大規模時系列観測データによるマルウェア感染予測

吉村　尚人[1,a)]　池上　雅人[2]　住田　裕輔[2]　木谷　浩[2]　白石　善明[1,b)]　森井　昌克[1,c)]

概要：近年インターネットの普及に伴い，サイバー攻撃が増加している．その中でもマルウェアによる被害は多く，対策の必要性も増している．マルウェアへの対策としては，マルウェア自体の解析や，脅威情報をまとめたレポートの発行などが行われている．一方でマルウェアがどのように感染拡大するかは認知されておらず，マルウェアへの事前の対策を困難としている．本稿では，データ取得期間中に新たに定義されたマルウェアの感染拡大傾向を分析し，その分析結果をもとにした新種マルウェアの発生予測手法を提案する．本研究では国別のマルウェア発生を記録した大規模時系列観測データを用いる．分析においては，マルウェア観測国数の時間変化からマルウェア感染拡大形態の分類を行い，その特性を考察する．また，クラスタ分析を行うことで類似した感染傾向を持つマルウェアを集約し，所属クラスタの推定による感染拡大予測を行う．実験の結果，提案予測モデルは単純な予測と比較し高い精度での予測が可能であることを確認し，マルウェア感染拡大予測への可能性が示された．

キーワード：マルウェア，時系列データ，クラスタ分析，予測

Predicting Malware Outbreak Using Large-scale Time-series Data

Naoto Yoshimura[1,a)]　Masato Ikegami[2]　Yusuke Sumida[2]　Hiroshi Kitani[2]
Yoshiaki Shiraishi[1,b)]　Morii Masakatu[1,c)]

Abstract: The damage caused by malware is significant and the need for countermeasures is increasing. To handle the threat of malware, many measures are performed. On the other hand, it is not well known how malware spreads around the world. In this paper, we analyze the spread of malware and propose a method for predicting malware outbreaks. In this study, we use large-scale time-series data that records malware observed by country. In our analysis, we classify the pattern of malware spreading and consider its characteristics. In addition, we aggregate the malware with similar infection routes by cluster analysis, and predict the outbreak by estimating the cluster to which they belong. The experimental results show that proposed prediction model performed better than naive prediction, indicating its potential for predicting the outbreak of malware infection.

Keywords: Malware, Time-series data, Cluster Analysis, Prediction

1. はじめに

近年インターネットの普及に伴い，サイバー攻撃が増加している．[1] その中でもランサムウェア等を中心としたマルウェアによる被害は多く，対策の必要性も増している．[2] マルウェアに対する対策としては，マルウェア自体の解析や，脅威情報をまとめたレポートの発行などが行われている．一方でマルウェアがどのように感染拡大するかは認知されておらず，マルウェアに対する事前の対策を困難としている．

本研究では国別のマルウェア発生を記録した大規模時系列観測データを用いる．本稿では，データ取得期間中新たにマルウェア定義データベースに追加されたマルウェアの

[1] 神戸大学
Kobe University
[2] キヤノンマーケティングジャパン株式会社
Canon Marketing Japan Inc.
[a)] yoshimura@stu.kobe-u.ac.jp
[b)] zenmei@port.kobe-u.ac.jp
[c)] mmorii@kobe-u.ac.jp

感染拡大傾向を分析し，その分析結果をもとにした新種マルウェアの発生予測手法を提案する．分析においては，マルウェア観測国数の時間変化からマルウェア感染拡大形態の分類を行い，その特性を考察する．また，得られた分析結果をもとに，クラスタ分析を行うことで類似した感染傾向を持つマルウェアを集約し，所属クラスタの推定による感染拡大予測を行う．

2. ESET Live Grid

ESET Live Grid[3] とは，ESET 社により提供されるマルウェアの早期警告システムである．ESET Live Grid は ESET 製品を通して世界各国に展開されており，マルウェアの検出及び，統計情報・疑わしいファイルの収集を行う．収集されたマルウェアの情報はポータルサイト ESET Virus Rader[4] に公開されている．

本研究ではこの ESET Live Grid により検出されたマルウェアの時系列観測データを用いる．データの各レコードは時刻（年月日時），発生国，検体名，発生数の 4 項目で構成されており，2018 年 5 月 1 日 0 時〜2018 年 10 月 14 日 22 時の期間における観測データを記録している．

本研究では観測期間内に ESET Live Grid のマルウェア定義データベースに追加されたマルウェアを対象に分析を行い，これを新種マルウェアと呼称する．また，データには 240 の国及び地域が存在するが，台湾やパレスチナ等国連には加盟していないが多くのマルウェアが観測される地域も存在するため，本研究では便宜上 240 の国及び地域を国として扱う．

3. マルウェアの発生予測に向けた研究

3.1 マルウェア間における発生相関分析

予測への活用を目的として，異なるマルウェア間における発生の相関を明らかにする研究が行われてきた．柏井ら (2013) は Nicter Open Network Security Test-Out Platform(NONSTOP) から得られたデータを用い，マルウェア間における発生数の時間的関連性を調査した．[5] 柏井らはマルウェア間で発生のラグを考慮した相関分析を行うことで時間的関連性を発見する手法を提案した．柏井らによる分析の結果，異なるマルウェア間で発生数に相関のある組み合わせが発見された．

