

探索的データ解析を目指す トラフィック解析についての一検討

宮本 大輔^{1,a)}

概要: 近年、通信トラフィックから特定のトラフィックを識別する、あるいは分類する手法の研究開発が盛んである。古典的には5タプルと呼ばれる送信元IPアドレス、送信先IPアドレス、送信元ポート番号、送信先ポート番号及びプロトコル番号から、トラフィックをパターン認識する手法が挙げられる。これらの研究ではデータセットの増加や機械学習の発展だけでなく、トラフィックの特徴を捉え数値化するという処理、いわゆる特徴量抽出が重要である。本研究では、トラフィック解析を行う先行研究について特徴量抽出という観点から調査を行う。また、Bag of Featuresを用いた深層学習による解析や、探索型データ解析を目指す上での可視化技術の取り組みについてプロトタイプ実装を用いて考察を行う。

キーワード: ネットワークトラフィック, 特徴量抽出, Bag of Features, 探索データ解析

A Consideration for Traffic Analysis toward Exploratory Data Analysis

DAISUKE MIYAMOTO^{1,a)}

Abstract: Recently, research of classification and/or detection using network traffic has attracted a lot of attention. The “five-tuple information” which comprises of source and destination IP addresses, source and destination port numbers and type of protocol has been widely used in traffic analysis. As for accurate characterization of traffic, it is important to extract features of the traffic as well as using advanced data and machine learning techniques. This paper shows a survey of traffic analysis aspect from the feature extraction. This paper then presents an issue and its possible solution, a bag of features for deep learning that does not depend on the development of new features. This paper finally shows my consideration toward exploratory data analysis and preliminary implementation of a visualization tool based on virtual reality.

Keywords: Network Traffic, Feature Extraction, Bag of Features, Exploratory Data Analysis

1. はじめに

近年、インターネットのトラフィックを解析する手法が数多く提案されている。例えば、クライアント端末におけるトラフィック解析は、トラフィックのパターン認識により、エンドユーザの意図しない通信を行うマルウェアを検知に役立てられる。また、また、サーバを保有するサービス事業者は、提供するサービスに対するトラフィックを解

析する、あるいは今後のサービスを充足させるためトラフィックの傾向を予測するなどの目的で用いられる。

トラフィック解析手法に共通して見られる点として、トラフィックデータの収集、収集されたデータから特徴量(Features)の抽出、抽出された特徴量を用いた機械学習などによる解析、といった3つの段階を経ることが挙げられる。解析者は、インターネットにおけるエッジや端末、あるいはコアなど場所を問わずトラフィックのデータを収集する。このデータは、PCAP形式[1]などのパケットをキャプチャしたデータ、IPFIX[2]などのフロー情報などであることが多いが、これに限らない。このデータをもとに、トラフィックを性格づけるような特徴を発見、これ

¹ 奈良先端科学技術大学院大学
Nara Institute of Science and Technology, Ikoma, Nara, 630-0101, Japan

^{a)} daisu-mi@is.naist.jp

を数値データとして抽出する。この手法は2節に詳述するが、多種多様な方法が開発されている。最後に、抽出によって得られた数値データなどをもとに機械学習などを用いてパターン認識による分類、あるいは良性が悪性かを検知するなど、目的に則した解析が行われる。今回調査の対象となった論文での解析手法には回帰、クラスタリング、決定木、ニューラルネットワークを利用するものも多く見られた。

なお、このようなフェーズを踏襲しなくても、トラフィック解析用のデータセットが配布されていることも多く、著名なデータセットとして KDD データセット [3] が挙げられる。データセットには、トラフィックのキャプチャデータであることもあるが、トラフィックデータから特徴量を抽出したものであることが多い。トラフィック、とりわけ各パケットのペイロードに含まれる内容には、エンドユーザにとって機微となる情報が含まれている可能性もあるため、プライバシーに配慮した措置と位置づけられている。一方で、トラフィックの収集や特徴量の抽出といった環境・設備・知識を持たなくても、データ解析が行えるという面もある。

一方で、近年はネットワーク資源、コンピュータ資源が増加の一途を辿っている。著者の研究グループでは、1日数 Gbyte オーダのネットワークデータが収集可能であり、このデータのリアルタイム解析を目的とした研究を行っている。このため、ビッグデータ解析基盤によるネットワーク集計基盤 [4] 及び GPU を用いた高性能解析環境 FENNEL [5] など、収集・解析を行うソフトウェア・ハードウェアの基盤技術に着目して研究開発を行ってきた。

本論文では、データからの特徴量抽出手法について調査を行う。調査方式では、トラフィック解析を行っている論文やそのデータセットにある特徴量を調べ、分類を試みた。一方で、新しい特徴量をどう発見していくかということは依然として課題であり、この問題に対するブレイクスルーとして深層学習で用いられる特徴量抽出のアルゴリズムとして、Bag of Features に着目した。一方で、データが増加している現状から確証的データ解析ではなく探索的データ解析が必要であるとも考える。この探索的データ解析を実施するための一助となるのが可視化技術であり、本論文では既存のトラフィック可視化ソフトウェアを VR 技術を用いて没入感ある環境で可視化を行う技法について考察を行った。

