動的解析ログを用いた特徴量の予測によるマルウェアの早期機能推定に関する検討

朝倉 紗斗至[1,a)]　中川 恒[2]　押場 博光[2]　吉浦 裕[1]　市野 将嗣[1]

概要：近年，巧妙化するマルウェアを用いたサイバー攻撃により，その侵入を検知することが難しくなっている．マルウェアの侵入後においてその被害を最小限に抑えるため，早期にそのマルウェアの機能を推定することが必要であると考える．そこで，本研究では動的解析ログにおいて，記録時間の短いログの特徴量から長いログの特徴量を予測し，予測した特徴量を機能推定に利用することを提案する．実際に，MWS Datasets の一部として提供されている Soliton Dataset 2019 に含まれる動的解析ログ（Mark Ⅱ ログおよび Cuckoo ログ）を使用し，提案手法を用いた機能推定実験を行った．その結果，Mark Ⅱ ログにおいて記録開始から 5 秒までのログに対し 2.6%，および Cuckoo ログにおいて記録開始から 1 秒までのログに対し 0.8%の推定精度の向上が見られた．

キーワード：マルウェア，動的解析ログ，機能推定，早期推定，特徴量予測

A Study on Early Function Estimation of Malware by Prediction of Features Using Dynamic Analysis Log

Satoshi Asakura[1,a)]　Ko Nakagawa[2]　Hiromitsu Oshiba[2]　Hiroshi Yoshiura[1]
Masatsugu Ichino[1]

Abstract: In recent years, cyber attacks using sophisticated malware have made it difficult to detect intrusions. In order to minimize the damage after a malware infiltration, we think it is necessary to estimate the function of the malware at an early stage. In this paper, we propose to use the predicted features of the dynamic analysis log for function estimation from the features of logs with a short recording time. We conducted function estimation experiments using the proposed method with the dynamic analysis logs (Mark Ⅱ logs and Cuckoo logs) included in the Soliton Dataset 2019, which is provided as part of MWS Datasets. The results showed 2.6% and 0.8% improvement in the estimation accuracy of Mark Ⅱ logs and Cuckoo logs from the start of recording to 5 seconds and 1 second, respectively.

Keywords: Malware, Dynamic Analysis Logs, Function Estimation, Early Estimation, Feature Prediction

1. はじめに

近年，マルウェアの巧妙化により検知が困難になっている [1]．そこで，最近ではマルウェアによる侵入を前提にした対策も必要となってきている [2]．

一般にマルウェアの挙動はいくつかの機能として分けられ，他のマルウェアをダウンロードしたり，データの暗号化を行うといった機能が存在する [3]．このようなマルウェアの機能をその挙動から推定することができれば，その機能に応じた対策を講じることができ，侵入後の被害を抑えることができると考える．例えば，挙動からダウンロードという機能が推定できた場合，外部との通信を遮断するなどして，これを防ぐことができる．

また，侵入後の被害を最小に抑えるためには，マルウェ

[1] 電気通信大学
The University of Electro-Communications
[2] 株式会社 FFRI セキュリティ
FFRI Security, Inc.
[a)] asakura@uec.ac.jp

アの機能の推定を早期に行う必要がある．マルウェアが既に目的を達成し終えた後に，その機能が推定できたとしても，被害を抑えることはできないためである．例えば，あるマルウェアがダウンロードの機能を持つと推定できた時点で，既に目的のダウンロードが完了していた場合，対策を講じても間に合わない．

マルウェアの挙動を記録したログに着目すると，記録時間が短いと挙動の差が少ないことが分かる．実際に 3 章の予備実験および 5 章の実験で使用したログ（Mark Ⅱ ログ）では，記録開始から 4 秒までの間では，プロセスの起動という挙動がほとんどであり，類似していた．一方，記録時間が長くなるにつれて，ファイルやレジストリの操作を行う挙動が増え，検体間の挙動の差が大きくなっていた．

つまり，マルウェアの機能を早期に推定するために短時間の挙動だけをそのまま用いるという方法では，長時間の挙動を用いた推定に比べ，推定精度が低下すると予想できる．実際に，3 章の予備実験の結果（表 4，5）では記録時間の短いログ（以降 “短いログ” と呼ぶ）を用いた方が記録時間の長いログ（以降 “長いログ” と呼ぶ）を用いた場合に比べ精度が低くなる傾向があった．マルウェアの早期推定に関する既存研究 [4], [5] では，推定に利用した分類モデルの学習データとテストデータに短いログのみを使用しており，先ほど述べた記録時間による精度の低下に対する工夫はされていない．

