ポート番号埋め込みベクトルを用いたダークネットスキャンパケット解析

石川 真太郎[1,a)]　小澤 誠一[1,2,b)]　班 涛[3,c)]

概要：近年, IoT デバイスの脆弱性を利用したサイバー攻撃による被害が深刻になっており, 対策が求められている. 本研究では, ダークネットで観測された TCP/SYN パケットに対し，宛先ポート番号の部分パターンの共起性に着目した埋め組みベクトル空間を FastText で学習し，得られたポート番号埋込ベクトルを用いて，スキャン活動の類似性に基づいたマルウェア亜種の分類を行う方法を提案する．具体的には，まず FastText でポート番号埋込ベクトルを学習し，スキャンパケットの宛先ポート番号から, 脆弱性探索のターゲットとなっているネットワークサービスの関係性を捉える. 次に, 非線形次元圧縮手法の UMAP を用いて, ホストを 2 次元空間に写像して可視化する. 最後に, DBSCAN によるクラスタリングを行い, 同様の攻撃パターンをもつマルウェア感染ホストのグループを識別する. 実験では, /16 のダークネットセンサーから収集された 1 か月間のダークネットトラフィックデータを使用して解析を行った. その結果, 提案手法によって, 同じボットネットに感染している可能性の高い Mirai 亜種のグループを検出できることを示した. また, 観測期間中に 9530/TCP の脆弱性を狙った, 新たな Mirai 亜種の出現を検知した.

キーワード：サイバーセキュリティ, 機械学習, ダークネット解析, 表現学習, IoT マルウェア

Darknet Scan Packet Analysis Using Port Embedding Vector

Shintaro Ishikawa[1,a)]　Seiichi Ozawa[1,2,b)]　Tao Ban[3,c)]

Abstract: In this work, we propose a method to track the attacks by malware-infected devices observed in the darknet and their changes using machine learning. First, feature extraction using FastText is executed, and the correlation among targeted network services is captured from the destination port numbers of scan packets. Then, we employ a nonlinear dimension reduction technique, UMAP, to project hosts into a 2-D embedding space for visualization purposes. Finally, DBSCAN is performed to automatically identify groups of malware infected hosts with the same attack patterns. In the experiments, we use a one-month darknet traffic trace collected from a /16 darknet sensor. As a result, we showed that a group of Mirai variants that could be controlled by the same botnet can be properly detected by using the proposed method. We also succeeded in detecting a new Mirai variant that targeted the vulnerability of 9530/TCP during the observation period.

Keywords: Cyber security, Machine learning, Darknet analysis, Representation learning, IoT malware

1. はじめに

近年, IoT デバイスの脆弱性を悪用するサイバー攻撃が増加している．マルウェアに侵入された IoT デバイスはボットネットを形成し, 重要なインフラストラクチャにサ

[1] 神戸大学大学院 工学研究科
Graduate School of Engineering, Kobe University
[2] 神戸大学 数理・データサイエンスセンター
Center for Mathematical and Data Sciences, Kobe University
[3] 情報通信研究機構
National Institute of Information and Communications Technology
[a)] 199t210t@stu.kobe-u.ac.jp
[b)] ozawasei@kobe-u.ac.jp
[c)] bantao@nict.go.jp

表 1　NICT ダークネットセンサーで観測されたパケットの年間統計.

年	2015	2016	2017	2018	2019
#パケット ($\times 10^8$)	545	1,281	1,504	2,121	3,279
#IP アドレス ($\times 10^3$)	280	300	300	300	300

イバー攻撃を実行することが問題となっており対策が急務となっている. 有名な例として, IP カメラやルーターなどのデバイスを標的とする IoT マルウェアである Mirai [1] が 2016 年に発見された. Mirai は, TCP/SYN パケットをランダムな IP アドレスに送信して, 実行中のサービスを探索するネットワークスキャンを実行し, デバイスに存在する脆弱性を悪用して侵入し大規模なボットネットを形成する. ボットネット形成後, C&C サーバーの指示に従って, 大量のパケットをターゲットサーバーに送信することにより, DDoS 攻撃が実行されるというマルウェアである. 2016 年 9 月に GitHub で Mirai のソースコードが公開されたことで, ボットネットを悪用する攻撃が急増し, 他の脆弱性を狙うように変更が加えられた多数の亜種の出現を招いている.