以下、2節にトラフィック解析における特徴量抽出手法についてまとめる。3節において Bag of Features アルゴリズムによる解析の試行、4節において探索的データ解析を行うための可視化について考察を行い、最後に最後に5節についてまとめと今後の課題を述べる。

表 1 5 タブルの特徴量

5 タブル情報		
内容	形式	単位
送信元 IPv4/IPv6 アドレス	数値/文字列	なし
送信先 IPv4/IPv6 アドレス	数値/文字列	なし
送信元ポート番号	数値	なし
送信先ポート番号	数値	なし
ICMP/ICMPv6 タイプ	数値	なし
ICMP/ICMPv6 コード	数値	なし
プロトコル種類	数値	なし

2. トラフィックの特徴量抽出方法

特徴量について、本論文では基本となる5タブルの情報、通信フローに関する特徴、アプリケーションプロトコル及びペイロードに関する特徴について述べる。

2.1 5 タブルの特徴量

5 タブルの特徴量は、IPv4 ヘッダ/IPv6 ヘッダに含まれる送信元及び送信先のアドレス (以下、IP アドレスと記す)、上位プロトコルの種別と、その上位プロトコルが TCP セグメント、UDP データグラムである場合は送信元及び送信先のポート番号となる。なお、上位プロトコルが ICMP/ICMPv6 プロトコルである場合は、ICMP タイプ及び ICMP コードを含める場合もある。また、IP アドレスは文字列としての情報を持ち、値の大小が何らかの分類に用いられることは少ない。一方で、IP アドレスの下位ビットの近さは、当該ホストが同一ネットワークに属するといった情報を示すため数値的に比較可能という側面も持つ。

TCP 及び UDP のポート番号は、ウェルノウンポート番号 (0-1023)、登録済みポート番号 (1024-49151)、動的・プライベートポート番号 (49152-65535) に分類される。特にウェルノウンポート番号は、HTTP やメールなど通信がどのようなサービスを利用しているかを特定できるため、特に重要である。ウェルノウンポート番号は著名なアプリケーションに用いられ、そのポート番号におけるサービスの起動には管理者の権限が必要となる。ウェルノウンポート番号は質的情報であるが、動的・プライベートポート番号などはクライアント端末がサーバ端末にコネクションを接続する際に利用されていないポート番号が順番に持ちられることから、通信の意味の近さを示す量的情報にもなり得る。

2.2 通信フローに関する特徴量

通信フローの解析では、アプリケーションによって動的に利用されるポート番号を集約して解析できる特徴がある。一般には、主に宛先ポート番号からサービスが類推される傾向にある。

通信フローに関する特徴量として、単一のフローで取得可能なフローの持続時間 (flow duration, connection time) を用いているものも多い。一方で、それらの最小、最大、平均、分散及び標準偏差が使われる傾向にある。以下、可読性を高めるため、最小、最大、平均、分散及び標準偏差を統計的情報として記載する。また、フロー情報はクライアント端末ではなくルータなどの中継装置で観測することが多く、一般的に outbound, outgoing, forward といった方向、あるいは inbound, incoming, backward といった方向など、フロー情報を採取する装置からみた方向性で定義されることが多い。可読性のため、本論文では外向けの通信を送信、内向けの通信を受信として記す。

通信フローに関する特徴量として、時間に関する情報としては、通信の間隔、アイドル時間 (データの送受信の両方がない時間)、パケット送信間隔、パケット受信間隔の統計的情報がある。

また、多数のフローを観測する場合、主となるフローに追従していくつもの TCP セッションが確立されることがある。本来はマルチパス TCP 用語であるが、本論文では参考文献 [6,7] に従いサブフローと記す。

通信時間以外にも、通信のパケットの数やパケットの長さの合計及びその統計的情報が用いられる。なお、Moore [8] では Ethernet, IP, 一般的な TCP セグメント及び Selective ACK などの特徴のある TCP オプションを用いている TCP セグメントの統計的情報や値をフーリエ変換した値が定義されており、249 種類の特徴が定義可能としている。

なお、厳密には 5 タプル情報に含まれるポート番号も TCP セグメントに関する特徴量であるが、それ以外の特徴量としてフローに含まれる TCP のフラグが付与されたパケットの数、フロー中における初期ウィンドウサイズの統計的情報も用いられている。

2.3 バイドロードに関する特徴量

一般にペイロードの内容まで解析することは、IDS などでは Deep Packet Inspection と呼ばれる。シグネチャや文字列パターンのマッチにより各パケットの内容を検査し、攻撃か否かを判定する際に用いられる。

HTTP トラフィックに関する特徴量は図 3 に挙げられる。ホスト名などある程度の数の HTTP トラフィックについて、特定のホスト名の文字列として編集距離などの調査に用いる特徴量とするもの、標準集団の中で当該ホストへの接続が何回行われているかを言わば Inverse Document Frequency (IDF) として特徴量とするものがある。なお、HTTP リクエストやレスポンスの到着間隔も用いられる [9] は、フローの到着間隔と同様であるため割愛した。なお、MIME タイプの有無とは、とりわけ、Windows における実行可能なファイル形式である EXE 形式などの有無の検査を示す。