そこで本研究では，短いログからその後の挙動を予測するによって，早期により高精度で機能推定する方法を提案する．また，この提案手法は，侵入後の被害を抑えるためだけでなく，長いログを保持している状態で，それを利用して短いログの分析に使用したいという状況であれば役立つと考えている．例えばマルウェアの動的解析において早期に機能を推定することにより，解析の補助および効率化を行うことができると考える．本稿では予測による精度向上の有効性を確認するための検討段階として，記録開始から十数秒までのログのみを利用して検証を行った．実際には様々な状況に応じて対応できるように，さらに長い時間のログを使用した予測も検証する必要があると考える．

本研究では，最初に提案手法を用いずにログの記録時間を変えた機能推定を予備実験として行った．そして，次に提案手法を用いた機能推定を行った．

以下 2 章で関連研究，3 章で予備実験，4 章で提案手法，5 章で提案手法による実験，6 章で実験結果と考察，最後に 7 章でまとめと今後の課題について記述する．

2. 関連研究

本章では，マルウェアの機能推定を行っている既存研究と，マルウェアの早期推定を行っている既存研究について述べる．また，それらを紹介した後，本研究の位置付けについて，紹介した研究に触れながら説明する．

2.1 マルウェアの機能推定

森ら [6] は動的解析ログを用いてマルウェアの役割推定に関する検討を行うために，マルウェアの特徴的な挙動の分析を行った．役割推定実験の結果，70.5%の推定精度を得た．Kawaguchi ら [4] は，近年複雑化するマルウェアの分析を効果的に行うために，マルウェアの動的解析ログである API コールログを用いてマルウェアの機能を推定する方法を提案した. 実験の結果，平均で 83.4%の推定精度を得た.

2.2 マルウェアの早期推定

Kawaguchi ら [4] は，マルウェアの初期の挙動として記録開始から 90 秒もしくは 120 秒までの間のログを用いて機能推定を行った．Rhode ら [5] は，記録開始から 5 秒までの間に記録された動的解析ログを用いてマルウェアの検知を行った．実験の結果 94%の推定精度を得た．

2.3 本研究の位置づけ

2.1 節で述べた文献 [6] ではマルウェアの機能推定は行っているが，早期推定を想定した推定手法ではなかった．2.2 節で述べた文献 [4] においても，使用しているデータは短いログだが，早期推定を意図した方法は行っていなかった．文献 [5] も記録時間を変えて精度を検証してはいるが，短いログのみを検知に使用するのに留まっていた．1 章でも述べたが，これらの方法では機能推定において，ログの記録時間が短いことによる検体ごとの挙動差が小さいままであり，推定精度が低下していると予想できる．実際に 3 章の予備実験でもその傾向は確認している．

そこで，本研究では短いログの情報量の少なさに起因する機能推定の精度低下を改善するための手法を提案する．そして，提案手法を用いて，短いログに対する機能推定の精度の向上を実験で検証する．

3. 予備実験

まず，機能推定に使用するログの記録時間を変えることによる機能推定の精度の変化を調査した．ここでは，あらかじめ定義した機能に決定木の分類モデルを用いて分類することで機能推定とした．予備実験として，分類モデルの学習データとテストデータのログの記録時間が同じ場合（3.6 節）と異なる場合 (3.7 節) の二つを行った．前者の実験概要を図 1 に，後者の実験概要を図 2 表す．

3.1 使用データ

MWS Datasets [7] の一部として提供されている Soliton Dataset 2019 を使用した．Soliton Dataset 2019 とはエンドポイントセキュリティ製品である “InfoTrace Mark I I for Cyber” [8]（以降 “Mark Ⅱ” と呼ぶ）が導入された環境でマルウェアを実行して得られた動的解析ログを中

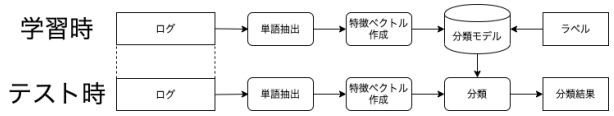

図 1　学習データとテストデータでログの記録時間が同じ機能推定

Fig. 1　Function estimation where the same recording time is used for training data and test data.

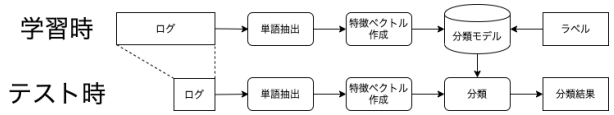

図 2　学習データとテストデータでログの記録時間が異なる機能推定

Fig. 2　Function estimation where different recording time is used for training data and test data.

