標的型攻撃の時系列データにおける1時間ごとの特徴と攻撃検知機械学習モデルの有用性検討

阿部 衛[1] 金岡 晃[1]

概要：標的型攻撃や APT（Advanced Persistent Threat）に対する検知技術の研究は, 機械学習の採用など様々なアプローチで行われている. それらの研究を評価する際に用いられるデータセットも多くの種類が存在する. 本研究では Los Alamos National Laboratory のデータセットに注目し, そのデータ特性調査と機械学習モデルへの適用の有用性の検討を行った. データ特性の調査では, 基礎調査により日ごとや週ごとの特徴がデータ量の推移から判明したため, 1 時間ごとに分割した時系列データのデータごとの関係性をクラスタリングを行い評価した. そしてそれらのデータを機械学習に適用する際の有用性を議論した.

Hourly characteristics in time-series data of Advanced Persistent Threats and investigation of the usefulness of attack detection machine-learning models

Mamoru Abe[1] Akira Kanaoka[1]

1. はじめに

特定のターゲットを狙い撃ちにした標的型攻撃や, その中でも高度な技術により長期間にわたり攻撃をする APT（Advanced Persistent Threat）は大きな脅威となっている. これら攻撃は, 明確な目的をもって特定のターゲットを標的とするサイバー攻撃であり, 攻撃者は特定の組織や団体から機密情報を窃取したり重要データを破壊することなどを目的としている. 攻撃者は主に標的を決めて電子メールを送信することを端緒に組織内の端末にマルウェアを感染させるなどの手段を利用して組織内ネットワークに侵入し, 組織内の上位権限を入手して機密情報にアクセスする. また, 機密情報へのアクセス権を利用してデータの破壊や脅迫を行うものもある.

標的型攻撃や APT の検知は学術研究として様々な研究が行われているが, 近年では機械学習を利用した研究が特に盛んである [9], [10], [11]. その際に用いられるデータセットは多種多様であり, それぞれの研究の著者らだけが持ちうるデータもあれば, オープンに共有されたデータセットもある. オープンに共有されたデータセットを用いることで提案された手法の比較が可能になることから, オープンに共有されたデータセットによる研究も盛んになっている.

本研究ではこれらオープンに共有されたデータセットの中から, Los Alamos National Laboratory（LAN-L）のデータセット [7], [8] に着目した. LAN-L のデータセットは大規模なデータセットであり, かつ多くの研究で用いられている. 一方で, LAN-L データセットを用いた研究ではすべてのデータを用いずに一部のデータに絞って学習や検証を行っているものが少なくない. その性能はその部分的なデータセットだけに特化されたものであることが否定できない.

そこで本研究では, LAN-L のデータセットを詳細に分析し, 機械学習モデルに適用した場合の有用性を議論する. まずデータセットの概観を基礎調査し, そこで得られた特徴から 1 時間ごとにデータセットを細分化し, 時間ごとのデータ特徴の推移をクラスタリングを用いて調査する. その結果, 週における同じ曜日の同じ時間の特徴推移や, 日ごとの同じ時間の特徴推移, 1 時間前と現在時間での特徴推移では週や日ごとの推移よりも 1 時間前のデータ特性に強い関連があることが示唆される結果を得た. このことから,

[1] 東邦大学
University of Toho

部分的なデータによる学習と検証は一定の精度があることを支える結果が示されたことに加え, 一方ではそのデータセットの特性が 1 時間単位という細かい単位で変動する可能性があることから, 時間経過を反映させる学習の重要性が伺えるものとなった.

2. 関連研究

サイバー攻撃の研究に用いられオープンに共有されるデータセットは多種にわたる. KDD Cup 99 データセットはそのうち最も有名なものと言ってよいだろう. 多くの研究で採用されているが, 一方でデータが古く, 現代の攻撃を反映しておらず, KDD Cup 99 データセットで高い精度を出したとしても現在の状況で高い精度が出ることを保証することはできないだろう. その他に Kyoto 2006+ Dataset や Kyoto 2016 データ [2], CMU Insider Threat Test Dataset[3], [4], DARPA Transparend Computing[5], [6], そして Los Alamos National Laboratory[7], [8] など, 近年になりそのデータセットが充実してきている. 本研究では Los Alamos National Laboratory のデータセットに着目した. データセットの詳細については後述するが, 本研究がターゲットとしている APT に関するデータセットであり, 数多くの研究に評価用として用いられていることが大きな理由である [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23]. そしてこれらの研究の多くがデータセットの多さからかデータの一部分だけを利用して提案手法を評価していることから, データ全体とその一部分による差がどれほどあるかにより精度が大きく変わる可能性があることから, 注目した.

それらのオープンに共有されたデータセットを用いた研究も多く, トップカンファレンスに採録される論文にも多く採用されている [9], [10].