村井ら（2016）は柏井らの方法における相関分析の前処理に改良を加え，データセットに ESET Live Grid の日毎データを用いた．[6] さらに，村井ら（2017）は前処理として移動平均と標準化による発生過程の平滑化を施すことで改良を加えた相関分析も行った．[7]

川原ら（2019）は ESET Live Grid の日毎データを用いてマルウェア間における発生の相関分析を行った．[8] 川原らは柏井らの手法から，前処理において休日データの削除や標準化の変更を加えた．

3.2 マルウェア発生傾向の分析

マルウェアの予測に関しては，マルウェア間の関係性のみではなく，国間の関係性に着目した研究も行われている．村井ら（2017）はマルウェアの発生過程と地域性における関係の分析を行った．マルウェア発生のピークとなる週を入力とした国のクラスタ分析の結果，マルウェアの発生には経済水準や使用言語が影響を与えていることが明らかになった．

川原ら（2020）は村井らの行ったクラスタ分析におけるクラスタリングの入力を変更し分析を行った．[9] 川原らはクラスタリングの入力として発生数の変化量を用いた．分析の結果，使用言語や地理が発生傾向に影響を与えることを明らかにした．また，川原らは感染の経路にパターンがあると考え，感染活動の推移に関する分析も行った．川原らは新種のマルウェアが発生してから初期段階，流行段階，駆除段階の順で感染活動が推移すると仮定し，この段階に基づいた発生形態の分類を行った．川原らの行った発生形態への分類はそれぞれのマルウェアを三角形モデル，ひし形モデル，直線形モデルに分けるものであり，ひし形モデルに関してはクラスタリングにより類似した感染経路を持つマルウェアを集約し，経路のパターン化も行った．

本研究ではデータセットを村井らや川原らの用いた ESET Live Grid の日毎データから時間毎データに変更し，川原らの手法に基づいた発生傾向分析を行う．そして，分析の結果を利用したマルウェアの発生予測手法を提案する．

4. 新種マルウェアの発生傾向分析

本章では新種マルウェアについて発生傾向分析を行う．本研究で新種マルウェアとは，データセットにおける観測期間中に ESET Live Grid のマルウェア定義データベースに新たに追加されたマルウェアをいう．対象となる新種マルウェアは 31,112 種であった．新種マルウェアのみを対象とするのはマルウェアの発生時期が把握でき，感染拡大の傾向をつかむのに適していると考えたためだ．

4.1 発生形態による分類

本節では川原らの提案した感染活動の推移モデルに基づいた，新種マルウェアの発生形態による分類を行う．発生形態への分類を行うにあたり，川原らは新種のマルウェアが発生してから初期段階，流行段階，駆除段階の順で感染活動が推移すると仮定した．これはマルウェアの感染活動が始まり，感染が拡大していくが，同時に対策も進められ，感染地域拡大の勢いも次第に弱くなると考えられるためだ．各段階の特徴を示す．

- 初期段階
  マルウェアが観測されたばかりで感染拡大が穏やかな期間．
- 流行段階

表 1　発生形態への分類結果.

Table 1　The number of each spreading pattern.

発生形態	マルウェア数
三角形発生モデル	6,849 種
ひし形発生モデル	407 種
直線形発生モデル	4,392 種
対象外	19,464 種
合計	31,112 種

マルウェアの活動が活発化し急速に感染が拡大する期間.

- 駆除段階

マルウェアに対策が打たれ，感染拡大が収束する期間.

そして，これらの段階を用いてマルウェアを以下 4 つの発生形態を表したモデルに分類する.

- 三角形発生モデル
  発生とともに流行段階が始まり，その後駆除段階に移行するモデル.
- ひし形発生モデル
  明確に初期段階，流行段階，駆除段階が存在するモデル.
- 直線形発生モデル
  流行段階が存在しないモデル.
- 対象外
  観測する国数が少なくモデルの判断が困難なマルウェア.

上記の感染段階及び発生形態の明確な定義を示す．データについては各マルウェア初観測から 90 日つまり 2,160 時間分のデータを利用する．そして，流行段階は 90 日間で観測した国の過半数が初観測する 3 日間と定義する．過半数を超える 3 日間が複数存在した場合は観測国数が最大の値をとる 3 日間を流行段階とする．そして流行段階の前後の期間をそれぞれ初期段階，駆除段階とする．また，流行段階を用いて，最初の 3 日間が流行段階となるものを三角形発生モデル，最初の 3 日間以外に流行段階が存在するものをひし形発生モデル，流行段階が存在しないものを直線モデルとする．ただし，90 日間での観測国数が 10 に満たないものに関しては対象外とする．新種マルウェア 31,112 種を発生形態に分類した結果を表 1 に示す.

4.2　発生形態ごとの特徴

4.1 節で行った分類の結果をもとに，各発生形態の特徴を分析する．本節では，発生形態への分類に用いた 90 日間のデータを分析するとともに，予測への活用を目的として発生後 3 日のデータについても分析を行う．図 1 は各発生形態において，90 日間で観測国数がどのように推移するかを，発生形態ごとの平均値で表している．また，図 2 は同じく 72 時間での推移を発生形態ごとの平均値で表している．最後に図 3 は新種マルウェアの発生から 3 日間で各国が初観測する割合をヒートマップで表している．これらをもとに各発生形態の特徴を見ていく.