心としたデータセットである．Cuckoo Sandbox [9]（以降 “Cuckoo”と呼ぶ）上で Mark Ⅱ を導入したゲスト環境の使用により，1 検体につき Mark Ⅱ ログと Cuckoo ログの二つがある．対象のマルウェアは 2018 年 1 月から 2019 年 3 月までに観測されたもので，VirusTotal [10] でマルウェア名（Emotet，Dreambot，Trickbot，ZACOM，TAIDOOR，IXESHE，DASERF，RedLeaves，CHCHES，ANEL，PLUGX，sorebrect，Oni，gandcrab）でサーチされた検体，および 10 以上のアンチウイルスエンジンでマルウェアと判定されている検体である。合計 485 検体のマルウェアで，1 検体につき，5 分間を基本として実行されている．

Cuckoo ログでは API 単位の詳細な挙動が記録されるため，より多くの情報が得られるが，僅かな時間でサイズが膨大になる．一方で，Mark Ⅱ ログでは挙動がイベント単位で保存されるため，Cuckoo ログほど詳細な情報が記録されるわけではないが，その分サイズが比較的小さい．この性質から，前者はリアルタイムに利用され，後者はログを多く保存できるため長期に渡っての解析が行える．以上の理由から，両者のログにおいて検証を行うことは，より状況に応じた対応ができるという点で有用だと考える．

本研究では，485 検体のうち Cuckoo ログのサイズが 1MB 以上かつ 3.2 節で説明するラベル付けをした 114 検体の Mark Ⅱ ログおよび Cuckoo ログの両方を使用した．

3.2　ラベル付け

マルウェア 1 検体ごとにラベル付けを行った．ここでのラベルとはマルウェアが持つ機能のことを指す．本研究ではラベルは “ダウンロード”，“暗号化”，“拡散”，“情報窃取”，“マイニング”，“バックドア”の 6 種類とした．また，1 検体につき一つのラベルを付与した．ラベル付けを行うにあたり，セキュリティベンダーが公開しているレポート [11], [12] を主に参考にした．ラベル付けを行った結果，

表 1　ラベルごとの検体数

Table 1　Number of samples per label.

ラベル（機能）	検体数
ダウンロード	21
暗号化	3
拡散	4
情報窃取	21
マイニング	40
バックドア	25
合計	114

ラベルごとの検体数の内訳は表 1 のようになった．

3.3　単語の抽出

ログから一定の規則に基づいて文字列を抽出する．本稿ではこの文字列を単語と呼ぶ．Mark Ⅱ ログと Cuckoo ログにおける単語抽出の規則を表 2 に示す．

Mark Ⅱ ログにおいては基本的にイベント (ログ中の evt) とサブイベント (ログ中の subEvt）をコロン（“:”）で繋げたものを単語とした．例外としてイベントが “file”，サブイベントが “close”の時，ファイルの読み込みおよび書き込みのバイト数によって “file:read”や “file:write”とした．これにより，単語数が増え，得られる挙動の情報量を増やした．

Cuckoo ログにおいては API コールを単語とした．

Mark Ⅱ ログからは 26 種類，Cuckoo ログからは 209 種類の単語が抽出された．

3.4　特徴ベクトルの作成

3.3 節で作成した単語ごとの出現回数を特徴ベクトルの要素とした（図 3）．また，ログの記録時間を制限したものから特徴ベクトルを作成した．例えば，記録開始から 5 秒後までのログ（以降このようなログを “5 秒のログ”と呼ぶ）から特徴ベクトルを作成する時，その 5 秒のログから抽出された単語の出現回数のみを使用した．ただし，5 秒後以降のログから抽出された単語が 5 秒より前に出現しなかった場合，出現回数 0 回として特徴ベクトルに追加した．本来は記録したログに対してはこの処理はできないため，実際に使用する場合は出現の可能性がある単語をあらかじめ定義しておきその出現回数を記録するなどの工夫が必要である．

1 検体および記録時間の制限一つにつき，Mark Ⅱ ログと Cuckoo ログの二つの特徴ベクトルを作成した．

3.5　分類

3.4 節で作成した特徴ベクトルを標準化したものを 3.2 節のラベルと共に決定木の分類モデルに使用して 1 個抜き交差検証を行った．式 (1) で標準化を行い，各特徴量の平均を 0，分散を 1 にした．ただし，式 (1) 中の x は標準化

表 2　単語の抽出規則
Table 2　Word extraction rule.

ログの種類	条件	単語の形	単語の例
Mark Ⅱ	evt=“file” かつ subEvt=“close” かつ read!=0 かつ write=0	“file” + “:” + “read”	file:read
	evt=“file” かつ subEvt=“close” かつ read=0 かつ write!=0	“file” + “:” + “write”	file:write
	上記以外の場合	evt + “:” + subEvt	ps:start
Cuckoo	全ての場合	API コール	NtReadFile

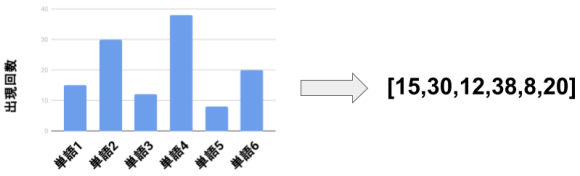

図 3　特徴ベクトルの作成
Fig. 3　Feature vector creation.

表 3　分類モデル（決定木）のハイパーパラメータのチューニング範囲
Table 3　Possible hyperparameter values of classification model (decision tree).