表 2 通信フローに関する特徴量

通信フローの時間に関するもの		
内容	形式	単位
フローの持続時間	数値	sec
アクティブ時間の統計的情報	数値	sec/なし
アクティブ時間の割合	数値	%
パケット送信間隔の統計的情報	数値	sec/なし
通信間隔の統計的情報	数値	sec/なし
アイドル時間の統計的情報	数値	sec/なし
アイドル時間の割合	数値	%
送信サブフローの持続時間の統計的情報	数値	sec
受信サブフローの持続時間の統計的情報	数値	sec
通信フローのパケット数に関するもの		
送信パケット数の合計	数値	パケット
送信パケット数の統計的情報	数値	パケット
単位時間あたりの送信パケット数	数値	pps
受信パケット数の合計	数値	パケット
受信パケット数の統計的情報	数値	パケット
単位時間あたりの受信パケット数	数値	pps
送受信パケット数の割合	数値	%
通信フローのパケット長に関するもの		
送信パケット長の合計	数値	Bytes
送信パケット長の統計的情報	数値	Bytes
単位時間あたりの送信パケット長	数値	Bps
受信パケット長の合計	数値	Bytes
受信パケット長の統計的情報	数値	Bytes
単位時間あたりの受信パケット長	数値	Bps
送受信パケット長の割合	数値	%
TCP コネクションに関するもの		
コネクションの数	数値	なし
TCP SYN パケットのエラー率	数値	%
TCP SYN パケットの数	数値	パケット
TCP PSH パケットの数	数値	パケット
TCP URG パケットの数	数値	パケット
TCP FIN パケットの数	数値	パケット
初期ウィンドウサイズの統計的情報 (送信)	数値	なし
初期ウィンドウサイズの統計的情報 (受信)	数値	なし
ホストで提供されるサービスの数	数値	
同じ送信元ポート番号の占める割合	数値	%
その他		
RTT 値の合計	数値	なし
サービス	数値	なし

表 3 HTTP トラフィックに関する特徴量

HTTP プロトコルに関する情報		
ホスト名	文字列/IDF	なし
リクエストメソッド	文字列/IDF	なし
ユーザエージェント	文字列/IDF	なし
URI	文字列/IDF	なし
特定の MIME タイプの有無	真偽値	なし
クッキーが含まれる割合	数値	%
リファラーが含まれる割合	数値	%
クライアント毎のリクエストの数	数値	なし
URL に関する情報		
URL に含まれる記号	数値	なし/%

一方、DNS トラフィックに関する特徴量は Bilge [10] (2014) らの研究に詳しい。マルチテナントの有無とは、ドメイン名の正引きによって得られた IP アドレスをさらに

表 4 DNS トラフィックに関する特徴量

ドメイン名に関する情報		
有意な文字の割合	数値	%
数字の割合	数値	%
NGram	IDF	
DNS プロトコルに関する情報		
DNS 応答の種類	数値	なし
DNS クエリの数	数値	なし
DNS クエリのシャノン情報量	数値	なし
DNS 回答 RR の数	数値	なし
DNS 権威 RR の数	数値	なし
DNS 追加 RR の数	数値	なし
DNS レコード長	数値	Bytes
権威ある回答か否か	真偽値	なし
NXDOMAIN レコードの内容	文字列	なし
SOA レコードの各時間	数値	sec
TXT レコードの長さ	数値	Bytes
応答に含まれる IP アドレスの種類	数値	なし
応答に含まれる異なる国	数値	なし
hline マルチテナントの有無	真偽値	なし
TTL に関する情報		
TTL の統計的情報	数値	sec/なし
異なる TTL 値の数	数値	なし
TTL の変更回数	数値	なし
特定 TTL が使われる割合	数値	%
パターンに関する情報		
1 日ごとの類似性の有無	真偽値	なし
hline 繰り返しの有無	真偽値	なし

逆引きするというもので、この IP アドレスに複数のドメインが登録されているかどうかを調べる際に用いられる。また、Fast flux などの攻撃手法では TTL が短く設定されるという傾向にあり、100 以下の TTL が使われる割合からサイバー攻撃を発見しようという試みもある。有意な文字列の割合は、ドメイン名の長さでドメイン名に含まれる文字列のうち、辞書などに掲載されている文字列集合の長さの、ドメイン名全体に占める割合から算出される。

2.4 その他の特徴量

その他の特徴量としては、パケットの IP アドレスから算出される AS 番号についての情報がある。悪性のサーバなどでは AS 番号による集約により分類を効果的に行えることが知られており [11]、トラフィック解析においても送信元 AS 番号や送信先 AS 番号を特徴量として用いる利点はあると考えられる。また、AS の隣接関係にはカスタマー AS、トランジット AS など様々な種類があり、これに基づいた解析も考えられる [12]。また、ファイル転送を行うプロトコルにおいて、コマンドによる制御を行う時間及びファイルをバルク転送する時間の入れ替わりなどが用いられている。なお、ダークネット IP アドレスへのアクセスは、多くはポートスキャン又は当該 IP アドレスが IP スプーフィング攻撃に用いられてその戻りとしてアクセスがあったなどと考えられるため、攻撃に起因するものである

