九州大学学術情報リポジトリ
Kyushu University Institutional Repository

Scalable and Fast Method for IoT Malware Clustering

何, 天祥

https://hdl.handle.net/2324/6787629

出版情報：九州大学, 2022, 博士（工学）, 課程博士
バージョン：
権利関係：

KYUSHU UNIVERSITY

（様式 2）

氏　　名　：何 天祥

論 文 名　：Scalable and Fast Method for IoT Malware Clustering
（IoT マルウェアクラスタリングのためのスケーラブルかつ高速な手法）

区　　分　：甲

論　文　内　容　の　要　旨

モノのインターネット(IoT)機器を悪用した大規模なサイバー攻撃が近年度々報告されており，IoT マルウェア（以下，単にマルウェアと記す）は多様化・高度化し，そのようなマルウェアの感染は世界的な社会問題となっている．そのため，感染挙動を早期に把握し，その情報を関連組織間で共有することが重要とされているが，現状は十分な対応が取れていない．その結果，セキュリティインシデント発生後の初動対応が遅れ，機密情報の漏洩などの重大なインシデントにつながるケースが多く発生している．こうした問題に対処するためには，大量に存在し，日々新しく生み出されている IoT マルウェアを整理し，解析する手段が有用である．特に，マルウェアには常に様々な改変が行われるため，多くの亜種が存在し，それらがグループを構成している．本論文は，これに着目し，マルウェアの実行ファイル（検体）の集合に対する効率的クラスタリング技術の確立を目指した研究について述べたものである．

マルウェアがオリジナルな検体から様々な亜種が生成される過程は，生物の進化の過程に類似していることから，進化系統樹（系統樹）を用いてマルウェア検体群を整理する手法が来嶋らによって提案されていた．進化系統樹を構成する手段としては，生物学への応用分野において古くから研究されており，種同士の類似性に基づいて系統樹を近隣結合法と呼ばれる手法が広く用いられている．この手法は種の全ての対の類似度を必要とするため，種の数の 2 乗オーダの計算時間がかかり，高々数千検体程度の対象にしか適用することができない．来嶋らの手法は，ごく一部の種の対の類似度だけに基づいて系統樹を作成するものであり，計算時間を大幅に削減できる可能性があった．

著者は，来嶋らのアイデアに基づき，高速に大規模な系統樹を作成するアルゴリズムの設計と実装を行い，マルウェアの大規模検体群に対し系統樹の作成を行った．さらに，作成した系統樹に基づいて，類似した亜種が形成するクラスターを求める手法の設計と実装を行い，実検体群に適用した．また，確立した系統樹作成法とクラスタリング手法を実運用するための改良を行い，実検体群を用いてその有効性を示した．これらの成果は，下記のようにまとめられる．

第一に著者は，近隣結合法を基に，大規模入力に対応可能な系統樹作成アルゴリズムの設計と実装を行い，7 万検体の実データでその有用性を示すことに成功した．著者の手法では，マルウェア検体群検体間の距離の計算に，来嶋らのアイデアに基づいて正規化圧縮距離（Normalized Compression Distance）を用いる．これは実行ファイルを圧縮したときの圧縮後のファイルサイズに基づいて距離を求めるもので，一つの対の距離の計算に秒単位の時間が必要となり，7 万検体に関する距離行列の計算には数か月の時間が必要であった．著者らが設計した手法（高速近似法）では，距離行列のごく一部だけを計算するだけで系統樹を作成することができる．著者らは，実データを用いた実験で，これが 3 日間程度に短縮できることを示した．さらに，記述長最小原理に基づいて系統樹を適切なノードでカットしクラスタリング手法の設計と実装を行った．その結果，オリ

ジナルの近隣結合法に基づく場合と同程度のクラスタリング精度が得られることを示した．

第二に，設計したクラスタリング手法を実運用に供することを目的とした改良を行った．マルウェアを解析する現場では，恒常的に新たな亜種が発生するため，一度得られた系統樹を徐々に更新していくオンライン手法が有用である．そこで著者は，高速近似法をオンライン形式に修正する改良を行い，7 万検体の実データを用いた実験により，その評価を行った．結果として，オンライン化してもクラスタリングの精度が保たれることを示した．また，このとき結果として計算時間が 7 割程度に削減できることが分かった．

第三に著者は，クラスタリングのさらなる高速化を目的に，能動学習の考え方に基づくクラスタリング手法の調査と性能評価を行った．近隣近似法は，距離行列を入力データとして検体間に辺を引く過程とみなすことができる．この見方においては，高速近似法はデータセットの一部のみをランダムに選択して用いている手法と考えられる．この過程は，データセットから情報量の高い部分を能動的に選択して用いる能動学習法に類似している．著者はこの考えたに基づく能動クラスタリング手法について調査し，Zoller らのアルゴリズムに着目した．この手法を実装し，3000 検体の実データを用いた評価を行った結果，高速近似法を用いたクラスタリングに比して，より観測数を削減できることを示した．同時に，最適化のために用いる平均場アニーリングの計算時間が膨大となる課題があることを明らかにした．