種類	範囲
criterion	“gini”, “entropy”
max_depth	5,7,9

前の各特徴量の値であり，x_{std} は標準化された値である．また，μ と σ はそれぞれ標準化前の特徴量の平均と分散である．

$$x_{\mathrm{std}} = \frac{x - \mu}{\sigma} \tag{1}$$

分類モデルの実装には scikit-learn 0.22.1 [13] を使用した．また，ハイパーパラメータのチューニングとして学習データで層化 3 分割交差検証を行った．パラメータの範囲は表 3 に示す．表 3 に記載のないパラメータにおいてはデフォルトのものを用いた．

3.6　学習データとテストデータに同じ記録時間のログを用いた機能推定実験

分類に使用する学習データとテストデータのログの記録時間を同じにして機能推定を行った．

Mark Ⅱ ログでは 4 秒のログ，5 秒のログ，…,15 秒のログの合計 12 の異なる記録時間のログごとに機能推定を行った．Cuckoo ログでは 1 秒のログ，2 秒のログ，…,10 秒のログの合計 10 の異なる記録時間のログごとに機能推定を行った．

3.7　学習データとテストデータに異なる記録時間のログを用いた機能推定実験

分類に使用する学習データのログの記録時間をテストデータのログの記録時間より長くして機能推定を行った．

実験は Mark Ⅱ ログのみを用いた．テストデータを 5 秒のログに固定し，学習データを 6 秒のログ，7 秒のログ,…,15 秒のログの合計 10 の異なる記録時間のログとし，異なる学習データごとに機能推定を行った．

3.8　結果

3.6 節の Mark Ⅱ ログの結果を表 4 に，Cuckoo ログの結果を表 5 に示す．3.7 節の Mark Ⅱ ログの結果を表 6 に示す．

表 4，5 の結果から，Mark Ⅱ ログと Cuckoo ログの両方において，短いログを用いた方が精度が低く，長いログを用いた方が精度が高くなる傾向があることが分かった．また，表 4 の 5 秒の時の精度と，表 6 の結果から，Mark Ⅱ ログにおいて，学習データとテストデータに使用するログの記録時間の差が小さいと精度が高くなり，大きくなると精度が低くなる傾向があることが分かった．特に，表 4 の 5 秒の時の精度が 66.7%であるのに対し，表 6 で最も良い精度が 8 秒の時の 41.2%であり，精度に大きい差があることから，ログの記録時間が同じである場合，精度が高いことが分かった．

4.　提案手法

表 4,5 より，短いログより長いログを用いた場合が精度が高くなることが分かった．早期推定に関する既存研究 [4], [5] ではこの特性を早期推定に利用していなかった．また，表 4 の記録時間が 5 秒の時の分類精度と，表 6 の結果より，分類モデルの学習データとテストデータに使用するログの記録時間の差は小さい方が精度が高いことが分かった．

そのためこれらの特性を利用して，本研究では早期に機能推定をより高精度に行うため，短いログから得られる特徴量から長いログから得られる特徴量を予測し，分類モデルの学習データとテストデータに使用するログの挙動を近づけて機能推定を行うことを提案する．提案手法の様子を図 4 に示す．

図 4 の分類モデルのテストデータには予測によって得られた特徴ベクトルを，学習データには予測を行わずに保有しているログから得た特徴ベクトルを使用している．学習時の長いログには 10 秒のログを使用し，テスト時の短いログに 5 秒のログを使用したとすると，最初に 10 秒のロ

表 4　学習データとテストデータに同じ記録時間の Mark Ⅱ ログを用いた機能推定実験の結果
Table 4　Results of function estimation experiments using Mark Ⅱ logs with the same recording time for training and test data.

記録時間 (s)	4	5	6	7	8	9	10	11	12	13	14	15
分類精度 (%)	39.4	66.7	74.6	78.1	76.3	77.2	78.9	77.2	76.3	77.2	76.3	76.3

表 5　学習データとテストデータに同じ記録時間の Cuckoo ログを用いた機能推定実験の結果
Table 5　Results of function estimation experiments using Cuckoo logs with the same recording time for training and test data.

