Computer Security Symposium 2017
23 - 25 October 2017

HTTP通信を活用した半教師あり機械学習による マルウェア感染端末の選別

西山 泰史[1]　熊谷 充敏[1]　神谷 和憲[1]　谷川 真樹[1]

概要：近年，機械学習を用いたマルウェア感染端末の分類手法の需要が高まっている．しかし，一般に機械学習を用いて汎用性の高い分類器を作るには，様々な環境で収集されたラベルありデータが必要となるが，マルウェア感染端末の分類においては，ラベル付けにかかる人件費や，様々な環境からログを取得することの難しさから，網羅的にラベルありデータを得ることは容易ではない．本稿では，グラフベースの半教師あり学習を用いて，HTTP 通信ログからマルウェア感染端末を検知する手法を提案する．これにより，ラベルありデータは少量しか得られないが，ラベルなしデータは大量に得られるような，セキュリティ分析者がよく直面する状況でも，汎用性の高い分類器を作成することが可能となる．また，評価実験により，提案法が従来の教師あり学習に基づく手法に比べて精度よく感染端末を検知できることを示す．

キーワード：半教師あり学習，ログ分析，マルウェア，感染端末，HTTP

Semi-supervised Machine Learning Approach for Detecting Malware Infected Host by Analyzing HTTP Traffic

Taishi Nishiyama[1]　Atsutoshi Kumagai[1]　Kazunori Kamiya[1]　Masaki Tanikawa[1]

Abstract: Nowadays, there are increasing needs in correctly detecting malware infected-hosts by machine learning. However, machine learning needs labeled data which covers various environments to generate versatile and sophisticated classifier. Since labels are manually set by security operators looking at limited numbers of enterprise logs, it is hard to always prepare good labeled data which have enough volume and cover various type of network environments. In this research, we propose to apply graph-based semi-supervised machine learning to HTTP traffic logs for detecting malware infected hosts. Through the proposed approach, versatile and sophisticated classifier for infected-hosts detection can be achieved in the case, security analysts often face the case, that small amount of labeled data and large amount of unlabeled data are available. In addition, our evaluations demonstrate the superiority of our approach in comparison to a conventional supervised learning approach.

Keywords: Semi-supervised Learning, Log Analysis, Malware, Infected Host, HTTP

1. はじめに

1.1 背景

セキュリティインシデントの報告は後をたたず，中でもマルウェア感染に起因したインシデントはその典型的な事例であると言える．2015 年に公法人で生じた個人情報漏洩事件 [1] や，2016 年に大手旅行代理店で生じた個人情報流出問題 [2] などはマルウェア感染に起因したインシデントとして記憶に新しい．このようなマルウェア感染被害を防ぐためには事前に感染を防ぐことが最善策であり，ウィルス対策ソフトがその典型的な対策手段として挙げられる．しかし，2016 年の 1 年間に第三者テスト機関の AV-Test[3] に寄せられた情報だけでも，新種のマルウェアが約 1.3 億種類発見されており，これらの新種のマルウェアをウイル

[1] 日本電信電話株式会社 セキュアプラットフォーム研究所
Nippon Telegraph and Telephone Corporation, Secure Platform Laboratories

ス対策ソフトなどの事前対策のみで完全に防ぐことは難しい．そのため，マルウェア感染はある程度避けられないものとして，感染後できるだけすみやかに感染した端末を発見し，攻撃の目標としている情報漏えいや情報破壊を食い止めるための出口対策の重要性が高まっている [4]．出口対策の手法としては，内部から外部に通信する際の通信ログを Firewall/IDS/Proxy サーバ/DNS サーバ等の機器から収集し，それらを分析することで，マルウェアに感染した端末を検知する手法が有効である．

実際，多くのセキュリティベンダは MSS(Managed Security Service）として，通信ログを監視/分析してインシデント情報を顧客企業に伝えるサービスを提供している．これらのセキュリティベンダの競争力の源泉となっているのは，顧客の通信ログから迅速かつ見逃しなく脅威情報を捉える能力であり，中でも新種のマルウェアを検知する技術のニーズは高い [5]．MSS 事業者は SOC（Security Operation Center）と呼ばれる組織を構築しており，専門のオペレータやセキュリティアナリストを常駐させ，顧客のネットワークの監視/分析を行っている．MSS/SOC の運用の流れを以下に記す．

MSS/SOC の運用の流れ

(1) ログ収集：顧客の網内の Firewall/IDS/Proxy サーバ/DNS サーバ等の機器からログを収集する
(2) 分類：収集したログの中から，正常な通信を取り除き，分析対象の疑わしい通信ログを抽出する
(3) 分析：分類後の疑わしいログをセキュリティアナリストが詳細に分析し，顧客環境に害のある攻撃を特定する
(4) 通知/対応：分析後，顧客環境に害のある攻撃の情報などを顧客に通知する．インシデントレスポンスのサポートなども行う