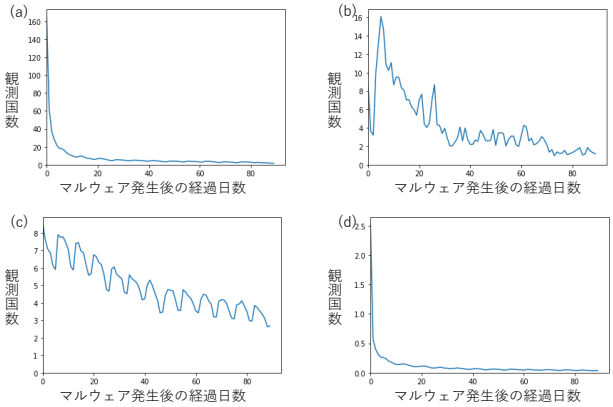

図 1　90 日間の観測国数の変遷.
(a) 三角形型発生モデル，(b) ひし形発生モデル，
(c) 直線形発生モデル (d) 対象外

Fig. 1　Changes in the number of observing countries within 90days. (a)triangle pattern, (b)diamond pattern, (c)straight-line pattern(d)out-of-scope

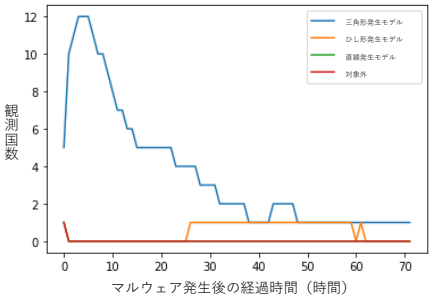

図 2　72 時間の観測国数の変遷.

Fig. 2　Changes in the number of observing countries within 72hours.

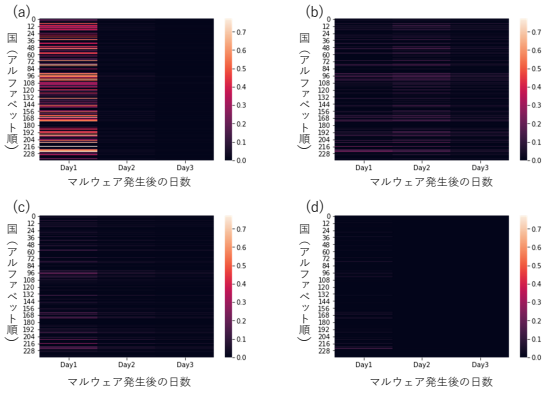

図 3　3 日間で初観測する割合のヒートマップ.
(a) 三角形型発生モデル，(b) ひし形発生モデル，
(c) 直線形発生モデル (d) 対象外

Fig. 3　Heat map of the ratio of first observation in 3 days. (a)triangle pattern, (b)diamond pattern, (c)straight-line pattern(d)out-of-scope

4.2.1 三角形モデル

三角形発生モデルは感染活動が流行段階から開始するモデルである．実際に図 1(a)，2 では，最初に観測国数が増加し，その後減少に転じている．また図 3(a) からも，多くの国が初日に観測することが分かる．このことから，三角形発生モデルでは，多くの国が初日に観測し，平均的に 3 日間以内に感染拡大が収束に転じることから，4 日目以降の感染拡大の恐れは小さいと考えられる．

4.2.2 ひし形モデル

ひし形発生モデルは初期段階，流行段階，駆除段階を経て感染活動が推移するモデルである．図 1(b) からは，いくつかのピークがみられるが，これはマルウェアにより流行段階の開始時期が異なるためだと考えられる．また，図 2 では発生より 25 時間から 60 時間あたりで，直線型発生モデル及び対象外と異なる傾向を示すが，これはひし形発生モデルの一部マルウェアにおいて流行段階が感染から 3 日以内に開始するためだと考えられる．図 3(b) では少し傾向が読み取れるが，大部分のマルウェアに共通する特徴はない．

4.2.3 直線形モデル

直線形発生モデルは顕著な流行期間を持たない発生モデルである．図 1(c) からは，平均的に観測国数は増減を繰り返しながら徐々に収束していくことが分かる．また，図 2，3(c) からも，発生から 3 日以内に多くのマルウェアに共通して感染が拡大する傾向は読み取れない．

4.2.4 対象外

90 日間での総観測国数が 10 未満のマルウェアを発生形態分類の対象外とした．定義からも推測できる通り，図 1(d) からはほとんど感染拡大が起こらないことを確認できる．また，発生後 3 日での特徴は図 2 ではおおよそ直線形モデルと一致し，図 3 ではひし形・直線型発生モデルがわずかに特徴を示すのに対し，対象外ではほとんど特徴を確認できない．

4.3 ひし形発生モデルのクラスタ分析

発生形態分類の中でもひし形モデルは 3 つの段階をもつため，比較的感染拡大の予測が容易であると考えた．しかし，図 3(b) からも，ひし形発生モデルに属する大部分のマルウェアに共通する特徴を挙げるのは困難である．そこで，本節では川原らの手法により，ひし形発生モデルのマルウェアにクラスタリングを行うことで類似した発生傾向を持つグループへの集約を行う．