記録時間 (s)	1	2	3	4	5	6	7	8	9	10
分類精度 (%)	71.1	70.2	72.8	75.4	67.5	71.9	81.6	79.8	76.3	73.7

グを学習データとして分類モデルを作成する．次に，5 秒のログから特徴量を作成し，その特徴量を 10 秒のログの特徴量に近づけるように予測し，予測された新たな特徴量を得る．この予測された特徴量を最初に作成した分類モデルのテストデータとして用いることで，分類を行う．

本提案手法により，マルウェアの短時間の挙動から機能を推定し，その機能に応じて対策を講じることで，被害が拡大する前に抑えることができると考える．また，マルウェアの動的解析においても，提案手法を用いることにより，早期にマルウェアの挙動を推定・分析することで，解析にかかるマルウェアの実行時間を減らすことができ，解析の効率化および補助が行えると考える．

5. 早期の機能推定に関する実験

実験の概要は提案手法で述べた，図 4 と同様である．使用データおよび，図 4 中の “単語抽出”，“特徴ベクトル作成”，“分類モデル” については 3 章で述べたものと同様である．

5.1 特徴ベクトルの予測

特徴ベクトルの予測の概要を図 5 に表す．図 5 は 5 秒のログの特徴ベクトルから 10 秒のログの特徴ベクトルを予測する様子である．

本研究では特徴ベクトルの予測は回帰で行った．回帰の説明のため，一般的な回帰式を例として式 (2) に示す．式 (2) の x_i を説明変数，y を目的変数という．a_i と b は係数であり，学習データを用いて求める．式 (2) の場合，x_i に各特徴量を入力し，y として予測された値が出力される．

$$y = a_1x_1 + a_2x_2 + \cdots + a_ix_i + b \tag{2}$$

ランダムフォレストを用いて回帰モデルを作成して一個抜き交差検証を行い，使用したテストデータの検体の特徴ベクトルを予測した．

説明変数には，特徴ベクトルにその特徴ベクトルの要素の和を新たな特徴量として加えたベクトルを標準化して検体間の分散が 0 の特徴量を除外したものを使用した．特徴ベクトルの要素の和を新たな特徴として追加したのは，予測性能を向上を図るためである．また，標準化は 3.5 節と同様に行った．説明変数に使用する特徴ベクトルは，Mark Ⅱ ログなら 5 秒，Cuckoo ログなら 1 秒のログから作成されたものに固定した．説明変数として，分散が 0 の特徴量を除外した結果，Mark Ⅱ ログの場合，27 個あった特徴量が全ての検証で 20 個になった．Cuckoo ログの場合，210 個あった特徴量が検証ごとの平均で 146.9 個になった．

表 7　回帰モデル（ランダムフォレスト）のハイパーパラメータのチューニング範囲
Table 7　Possible hyperparameter values of regression models (random forests)

使用ログ	種類	値
Mark Ⅱ	n_estimators	14,15,16,17,18
	max_depth	10,11,12,13,None
	max_samples	0.9
Cuckoo	n_estimators	5,7,9,11,13
	max_depth	6,7,8,9,10,11,None
	max_samples	None

目的変数には，Mark Ⅱ ログでは 6 秒，7 秒，⋯,15 秒のログの合計 10 個のログの特徴ベクトルを，Cuckoo ログでは 2 秒，3 秒，⋯,10 秒のログの合計 9 個のログの特徴ベクトルを用いた．また，目的変数の特徴ベクトルごとに回帰モデルを作成した．

実装には scikit-learn 0.22.1 [13] を使用した．回帰モデルのハイパーパラメータのチューニングは，最初に学習データを回帰モデルで学習させたあと, 同じ学習データで特徴ベクトルの予測を行った．その後，予測された特徴ベクトルを 5.2 節と同様の方法で分類し，その精度が最も良くなるものをパラメータとして選択した．パラメータのチューニングの範囲を表 7 に示す．表 7 に記載のないパラメータにおいてはデフォルトのものを用いた

5.2 分類

分類は 3.5 節と同様に行った．ただし，図 4 のように，分類モデルに使用する学習データには実際に取得したログの特徴ベクトル（図 4 中の長いログ）を，テストデータには短いログから予測された特徴ベクトルを用いた．

表 6　記録時間を固定したテストデータ（5 秒）に対し異なる記録時間の学習データを用いた機能推定実験の結果（Mark Ⅱ ログ使用）

Table 6　Results of function estimation experiments using test data with a fixed recording time (5 seconds) and training data with different recording times (using Mark Ⅱ logs).

学習データの記録時間 (s)	6	7	8	9	10	11	12	13	14	15
分類精度 (%)	35.1	39.5	41.2	31.6	32.5	29.8	32.5	28.9	38.6	30.7

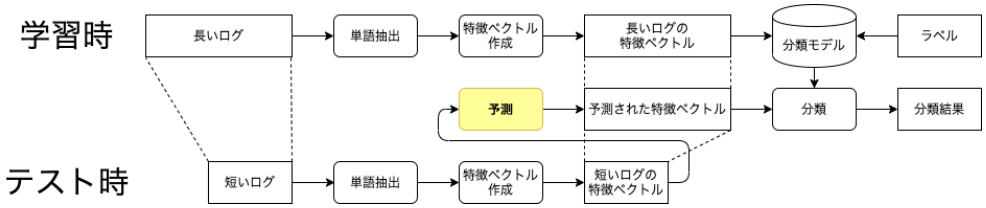

図 4　特徴量を予測して学習データにテストデータの挙動を近づけた機能推定

Fig. 4　Function estimation that approximates the behavior of test data to training data by predicting features.