3. LAN-L Dataset の予備調査

本研究では, ロスアラモス国立研究所が公開している Los Alamos National Laboratory’s(以後 LAN-L と記載)Dataset[7] を調査対象のデータとする. このデータはセキュリティの研究で利用されているデータセットの中でも KDD CUP ’99 Dataset などと比較して長期間にわたる攻撃が含まれること, 99 年にはない高度な標的型攻撃が含まれていることにより標的型攻撃の検知に適している. また, 総イベント[*1]数 1,648,275,307 イベントに対し, 攻撃イベントは 749 イベントとなっている. このデータは総イベント数に比べ攻撃イベント数が 0.1%以下になっており正常なデータとの大きな差があり, より少ない割合の攻撃検知に向けたデータとして本研究に適している. さらに, 攻撃データ内で見られるまとまった時間が偏っており, 時系列による攻撃の変化に対応する点で本研究に利用しやすいと考えた.

[*1] イベント：コンピュータ内のログ 1 行のこと

データ名	データ数 (行)	次元数	主な固有データ
auth.txt	1,051,430,459	9	認証方法, ログオンログオフ
proc.txt	426,045,096	5	プログラムの スタートと終了
flows.txt	129,977,412	9	プロトコル, パケット数
dns.txt	40,821,591	3	解決 CPU. 送信元 CPU
redteam.txt	749	4	攻撃元 CPU, 攻撃者, 攻撃対象

表 1: LAN-L Dataset の構成と各データの概要

3.1 LAN-L Dataset[7] について

このデータセットは以下の 5 つのデータで構成されている.

- ログインの情報である auth.txt
- プログラムの情報である proc.txt
- ネットワークフローの情報である flow.txt
- dns 情報である dns.txt
- 攻撃者が行った行動の情報である redteam.txt

LAN-L のデータ構成は,1 行がカンマで区切られたデータになっており, このカンマまで 1 つ分を 1 次元とするベクトルになっている. よって,LAN-L の各データは n 次元のベクトルとなっている. また, LANL-L のデータについて, それぞれが表 1 のとおりとなっている.

また, redteam.txt のイベントは auth.txt から抽出したものであるとしてあった.

3.2 データの分割

LAN-L Dataset のどのデータにおいても 1 次元目に存在している”time”の情報を利用する. この”time”の情報は 1 秒ごとに 1 加算されるデータになっており, 1 時間で 3,600,1 日で 86,400 加算される. したがって, 1 日後の同時間帯は”time”で 86,400 の差があるデータとなる. そこで, ”time”1〜3,600 を”1 日目の 1 時間目”と定義し”time”の変化が 3,600 起こるごとにデータを分割する. 最終的に 58 日目の 24 時間目までの分割となり, 1,392 個のデータに分割した. また, その際どの日, 時間帯に属しているかを示すため, 1 日目の 1 時間目を”01d01h”と定義し, 1,392 個のデータを”01d01h”〜”58d24h”と表現する.

3.3 各時間ごとのイベント数

各時間ごとのイベント数の推移を図 1 に示す. この図は, 横軸が時間の変化, 縦軸が各”time”におけるイベント数で

ある．また, それぞれの色は図に示されている通り各データの変化を表している．auth のデータに注目してみると, 例えば時間が 1〜70 の時ははイベント数の変化が顕著であるが 70〜116 大きく差がなく, そして似たような繰り返しがその後の時間にも続いていく．このように, イベント数の変化が一定で存在しており, また, 平均的にイベントが少ない部分と大きい部分から, 平日および休日が推測できる．また, イベント数は 58 日間のうちで後半になるにつれどんどん大きくなっていく傾向がある．

4. 時系列変化の調査方法

4.1 時系列データに現れる傾向の調査

時系列データに現れる傾向の調査を行うため, k-menas クラスタリングを用いて傾向分析を行った．具体的には, 1 時間ごとの分割に対してクラス数 4〜19 で分割する．さらに, クラスタごとの中心を出力する．次の 1 時間でも同じ作業を行い, 全時間帯のデータに対してクラスタリング結果と中心を求める．最後に, 1 時間のクラスタ中心を比較する．クラスタ中心に対しては”一致率”を定義する．今回, 一致率とするのは

(1) クラスタごとに中心座標の各次元の差の 2 乗を計算する
(2) 各値を比較し, 一番低い値同士で組み合わせを探す.
(3) 1 対 1 対応ができる組み合わせを数え, 総クラスタ数で割る

とする．この時, 総クラスタ数は各クラスタリングごとの総数とする．(クラスタ数 4 の時は 4, クラスタ数 19 の時は 19 といった具合) この評価手法は, 異なる時間帯でのクラスタに着目し, 中心距離が近いクラスタ同士を 1 対 1 対応させることでクラスタの性質が似ていると判断される. そして, そのクラスタが別の時間にも類似した性質と判断されることで, 連続した時間において存在していると考えることもできる. 今回の検証ではこの評価手法を用いて評価する.

