平成26年度　情報工学コース卒業研究報告要旨

高倉 研究室	氏　　名	蛭 田　将 平
卒 業 研 究 題 目	トラフィックデータとバイナリのFuzzy Hash値に基づくマルウェア分類に関する研究	

近年急増しているサイバー攻撃には，既存のマルウェアから派生した亜種が使用される場合が多い．またマルウェアの機能も多様化し，数年前までは感染した相手を驚かすだけで直接的な被害がない迷惑行為を行うマルウェアが主流であったが，最近は特定の企業や組織を標的とした金銭的利益を目的としたマルウェアが主流になりつつある．

日々増加するマルウェアに対し解析者の人数は不足しているため，解析のコストは大きい．少ない解析者は新種のマルウェアの精査に専念できるように，解析のコストを軽減する高速なマルウェア分類手法が求められている．

そこで，本研究ではマルウェアのトラフィックデータとバイナリのFuzzy Hash値の類似度計算による多段マルウェア分類手法を提案する (図 1)．提案手法では，まずマルウェアのトラフィックデータから特徴を抽出し，クラスタリングを行う．得られたクラスタにラベルを割り当て，その結果を元にトラフィックデータからクラスタ列に変換する．生成したクラスタ列に対して既知のマルウェアとの類似度を計算して分類を行う．分類結果の候補が複数現れたマルウェアについては，バイナリの Fuzzy Hash 値を求め，得られたハッシュ値に基づいて類似度計算を行い，再度分類する．

提案手法に対する評価実験を行うため，独自環境で収集したマルウェア 340 検体とそれらを実際に動作させて得られたパケットキャプチャファイルを用いた．収集したマルウェアのファミリの数は 48 個であった．340 検体のうち 113 検体をテストデータとし，残りを学習データとした．分類結果は，類似度計算で最も高い値を示したマルウェアが属していたファミリとした．本手法を適用し分類を行った結果，トラフィックデータのみによる分類ではテストデータ 113 個中 50 個が正しく分類され，正解率は 44.2%であった．また，トラフィックデータによる分類結果の候補が複数存在したテストデータ 75 個について，バイナリの Fuzzy Hash 値による分類を行った．その結果，テストデータ 75 個中 46 個が正しく分類され，正解率は 61.3%であった．本手法により正しく分類できたテストデータは 113 個中 66 個で正解率は 58.4%であり，トラフィックデータのみによる分類と比べ改善率は 32.1%であった．正しく分類できなかったマルウェアが存在する原因として，そのマルウェアに特有なトラフィックがなく，またマルウェアのバイナリが暗号化・難読化されていたことが挙げられる．そのため，新たなクラスタリングや類似度計算アルゴリズムを適用し，トラフィックデータによる分類の精度を向上させることが今後の課題である．

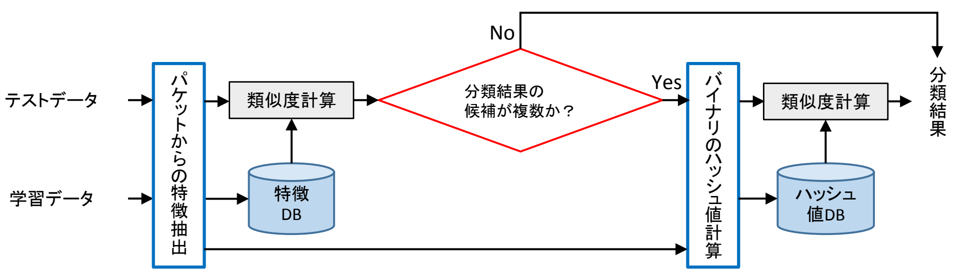

図 1　提案手法の流れ