Mirai の亜種やその他の新しいマルウェアによる被害を軽減するために, 攻撃対象となっている脆弱性などの特徴的な情報を早期に得ることは有効である. マルウェアに感染したデバイスのネットワークトラフィックを分析してサイバー攻撃を特定することは動的分析に属しており, マルウェアのプログラムが判明していない場合にマルウェアの動作機能を取得するのに効果的である.

動的分析を行うために必要となるネットワークトラフィックデータを収集する方法の一つとして, 未使用の IP アドレス空間であるダークネットに届くパケットの収集が挙げられる. ダークネットでは, マルウェアに感染したデバイスと密接に関連しているパケットを収集できるため, 通常の通信により発生するノイズの影響をほとんど受けずにマルウェアのスキャン活動の解析を行える. また, ダークネットはインターネット上に存在するあらゆるホストから届くパケットを観測できるため, 世界中のサイバー攻撃の傾向を反映できる.

表 1 に示すように, NICT [2] が運用しているダークネットセンサーでは, 大量のパケット (2019 年では 3279 億パケット) が観測されており, この膨大なネットワークトラフィックから, 絶えず変化する攻撃の傾向を特定し新たなマルウェアを検出するには, 機械学習が重要な役割を果たす. これまでに, ダークネットトラフィック分析を行うための, さまざまな機械学習ベースのアプローチが考案されてきた [3]. 小澤ら [4] は, 相関ルール解析を使用してマルウェアの動作をルールとして抽出することを提案した. Ring ら [5] は, テキストマイニング手法である Word2Vec [6] を使用して, パケットデータ内の IP アドレス, 宛先ポート番号, およびプロトコルの共起性から IP アドレス間の類似性をモデル化する IP2Vec を提案した.

本研究では, マルウェアに感染したホストからのポートスキャンの傾向を把握するために, ダークネットで観測された TCP/SYN パケットを収集して分析を行う. 我々の分析は, スキャンの標的となっているネットワークサービスの識別子である宛先ポート番号の相関関係に基づいている. IoT 関連のポートスキャンにおいて, (23/TCP, 2323/TCP) や (80/TCP, 8080/TCP) のような宛先ポート番号のペアが, 関連するマルウェアの亜種の間で共通して見られる. これは, 既存のポート番号を置き換えるために新しく割り当てられたポート番号で語彙の類似性を維持する傾向があるという人間の認知的習慣に起因する. この特徴を利用することで, マルウェアの亜種間の類似性を宛先ポート番号間で共通するサブワードから推測でき, マルウェアの亜種間の関係をモデル化して評価することが可能となる. サブワードを考慮するために, 宛先ポート番号に FastText [7] を用いて, スキャンの対象となりやすい類似ポート番号のベクトル表現を作成し, この表現に基づいて 2 つの宛先ポートセット間の類似性を測定する. また, 可視化とクラスタリングによる分析を行うことで, 新しいマルウェアの亜種の出現を検出する.

2. 提案手法

本節では, マルウェアのスキャン活動を迅速かつ正確に把握する方法を提案する. マルウェアが更新される際, 新しく割り当てられたポート番号は人間の認知的習慣により, 語彙の類似性を維持する傾向があり, この類似性を維持する宛先ポート番号のペアは, 関連するマルウェアの亜種に共通して見られる. この性質を利用するために, ポート番号を分解して得られるサブワードの関係と, ポート番号間の共起性から得られるポート番号埋め込みベクトルの学習を提案する. このポート番号埋め込みベクトルにより, マルウェア亜種間の類似性に基づいたクラスタリングを試みる.

2.1 ポートセットの作成

マルウェアに感染したホストを類似したスキャン活動でグループ化するために, ダークネットで収集された TCP/SYN パケットから送信元 IP アドレスごとに宛先ポート番号を抽出し, これをポートセットと呼ぶことにする. それぞれのポート番号にネットワークサービスが割り当てられているため, ダークネットで観察されるパケットの宛先ポート番号は, マルウェアの標的となるネットワークサービスを表す. 従って, ポートセットはマルウェアに感染したホストによるスキャン活動を表現するデータとみなすことができる. 作成したポートセットの類似性に基づいて, マルウェアに感染したホストのクラスタリングを試みる.