4.1.1 各時間ごとのクラスター致率調査

図 2 は, 先述した一致率をクラスタリング後に求め, その一致率を各クラスタ数及び各データに応じてグラフ化したものである.hour,day,week はそれぞれ, 比較するデータがちょうどそれぞれ 1 時間 (hour) ,24 時間 (day) ,168 時間 (week) 違いの関係にあるデータに対し一致率を計算したものである. また, それぞれ横軸はクラスタ数を表しており, 右に行くほどクラスタ数が大きくなっている. 縦軸は一致率の計算を表しており, 上に行くほど一致率が高いことを示している. このグラフでは, 主に hour の一致率に関して総クラスタ数が小さいときに一致率が 1 に向かって安定している. 次に, 図 3 は, 一致率が 1 になった組み合わせの割合の変化をクラスタごとに記載したものである. これは, 総クラスタ数が多くなるほど一致率が 1 になる組み合わせが少なくなっている. このことから, クラスタ数がより少ないほうが, 安定して次の時間への関連性が発生することがわかる.

5. 考察

今回の調査では,LAN-L Ddataset について,1 時間ごとに分割して調査を行った. イベント数に関しては, その傾向から, どの時間帯かや, 平日・休日の分類, 曜日の推測, イベント数の多い週など, 時間ごとに見ることである程度の推測が行えることが分かった. また, その中でも 1 日,1 週間の関連性に着目したが, その期間が長く空くにつれ関連性が低くなっていくことが分かった. これは, 最初の予想でもある, 時間が経つごとにデータに変化があることを裏付けていると考えられる. そして, この変遷をとらえることで, より侵入検知を行いやすくすることができると考える．これにより, 時系列により着目した検知技術の提案が考えられる．

6. まとめ

特定のターゲットを狙い撃ちにした標的型攻撃や, その中でも高度な技術により長期間にわたり攻撃をする APT は大きな脅威となっており, 対策は必須である．これら攻撃の検知は学術研究として様々な研究が行われているが, 近年では機械学習を利用した研究が特に盛んである. またその際に用いられるデータセットは多種多様である．本研究では, Los Alamos National Laboratory（LAN-L）のデータセット [7], [8] に着目した．LAN-L のデータセットは大規模なデータセットであり, かつ多くの研究で用いられている．一方で, LAN-L データセットを用いた研究ではすべてのデータを用いずに一部のデータに絞って学習や検証を行っているものが少なくない．本研究では, LAN-L のデータセットを詳細に分析し, 機械学習モデルに適用した場合の有用性を議論した．1 時間ごとに LAN-L Dataset を分割し, ここに k-menas クラスタリングを使用して前述の一致率を求めた．結果として, 連続した 1 時間の場合かつ総クラスタ数が少ないときは関連性が高く安定し, より期間が空くにつれ関連性が低く, また総クラスタ数が少なくなると関連性が低くなっていくことが分かった. この結果から, 時間がたつごとにデータ全体に変化があることが考えられる．今後はこの変化に対応する検知方法を考案する必要がある．

参考文献

[1] ”FARE:Enabling Fine-grained Attack Categorization under Low-quality Labeled Data” CCS ’ 21, November 15–19, 2021, Republic of Korea
[2] ”Traffic Data from Kyoto University’s Honeypots”, http://www.takakura.com/Kyoto_data/
[3] Software Engineering Institute, Carnegie Mellon University, ”Insider Threat Test Dataset”,

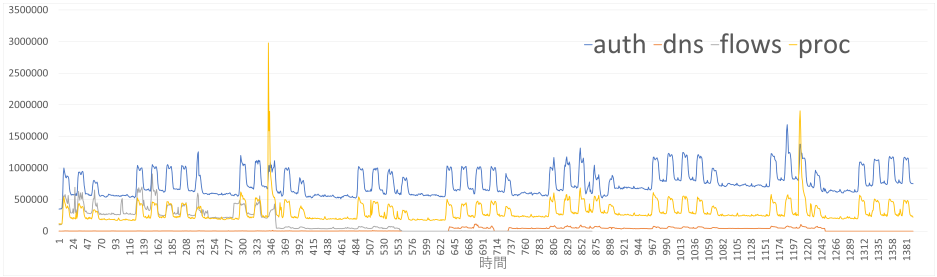
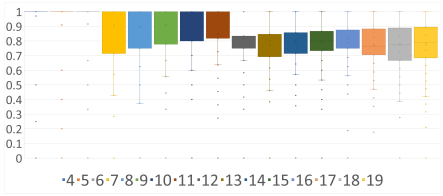
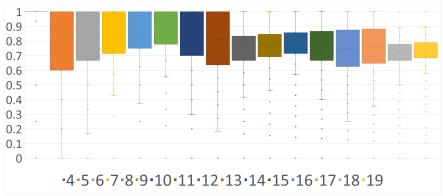