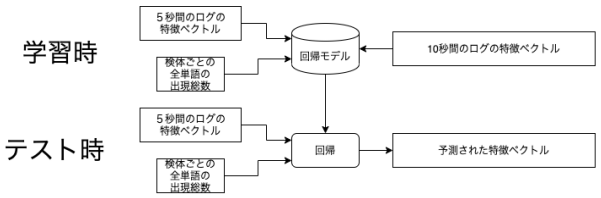

図 5　特徴ベクトルの予測

Fig. 5　Feature Vector Prediction.

6. 実験結果と考察

6.1 実験結果

記録時間ごとの分類精度について，5 章の Mark Ⅱ ログを用いた実験の結果を表 8 に，Cuckoo ログを用いた実験の結果を表 9 に示す．表 4 の記録時間が 5 秒の時および表 5 の記録時間が 1 秒の時の分類精度に対し，精度が向上しているものは表 8 と表 9 にて太字として表した．

6.2 考察

6.2.1 特徴量の予測による機能推定の精度向上

最初に，テストデータとして短いログが得られた時，提案手法により精度が向上するかを確認する．まず，Mark Ⅱ ログ使用時において，表 4 の記録時間が 5 秒の時の分類精度は 66.7%であった．提案手法を用いた結果である表 8 と比較すると，学習データの記録時間によっては精度が向上しない場合もあったが，テストデータの記録時間が 5 秒で，学習データの記録時間が 8 秒，10 秒，12 秒の時 66.7%を上回り，最大で 8 秒の時，2.6%精度が向上した．次に Cuckoo ログ使用時は，表 5 の記録時間が 1 秒の時の分類精度は 71.1%であった．提案手法を用いた結果である表 9 と比較すると，テストデータの記録時間が 1 秒で，学習データの記録時間が 7 秒の時のみ 71.1%を上回り，0.8%精度が向上した．以上より，学習データの記録時間により精度は左右されてはいるが，いくつかの記録時間においては Mark
Ⅱ ログと Cuckoo ログの両方において精度が向上し，提案手法の有効性を確認した．Mark Ⅱ ログ使用時は複数の場合で精度が向上したが,Cuckoo ログ使用時では，精度が向上したのは学習データの記録時間が 7 秒の時のみであり，精度の向上も比較的少なかった．これは，学習データとテストデータに同じ記録時間のログを用いた時の結果（表 4, 5）において，Mark Ⅱ ログでは記録時間が 5 秒の時の分類精度が 66.7%であり，その後の記録時間での分類精度と比較して低かったのに対して，Cuckoo ログでは記録時間が 1 秒の時の分類精度が 71.1%とその後の記録時間での分類精度比較して，そこまで低くない数値だったことに関係していると考える．つまり，高精度に特徴量の予測できていたとしても，予測目的対象の特徴量を使用した時の分類精度が相対的に高いものでなければ，予測による精度向上は望めないと考えられる．Cuckoo ログを用いた今回の提案手法によるさらなる精度向上を行うための方法の一つとして，テストデータとして記録時間が 1 秒未満のログを用いるということが考えられる．1 秒未満のログでは挙動情報がさらに少なくなり，検体間で挙動に差が出ず，その結果同じ時間同士の機能推定では分類精度が低くなると予想できる．そして，予測目的対象の特徴量を使った時の分類精度との差が大きくなるため，より予測による精度向上の効果が出やすいと考える．言い換えれば，Cuckoo ログは Mark Ⅱ ログに比べ，より早期の推定に適応していると考えられる．

次に，テストデータの記録時間を固定し，学習データの記録時間を変化させた時に，提案手法を使用した場合に精度が向上するかを確認する．Mark Ⅱ ログを使用した結果である表 6 と表 8 を比較すると，表 6 の全ての分類精度に対し，表 8 の結果が上回っていることがわかる．このこと

表 8　Mark Ⅱ ログを用いたテストデータ（5 秒）の予測による機能推定実験の結果
Table 8　Results of experiments to estimate functions by predicting test data (5 seconds) using Mark Ⅱ logs.

学習データの記録時間 (s)	6	7	8	9	10	11	12	13	14	15
分類精度 (%)	65.8	60.5	**69.3**	63.2	**67.5**	66.6	**67.5**	64.9	64.0	64.0

表 9　Cuckoo ログを用いたテストデータ（1 秒）の予測による機能推定実験の結果
Table 9　Results of experiments to estimate functions by predicting test data (1 second) using Cuckoo logs

学習データの記録時間 (s)	2	3	4	5	6	7	8	9	10
分類精度 (%)	69.3	68.4	66.7	59.6	68.4	**71.9**	65.8	62.3	67.5