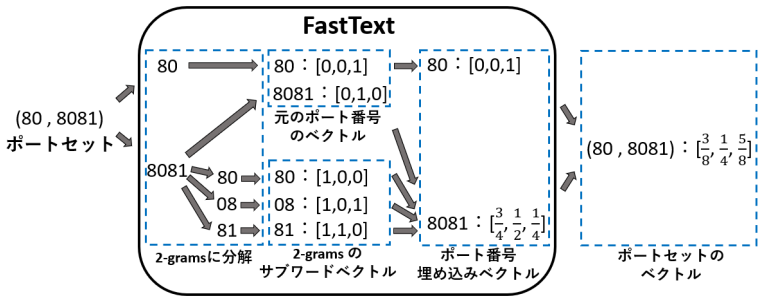


図 1 2-gram の FastText を用いてポートセットからポート番号埋め込みベクトルを取得する流れ.

2.2 ポート番号埋め込みベクトル

(23/TCP, 2323/TCP) と (80/TCP, 8080/TCP) は, 関連するマルウェアの亜種間で共通して見られるポート番号である. これは, 同じ目的で攻撃するネットワークサービスをポート番号から容易に識別できるように, ある程度のポート番号の類似性を維持しようとする人間の慣習に起因している. この性質を利用することで, 類似したマルウェアの亜種を同じグループに分類することが可能となる.

ポートセットにテキストマイニング手法の一つである FastText を適用することで, 宛先ポート番号のサブワードを考慮した埋め込みベクトル取得できる. 図 1 に, 提案手法であるポート番号埋め込みベクトルを取得する流れを示す. 本手法では, ポートセットはドキュメント分析における文と見なされ, このポートセットに FastText を適用することで, 各ポート番号が n -gram のサブワードに分割され, 次に, n -gram サブワードと元の宛先ポート番号のベクトルが取得される. 各宛先ポート番号とそれを構成する全てのサブワードの平均をとることで, ポート番号埋め込みベクトルを取得できる. 最終的なポートセットのベクトルは, ポートセット内に出現する宛先ポート番号のポート番号埋め込みベクトルを平均化することで得られる. これにより得られるベクトル表現をポートセットベクトルと呼ぶことにする.

2.3 スキャン活動の可視化

スキャン活動のクラスタリングと新たなスキャン活動を検出をするために, ポートセットベクトルを 2 次元の埋め込み空間で可視化する. ポートセットベクトルの次元を削減する方法として, 非線形次元削減法である UMAP [8] を使用する. UMAP では, 最初に高次元空間のデータをグラフとして表し, 次に, 対応する低次元空間のグラフ構造が, 高次元空間のグラフ構造にできるだけ類似するように最適化される. これにより, UMAP は高速で高性能の次元削減を実現している.

本手法では, マルウェアに感染したホストの分布を観測するために UMAP を使用し, 提案手法であるポート番号埋め込みベクトルがマルウェアによるスキャン活動を適切に反映できているかどうかを確認する.

2.4 スキャン活動のクラスタリング

ポートセットベクトルをクラスタリングすることにより, 同じスキャンパターンを持つマルウェアに感染したホストのグループを自動的に識別する. クラスタリング手法は, 密度ベースでクラスタリングを行う DBSCAN [9] を用いる. DBSCAN は 2 つのステップで構成される. まず, 全てのデータ点を, 指定された半径内に指定された数を超える隣接点を持つコア点, コア点から指定された半径内に存在するコア点同士を全て結んだネットワークに属する到達可能点, どちらにも当てはまらない外れ値に分ける. 次に, コア点と到達可能点の集合をクラスタとすることでクラスタリングを行う.

DBSCAN は事前にクラスタ数を決定する必要がないため, 日々特徴が変化しクラスタ数を定義することが困難であるダークネット分析に有効なクラスタリング手法である. ポートセットベクトルを DBSCAN に入力することで, 類似したスキャンパターンを持つマルウェア感染ホストを自動的にクラスタリングすることができる.

3. 評価実験

本節では, 提案手法であるポート番号埋め込みベクトルの有効性を評価するために, 有名な IoT マルウェアである Mirai のスキャン活動に注目した実験を行う. 特定のホストから送信されたパケットの 90 %以上が, Mirai のシグネチャである, 「宛先 IP アドレス=シーケンス番号」, 「宛先ポート番号=23」, 「送信元ポート番号> 1024」を満たすとき, そのホストから抽出されるポートセットを Mirai 感染ポートセットとする.