4.3.1 ひし形発生モデルのクラスタリング方法

クラスタリングを行うにあたり特徴量の定義を行う．川原らはひし形発生モデルの初期段階，流行段階，駆除段階，それぞれにおける各国の観測を表すベクトル D0，D1，D2 を定義した．D0 は初期段階で観測した国に関する二値ベクトルとし，すべての国で観測されれば要素が全て 1 のベクトルとなる．要素数は 2 章で述べた通り便宜上 240 の国及び地域全てを国として扱うため 240 次元のベクトルとなる．同様に D1，D2 においても二値ベクトルを作成し，D0，D1，D2 を順に結合した 720 次元のベクトルを各検体の特徴量とする．

クラスタリングは階層型クラスタリングにより行う．階層型クラスタリングは距離の近いものから結合させていくことでクラスタリングを行う方法であり，本研究では川原らの手法に従い，検体間の距離にユークリッド距離，クラスタリング手法に Ward 法を用いる．ユークリッド距離はベクトル間の各要素で差をとり，その 2 乗和の平方根を計算することで求められる．また，Ward 法はクラスタ内の分散が最小になる方法で分類感度が良いとされている．今回は以上の方法を用いてひし形発生モデルのマルウェア 407 種を 9 つのクラスタに集約する．

4.3.2 ひし形発生モデルのクラスタリング結果

ひし形発生モデルのクラスタリング結果を表 2 に示す．表 2 では発生段階ごとに各クラスタで出現頻度の高かった上位 5 か国を示している．ただし，クラスタ番号横の（）内の値はそのクラスタに属するマルウェア数を表す．

クラスタリングの結果を見ていくと，各クラスタにおいて，出現頻度の高い国に特徴があることが分かる．また，クラスタに含まれるマルウェア数と実際に観測したマルウェア数間の関係，つまりクラスタの当てはまりの良さはクラスタや発生段階によって異なることが読み取れる．初期段階・流行段階では当てはまりの良いクラスタが多く，駆除段階では当てはまりの良いクラスタは見られなかった．

5. マルウェアの発生予測

本章では，4 章で行った新種マルウェアの発生傾向分析の結果をもとに 2 つのマルウェアの発生予測手法を提案する．1 つ目が新種マルウェア発生後 3 日間のデータをもとに三角形発生モデル及び対象外となるマルウェアを予測する手法である．これは，4 日目以降感染拡大の恐れが少なく，対策の必要性の小さいマルウェアを判別することを目的としている．2 つ目が同じく新種マルウェア発生後 3 日間のデータをもとに，ひし形発生モデルの感染拡大を予測するモデルである．この手法では，明確に初期段階，流行段階，駆除段階が存在するひし形発生モデルについて，ある時点での発生段階と感染拡大の恐れがある国の把握を目的としている．

5.1 三角形発生モデル及び対象外の予測

本研究で用いている観測データでは，およそ 5 か月半で ESET の検知データベースに新たに加えられた新種マルウェアの内，31,112 種が実際に観測されていた．しかし，このすべてに対して対策を行うことは現実的ではない．そこでここでは対策が必要となる恐れの少ないマルウェアを

表 2　ひし形発生モデルのクラスタリング結果.
Table 2　Clustering result of diamond pattern.

cluster1 (202)						cluster2 (10)						cluster3 (30)					
D0		D1		D2		D0		D1		D2		D0		D1		D2	
United States	43	Turkey	92	Mexico	15	Turkey	10	Austria	10	Paraguay	4	Ukraine	11	Bangladesh	30	Ghana	10
Russia	34	Thailand	84	Italy	13	Thailand	9	Palestinian	10	Cyprus	3	Turkey	11	Israel	30	Yemen	10
Turkey	30	France	68	France	13	Indonesia	8	China	10	Moldova	3	United States	9	Serbia	30	Namibia	9
Iran	30	Poland	65	India	12	Mexico	7	United Kingdom	10	Madagascar	3	Russia	6	Sri Lanka	30	Panama	8
Ukraine	27	Germany	64	Iran	11	Romania	7	Albania	9	Trinidad Tobago	3	Arab	6	India	29	Botswana	8
cluster4 (20)						cluster5 (11)						cluster6 (14)					
D0		D1		D2		D0		D1		D2		D0		D1		D2	
Turkey	18	Singapore	18	Guatemala	6	Germany	10	Argentina	10	El Salvador	3	Egypt	13	Argentina	12	Malaysia	3
United Kingdom	17	Belgium	16	Egypt	3	Hong Kong	10	Chile	10	Norway	3	India	13	Morocco	11	Bulgaria	2
Hong Kong	16	Canada	16	Honduras	3	Iran	10	Belgium	9	Belarus	2	Russia	13	Algeria	10	Georgia	2
Iran	15	Russia	16	Costa Rica	3	Poland	10	Ireland	9	Bolivia	2	Thailand	11	Brazil	9	Honduras	2
Greece	14	Hungary	15	Macedonia	3	Turkey	10	Luxembourg	9	Mauritius	2	Ukraine	11	Sri Lanka	9	Sweden	2
cluster7 (55)						cluster8 (28)						cluster9 (37)					
D0		D1		D2		D0		D1		D2		D0		D1		D2	
Egypt	16	Italy	37	Belgium	9	Russia	17	Philippines	26	Cuba	7	Turkey	16	Croatia	34	Colombia	6
Turkey	15	Greece	36	Israel	9	Ukraine	14	Hungary	25	Belgium	6	Taiwan	14	Spain	33	Saudi Arabia	6
United States	14	Spain	34	Egypt	8	Spain	7	Italy	25	Lao	6	Hong Kong	12	Greece	32	Sri Lanka	6
Thailand	13	Turkey	34	Slovakia	8	Belarus	6	Thailand	25	Tunisia	6	France	11	Israel	31	Cyprus	6
Iran	13	India	33	Netherlands	8	Kazakhstan	6	France	24	Honduras	5	Arab	11	Netherlands	31	Tunisia	5