より，本提案手法により，分類精度が向上したことがわかる．これは，予備実験の結果（3.8 節）と同様に，異なる時間の間に記録された挙動同士を学習データとテストデータとして比較するより，同じ時間同士の挙動に予測して近づけてから比較する方が精度が向上するためだと考えられる．

6.2.2　予測性能に関する評価

最初に，予測目的対象の記録時間を変えた時の予測性能と分類精度の関係について考察する．5.1 節での回帰モデルの予測性能の評価のため，全検体の予測した特徴ベクトルと予測目的対象である実際に取得したログから得た特徴ベクトルの平均二乗誤差を求めた．Mark Ⅱ ログおよび Cuckoo ログを用いた時の特徴ベクトルの平均二乗誤差を表 10，11 に示す．表 10，11 より，予測先の時間が大きくなればなるほど平均二乗誤差が大きくなる傾向があることがわかる．これは近い未来の予測は誤差が小さくて済むが，遠くの未来の予測は誤差が大きくなるという直感に当てはまる．しかし，表 8，9 の分類精度と，表 10，11 の平均二乗誤差を比較すると，必ずしも平均二乗誤差が大きいからといって分類精度が低いわけではないことがわかる．これは，遠くの未来を予測することで誤差は大きくなったが，同時に予測により得られた挙動の情報量も増えたからだと考えられる．実際に Mark Ⅱ ログ使用時の予測された特徴量の合計の 1 検体あたりの平均は，記録時間が 5 秒から 6 秒に予測されたものでは 98.1，5 秒から 15 秒に予測されたものでは 126.2 であり，増加していた．

次に，予測により精度が向上した結果において，どのくらい特徴量が予測できていたかを視覚的に確認する．提案手法を用いた Mark Ⅱ ログ使用時の結果である表 8 において，5 秒のログの特徴量から，8 秒のログの特徴量を予測した時，分類精度が最も向上した．まず，この時の予測による分類結果を確認するため，(a) 実際に取得しているログである 5 秒のログと (b)8 秒のログ，および (c)(a) から (b) を予測することにより得られた特徴量，のそれぞれを用いた時の識別の正誤ごとの検体数を表 12 に示す．本提案手法である予測が高精度にできた場合，理論上分類ができるようになる検体は，5 秒間のログでは誤って分類されていて，8 秒間のログでは正しく分類された表 12 における 14（5 + 9）検体である（表 12 にて太文字で表している検体）．言い換えれば，この 14 検体が正しく分類されていれば予測が高精度に行われたと考えられる．今回，この 14 検体のうち，予測された特徴量を用いて正しく分類できたのは 5 検体あり，誤って分類されたのは 9 検体あった．この 5 検体が今回の予測による精度向上に大きく貢献したと考えられる．この予測により正しく分類された 5 検体のラベルは，ダウンロード，マイニング，バックドアのみであり，この 5 検体からラベルごとに 1 検体ずつ特徴量のヒストグラムを作成した（図 6）．図 6 の横軸は単語，縦軸は単語の出現回数となっており，青色，赤色，黄色の棒はそれぞれ，先ほど述べた (a)，(b)．(c) の特徴量である．つまり，赤色と黄色の棒が類似していれば予測が高精度に行われたということである．実際に，図 6 における赤色と黄色の棒は類似しており，予測が高精度で行われたと考えられる．同様に，先ほど述べた 9 検体についてのヒストグラムも図 7 に示す．図 6 に比べて赤色と黄色の棒に差が生じている特徴量が目立っている．このことから，これらの検体に対する予測性能が十分でなかったと考えられる．以上のことから，予測性能をあげることができれば，分類の精度も向上する可能性があると考える．

表 12　予測された特徴量と予測対象の元のログのそれぞれを用いた分類結果の正誤ごとの検体数
Table 12　Number of samples per correct or incorrect classification results using each of the predicted features and the original logs of the prediction.

		5 秒間のログの分類結果			
		正		誤	
		8 秒間のログの分類結果			
		正	誤	正	誤
5 秒間のログの特徴量から 8 秒間のログの特徴量を予測して行った分類結果	正	69	2	**5**	3
	誤	4	1	**9**	21

7.　まとめと今後の課題

本研究では早期に機能推定をより高精度に行うため，記

表 10　Mark Ⅱ ログを用いて予測された特徴ベクトルと予測目的対象の元のログの特徴ベクトルの平均二乗誤差
Table 10　Mean squared error of the predicted feature vectors and the original log feature vectors of the target of prediction using Mark Ⅱ logs.

学習データの記録時間 (s)	6	7	8	9	10	11	12	13	14	15
平均二乗誤差	4.22	6.58	6.53	6.51	7.07	6.56	7.15	7.10	7.19	7.07

表 11　Cuckoo ログを用いて予測された特徴ベクトルと予測目的対象の元のログの特徴ベクトルの平均二乗誤差
Table 11　Mean squared error of the predicted feature vectors and the original log feature vectors of the target of prediction using Cuckoo logs.