以下の実験では, 情報通信研究機構 (NICT) が運用する/16 ダークネットセンサーで観測された 2020 年 2 月 1 日から 2020 年 2 月 29 日までの 9,875,671,868 パケットを使用した.

3.1 ポートセット間の類似度

一般的に, 埋め込み空間が適切に表現されていると, 観

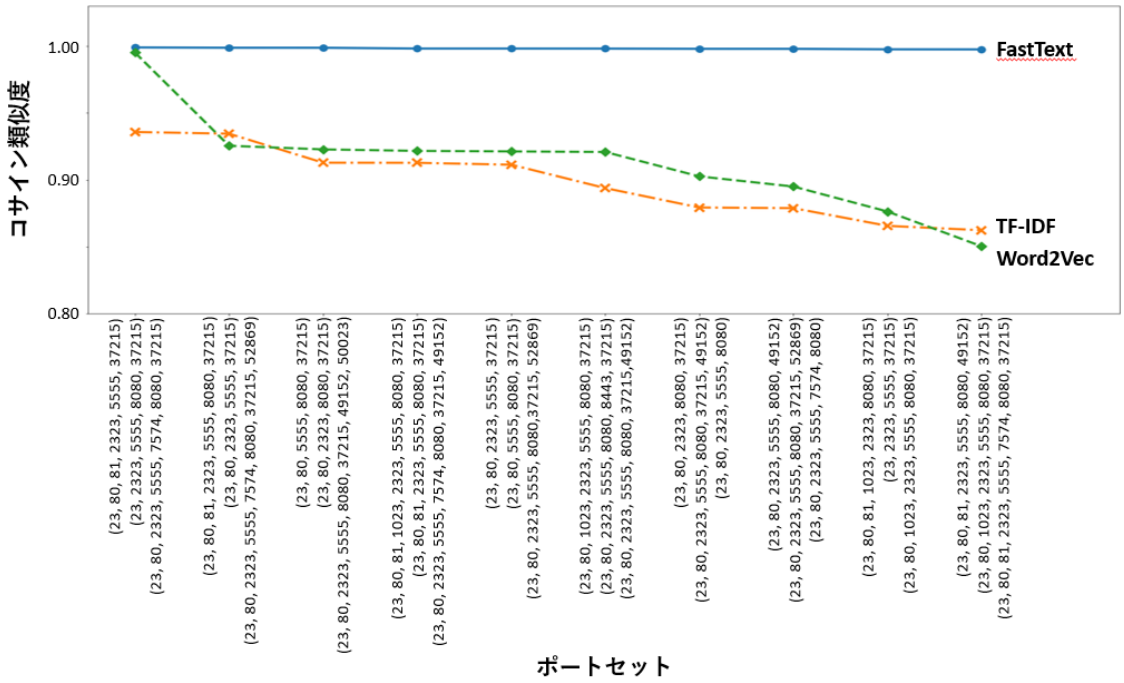


図 2 ポートセット (23, 80, 2323, 5555, 8080, 37215) とのコサイン類似度が高い上位 10 個のポートセット. x 軸は, 1 目盛ごとに左側が FastText のポートセット, 中央が TF-IDF のポートセット, 右側が Word2Vec のポートセットである.

測された現象を理解する際に有意義な解釈をもたらすことができ, 分類精度やクラスタリング結果の性能向上につながる. 従って, 提案手法により得られた Mirai 感染ポートセットの埋め込みベクトルが, Mirai 亜種間の類似性を適切に示す指標になっているのかを評価する.

以下の実験では, TF-IDF [10], Word2Vec, FastText の 3 つの手法により得られるポート番号埋め込みベクトルの性能を比較する. 2020 年 2 月 9 日から 2020 年 2 月 15 日の TCP/SYN パケットにおいて, 62,817 ポートが観測されるため, TF-IDF によるベクトルは 62,817 次元となる. Word2Vec による埋め込みベクトルは, 学習データとして 2020 年 2 月 9 日から 2020 年 2 月 15 日の TCP/SYN パケットを使用し, この期間に観測された 62,817 ポートそれぞれに 62,817 次元の one-hot ベクトルを設け, それを skip-gram [11] を用いて学習し 200 次元にすることで得られる. 提案手法である FastText による埋め込みベクトルは, 学習データとして 2020 年 2 月 9 日から 2020 年 2 月 15 日の TCP/SYN パケットを使用し, この期間に観測された 62,817 ポートとそのサブワードから成る 65,533 個の語彙に, 65,533 次元の one-hot ベクトルを設け, それを skip-gram を用いて学習し 200 次元にすることで得られる.