(3) ではセキュリティアナリストの手で分析を行っているが，顧客のネットワーク内のログを全て分析することはコストの観点から難しい．そこで，あらかじめ (2) で正常なログか疑わしいログかの分類を機械的に行って，疑わしいログのみをセキュリティアナリストが分析している．新種のマルウェアを検知できるかどうかが MSS の競争力の源泉となっているため，いかに誤検知を減らしつつ，新種のマルウェアを見逃さない分類を行うかが重要となっている．現状，分類のフェーズで用いる分類器は様々な情報ソースを用いてオペレータ/セキュリティアナリストが手動で作成しているが，新種のマルウェアに対応するためには，マルウェアの進化に応じて更新を行う必要があり，オペレータ/セキュリティアナリストの負担となっている [6]．

近年，その分類器を機械学習を用いて作る技術が注目されている．新種のマルウェアは日々大量に作成されているものの，それらのマルウェアは完全に新しいものではなく，ソースコードを再利用して一部だけ変更しているケースや，リパッケージし直したりして作成された亜種のマルウェアであるケースが多い [7]．つまり，全体の特徴はあまり変わらないため，既知のマルウェアの通信パターンと類似する場合が多い．そこで，通信ログに対して機械学習を適用して分析することで，既知のマルウェアの通信と類似した特徴をとらえて検知できることが期待されている．

分類器を機械学習を用いて作ることによるメリットは大きく 2 点ある．1 点目はオペレータ/セキュリティアナリストの負担軽減になることである．現状，分類器は手動で作成しているが，これを機械学習で行うと自動でデータから分類器を作成できるため，分類器作成にかかる負担を軽減できる．また，分類器の精度が上がれば，セキュリティアナリストが解析すべき通信ログの候補を削減できるため，セキュリティアナリストの負担軽減につながる可能性も考えられる．2 点目は新種のマルウェアの進化に対応したタイムリーな更新が可能となる可能性があることである．先述したように，新種のマルウェアは爆発的に増加しているため，出現するたびに新種のマルウェアの通信ログに対応したシグネチャを人の手で追加していくには限界がある．機械学習を用いてデータから自動的に分類器を作ることができれば，人手よりも早くかつ幅広く新種のマルウェアに対応できることが期待できる．

そこで，本稿では，感染端末の検知において効果が高い，Proxy サーバから取得できる Proxy ログに着目して，新種のマルウェアに対応したマルウェア感染端末の分類器を機械学習で作成することについて考える．

1.2 課題

機械学習は大きく分けて，教師なし学習と教師あり学習に二分される．教師なし学習はラベルありのデータが不要であるといった長所があるものの，分類精度が低いという短所がある．今回のような正常な通信か疑わしい通信かを分類するタスクで精度が低いと，「新種のマルウェアが見つからない」「誤検知が多くセキュリティアナリストの分析にかかる稼動が増える」といった問題が生じる．そのため，本研究のようなタスクでは教師なし学習は不適であると考えられる．一方，教師あり学習は分類精度が高いという長所があるものの，学習に大量のラベルありのデータが必要であるという短所がある．正常な通信かマルウェアの通信かを分類した後のデータは一定数存在しているため，これらを学習用データとして教師あり学習を行うことは可能である．しかし，新種のマルウェアに対応するためには，マルウェアの進化にあわせてラベルありデータも定期的に更新しなければならないが，ラベルを付与するためには SOC

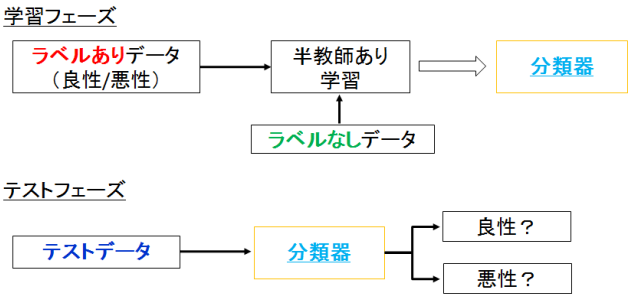

図 1　半教師あり学習の適用方法
Fig. 1　Method of applying semi-supervised learning

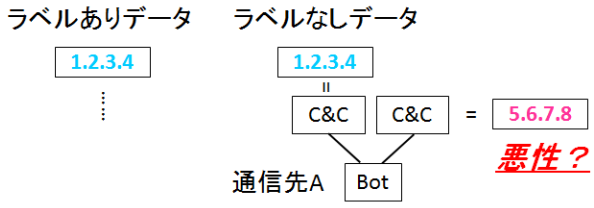

図 2　半教師あり学習による効果の例
Fig. 2　Example of semi-supervised learning effects

のアナリストなどの専門家が分析する必要があり，コストの面から難しい．

1.3　取り組み

前節の課題を解決するため，本研究では半教師あり学習を適用することで，ラベル付与の負担を軽減しつつ分類器を更新することを考える．半教師あり学習とは，ラベルありデータに加えて，ラベルなしデータからも学習することで，ラベルありデータだけで学習した場合に比べて，より予測精度の高い分類を実現することを目的とする機械学習手法である．本研究で取り扱う問題の場合では，正常なログが良性のラベルありデータ，既知のマルウェア由来の疑わしいログが悪性のラベルありデータ，未解析な通信ログがラベルなしデータにそれぞれ対応する．図 1 に手法の概要を示す．ラベルなしデータとして新種のマルウェアに関連した情報を含みうる未解析な通信ログを使用することで，ラベル付けの負担を軽減しつつ，新種のマルウェアに対応した分類器ができる．