学習データの記録時間 (s)	2	3	4	5	6	7	8	9	10
平均二乗誤差	144664	373789	902543	1736055	2833127	3800627	5738236	8279703	10045119

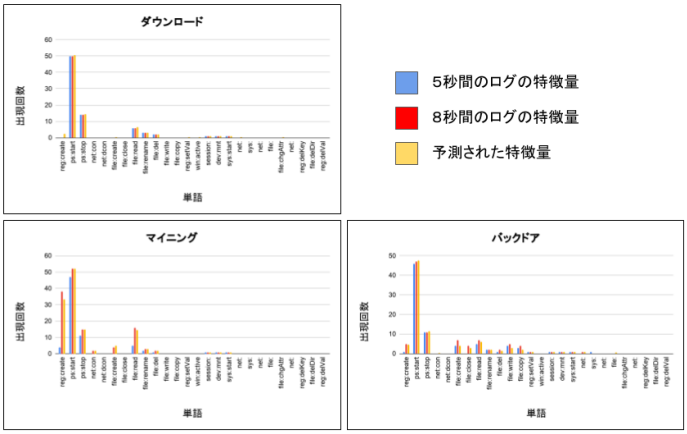

図 6　予測により正しく分類できた検体の特徴量のヒストグラム
Fig. 6　Histogram of samples' features correctly classified by prediction.

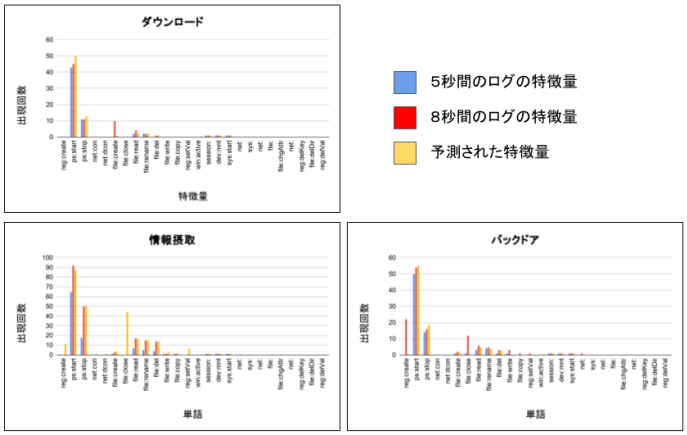

図 7　予測したが誤って分類された検体の特徴量のヒストグラム
Fig. 7　Histogram of samples' features incorrectly classified by prediction.

録時間の短いログから得られる特徴量から記録時間の長いログから得られる特徴量を予測し，分類モデルの学習データとテストデータに使用するログの挙動を合わせて機能推定を行った．その結果，Mark Ⅱ ログでは記録時間が 5 秒のログに対し 2.6%，Cuckoo ログでは記録時間が 1 秒のログに対し 0.8%精度が向上した．

今後の課題として，異なるデータでの評価や，さらなる予測性能の向上のための予測手法の検討が必要である．また，早期の機能推定後の被害を抑えるための対応も手動では間に合わないことも考えられるため，その対応の自動化の検討も必要である．

参考文献

[1] FFRI セキュリティ: 未知の脅威に対抗する「先読み対策」とは, https://www.ffri.jp/special/special_1.htm（2020/08/16 参照）.
[2] ITmedia: “侵入されて当たり前”の時代に効く、進化する次世代セキュリティ対策とは, https://www.itmedia.co.jp/enterprise/articles/1801/09/news001.html （2020/08/16 参照）.
[3] 佐々木良一, 他: ネットワークセキュリティ, オーム社 (2014).
[4] Naoto, K. et al.: Malware Function Estimation Using API in Initial Behavior. IEICE Transactions on Fundamentals, Vol. E100-A, No. 1, pp. 167—175 (2017).
[5] Rhode, M. et al.: Early-stage malware prediction using recurrent neural networks. computers & security 77, pp. 578–594.(2018)
[6] 森優輝, 他: マルウェアによる感染活動の目的推定に向けた動的解析ログに基づく分析, コンピュータセキュリティシンポジウム (2018).
[7] 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, Vol.2020-IFAT-139 No.8, (2020 年 7 月).
[8] InfoTrace Mark II for Cyber, https://www.soliton. co.jp/mark2/（2020/08/16 参照）.
[9] Cuckoo Sandbox, https://cuckoosandbox.org/（2020/08/16 参照）.
[10] VirusTotal, https://www.virustotal.com/gui/（2020/08/20 参照）.
[11] トレンドマイクロ：マルウェア, https://www.trendmicro.com/vinfo/jp/threat-encyclopedia/malware（2020/08/16 参照）.
[12] kaspersky: Threats, https://threats.kaspersky.com/en/threat/（2020/08/16 参照）.
[13] scikit-learn, https://scikit-learn.org/stable/（2020/08/16 参照）.