3 つの特徴表現方法を評価するために, 次のポートセットを例として挙げる: (23, 80, 2323, 5555, 8080, 37215) . このポートセットを持つホストは 2 月 15 日に観測され, そのパケットのトラフィックパターンは Mirai のシグネチャと一致した. 従って, 上記のポートセットと類似したスキャンパターンを持つホストは, Mirai の亜種に感染している可能性が高いと考えることができ, 類似したベクトルで表現されている必要がある.

3 つのベクトル埋め込み手法のそれぞれで, ポートセット (23, 80, 2323, 5555, 8080, 37215) と類似性の高いポートセットをコサイン類似度の算出により求めた. 図 2 は, コサイン類似度が最も高い上位 10 個のポートセットを示している. 提案手法である Fasttext を用いた埋め込みベクトルは, 上位 10 個のポートセットに対して 0.99 以上のコサイン類似度を示しており, 類似したのポートセットが類似したベクトルで表現できていることがわかる. 一方, 他の埋め込み手法では, コサインの類似度が大幅に低下していることがわかる. この結果から, 提案手法がポート番号を適切にベクトルとして表現できていることが示された.

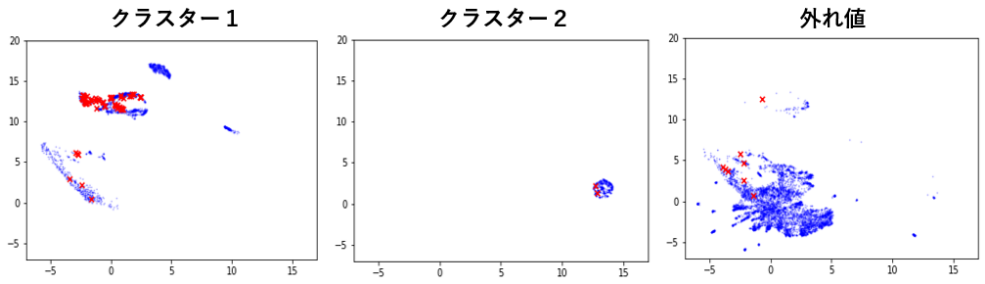
3.2 Mirai 感染ポートセットのクラスタリング

3.1 節と同じデータと手法を用いてスキャンパターンをクラスタリングし, その精度を比較した. スキャンパターンのクラスタリング精度を比較するために, 2020 年 2 月 15 日に観測された 9,581 個のポートセットから, Mirai 感染ポートセットを抽出したところ, 75 個確認できた.

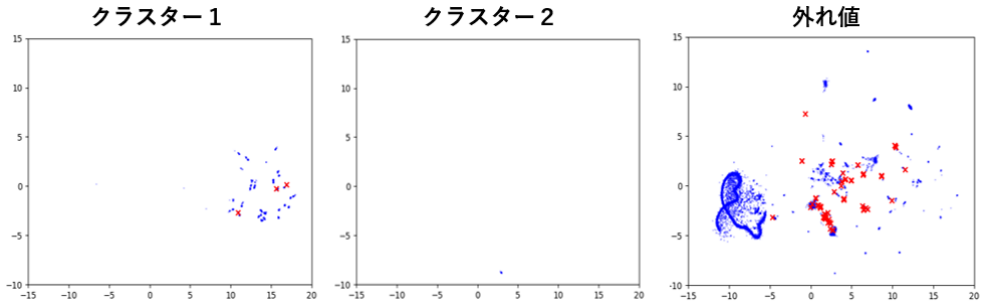
表 2 は, FastText, TF-IDF, Word2Vec の各手法により得られたポートセットベクトルを, DBSCAN を用いてクラスタリングした結果を示し, 図 3 は, 各手法により得られたポートセットベクトルに UMAP を用いて次元圧縮を行い, 可視化した結果を示している. 外れ値とは, DBSCAN のコ

表 2　FastText, TF-IDF, Word2Vec により得られたポート番号埋め込みベクトルを用いて，2020 年 2 月 15 日に抽出された 75 個の Mirai 感染ポートセットを含む 9,581 個のポートセットをクラスタリングした結果.