マルウェア発生から 3 日間のデータを用いて予測する手法を提案する．前提として対策が必要となる恐れの少ないマルウェアに，三角形発生モデル及び対象外のマルウェアを設定した．これらのマルウェアを設定したのは 4 章での分析結果から，平均的に発生から 3 日間で観測国数が減少に向かい，4 日目以降の感染拡大の恐れが少ないと判断したためだ．三角形発生モデル及び対象外のマルウェアは合計で 26,313 種となり，データセットに含まれる新種マルウェア全体の約 85 %を占める．これらを対策の必要がないマルウェアとして判断できれば対策への負担を大幅に軽減できると考える．

これより，予測の方法を述べる．提案手法では 2 つの特徴量と分類器を用いて予測を行う．1 つ目の分類器では三角形発生モデルとその他の発生形態の判別を行い，三角形発生モデルと判断されなかったマルウェアに関しては 2 つ目の分類器により対象外であるかを判断する．特徴量に関しては，1 つ目の分類器では，発生から 72 時間の観測国数を表す 72 次元のベクトルを特徴量として用いる．これは，4 章での分析の結果，発生後 72 時間での観測国数の推移が三角形発生モデルとその他の発生形態との間で大きく異なったためだ．2 つ目に関しては対象外とひし形・直線形発生モデルの間で比較的違いが大きかったため，マルウェアが発生してから各国観測までのラグを表した 240 次元のベクトルを特徴量に用いる．この特徴量では 3 日間のうちに観測しなかった国については-1 を要素とする．分類器にはともに，ランダムフォレストを用いる．ランダムフォレストを用いる理由としては 2 つ目の特徴量として観測しなかった国を−1 であらわしており順序を反映した特徴量となっていないため，カテゴリ変数を扱えるランダムフォレストを採用した．

予測手法の精度を確かめるために実験を行った．新種マルウェアを訓練データとテストデータに分割し，テストデータで評価を行った．評価には Accuracy，Recall，Precision を用いた．これらの評価値は表 3 に示す行列の各要素を用いて以下のようにあらわされる．

表 3　混同行列.
Table 3　Confusion Matrix.

		予測値	
		Positive	Negative
真値	Positive	TP	FP
	Negative	FP	TN

表 4　三角形発生モデル及び対象外の予測結果.
Table 4　Prediction Result of triangle pattern and out-of-scope.

発生形態	Accuracy(%)	Recall(%)	Precision(%)
三角形発生モデル	95.2	81.3	91.2
対象外	86.5	85.1	92.7
ひし形・線形発生モデル	82.8	74.4	54.7

$$\begin{aligned} \text{Accuracy} &= \frac{TP+TN}{TP+FN+FP+TN} \\ \text{Recall} &= \frac{TP}{TP+FN} \\ \text{Precision} &= \frac{TP}{TP+FP} \end{aligned} \tag{1}$$

実験の結果，表 4 に示す結果が得られた．結果から，高い精度で予測が行えたことが読み取れ，感染拡大の恐れの少ないマルウェアを判断することで対策への負担軽減への可能性が示された．

5.2　ひし形発生モデルの感染拡大予測

本節ではひし形発生モデルの感染拡大予測手法を提案する．4.3 節で行ったクラスタ分析の結果から，クラスタごとに主に初期段階と流行段階において初観測する国に特徴があることが明らかになった．そこで，新たにマルウェアが発生した際に発生後 3 日間のデータを用いてそのマルウェアと類似するクラスタを推測することで，その後の感染拡大が可能であると考えた．提案手法ではクラスタの予測の当てはまりの良さや，クラスタが示す特徴の度合いから，ある国がある段階でマルウェアを観測する確信度を出力する予測モデルを考案した．提案した予測モデルの学習フェーズと予測フェーズについてそれぞれ述べる．

5.2.1 予測モデルの学習

予測モデルの学習フェーズでは，クラスタリング及び 4 つの関数 TF，IDF，CP，FCC と 2 つのパラメータ α，β の学習を行う．学習では発生後 3 日間に観測した国のリストと各発生段階での国の観測を 2 値で表したベクトルを用いる．2 値ベクトルは 4.3 節でクラスタリングに用いたベクトルと同様のものを用いる．