半教師あり学習を用いた際の効果を説明するため，図 2 のような簡単な例を考える．図 2 は IP アドレス情報を拡張する場合を示している．まず，ラベルありデータとして，IP アドレス 1.2.3.4 が C&C サーバ（悪性）の IP アドレスで 1.2.3.4 と通信を行った端末は感染端末であるという情報が与えられたとする．また，ラベルなしデータとして，IP アドレス 1.2.3.4 と頻繁に通信を行っている通信先 A が別の IP アドレス 5.6.7.8 と通信をしている，という情報が与えられたとする．これらを半教師あり学習すると，ラベルありデータに含まれていない 5.6.7.8 を C&C サーバ（悪性）である可能性が高い IP アドレスとして学習できる．C&C サーバと Bot の特徴として，「C&C サーバと Bot は定期的かつ大量に通信を行っている」「Bot は複数の C&C サーバと通信を行う場合がある」ことなどが知られているため，このような状況は実際のサイバーセキュリティの運用でも生じうるケースである．

ここでは，C&C サーバの IP アドレス情報の拡張の例を考えたが，本稿で取り扱う Proxy ログは IP アドレス情報の他に，URL，HTTP ステータスコード，HTTP ユーザエージェント，宛先ポート番号，プロトコルなどの情報を含んでいるため，半教師あり学習をこれらの情報にうまく適用することができれば，より汎用的な分類器が作成できる．

2.　関連研究

本章では半教師あり学習を用いたログ分析技術における関連研究について述べる．

Gabriel ら [8] は半教師的なアプローチで，未知の悪性の URL を検知する方法について述べている．One Side Class Perceptron Algorithm[9] という教師あり学習を用いて分類器を生成している．One Side Class Perceptron とは誤検知をできるだけ抑えた Perceptron のことであり，検知率の低下は生じるが低誤検知率であるため，実運用に適したアルゴリズムであるとしている．また，その分類器を用いてテストデータを判定した結果をデータベースにキャッシュし，外部の情報を用いてその正誤を確かめ，分類器の更新に使うという半教師的な機構も提案している．ただし，この方法で分類器を更新するためには，外部の情報などを用いてラベル付けをする必要がある．

Shi ら [10] は企業網内で得られる Proxy ログに対して，グラフベースの半教師あり学習と教師あり学習（Random Forest）を独立に適用し，それらから出力された悪性度を表すスコアを活用して，未知の悪性のドメインを検知する手法について述べている．半教師あり学習部分ではドメイン情報のみを用いて，送信元と宛先の通信関係を示す二部グラフを作成し，悪性ドメイン情報の拡張を行っている．一方，教師あり学習部分では URL の特徴，接続の特徴（送信元/宛先 IP アドレス，送信元/宛先ポート番号，タイムスタンプ，HTTP Method など），IP アドレスの Whois 情報などを用いて分類を行っている．ただし本手法は二部グラフを用いているので，複数の特徴量を用いるケースに拡張することは難しい．本研究では，Shi らと別の手法を用いることで，ドメイン情報の他に URL や HTTP ステータスコード，HTTP ユーザエージェントなど，複数の情報を加えた半教師あり学習を行う．

長田ら [11] は VAE をベースとした半教師あり学習手法を提案し，それをネットワーク IDS の検知アルゴリズムとして用いる方法について述べている．学習に必要となるラ

ベル付けされたトラヒックデータを得ることはコストの観点から難しいため，少数のラベルありデータと比較的大量に得られるラベルなしデータを用いた半教師あり学習を行うことでその課題を解決しようとしている．ただし，ここで用いている VAE は一般に，どういう特徴量で分類したのかが不明瞭な手法である．実運用を考えると，最終的な分析は SOC のセキュリティアナリストなどの人手で行うため，検知できた/検知できなかった理由が不透明だと分析に時間がかかり非効率となる．また，VAE は一般に計算コストが高いため，本研究のような大規模なデータを処理するケースに適用するのは難しいと考えられる．