表 5 その他の特徴量

送信先 AS 番号	数値	なし
送信元 AS 番号	数値	なし
隣接 AS 数	数値	AS
制御と転送の入れ替わり時間	数値	sec
ダークネット宛のトラフィックか否か	真偽値	なし

と想定される。

2.5 関連研究

トラフィック分類を試みる 2000 年台前半の研究には、インターネットトラフィックの分類を試みるものも多く見られる。特に、動画再生などに用いられるのが HTTP ではなく RTSP などのプロトコルが使われていた時代背景も影響していたと見られる。Roughan [13] (2004) らは、データセットに Auckland dataset [14] を用いて、telnet のようなインタラクティブな通信、ftp-data のようなバルクデータ転送、RTSP のようなストリーミングメディアなどのトラフィックを、Linear Discriminant Analysis (LDA) 及び K 近傍法で分類している。McGregor [15] (2004) は、パケットサイズの統計量、受信間隔の統計量、送信バイト数、フロー時間、通信モードとバルク転送モードの入れ替わりの時間を調査している。

Moore [16] (2005) らは、ポート番号、パケットヘッダ、プロトコル特有の挙動（最初の送信バイト数）などから、バルクデータ転送、データベースへのアクセス、インタラクティブな通信、メール、サービス、www、P2P、悪意ある通信、ゲーム、ストリーミングメディアのトラフィックの分類を試みている。さらに、この特徴量を改訂し、[17] (2005) では、フロー持続時間、ポート番号、パケット到着間隔時間・ペイロードの統計的情報、及びパケット到着間隔時間のフーリエ変換した値を用い、Naive Bayes による分類を行った。

Dainotti [18] (2009) らは分類のためのフレームワークを提案しており、このフレームワークにおけるトラフィック分類の方法として、ペイロードを CAIDA CoralReef [19] を用いた他、パケット長及び到着間隔による解析を行っている。

Este [20] (2009) らは CAIDA などのデータセットからトラフィック分類を試みるため、one class SVM の改造を用いて TCP トラフィックを分類している。Keshapag [21] (2012) はさらにこの方法を Lagrangian SVM (LSVM) によって高精度化を試みている。

Wheelus [22] (2014) は、ビッグデータ観測環境においてトラフィック解析を行う研究を進めている。トラフィック抽出については [23] で述べるように、5 タプル情報とフローの時間に関する情報、送信・受信のパケット長の統計的情報（特に類似性を分散によって推測できると主張している）、送信・受信の時間あたりのパケット数、パケット長

及び送受信における割合, AS 番号, トラフィックがデータネットワーク IP アドレス宛か否か, TCP フラグを特徴量として用い, C4.5 や Naive Bayes, Multi Layer Perceptron (MLP), Random Forest を用いて解析している。

Divyatmika [24] (2016) は, NSL-KDD データセットをもとに K 平均法で分類, 後に MLP で検知を行っている。一方で, Potluri [25] (2016) は NSL-KDD データセットをもとに Deep Neural Network による検知を試みている。

通信フローに関して, Mai [26] (2016) らは Tranalyzer [27] を用い, 各フローから 93 個の特徴量を抽出し, 特に送信元ポート番号, 送信先ポート番号, パケット数の統計的情報及び 1 秒あたりのパケット送信数・送信長, コネクションの数を用い, K 平均法及び Random Forest によりボットネットの検知を行っている。Borger [28](2016) らは, フローの持続時間及びサービス, プロトコル種別と, ホストごとあるいはサービスごとにコネクション数, 送信バイト数, TCP SYN パケットに対するエラー率を計算し, K 平均法を用いて分類を試みている。S. Kumar [29] (2016) らはフローの持続時間, サービス, 送受信のパケット数・パケット長の統計的情報を, C4.5, Random Forest, Ripple Down Rule learner, JRIP [30] などを用いた攻撃検知を行っている。Azab [31] (2016) の用いた特徴量は [6] が提案したものが基礎であり, 5 タプル情報, パケット送信数・送信長・受信数, 受信長の合計及び統計的情報量, 間隔・通信時間・アイドル時間の統計的情報量, サブフローの時間及び 送信・受信ごとの TCP PSH セグメントの数から C4.5 による分類が行われている。

Z. Wang [32] (2016) は, HTTP トラフィックに含まれる URL の隣接構造をグラフ化し, K 平均法で分類を行っている。また, S. Wang [33] (2016) は, HTTP トラフィックをホスト名, URI, HTTP リクエストメソッド, ユーザエージェントの文字列, 及び通信フローの送受信のバイト数の統計的情報を特徴量として, C4.5 により悪性トラフィックの検知を行っている。

マルウェアの発する通信に注目した検知も行われている。Android マルウェアの送出する HTTP のパケット長, 到着間隔についての統計的情報を既存のソフトウェアを使ったシグネチャと合わせて解析する Chen [34] (2015) らの方法の他, Morichetta [35] (2016) らはマルウェアの送出する HTTP パケットに含まれる URL を DBSCAN による文字列距離に基づいて検知を試みている。Lashkari [7] (2017) は, Android マルウェアの発する通信に着目した解析を行っている。フローの時速時間, 送受信パケット長の合計及び統計的情報, アイドル時間及び送受信毎の時間間隔の統計的情報, 送受信のサブフローの時間, フローの TCP FIN パケット数, 送受信の初期ウィンドウサイズ及びセグメントサイズの統計的情報を C4.5, Random Forest, K 近傍法及び回帰モデルによって解析し, マルウェアの通