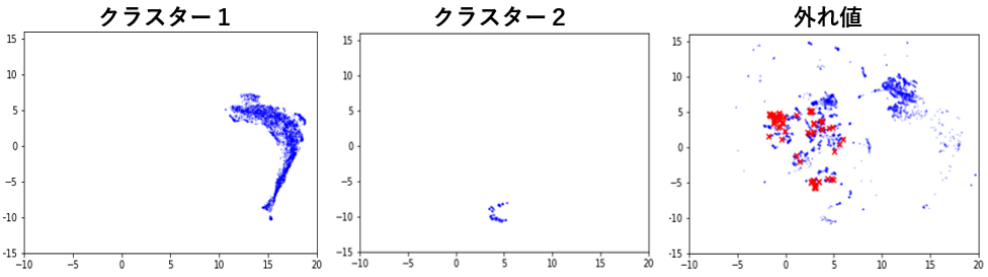
	FastText		**TF-IDF**		**Word2Vec**	
Cluster No.	#ポートセット	#Mirai ポートセット	#ポートセット	#Mirai ポートセット	#ポートセット	#Mirai ポートセット
1	**2,428**	**65**	671	3	4,439	0
2	494	2	27	0	237	0
3	7	0	449	0	444	0
4	40	0	87	0	23	0
5	110	0	46	0	4	0
6	9	0	1	0	3	0
外れ値	6,763	8	**8,570**	**72**	**4,701**	**75**



(a)FastText



(b)TF-IDF



(c)Word2Vec

図 3　(a) FastText, (b) TF-IDF, (c) Word2Vec を使用した Mirai 感染 (×) および Mirai 非感染 (○) ポートセットの可視化.

ア点でも到達可能点でもない点を指す．可視化結果は，全手法で Mirai 感染ポートセットを含むクラスタ 1，クラスタ 2，および外れ値のみを示している．

提案手法である FastText のクラスタリング結果は，ほとんどの Mirai 感染ポートセットをクラスタ 1 にクラスタリングできていることを示している．図 3 (a) からは，クラスタ 1 内でも Mirai 感染ポートセットが集まって可視化されていることが分かる．クラスタ 2 にクラスタリングされた Mirai 感染ポートセットには，クラスタ 1 には出現しない 7547/TCP が含まれているため，別の攻撃グループとしてクラスタリングされていると考えられる．7547/TCP は Mirai の亜種が標的とするポート番号で，2019 年 12 月以降のアクセス増加が報告されている．外れ値に含まれる Mirai 感染ポートセットには，0/TCP, 654/TCP, 55717/TCP などのポートセット全体でほとんど出現しないポート番号が含まれているため，外れ値になっていると考えられる．

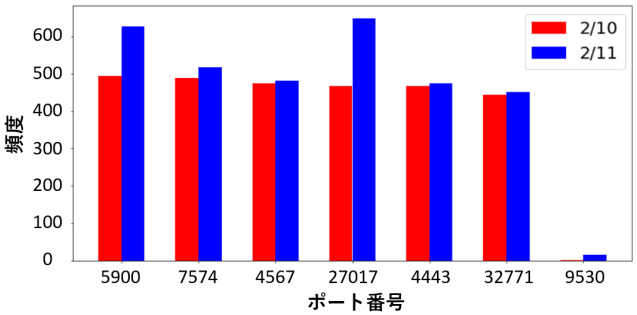


図 4　2 月 10 日と 2 月 11 日にクラスタ 1 にクラスタリングされたポートセット内に存在し, 2 月 10 日時点で出現頻度が 500 以下, 出現頻度が前日より増加した 5 以上のポート番号とその頻度

TF-IDF と Word2Vec については, ポートセットの半分以上が外れ値に含まれており, 図 3 (b) (c) の可視化結果においても, Mirai 感染ポートセットが提案手法よりも分散して可視化されていることが分かる.