3. 機械学習

本章では，本稿で用いた機械学習手法である Logistic Regression と Label Spreading の概要について述べる．本研究で適用する機械学習手法としては次の 2 要件を満たしていることが望ましい．第 1 に計算コストが小さいことである．本研究で扱う HTTP 通信ログは一般に大規模なデータであるため，実運用では，計算コストの大きい機械学習手法を用いることは難しい．第 2 に結果の解釈性があることである．実運用では最終的に SOC のセキュリティアナリストなどの人手で行うため，検知できた/検知できなかった理由は明確であることが望ましい．これらの要件から，本研究では Logistic Regression と Label Spreading を採用した．これらは比較的計算コストが小さく，重みなどを算出することで結果の解釈がある程度可能な手法である．

3.1 Logistic Regression

Logistic Regression を用いた 2 クラス分類について述べる [12]．2 クラスをそれぞれ，D_a，D_b とすると，これらの事後確率は，以下のように書ける．

$$p(D_a|\mathbf{x}) = \sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x}) \tag{1}$$

$$p(D_b|\mathbf{x}) = 1 - \sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x}) \tag{2}$$

$$\sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x}) = \{1 + \exp(-\mathbf{w}^{\mathrm{T}}\mathbf{x})\}^{-1} \tag{3}$$

ただし，$\mathbf{w} \in \mathbb{R}^f$ は重みベクトル，$\mathbf{x} \in \mathbb{R}^f$ は特徴ベクトルである．Logistic Regression を用いた 2 クラス分類では，前式において $p(D_a|\mathbf{x}) > 0.5$ ならばクラス D_a，$p(D_b|\mathbf{x}) > 0.5$ ならばクラス D_b に分類する．

次に，適切な $\mathbf{w}$ を決定する．学習用データが N 個あり，その n 番目の学習用データの特徴ベクトルが $\mathbf{x}_n$，ラベルが y_n で表されるとする．ただし，y_n は $y_n = 1$（D_a のクラス）か $y_n = 0$（D_b のクラス）をとる．このとき，尤度の負の対数をとって誤差関数を定義すると，この誤差関数は，

$$\begin{aligned} E_{LR} = \; & -\sum_{n=1}^{N}\{y_n \log \sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x}_n) \\ & +(1-y_n)\log(1-\sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x}_n))\} \end{aligned} \tag{4}$$

と書ける．この式は交差エントロピー誤差関数と呼ばれており，これを最小化する $\mathbf{w}$ を決定することが Logistic Regression における学習に相当する．

ただし，入力するデータの数に比べて特徴量が多すぎると，過学習が生じてしまう場合がある．そのような場合，十分な精度を出す方法として，L_2 正則化などが用いられている．L_2 正則化を加えた場合，式 (4) は重み C を用いて以下のように書ける．

$$E_{LR-L2} = CE_{LR} + \frac{1}{2}||\mathbf{w}||^2 \tag{5}$$

3.2 Label Spreading

Label Spreading を用いてラベルありデータからラベルなしデータにラベルを伝搬させる方法について述べる [13]．$\mathbf{x}$ をデータ点，y をラベルとする．このとき，l 個の組 $(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_l, y_l)$ をラベルありデータ，u 個の組 $\mathbf{x}_{l+1}, \cdots, \mathbf{x}_{l+u}$ をラベルなしデータと定義する．ラベルなしデータが D_a，D_b の 2 クラスに分類されるとすると，$y_i \in \{0, 1\}$ $(1 \leq i \leq l)$ となる．最終目標はラベルなしデータ $\mathbf{x}_k (l+1 \leq k \leq l+u)$ にラベルを付与することである．

ここで，$\mathbb{F}$ を $n \times 2$ の非負行列の集合とする（$n = l+u$）．このとき，行列 $F = [F_1^{\mathrm{T}}, \cdots, F_n^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{F}$ はベクトル関数で，各データ点 x_i に対してベクトル $F_i \in \mathbb{R}^2$ が割り当てられており，$y_i = \mathrm{argmax}_{j \leq 2} F_{ij} (1 \leq i \leq l)$ という関係が成立しているものとする．また，ラベル行列 $Y \in \mathbb{F}$ の第 i 行は $y_i = j$ となるとき $Y_{ij} = 1$，それ以外のときは $Y_{ij} = 0$ をとるものと定義する．このとき，Label Spreading のアルゴリズムは以下のようになる．

(1) アファイン行列 W を以下のように定義する．

$$\begin{aligned} & i \neq j \quad W_{ij} = \exp(-||\mathbf{x}_i - \mathbf{x}_j||^2/2\sigma^2) \\ & i = j \quad W_{ij} = 0 \end{aligned} \tag{6}$$

ただし，σ はハイパーパラメータである．なお，本稿では，$\gamma = 1/2\sigma^2$ と γ を定義する．

(2) (i, i) 成分が W の i 行の合計と等しい対角行列 D を用いて，$S = D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$ と S を定義する．

(3) 収束するまで，$F(t+1) = \alpha S F(t) + (1-\alpha) Y$ という計算を繰り返す．ここで，α は区間 $(0, 1)$ で定義されているものとする．このとき，α は初期ラベルの変更のしやすさを調整するパラメータで，この値が大きいほど初期ラベルが変わりやすくなる．