信の検知を行っている。

He [36] (2016) は VM に対するトラフィックの分類を, フローにおけるプロトコルの種類, サービス (ポート番号), 送信バイト数, 受信バイト数を基礎的な特徴量として抽出する。また, フローまたは複数フローに関して特定ホストに対するコネクションの数, エラー率, 異なるサービスへのアクセス回数及びその割合, 送信先ホスト数の種類, あるいは VM へのログインであれば成功・失敗の判定を特徴量とし, LSVM, RSVM, Multinomial Naive Bayes により分類している。

Heuer [37] (2016) は DNS に基づく特徴量として Bilge [10] らの特徴量を, AS に基づく特徴量としては Noroozian [12] らの特徴量を使い, ボットネットの検知を行った。同様に DNS と検知という文脈では, Watkins [38] (2017) は DNS クエリのシャノン情報量, DNS レコードの TTL, DNS クエリ数, 回答 RR 数, 権威 RR 数, 追加 RR 数, 権威ある回答か否か, レコードの種類を用いている。

Mishra [39] (2017) は, UNSW-NB [40] データセット, ITOC [41] を用い, それらを Decision Tree, Naive Bayes, Logistic Regression, ANN, EM Clustering, Random Forest を用いて解析されている。解析に用いられているアルゴリズムは Decision Tree, ANN, Bayes, SVM, K-NN である。

2.6 特徴量抽出の問題点

トラフィックの分類・検知の精度を高めるにはどうすればよいか。著者が考察する点は以下の 3 つである。

- データの量, データの質, データの種類増加
- 効果的な特徴量の発見及び利用
- 機械学習理論の高度化
- 計算機資源の増加

データの量の増加は, すなわち過剰学習を防ぐ効果を持ち, 予測の精度を高める直接的な効果を持つ。インターネットトラフィックは量が劇的に増加しており, また多種多様なアプリケーションのトラフィックも内包され多様性も持つことからビッグデータであると言える。また, 機械学習理論が高度化していること, それを支える計算機資源の増加も行われ続けていると言える。

問題となるのは特徴量の発見及び利用である。よりよい精度を得るためには効果的な特徴量を発見する必要がある。効果的な特徴量を選択的に用いることが必要であると思われる。一方で, 思いつく特徴量の発見には限りがあるとも考えられる。この問題については 3 節にて考察する。

また, これらのデータ解析がいわゆる確証的データ解析になっているのではないかと懸念がある。確証的データ解析とは, 統計的仮説検定に顕著であるが, 限られたサンプルにおけるデータの分布がどのような分布に従うかを予想し, データの母集団を予想するという方式である。例

例えば EM クラスタリングを行う際、特徴量の分布が混合ガウス分布に従っていることを示すべきであり、先に述べた文献 [15] ではこの検証を行っている。

確証的データ解析手法の長所は、適切なデータでは高い予測精度が得られること、及び理論や手法がいわばこなれており、誤差の少ない予想を行いやすく、その理由を説明しやすい。一方で、適切でないデータに対して精度が劣化する、先入観により誤った解析が行われる、予期しない結果に気づくのが困難といった短所もある。この問題について 4 節にて考察する。

3. Bug of Features による解析

トラフィックの特徴量抽出が人間の知見をもとに開発されている点について、本論文では近年の深層学習の分野にて頻出する、特徴量を自動的に発見する手法に着目して考察する。

例えば、コンピュータチェスや将棋の分野では、駒や陣形に人間が発見した経験則から特徴量を抽出していた方式から、三つの駒の位置関係の特徴量として抽出することにより飛躍的に性能が向上するというブレイクスルーがあった。この点を鑑みると、これまでのように人間が特徴量は開発して特徴量抽出を行うのではなく、数多くの特徴量を自動的に抽出し、深層学習と計算機資源で解析するという研究のアプローチも考え得る。

そこで、パケットのバイト列の羅列として Bag of Features (BoF) による特徴量の自動抽出を試みる。例えばパケットの IP ヘッダが図 1 のような PCAP 形式で与えられたとする。

```
0x0000: 4500 05dc...
```

図 1 パケットの形式

通常、このような形式のパケットの場合、4 は IPv4、5 はパケットが 5 ワード (160 オクテットまたは 160 バイト) であると認識される。これを 0x45 からなるバイト列が 1 回、0x50 が 1 回、0x00 が 2 回として特徴量の抽出を行う。次に、この 256 種類の特徴量を用いて、著者の研究グループで採取したトラフィックと IDS の出力を IP アドレス及びポート番号から結合した。データセットでは 27000 パケットのアラートをあげた通信を異常な通信として、同じ数のパケットを正常な通信として扱う。このデータセットを 256 種類の特徴量を用い、Deep Learning を用いて 10 分割交差検定により解析したところ、エラー率は 22.3%、 f_1 値は 0.773、AUC は 0.819 となる結果が観測された。