(1) クラスタリング

学習の初めの段階としてクラスタリングを行う．クラスタリングには各発生段階における国の観測を 2 値で表したベクトルを用い，4.3 で述べた方法で行う．

(2) TF-IDF

文書の中の重要単語を表す手法に Term Frequency-Inverse Document Frequency（TF-IDF）がある．TF - IDF では文書内の単語の出現頻度を表す TF と単語の出現する特異性を表す IDF の積で文章の重要性を数値として表す．これにより，特定の文書のみに多数出現する単語の重要度が大きくなる．提案手法ではクラスタを文書，国を単語と見立てることで，TF-IDF を用いて，マルウェア発生後から 3 日間のうちに観測した国からクラスタの推定を行う．そのために，国 c_i とクラスタ clust_j に対する関数 $\mathrm{TF}(c_i, \mathrm{clust}_j)$ と国 c_i に対する関数 $\mathrm{IDF}(c_i)$ の学習を行う．$n_{c_i,\mathrm{clust}j}$ をクラスタ clust_j における国 c_i の出現回数，$|C|$ を総クラスタ数とすると，TF,IDF は以下のようにあらわされる．

$$\mathrm{TF}(c_i, \mathrm{clust}_j) = \frac{n_{c_i,\mathrm{clust}j}}{\sum_{k=1}^{240} n_{c_k,\mathrm{clust}j}} \tag{2}$$

$$\mathrm{IDF}(c_i) = \log\left(\frac{|C|+1}{|\{\mathrm{clust} : \mathrm{clust} \ni c_i\}|+1}\right)+1 \tag{3}$$

そして予測されるクラスタは TF と IDF を用いて，以下のように予測する．

$$obs(c_i) = \begin{cases} 1\ (if\ c_i\ observed\ within\ 3days) \\ 0\ (else) \end{cases}$$

$$\mathrm{predClust} = \arg\max_{\mathrm{clust}_j} \sum_{k=1}^{240} (\mathrm{TF}(c_k, \mathrm{clust}_j) \times \mathrm{IDF}(c_k) \times \mathrm{obs}(\mathrm{c_k})) \tag{4}$$

(3) CP(Cluster Precision)

CP(Cluster Precision) は TF-IDF による予測がどの程度正しいかをクラスタごとに求めるために定義した関数であり，訓練データで予測を行った際の precision を用いる．clust_j と判定されたマルウェアを m_j, その総数を $|M_j|$ とすると $\mathrm{CP}(\mathrm{clust}_j)$ は以下であらわされる．

$$\mathrm{CP}(\mathrm{clust}_j) = \frac{|\{m_j : m_j \in \mathrm{clust}_j\}|}{|M_j|} \tag{5}$$

(4) FCC(Frequency of Country in Cluster)

FCC(Frequency of Country in Cluster) は実際に国がどの程度観測するかを求めるために定義した関数であり，クラスタリングの結果を用いて求める．クラスタリングの結果 clust_j となったマルウェアを mc_j，その総数を $|\mathrm{MC}_j|$，発生段階を phase_D とすると，$\mathrm{FCC}(c_i, \mathrm{clust}_j, \mathrm{phase}_D)$ は以下のように表される．

$$\mathrm{FCC}(c_i, \mathrm{clust}_j, \mathrm{phase}_D) = \frac{|\{\mathrm{mc}_j : \mathrm{mc}_j\ is\ observed\ by\ c_i\ in\ \mathrm{phase}_D\}|}{|MC_j|} \tag{6}$$

そして，国が現れると予測される段階は，予測されたクラスタ内で FCC が最大となる段階とする．

$$\mathrm{predPhase}(c_i, \mathrm{clust}_j) = \arg\max_{\mathrm{phase}_D} \mathrm{FCC}(c_i, \mathrm{clust}_j, \mathrm{phase}_D) \tag{7}$$

(5) 線形回帰

ここまで定義した関数を用いて国が予測した段階に現れる確信度を以下のように表す．

$$\text{confidence seed}(c_i, \mathrm{predClust}, \mathrm{predPhase}) = \mathrm{CP}(\mathrm{predClust}) \times \mathrm{FCC}(c_i, \mathrm{predClust}, \mathrm{predPhase}) \tag{8}$$

最後に confidence seed と実際に予測した段階で国が観測する割合を対応付けるために線形回帰を行う．つまり，実際に予測した段階で国が観測する割合を confidence seed の 1 次式で表す．こうして得られた傾きを α，切片を β とし，最終的な観測の確信度を以下であらわす．

$$\mathrm{confidence}(c_i, \mathrm{predClust}, \mathrm{predPhase}) = \alpha \times \text{confidence seed}(c_i, \mathrm{predClust}, \mathrm{predPhase}) + \beta \tag{9}$$

ただし，マルウェア発生から 3 日以内に観測した国に関しては，confidence 1 で初期段階の予測結果に加える．

5.2.2 モデルを用いた予測

5.2.1 節で述べた方法で学習を行った予測モデルを用いて，実際に予測を行う方法を示す．予測では入力としてマルウェア発生後 3 日間で観測した国のリストを与える．これに対して，出力としては各国の観測予測段階とその確信度が得られる．本研究では得られた出力から確信度に対して閾値を設けることで，閾値以上の国が観測すると予測するとした場合の精度を評価した．

5.2.3 ひし形発生モデルの予測実験

ひし形発生モデルによる予測の精度を評価するために実験を行った．実験ではひし形発生モデルのマルウェア 407

種を訓練データ：テストデータ＝ 6 : 4 で分割し，テストデータで閾値ごとの精度を Recall と Precision により評価した．また，比較対象として単純な予測を行った．単純な予測では提案予測モデルで観測されると予測された総国数と同数を，マルウェア総観測数上位の国から順に取り出し，初期段階：流行段階：駆除段階＝ 15 : 70 : 15 で当てはめた．単純な予測にこのような方法を採用したのは，一般に観測数の多い国はその出現頻度も高く，長期間観測すると考えられ，ひし形発生モデルマルウェア全体の段階ごとの発生比がおよそ 15 : 70 : 15 であったためだ．