(4) $F^* = \lim_{t \to \infty} F(t)$ と定義する．このとき，$y_i = \mathrm{argmax}_{j \leq 2} F^*_{ij}$ となる．

4. 提案法

本章では，Label Spreading と Logistic Regression を用いた半教師あり学習により，Proxy ログからマルウェア感染端末を検知する分類器を作る手法について示す．手順の概要を以下に示す．

表 1　特徴量の候補
Table 1　Feature candidate

Destination IP Address
Destination Port Number
HTTP Method
HTTP User Agent
HTTP Status Code
Protocol
URL Scheme
URL Netloc
URL Path
URL Parameters
URL Query

手順の概要

(1) 良性/悪性の判断がすでにされている Proxy ログ（ラベルありデータ）と未解析の Proxy ログ（ラベルなしデータ）を学習用データとして用意する．
(2) ラベルありデータに良性=0，悪性=1 としたラベルを付与する．このラベルが y に相当する
(3) 特徴量を決定する
(4) 端末ごとに特徴量を抽出し，特徴ベクトルへ変換する．この特徴ベクトルが $\mathbf{x}$ に相当する
(5) Label Spreading を適用し，良性/悪性の判断がすでにされている Proxy ログを基準に，未解析の Proxy ログの良性度/悪性度のスコアに相当する F_{ij} を算出する
(6) 算出した悪性度がある閾値を下回ったもの/超えたものに良性/悪性のラベルを付与し，これらのデータと元々良性/悪性の判断がすでにされていた Proxy ログ（ラベルありデータ）をあわせて新たな学習用データとする
(7) 新たな学習用データに対して Logistic Regression を適用して分類器を作成する
(8) テストデータを先述と同じ特徴量で特徴ベクトル化し，分類器に代入してスコアを算出する
(9) スコアからテストデータの良性/悪性の判定を行う

(3) の特徴量について，本研究では表 1 のような特徴量の候補を考える．これらの特徴量の候補が良性/悪性の判断にどの程度影響するかについては検討が必要である．各特徴量の寄与度を算出するため，本研究では各候補 1 個 1 個に対して教師ありの L_2 正則化を含む Logistic Regression を適用し，AUC という指標を用いた．ここで求めた AUC が高いものは良性/悪性をうまく分別できることを意味しているため，特徴量として採用した．

(6) のラベル付与では，閾値をもうけて新たな学習用データに加えるラベルなしデータに制約を加えている．この閾値は調節されるべきハイパーパラメータである．あまり信頼できないままラベルを付与されたラベルなしデータを学習用データに組み入れてしまうと，分類器の精度が悪化してしまう可能性があるためこのような措置をとっている．本稿では，良性=0，悪性=1 とラベル付けしているので，閾値 T_u，T_l を設け，悪性度のスコアが閾値 T_u 以上のものを悪性，閾値 T_l 以下のものを良性と判定することとした．

5. 評価実験

本章では，先に示したマルウェア感染端末の検知手法の有効性を確認するため，「正常通信を行う端末」と「マルウェア由来の通信を行う端末」を分類する問題に対して，半教師あり学習を適用した結果と教師あり学習を適用した結果を比較する．

5.1 データセット

評価実験で用いたデータセットについて述べる．本稿では，「正常なログ=良性ログ」「マルウェア由来の疑わしいログ=悪性ログ」と定義する．良性ログとしては企業網の Proxy ログの一部を用いた．また，悪性ログとしてはマルウェア共有サイトで収集されたマルウェア検体を動的解析して得た HTTP トラヒックログの一部を用いた．なお，マルウェア検体は SHA1 Hash が異なり，各種ウイルス対策ソフトによる検知傾向などがなるべくランダムになるよう，日々最新のものを一定数収集し，動的解析している．

この実験では，ラベルなしデータを加えることによる効果を確認するため，ラベルなしデータの数を固定して，ラベルありデータのデータセットを 4 種類作成した（Dataset A〜D）．Dataset A から Dataset D に遷移するにつれて，ラベルありデータの数が多くなるようにデータセットを用意している．それぞれのラベルありデータセットの端末数を表 2 として示す．ただし，表 2 中，Train-leg は正常通信を行う端末，Train-mal はマルウェア由来の通信を行う端末をそれぞれ意味しており，表中の括弧内の数字は通信ログの数を意味している．なお，ラベルなしデータセットとしては，正常通信を行う端末：1000 端末（263037 ログ），マルウェア由来の通信を行う端末：1000 端末（24229 ログ），テスト用データセットとしては，正常通信を行う端末：200 端末（40554 ログ），マルウェア由来の通信を行う端末：200 端末（2502 ログ）とした．