上記の実験結果から, 提案手法が Mirai 感染ポートセットを他の埋め込み手法よりも適切に検出できており, スキャンパターンのクラスタリングに効果的であることを示した.

3.3　新たな脆弱性の検知

1 節で述べた通り, Mirai のソースコードの公開により, 新たな脆弱性を攻撃対象とするように更新された Mirai の亜種の出現が問題となっている. そこで, 新たな脆弱性を狙う Mirai 亜種の出現を検知するため, 提案手法を用いてクラスタリングを行い, Mirai が最も多く含まれるクラスタ内に存在する宛先ポート番号の頻度を数えた. その結果, 2020 年 2 月 11 日から増加した新たな Mirai の亜種によるものと思われる 9530/TCP への通信を検知した.

2 月 10 日と 2 月 11 日に抽出されたポートセットをクラスタリングすると, 外れ値とクラスタ 1 から 7 までの 7 つのクラスタにクラスタリングされた. クラスタ 1 のポートセットには, 23/TCP, 2323/TCP, 80/TCP, 8080/TCP が非常に多く含まれているため, クラスタ 1 は Mirai と Mirai の亜種で構成されているクラスタであると推測できる. 新たな脆弱性を標的とした Mirai 亜種の出現を検知するために, クラスタ 1 に存在する宛先ポート番号の頻度を求めた (図 4) . 2 月 10 日の時点で, 9530/TCP への通信は 1 回だけであったが, 2 月 11 日では 16 回に増加していることが確認できた. ただし, 2 月 10 日時点で出現頻度が 500 を超えている宛先ポート番号と, 増加した出現頻度が 5 以下の宛先ポート番号は新たに狙われ始めている脆弱性ではないとして省略している. 図 5 の可視化結果からも, 9530/TCP を含むポートセットが集合を形成しながら増加していることを確認することができた.

2020 年 2 月からの 9530/TCP への通信の増加は, 国立情報通信技術研究所が発行する NICTER ブログ [12] において Mirai の亜種であると報告されている. 9530/TCP への通信の増加が一時的なものではなく, Mirai に新たな攻撃対象として組み込まれた脆弱性であることを確認するため, 9530/TCP を含むポートセット数の推移を図 6 に示す. 2 月 11 日以降もポートセットの数が増加していることが確認できたので, 提案手法により, 新たな脆弱性を標的とするように更新された Mirai 亜種の出現の検知が可能であることを示した.

4.　結論

本論文では, FastText を用いて, マルウェアの亜種間の宛先ポート番号の部分パターンの類似性を学習するポート番号埋め込みベクトルを提案した. ダークネットで観測される TCP/SYN パケットに FastText を適用することにより, ポート番号において頻繁に共起する部分文字列の関係に基づいてポート番号のベクトル表現を取得した。さらに、このポート番号埋め込み空間でスキャナーのクラスタリングを行うことで, マルウェアの傾向を把握した.

実験では, /16 ダークネットセンサーで 2020 年 2 月 1 日から 29 日までに収集されたパケットに分析を行った. その結果, 提案手法である FastText を用いた手法は, TF-IDF と Word2Vec によって得られたポート番号埋め込みベクトルよりも Mirai 亜種のスキャン活動をよりよくクラスタリングできることが示された. さらに, Mirai により構成されているクラスタ内の宛先ポート番号の頻度を調べたところ, 2020 年 2 月 11 日から増加し始めた新たな Mirai 亜種による 9530/TCP への通信を検知した. 以上の結果から, 提案手法はマルウェアによるスキャン活動の監視に有効であると考えられる.

今後の取り組みでは, ポート番号埋め込みベクトル表現を改善して, マルウェアのクラスタリングと検出の精度を高める. 本論文では, Mirai のシグネチャを利用してクラスタリングの精度を評価したが, Mirai 以外のマルウェアにおいても正確に特定する手段を見つけ, 提案手法であるポート番号埋め込みベクトルの有効性を引き続き検証する必要がある. マルウェアを正確にクラスタリングできれば, 同時に動作する同じマルウェアに感染したホストを特定でき, 大規模攻撃を引き起こすボットネットの検出につながる. また, これらのボットネットの活動を監視することで, 新たなサイバー攻撃の兆候を検知できるシステムの開発が可能になると考えている.

5.　謝辞

本研究は, 科研費基盤研究 (B) (課題番号 16H02874) の助成を受けたものである.