実験の結果を図 4 に示す．グラフでは実線で提案手法，点線で単純な予測の精度が示されており，青が Recall，赤が Precision を表す．得られた結果から，初期段階では閾値の大きい領域において，Recall，Precision ともに高い値を示しているが，単純な予測との差は小さく，単純な予測の精度が上回る領域も見られる．このことから，初期段階では最初の 3 日間で観測する国が支配的であり，その他に観測される国ではマルウェアの総観測国数が多い国が観測する傾向にあることが分かる．次に，流行段階では，閾値の大きい領域では Precision が高く，閾値の小さい領域では Recall が高いことが読み取れる．また，総じて提案予測モデルの精度は単純な予測より高い結果となった．このことから，提案予測モデルは単純な予測と比較しより適切な予測を行っており，単純な予測では予測できなかったマルウェア観測数の少ない国でも特徴から予測が出来ていると考察できる．最後に駆除段階ではクラスタ分析で当てはまりが良くなかったことからも予想できた通り，精度は高くなかった．しかし，駆除段階はマルウェアの感染拡大が収束に向かう段階であるため，予測への影響は小さいと考えられる．

以上の結果から，提案予測モデルを用いた予測の活用シナリオを提案する．まず，初期段階においては精度が高いため，初期段階での観測が予測された国でマルウェアの初観測が起こっている間は初期段階だと推測できる．流行段階に移行したことを推測するためには高い閾値を用いる．流行段階での予測において，高い閾値での Precision の値は高かったため，高い閾値で観測が予測される国が観測し始めると流行段階に移行した可能性があると予測する．一方で高い閾値では Recall が低い問題があるため，低い閾値を用いて Recall の低さを補い，感染拡大の恐れのある国を列挙する．これにより，提案予測モデルでは感染拡大が一気に広がる流行段階の開始を捕捉し，感染の恐れのある国を予測できると考える．

図 5 に提案予測モデルを用いて予測を行った例を示す．図の上段は実際に各段階で観測した国を示しており，国は観測した順番に並べてある．一方下段は実際に観測した国に予測されたが観測されなかった国を加えたものである．下段の色分けについて説明する．緑はマルウェア発生後最

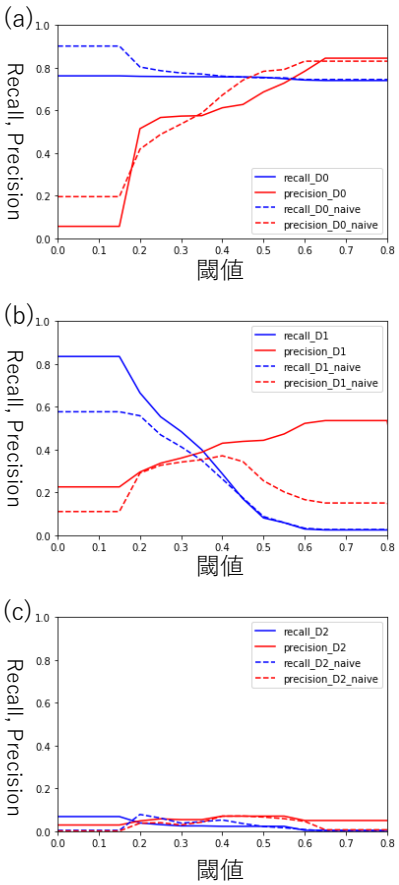

図 4　提案予測モデルの精度．
(a) 初期段階，(b) 流行段階，
(c) 駆除段階

Fig. 4　Recall and Precision of proposed prediction model.
(a)initial stage, (b)epidemic stage,
(c)removal stage

初の 3 日間で観測した国を表す．初期段階の予測には閾値 0.5 を用いた．この例では最初 3 日間で観測した国と初期段階で観測した国が一致している．次に赤は閾値 0.5 で観測すると予測された国である．赤で示す国は実際に現れる可能性が高いと考えられ，流行段階への移行の判断において重要である．次に橙は閾値 0.25 で観測すると予測された国であり，観測の恐れのある国をもれなく予測することが期待される．青は，観測すると予測されたが，実際には観測されなかった偽陽性の国を表す．紫では予測された段階とは異なる段階で観測された国を表す．最後に灰色は観測が予測されなかったが，実際には観測された国を表す．これより，示した例では，観測が予測された国は実際に観測した国の大部分をカバーしており，偽陽性が少なく，流行段階への移行の判断も可能だと考えられる．このことから提案予測モデルはマルウェアの感染拡大予測への可能性を示した．