また，これらのログが収集された時期を表 3 として示す．ただし，Train-leg-semi はラベルなしデータのうち正常通信を行う端末，Train-mal-semi はラベルなしデータのうちマルウェア由来の通信を行う端末，Test-leg はテストデータのうち正常通信を行う端末，Test-mal はテストデータのうちマルウェア由来の通信を行う端末を意味する．表 3 のように，テスト用データセットは学習用データセットより

表 2　ラベルありデータの端末数と通信ログ数
Table 2　Number of terminals and traffic logs of labeled data

Dataset	Train-leg	Train-mal
Dataset A	10 (80)	10 (113)
Dataset B	50 (4180)	50 (1731)
Dataset C	200 (15738)	200 (3928)
Dataset D	1000 (64812)	1000 (24492)

表 3　各通信ログの取得時期
Table 3　Acquisition period of traffic logs

	Acquisition period
Train-leg	Feb/22/2014 - Feb/26/2014
Train-mal	Mar/1/2015 - Mar/7/2015
Train-leg-semi	Feb/27/2014
Train-mal-semi	Mar/21/2015 - Mar/31/2015
Test-leg	Mar/4/2014 - Mar/5/2014
Test-mal	July/31/2015

も遅い時期に取得されたログを用いている．

5.2　特徴量

特徴量の候補として，表 1 のような量を考え，これらの候補の中から特徴量を決定した．特徴量を決定するため，各候補 1 個 1 個に対して教師ありの L_2 正則化を含む Logistic Regression を適用し，AUC という指標を用いて比較した．詳細は付録にて述べるが，AUC とは分類器の性能を表す指標で，0〜1 の間の値をとり，1 に近いほどよい分類器とされる．ここで求めた AUC が高いものは良性/悪性をうまく分別できることを意味しているため，特徴量として採用した．そのときの実験結果を表 4 として示す．なお表中の太字は各候補の AUC が 0.7 以上であることを示している．Logistic Regression のハイパーパラメータ C はそれぞれの AUC が最大となるように調整した．なお，実験の際は Dataset D とテスト用データセットを用い，これらの量をベクトル化する際には，通信端末単位で各特徴量において存在する全てのパターンを 1 つの要素とみなして，その要素が当該通信端末に出現したかどうかで 0/1 を割り当てた（Bag-of-words により特徴量を抽出した）．具体的に説明するため，元のデータセットの例を表 5 とし，それをベクトル化したものを表 6 として示す．ただし，表 5 中の Log1，Log2，Log3，Log4 は Terminal A の通信ログであり，Log5，Log6 は Terminal B の通信ログである．この例の場合，表 6 から，Terminal A は $[1,1,1,0,1,0]^{\mathrm{T}}$，Terminal B は $[1,0,0,1,1,1]^{\mathrm{T}}$ とベクトル化できる．本実験では，このような操作を全ての特徴量，各特徴量の全ての要素に対して行うことで特徴量をベクトル化している．

表 4 において，本稿では AUC が 0.7 以上であれば良性/悪性ログをうまく分類する特徴が抽出できていると考え，表 7 のような特徴量を採用した．

表 4　各候補に対する AUC
Table 4　AUC of each candidate

Feature	C	AUC
Destination IP Address	0.1	**0.993**
Destination Port Number	1	0.518
HTTP Method	0.0001	0.644
HTTP User Agent	10	**0.924**
HTTP Status Code	100	**0.846**
Protocol	1	0.5
URL Scheme	1	0.5
URL Netloc	5	**0.992**
URL Path	50	**0.982**
URL Parameters	1	0.5
URL Query	1	**0.748**

表 5　通信ログの中身の例
Table 5　Example of contents of traffic logs

		Dst IP	Dst Port
Terminal A	Log1	1.1.1.1	80
	Log2	2.2.2.2	80
	Log3	3.3.3.3	80
	Log4	2.2.2.2	80
Terminal B	Log5	4.4.4.4	8080
	Log6	1.1.1.1	80

表 6　ベクトル化した通信ログの例
Table 6　Example of vectorized traffic logs

	Dst IP				Dst Port	
	1.1.1.1	2.2.2.2	3.3.3.3	4.4.4.4	80	8080
Terminal A	1	1	1	0	1	0
Terminal B	1	0	0	1	1	1

表 7　採用した特徴量
Table 7　Adopted feature

Destination IP Address
HTTP User Agent
HTTP Status Code
URL Netloc
URL Path
URL Query

5.3　実験結果

半教師あり学習と教師あり学習による結果の差を比較するため，AUC を用いて比較する．それぞれを適用した結果を図 3 として示す．赤の実線が半教師あり学習による結果，青の点線が教師あり学習による結果である．本実験で用いたハイパーパラメータの値については付録 A.2 に記載している．図 3 から，Dataset A〜Dataset D の全てのパターンで半教師あり学習の精度は教師あり学習の精度を上回っており，ラベル密度が低い（ラベルなしデータがラベルありデータに比べて多い）ほど半教師あり学習と教師あ