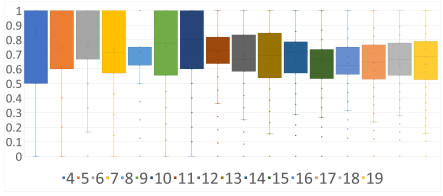
図 1: 各時間におけるイベント数の推移グラフ



(a) hour



(b) day



(c) week

図 2: 総クラス数ごとの中心座標一致率の分布

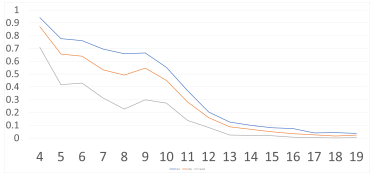


図 3: 各クラスタ数に応じた一致率が 1 になった組み合わせの割合

https://resources.sei.cmu.edu/library/asset-view.cfm?assetid=508099

[4] Nijhawan, Subin. "Bridging the gap between theory and practice with design-based action research." Studia paedagogica 22.4 (2017): 9-29.

[5] DARPA, "Transparent Computing (Archived)", https://www.darpa.mil/program/transparent-computing

[6] DARPA I2O, "Transparent Computing ENgagement 5 Data Release", https://github.com/darpa-i2o/Transparent-Computing

[7] A. D. Kent, "Comprehensive, Multi-Source Cybersecurity Events," Los Alamos National Laboratory, http://dx.doi.org/10.17021/1179829, 2015.

[8] A. D. Kent, "Cybersecurity Data Sources for Dynamic Network Research," in Dynamic Networks in Cybersecurity, 2015.

[9] Liu, Fucheng, et al. "Log2vec: A heterogeneous graph embedding based approach for detecting cyber threats within enterprise." Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security. 2019.

[10] Milajerdi, Sadegh M., et al. "Poirot: Aligning attack behavior with kernel audit records for cyber threat hunting." Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security. 2019.

[11] van Ede, T., Aghakhani, H., Spahn, N., Bortolameotti, R., Cova, M., Continella, A., van Steen, M., Peter, A., Kruegel, C. & Vigna, G. (2022, May). DeepCASE: Semi-Supervised Contextual Analysis of Security Events. In 2022 Proceedings of the IEEE Symposium on Security and Privacy (S&P). IEEE.

[12] Kumar, Pradeep, and H. Howie Huang. "Graphone: A data store for real-time analytics on evolving graphs." ACM Transactions on Storage (TOS) 15.4 (2020): 1-40.

[13] Siadati, Hossein, Bahador Saket, and Nasir Memon. "Detecting malicious logins in enterprise networks using visualization." 2016 IEEE Symposium on Visualization for Cyber Security (VizSec). IEEE, 2016.

[14] Guo, Xiaojie, Lingfei Wu, and Liang Zhao. "Deep graph translation." IEEE Transactions on Neural Networks and Learning Systems (2022).

[15] Heard, Nick, and Patrick Rubin-Delanchy. "Network-wide anomaly detection via the Dirichlet process." 2016 IEEE Conference on Intelligence and Security Informatics (ISI). IEEE, 2016.

[16] Bowman, Benjamin, et al. "Detecting lateral movement in enterprise computer networks with unsupervised graph AI." 23rd International Symposium on Research in Attacks, Intrusions and Defenses (RAID 2020). 2020.

[17] Zago, Mattia, Manuel Gil Pérez, and Gregorio Martínez Pérez. "UMUDGA: A dataset for profiling DGA-based botnet." Computers & Security 92 (2020): 101719.

[18] Smith, Shaden, et al. "Streaming tensor factorization for infinite data sources." Proceedings of the 2018 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, 2018.

[19] Bohara, Atul, et al. "An unsupervised multi-detector approach for identifying malicious lateral movement." 2017 IEEE 36th Symposium on Reliable Distributed Systems (SRDS). IEEE, 2017.

[20] Brown, Andy, et al. "Recurrent neural network attention mechanisms for interpretable system log anomaly detection." Proceedings of the First Workshop on Machine Learning for Computing Systems. 2018.

[21] Rubin-Delanchy, Patrick, et al. "A statistical interpreta-

tion of spectral embedding: the generalised random dot product graph." arXiv preprint arXiv:1709.05506 (2017).
[22] Tuor, Aaron Randall, et al. "Recurrent neural network language models for open vocabulary event-level cyber anomaly detection." Workshops at the thirty-second AAAI conference on artificial intelligence. 2018.
[23] Yuan, Yali, et al. "Ada: Adaptive deep log anomaly detector." IEEE INFOCOM 2020-IEEE Conference on Computer Communications. IEEE, 2020.