今回は試験的な実装により 2 バイトずつの BoF から特徴量抽出を行ったが、4 バイトずつの特徴量抽出も可能であり、この組み合わせにより効果的な特徴の発見も可能ではないかと考える。なお、本抽出を行う

アプリケーションはオープンソースソフトウェアとして <https://github.com/daisu-mi/pcap2csv> に公開している。

4. 探索型データ解析

Tukey の提唱した探索型データ解析 [42] は、データの可視化・要約を繰り返すことにより、解析者がデータについての知見を得て、解析モデルの適応を繰り返すことにより、より予測の精度を高める方式といえる。これは仮説検定などを必要としないため、データ以外のものは必要としない。一方で、決定論的な予測が提供されない、過剰適合が生じやすいといった問題がある。

しかし、ビッグデータの時代において、データが大量にあるということは探索型データ解析における過剰適合の問題は緩和されるのではないかと言える。一方で、データが大量にあるがゆえに仮説をたてるのが難しいといったビッグデータの問題は回避され得るのではないかと予想される。

故に重要なのは、データの可視化技術であり、本論文においてはトラフィックデータの可視化技術である。古くは MRTG から始まり、多くの可視化技術は二次元平面・三次元空間において先述の特徴量をグラフ構造で表示する。一方で、ヒューマンインタラクション技術は日々進化しており、このような技術とトラフィック可視化ソフトとの融合によるブレイクスルーが期待される。

著者らの開発しているトラフィック可視化ソフト PACTER [43] は、本来、サイバー脅威をリアルタイムに三次元空間上に可視化することを目的としたソフトウェアであった。これは、可視化によりサイバー攻撃の存在に気づくだけでなく、サイバー攻撃の対策が有効かを決定しやすくなるという試みである。PACKTER を開発するグループが所属していた研究室には、後にサイバー脅威情報をリアルタイムに可視化する NICTER [44] の開発に関与するメンバーもあり、一部の可視化には類似性も見られる。例えば、IP アドレス (0.0.0.0~255.255.255.255) を x 軸にとり、ポート番号を y 軸にとり、送信側及び受信側のトラフィックを三次元平面で表示する点が挙げられる。

本論文では、この PACKTER を Unity 及び VR 機能を用いることにより [45]、VR 画面上にトラフィック可視化を行う技術も試みた。その動作画面のうち、解析者の主観画面を図 2 に、両目に投影される画面を図 3 に示す。トラフィックを閲覧するユーザを三次元空間上に没入させることにより、新しい発見を得ようという取り組みである。VR コントローラとして HTC Vive が利用可能であり、実装はオープンソース化することを予定している。VR に対応した PACKTER のプログラムは <http://www.packter.jp> から利用可能である。

現時点では、トラフィックを没入感を持って閲覧する、トラフィックのフローを擬似的に体感するといった長所が



図 2 PACKTER VR 出力の主観画面

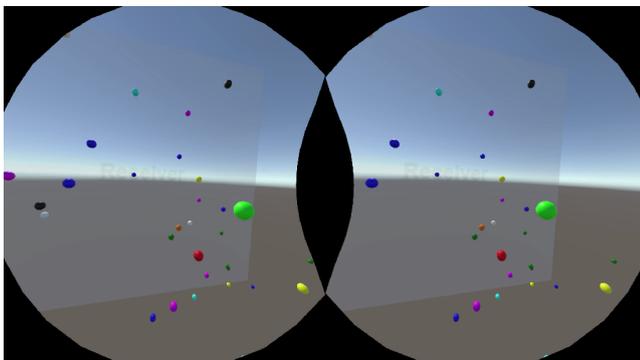


図 3 PACKTER VR 出力の両目に投影される画面

ある。一方で、VR ゴーグル着用による解析者の身体的な疲労といった短所もある。ユーザインタラクションの研究であればユーザエクスペリエンスを評価するものと思われるが、探索型データ解析を目指すのであれば、本手法がどの程度、既存の可視化では分からなかったトラフィックの特徴を発見できたかが評価されるべきであり、今後の課題である。

5. おわりに

本論文では、トラフィック解析における重要な要素である特徴量抽出について調査し、様々な特徴量が考案されている現状について解説した。その上で、トラフィックの特徴量抽出とは異なるアプローチとして、深層学習で用いられるような自動的に効果的な特徴量を発見する Bag of Features (BoF) の方式がどの程度利用可能について考察を行った。また、探索的データ解析を目指す上で必要不可欠なトラフィック可視化について、VR 技術を用いて可視化するソフトウェアのプロトタイプ実装を行った。今後の課題として、BoF 処理における抽出可能な特徴量の増加、探索的データ解析者の支援及び評価方法の確立が挙げられる。本論文のために開発したソフトウェアは本文に示した通り公開しており、より高度なトラフィック解析に役立てられることを期待している。