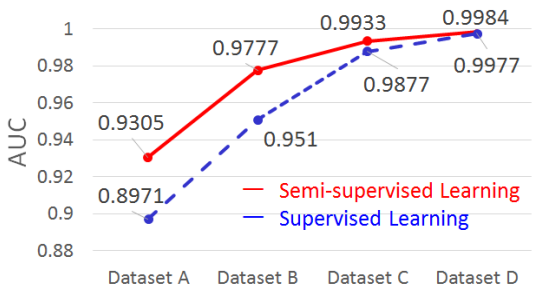

図 3　データセットごとの AUC
Fig. 3　AUC of each dataset

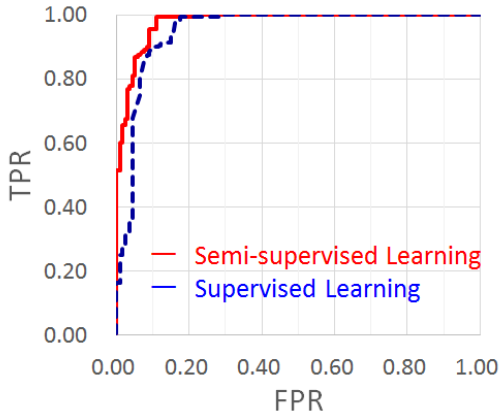

図 4　Dataset B に対する ROC 曲線
Fig. 4　ROC curve of Dataset B

り学習の AUC の差が大きいことが読み取れる．一般に，半教師あり学習はラベルなしデータの数が多くなればなるほど精度の向上が見込めることが知られており，本実験結果でも同様の傾向が見られたと言える．

また，Dataset B の場合で描いた ROC 曲線を図 4 として示す．図 4 から，同じ FPR（誤検知率：False Positive Rate）で運用した場合，半教師あり学習の検知率は教師あり学習の検知率を上回ることがわかる．

次に，誤検知率を抑制した際の検知率についても比較する．セキュリティ分野では，誤検知が多いと可用性が失われるため，誤検知が少ない場合の検知率が重視される．そこで，テストデータの正常通信を行う端末に対しての誤検知率が 1%になるように閾値を調整した際の検知率を，半教師あり学習および教師あり学習のそれぞれに対して算出した結果を表 8 に示す．表 8 から，いずれの場合も半教師あり学習による結果が教師あり学習による結果を上回っていることが読み取れる．特に，Dataset B では FPR=1%のときの TPR が約 35 ポイント向上していることがわかる．

最後に計算時間について記す．実運用では膨大なログの数を処理しなければならないため，計算コストが小さい機械学習手法を用いる必要がある．Dataset D を用いて，本実験で要した時間を計測すると，教師あり学習の場合が約

表 8　FPR=1%の際の半教師あり学習と教師あり学習の TPR
Table 8　TPR of semi-supervised learning and supervised learning when FPR=1%

	Supervised Learning	Semi-supervised Learning
Dataset A	0.07	**0.095**
Dataset B	0.165	**0.515**
Dataset C	0.54	**0.765**
Dataset D	0.92	**0.98**

30 秒，半教師あり学習は約 105 秒となり，十分に計算コストが小さいことが確かめられた．なお，本実験で用いた実験環境は CPU：Intel(R) Xeon(R) E5-2660 v3 2.60GHz，OS：Ubuntu 14.04 である．

6.　まとめ

本稿ではグラフベースの半教師あり学習を用いて，HTTP 通信ログからマルウェア感染端末を検知する手法を提案した．また，AUC や誤検知率が 1%になるように閾値を調整した際の検知率，という指標を用いて，提案法と従来の教師あり学習に基づく手法の精度を比較した．AUC の比較では，用いたデータセットの全てのパターンで提案法の優位性が確認され，ラベル密度が低い場合ほど精度の向上が見られることを示した．また，誤検知率が 1%になるように閾値を調整した際の検知率の比較でも，用いたデータセットの全てのパターンで提案法の優位性が確認された．

本結果から，半教師あり学習を用いることで，ラベル付けの負担を軽減しつつ汎用性の高いマルウェア感染端末の分類器を作成できることが示せた．このことから，ラベルなしデータとして新種のマルウェアに関連した情報を含みうる未解析な通信ログを活用して半教師あり学習を行うと，新種のマルウェアの一部に対応した分類器が作成できることが期待できる．

今後の課題として，HTTP 通信ログの特徴量の種類/とり方のさらなる検討が挙げられる．特徴量の種類としては，URL の長さ，ドメイン名に IP アドレスのような数字を含むか否か，URL に拡張子が入っているか否か，など様々なものが候補として考えられる．また，外部の情報源を用いて，GeoIP 情報やドメインのレピュテーションなどを取得し，特徴量として加える方法なども考えられる．特徴量のとり方も検討すべき事項の一つである．例えば，Daeef ら [14] は URL を特徴ベクトル化する際に本稿と同じくホスト名，パス，クエリに着目しているが，Bag-of-words ではなく，N-gram 法を用いて特徴ベクトル化している．

