e-15≦γ≦1)の最適化を行った．(C=10， γ=1e-10 が最適値となった)

表 1 抽出した特徴量

パケット数	パケット平均サイズ	パケット最大サイズ
パケット総サイズ	平均パケット毎秒	平均ビット毎秒
54Bのパケット割合	ACKパケットの割合	SYN,ACKパケット数

5. 実験結果

図 1 に示す分類手法で得られた実験結果の中で，Accuracy，Precision，Recall，F 値の測定値を表 2 に示す．Accuracy が 88%以上の数値で分類が可能であると判明した．特に Underminer が他と比べ Accuracy が 94%と高い結果となった．また，Precision も 96%と高く，検知率が高く誤検知が少ない結果となった．

表 2 実験結果

EK name	Accuracy	Precision	Recall	F値
Underminer	0.94	0.96	0.93	0.94
Fallout	0.88	0.87	0.9	0.88
GrandSoft	0.87	0.89	0.84	0.87

6. まとめ

Scikit-Learn を活用し，GridSearch と交差検証を組み込んだ SVM モデルでは，Exploit Kit を使用した悪性通信の分類に表 1 に示した特徴量は有効であると考える．また，実験結果から Exploit Kit を使用した通信には表 1 に示した特徴量が分類に有効であることが分かった．今後は検討対象外とした Exploit Kit のうち RIG や Spelevo を追加し，表 1 の特徴量が分類に有効であるかを検証する予定である．augmaDataSet2020 は頻繁に使われる Exploit Kit のデータセットであるため，Exploit Kit を使用した通信の検知にも今後は役立つと考える．

7. 参考文献

[1]寺田真敏，他:マルウェア対策のための研究用データセット MWS Datasets～コミュニティへの貢献とその課題，情報処理学会，Vol2020-IFAT-139No8，2020．7
[2]Amazon ALEXA/The top500 sites on the web，https://www．alexa．com/topsites(2020．10．1)