謝辞

本研究は、JST, CREST, JPMJCR1783 の支援を受けたものである。

参考文献

- [1] Jacobson, V., Leres, C. and McCanne, S.: libpcap and tcpdump, Online Available at: <http://www.tcpdump.org> (accessed 2018-01-19).
- [2] Aitken, P., Claise, B. and Trammell, B.: Specification of the IP Flow Information Export (IPFIX) Protocol for the Exchange of Flow Information, RFC 3954 (2013).
- [3] UCI KDD: KDD Cup 1999 Data, Online Available at: <http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html> (accessed 2018-01-19).
- [4] Tazaki, H., Okada, K., Sekiya, Y. and Kadobayashi, Y.: MATATABI: Multi-layer Threat Analysis Platform with Hadoop, *Proceedings of the 3rd International Workshop on Building Analysis Datasets and Gathering Experience Returns for Security* (2014).
- [5] Information Technology Center University of Tokyo: FENNEL, Online Available at: <https://www.cc.u-tokyo.ac.jp/service/wakate> (accessed 2018-01-19).
- [6] Arndt, D.: NetMate, Online Available at: <https://github.com/DanielArndt/flowtbag> (accessed 2018-01-19).
- [7] Lashkari, A. H., A.Kadir, A. F., Gonzalez, H., Mbah, K. F. and Ghorbani, A. A.: Towards a Network-Based Framework for Android Malware Detection and Characterization, *Proceedings of the Privacy, Security and Trust 2017 Conference* (2017).
- [8] Moore, A. W., Zuev, D. and Crogan, M.: Discriminators for use in flow-based classification, Technical report, Queen Mary University of London, RR-05-13 (2005).
- [9] Bhole, Y. and Popescu, A.: Measurement and Analysis of HTTP Traffic, *Journal of Network and Systems Management*, Vol. 13, pp. 357–371 (2005).
- [10] Bilge, L., Sen, S., Balzarotti, D., Kirada, E. and Kruegel, C.: EXPOSUER: a Passive DNS Analysis Service to Detect and Report Malicious Domains, *ACM Transaction on Information and System Security*, Vol. 16, No. 4, pp. A1–A28 (2014).
- [11] Whittaker, C., Ryner, B. and Nazif, M.: Large-Scale Automatic Classification of Phishing Pages, *Proceedings of the Network and Distributed System Security Symposium* (2010).
- [12] Noroozian, A., Korczynski, M., Tajalizadehkhoo, S. and van Eeten, M.: Developing security reputation metrics for hosting providers, *Proceedings of the 8th Workshop on Cyber Security Experimentation and Test* (2015).
- [13] Roughan, M., Sen, S., Spatscheck, O. and Duffield, N.: Class-of-Service Mapping for QoS: A Statistical Signature-based Approach to IP Traffic Classification, *Proceedings of the ACM Internet Measurement Conference*, pp. 135–148 (2004).
- [14] Micheel, J., Graham, I. and Brownlee, N.: The Auckland data set: an access link observed, *Proceedings of the 14th ITC Specialist Seminar on Access Networks and Systems* (2001).
- [15] McGregor, A., Hall, M., Lorier, P. and Brunskill, J.: Flow Clustering Using Machine Learning Techniques,