また，提案法を用いて検知できたマルウェアに関してはなぜ検知できたのか，どういったマルウェアが検知できたのか，などの考察は今後行う予定である．Label Spreading は類似度（スコア），Logistic Regression は各特徴量の寄与率にあたる数値を算出することができるため，これらを用

いることで結果の解釈が可能であると考えられる．

参考文献

[1] NISC サイバーセキュリティ戦略本部，日本年金機構における個人情報流出事案に関する原因究明調査結果，2015.
[2] 国土交通省，旅行業界情報流出事案検討会中間とりまとめ-旅行業情報セキュリティ向上のため早急に構ずべき対策-，2016.
[3] AV-Test, https://www.av-test.org/en/statistics/malware/.
[4] 独立行政法人情報処理推進機構，「新しいタイプの攻撃」の対策に向けた設計・運用ガイド，2011.
[5] Gartner, Magic Quadrant for Mangaed Security Services, Worldwide, 2015.
[6] Intel Security，SOC 運用はどんな業務で成り立っているのか？，2015.
[7] Jiyong Jang ei al., BitShred: Feature Hashing Malware for Scalable Triage and Semantic Analysis，Proceedings of the 18th ACM Conference on Computer and Communications Security, CCS 2011，309/320，2011.
[8] Anton Dan Gabriel et al., Detecting malicious URLs. A semi-supervised machine learning system approach, 18th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, IEEE, 233/239, 2016.
[9] Dragos Gavrilut et al., Optimized zero false positives perceptron training for malware detection, In 14th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, SYNASC, 247/253, 2012.
[10] Liang Shi et al., A hybrid learning from multi-behavior for malicious domain detection on enterprise network, 15th International Conference on Data Mining Workshops, IEEE, 2015.
[11] 長田 他，Variational Auto-Encoder を用いた半教師あり学習によるネットワーク侵入検知，2017 Symposium on Cryptography and Information Security, SCIS, 2017.
[12] C.M. ビショップ et al.，パターン認識と機械学習-上，丸善出版，2012.
[13] D. Zhou et al.，”Learning with Local and Global Consistency”，Advances in Neural Information Processing Systems 16，NIPS，2003.
[14] Ammar Yahya Daeef et al., Wide scope and fast websites phishing detection using urls lexical features, 2016 3rd International Conference on Electronic Design, IEEE, 410/415, 2016.

付　　録

A.1　性能評価指標

ここでは性能比較に用いた TPR，FPR や AUC といった指標について説明する．まず，以下の表のように，真陽性（TP: True Positive），偽陽性（FP: False Positive），偽陰性（FN: False Negative），真陰性（TN: True Negative）を定義する．

このとき，検知率（TPR: True Positive Rate）および誤検知率（FPR: False Positive Rate）は次のようになる．

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{A.1}$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} \tag{A.2}$$

表 **A·1**　混合行列
Table A·1　Confusion matrix

		Actual	
		Malicious	Legitimate
Test	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

次に，AUC について説明する．AUC とは Area Under the Curve の略で，分類器の性能を表す指標である．縦軸に TPR，横軸に FPR を二次元プロットして連結した曲線のことを ROC 曲線（Receiver Operating Characteristic）と呼んでいるが，AUC は ROC 曲線の下の部分の面積を算出することで得られる．AUC は必ず 0〜1 の間の数字をとり，1 に近ければ近いほど性能が高い分類器と言える．

A.2　ハイパーパラメータ

実験で用いたハイパーパラメータを示す．教師あり学習（Logistic Regression）で用いたハイパーパラメータを表 A – 2，半教師あり学習（提案法）で用いたハイパーパラメータを表 A – 3 とする．また，提案法の途中で Label Spreading を適用して悪性度が閾値を下回った/超えたラベルなしデータにラベルを付与しているが，その際に付与したラベルの数を表 A – 4 として示す．表 A – 4 中の Sum はラベルを付与した総計である．

表 **A·2**　教師あり学習のハイパーパラメータ
Table A·2　Hyper-parameter of supervised learning

	C
Dataset A	0.001
Dataset B	0.03
Dataset C	50000
Dataset D	50000

表 **A·3**　半教師あり学習のハイパーパラメータ
Table A·3　Hyper-parameters of semi-supervised learning

	C	γ	α	T_u	T_l
Dataset A	0.006	1	0.1	0.99	0.01
Dataset B	10	2.3	0.2	0.99	0.01
Dataset C	900	3	0.2	0.99	0.01
Dataset D	300	4	0.99	0.99	0.01

表 **A·4**　Label Spreading によるラベル付与数
Table A·4　Number of label assignment by Label Spreading

	TN	FP	FN	TP	Sum
Dataset A	286	0	0	93	379
Dataset B	504	3	0	203	710
Dataset C	564	2	16	440	1022
Dataset D	608	1	0	691	1300