クラウドを活用したマルウェアのリアルタイム検出システム

A-7　Real-Time Malware Detection System Using Cloud System

小出 遼† 笠間 貴弘‡ 宮保 憲治†
Ryo KOIDE† Takahiro KASAMA ‡ Noriharu MIYAHO†
† 東京電機大学大学院 情報環境学研究科 ‡ 国立研究開発法人情報通信研究機構
† Graduate School of Information Environment. Tokyo Denki University ‡ National Institute of Information and Communications Technology

1. はじめに

近年，多種多様なマルウェアがサイバー社会に蔓延し，その被害は増加している．そのため，マルウェアを含んだ通信の態様を正確に把握する必要がある．ハニーポットを活用し，大量のマルウェアデータを含む通信の観測により取得したデータを分析することでマルウェア通信の特徴を把握する方法が有望と考えられる．

具体的には，クラウド(Microsoft Azure)上に構築したハニーポットを使用した上で，マルウェア感染時に行われる悪性通信ログの収集結果を活用し，当該データと関連性の高いデータセットを併用することで，マルウェア感染時の悪性通信の特性を定量的に評価した．

本論文では，各種の機械学習アルゴリズムを活用して，悪性通信の特徴を指定パラメータで抽出することより，マルウェア検出をリアルタイムで行う評価システムの構築法，並びに，性能実験により分析した評価結果を報告する．

2. 機械学習を活用したマルウェア検出システムの提案

2.1 ハニーポットを用いた攻撃観測のメカニズム

Microsoft Azure 上に構築したハニーポット(T-Pot)を運用（2020/10/1～2020/10/31）し，第三者がインターネット経由で実施した攻撃活動を観測した．当該の通信ログは合計 555 セッションであり，その内 100 セッションを使用してマルウェア感染時における悪性通信の特徴を分析した．これらの通信ログでは全て Eternal Blue と呼ばれる攻撃ツールが使用されていた．また，当該悪性通信ログデータと関連性の高いデータセットとして Augma Datasets[1]を活用した．本データセットに収録されているログデータの中から，攻撃ツールが行う悪性通信ログデータ（7 種類）に対して，各 100 セッションずつのデータ抽出を行った．当該悪性通信ログ(合計 8 種類，800 セッション)から得られる通信特性の特徴を評価した[2]．

悪性通信と定量的な比較を行うために良性通信データとして独自にブラウジングを行い，当該通信ログを収集した．

2.2 性能評価実験の概要

悪性通信と良性通信を比較する特徴量として，合計パケット数，合計送受信バイト数，受信パケット数，受信バイト数，送信パケット数，送信バイト数，平均受信フレーム長，平均送信フレーム長，1 パケットあたりの平均データ量，1 秒あたりのパケット数，総パケット数に対する受信パケット数の比率，総パケット数に対する送信パケット数の比率，平均送受信バイト数の 13 つの特徴量を抽出した．システムの性能評価を実施する際に使用した学習アルゴリズムとして，SVM，Random Forest，ロジスティック回帰の 3 種類を使用した．機械学習アルゴリズムによる特徴量最適化の手法を適用し，システム全体の性能向上を評価する実験も併せて実施した．また，交差検証として 5 分割交差検証を使用し，評価指標としては適合率と再現率の調和平均である F 値を使用することとした．

2.3 性能評価実験の結果

マルウェア検出システムを用いた性能評価実験を実施した．表 1 に性能評価実験の結果を示す．

表 1.　性能評価実験の結果

	アルゴリズム	F値(特徴量最適化なし)	アルゴリズム	F値(特徴量最適化あり)
良性通信	ロジスティック回帰	0.99	ロジスティック回帰	0.99(±0)
EternalBlue	RandomForest	0.98	ロジスティック回帰	0.98(±0)
Fallout	RandomForest	0.85	SVM	0.93(+0.8)
RIG	ロジスティック回帰	0.89	RandomForest	0.98(+0.9)
UnderMiner	RandomForest	0.98	SVM	0.98(±0)
Spelevo	ロジスティック回帰	0.91	ロジスティック回帰	0.93(+0.2)
PurpleFox	RandomForest	0.90	RandomForest	0.96(+0.6)
Bottle	ロジスティック回帰	0.90	RandomForest	0.93(+0.3)
GrandSoft	RandomForest	0.88	RandomForest	0.93(+0.5)
平均		0.92		0.96(+0.4)

3. 考察

特徴量の最適化を施すことによって，検出精度を維持し，かつ検出精度の向上化が可能であることを確認した．

このことは，特徴量の種別数を最適化（特徴量を 8 種類に絞り込み）することにより，悪性通信の検出が効果的になり，検出精度の向上に繋がったと考えられる．また，マルウェア検出システムに複数の機械学習アルゴリズムを活用したことも，検出精度向上には有効であったと考えられる．

4. まとめと今後の課題

本稿では，クラウド上に構築したハニーポットを活用した上で，悪性通信の通信ログを収集し，構築したマルウェア検出システムを使用した性能評価を定量的に実施した．今後はデータ量を増大させ，大規模な検証を行うと共に，マルウェアが行う悪性通信の検出精度向上に寄与するパラメータを更に調査・抽出する予定である．

5. 参考文献

[1] Rintaro Koike and Yosuke Chubachi, "Finding drive-by rookies using an automated active observation platform", VirusBulletin 2019, Oct. 2019.
[2] 小出遼，笠間貴弘，宮保憲治: クラウドを活用したマルウェアのリアルタイム検出システム，コンピュータセキュリティシンポジウム 2021，デモンストレーションセッション(DS-01)，Oct. 2021.