6. まとめ

本稿ではマルウェアへの対策を目的とし，大規模時系列

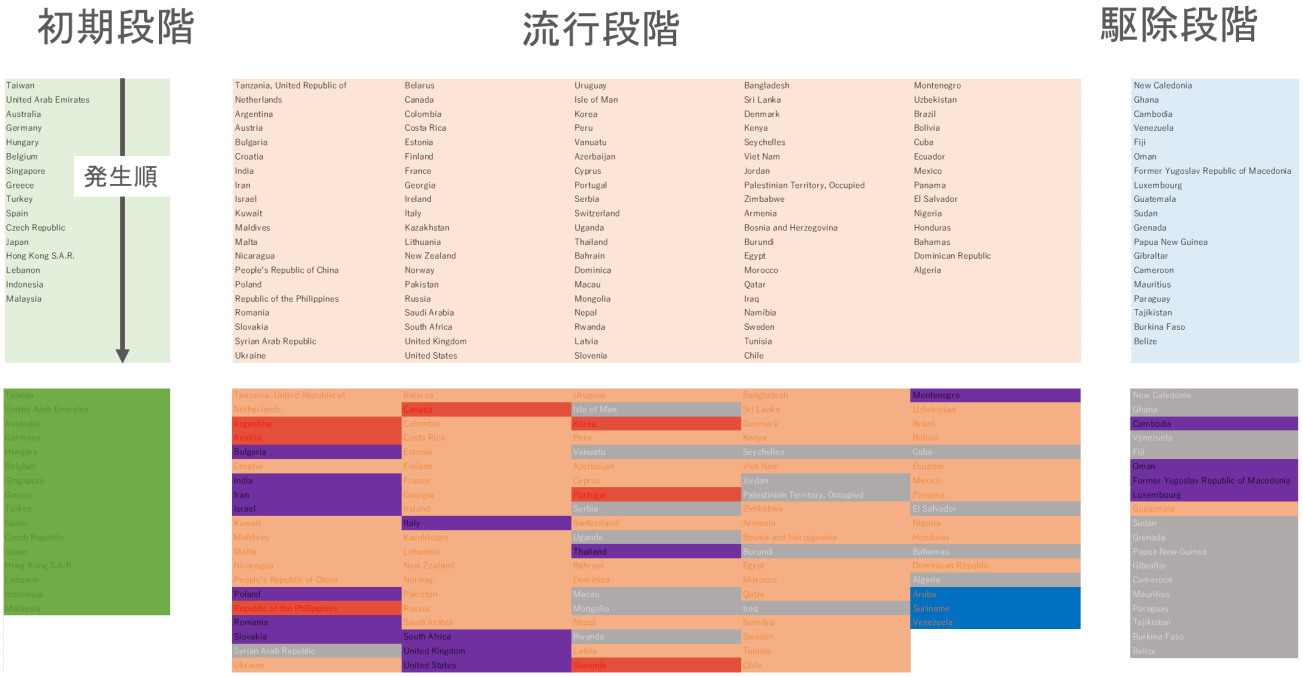

図 5 提案モデルを用いた予測例.

Fig. 5 Example of prediction.

観測データを用いたマルウェアの発生傾向分析を行い，その結果をもとにマルウェアの発生予測手法を提案した．

1つ目の提案手法では，日々発生する膨大なマルウェアから，感染拡大の恐れが少ないマルウェアを予測する手法を提案した．発生初期のデータを用いた予測の結果，感染拡大の恐れが少ないマルウェアを高精度で分類した．これにより，マルウェアへの対応に要する負担が軽減できると考える．2つ目の提案手法ではひし形発生モデルのマルウェアを対象に感染拡大予測を行う予測モデルを提案した．実験の結果は，感染拡大の可能性を示し，対策に有用であると考えられる．本研究では，提案した感染予測手法をさらに広く適用できるモデルへと修正し，一般的な感染予測手法に拡張する予定である．

参考文献

[1] 独立行政法人情報処理推進機構 (IPA) : 情報セキュリティ白書 2019, 独立行政法人情報処理推進機構 (IPA)(2019).

[2] 独立行政法人情報処理推進機構セキュリティセンター：情報セキュリティ 10 大脅威 2020 ～セキュリティ対策は一丸となって、Let's Try!!～, 入手先 ⟨https://www.ipa.go.jp/files/000080871.pdf⟩(2020.08.04).

[3] ESET : Malware Protection & Internet Security, 入手先 ⟨https://www.eset.com⟩(2020.08.04).

[4] ESET:Virus Rader, 入手先 ⟨https://www.virusradar.com/⟩(2020.08.04).

[5] 柏井祐樹，森井昌克，井上大介，中尾康二：*NONSTOP* データを用いたマルウェアの時系列分析, コンピュータセキュリティシンポジウム 2013（CSS2013） (2013).

[6] 村井健祥，森井昌克，池上雅人，長谷川智久，石川堤一：大規模ライブネット観測データを用いたマルウェアの時系列解析に関して, コンピュータセキュリティシンポジウム 2016（CSS2016） (2016).

[7] 村井健祥，森井昌克，池上雅人，長谷川智久，石川堤一：大規模マルウェア時系列観測データの解析結果について, 暗号と情報セキュリティシンポジウム 2017（SCIS2017） (2017).

[8] 川原大弥，池上雅人，木谷浩，長谷川智久，森井昌克：大規模マルウェア観測データに基づく時系列解析, 暗号と情報セキュリティシンポジウム 2019（SCIS2019） (2019).

[9] 川原大弥，近藤暖，池上雅人，長谷川智久，原田隆史，木谷浩，森井昌克：時系列観測データによるマルウェア発生分布の解析, 暗号と情報セキュリティシンポジウム 2020（SCIS2020） (2020).