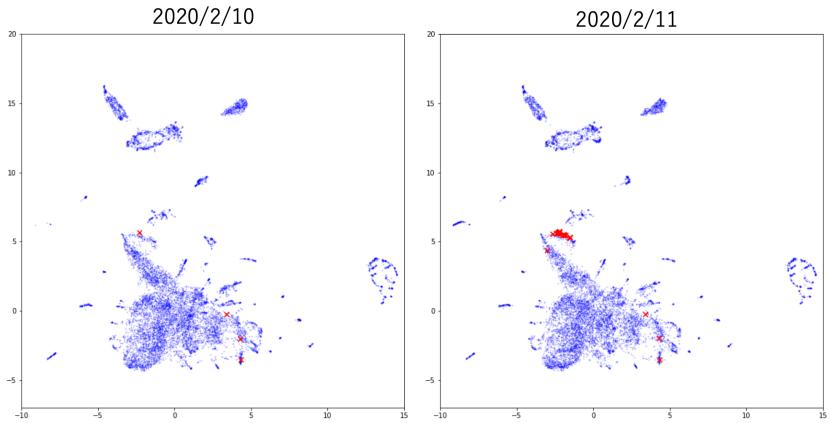


図 5　2 月 10 日と 2 月 11 日における 9530/TCP（×）を含むポートセットと 9530/TCP（○）を含まないポートセットの可視化.

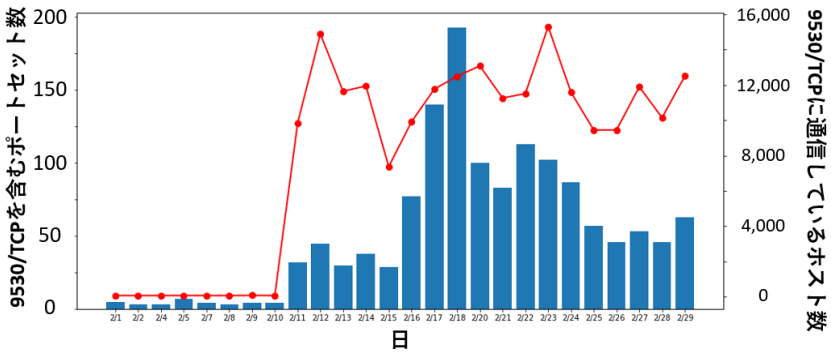


図 6　9530/TCP を含むポートセット数の変化と, ポートセットを満たすホスト数の変化.

参考文献

[1] Kolias, C., Kambourakis, G., Erhan., Stavrou, A., Voas, J.: DDoS in the IoT: Mirai and Other Botnets. *IEEE Computer*. 50, 80 - 84 (2017)

[2] National Institute of Information and Communications Technology. NICTER Observation report 2019, Retrieved June 20, 2020, from https://www.nict.go.jp/cyber/report/NICTER_report_2019.pdf

[3] Buzak, L.A., G. Erhan.: A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection. *IEEE Communications Surveys & Tutorials*. 18, 1153 - 1176 (2016)

[4] Ozawa, S., Ban, T., Hashimoto, N., Nakazato, J., Shimamura, J.: A Study of IoT Malware Activities Using Association Rule Learning for Darknet Sensor Data. *Int. J. Inf. Secur.* 19, 83–92 (2020)

[5] Ring, M., Dallmann, A., Landes, D., Hotho, A.: IP2Vec: Learning Similarities between IP Addresses. *IEEE International Conference on Data Mining Workshops (ICDMW)* (2017)

[6] Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781* (2013)

[7] Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*. 5, 135–146 (2017)

[8] Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv:1802. 03426* (2018)

[9] Ester, M., Kriegel, P.H., Sander, J., Xu, X.: A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *KDD'96: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining,*. 226–231 (1996)

[10] Ramos, J.: Using TF-IDF to Determine Word Relevance in Document Queries (2003)

[11] Mikolov, T., Sutskever. I , Chen. K , Corrado. G , Dean. J.:Distributed Representations of Words and Phrases and their Compositionality. *Advances in Neural Information Processing Systems*. 26 (2013)

[12] National Institute of Information and Communications Technology. NICTER Blog, Retrieved June 20, 2020, from https://blog.nicter.jp/