- Proceedings of the 5th International Workshop on Passive and Active Network Measurement*, pp. 205–214 (2004).
- [16] Moore, A. W. and Papagiannaki, K.: Toward the Accurate Identification of Network Applications, *Proceedings of the 6th International Workshop on Passive and Active Network Measurement*, pp. 41–54 (2005).
- [17] Moore, A. W. and Zuev, D.: Internet Traffic Classification Using Bayesian Analysis Techniques, *Proceedings of the ACM SIGMETRICS international conference on Measurement and modeling of computer systems* (2005).
- [18] Dainotti, A., de Donato, W., Pescapé, A. and Ventre, G.: TIE: a Community-oriented Traffic Classification Platform, *Proceedings of the 1st International Workshop on Traffic Monitoring and Analysis*, pp. 64–74 (2009).
- [19] CAIDA: CoralReef Software Suite, Online Available at: <https://www.caida.org/tools/measurement/coralreef> (accessed 2018-01-19).
- [20] Este, A., Gringoli, F. and Salgarelli, L.: Support Vector Machines for TCP Traffic Classification, *Computer Network*, Vol. 53, No. 14, pp. 2476–2490 (2009).
- [21] Keshapagu, S. and Suthaharan, S.: Analysis of Datasets for Network Traffic Classification, *Proceedings of the 8th Annual UNGG Regional Mathematics and Statistics Conference*, pp. 155–168 (2012).
- [22] Wheelus, C., Bou-Harb, E. and Zhu, X.: Towards a Big Data Architecture for Facilitating Cyber Threat Intelligence, *Proceedings of the 8th IFIP International Conference on New Technologies, Mobility and Security* (2016).
- [23] Wheelus, C., Zuech, R. and Najafabadi, M. M.: A Session Based Approach for Aggregating Network Traffic Data – The SANTA Dataset, *Proceedings of the IEEE Conference on Bioinformatics and Bioengineering*, pp. 369–378 (2014).
- [24] Divyatmika and Sreelesh, M.: A Two-tier Network based Intrusion Detection System Architecture using Machine Learning Approach, *Proceedings of the IEEE International Conference on Electrical, Electronics, and Optimization Techniques*, pp. 42–47 (2016).
- [25] Potluri, S. and Diedrich, C.: Accelerated Deep Neural Networks for Enhanced Intrusion Detection System, *Proceeding of the 21th IEEE International Conference on Emerging Technologies and Factory Automation* (2016).
- [26] Mai, L., Kim, Y., Choi, D., Bao, N. K., V.Phan, T. and Park, M.: Flow-Based Consensus Partitions for Botnet Detection, *Proceedings of the International Conference on Information and Communication Technology Convergence*, pp. 1253–1255 (2016).
- [27] Burschka, S., Dupasquier, B. and Fiaux, A.: Tranalyzer, Online Available at: <https://tranalyzer.com> (accessed 2018-01-19).
- [28] Boger, M., Liu, T., Ratliff, J., Nick, W., Yuan, X. and Esterline, A.: Network Traffic Classification for Security Analysis, *Proceedings of the IEEE SoutheastCon* (2016).
- [29] Kumar, S., Viinikainen, A. and Hamalainen, T.: Machine Learning Classification Model For Network Based Intrusion Detection System, *Proceedings of the 11th International Conference for Internet Technology and Secured Transactions*, pp. 242–249 (2016).
- [30] Cohen, W. W.: Fast Effective Rule Induction, *Proceedings of the 12th International Conference on Machine Learning*, pp. 115–123 (1995).
- [31] Azab, A., Alazab, M. and Aiash, M.: Machine learning based Botnet Identification Traffic, *Proceedings of the IEEE Trustcom/BigDataSE/ISPA*, pp. 1788–1794 (2016).
- [32] Wang, Z., Zou, F., Pei, B., He, W., Pan, L., Mao, Z. and Li, L.: Detecting Malicious Server based On Server-to-Server Relation Graph, *Proceedings of the 1st International Conference on Data Science in Cyberspace*, pp. 698–702 (2016).
- [33] Wangyz, S., Chen, Z., Zhangyz, L., Yanx, Q., Yangz, B., Pengyz, L. and Jiayz, Z.: TrafficAV: An Effective and Explainable Detection of Mobile Malware Behavior Using Network Traffic, *Proceedings of the 24th IEEE/ACM International Symposium on Quality of Service* (2016).
- [34] Chen, Z., Han, H., Yany, Q., Yang, B., Peng, L., Zhang, L. and Li, J.: A First Look at Android Malware Traffic in First Few Minutes, *Proceedings of the IEEE Trustcom/BigDataSE/ISPA*, pp. 206–213 (2015).
- [35] Morichetta, A., Bocchi, E., Metwalley, H. and Mellia, M.: CLUE: Clustering for Mining Web URLs, *Proceedings of the 1st International Conference in Networking Science & Practice*, pp. 286–294 (2016).
- [36] He, L., Xu, C. and Luo, Y.: vTC: Machine Learning Based Traffic Classification as a Virtual Network Function, *Proceedings of the ACM International Workshop on Security in Software Defined Networks & Network Function Virtualization*, pp. 53–56 (2016).
- [37] Heuer, T., Schiering, I., Klawonn, F., Gabel, A. and Seeger, M.: Recognizing Time-Efficiently Local Botnet Infections - A Case Study, *Proceedings of the 11th International Conference on Availability, Reliability and Security*, pp. 304–311 (2016).
- [38] Watkins, L., Beck, S., Zook, J., Buczak, A. and Robinson, W. H.: Using Semi-supervised Machine Learning to Address the Big Data Problem in DNS Networks, *Proceedings of the 7th IEEE Annual Computing and Communication Workshop and Conference* (2017).
- [39] Mishra, P., Pilli, E. S., Varadharajany, V. and Tupakulay, U.: Out-VM Monitoring for Malicious Network Packet Detection in Cloud, *Proceedings of the Asia Security and Privacy (ISEASP)* (2017).
- [40] ACCS: The UNSW-NB15 dataset description, Online Available at: <https://www.unsw.adfa.edu.au/australian-centre-for-cyber-security/cybersecurity/ADFA-NB15-Datasets> (accessed 2018-01-19) (2015).
- [41] Sangster, B., O’Connor, T. J., Cook, T., Fanelli, R., Dean, E., Morrell, C. and Conti, G.: ITOC dataset, Online Available at: <https://www.usma.edu/crc/sitepages/datasets.aspx> (accessed 2018-01-19) (2009).
- [42] Tukey, J. W.: *Exploratory Data Analysis*, Addison-Wesley (1977).
- [43] Miyamoto, D. and Imura, T.: PACKTER: implementation of Internet traffic visualizer and extension for network forensics, *Journal of Computing*, Vol. 96, pp. 79–80 (2014).
- [44] Inoue, D., Eto, M., Yoshioka, K., Baba, S., Suzuki, K., Nakazato, J., Ohtaka, K. and Nakao, K.: nictcr: An Incident Analysis System toward Binding Network Monitoring with Malware Analysis, *Proceedings of WOM-BAT Workshop on Information Security Threats Data Collection and Sharing*, pp. 58–66 (2008).
- [45] Unity Technologies: Virtual Reality, Online Available at: <https://unity3d.com/jp/learn/tutorials/s/virtual-reality> (accessed 2018-01-19).