# IoT機器へのTelnet接続ログのコマンドに着目したクラスタリングによる分析

馬場 隆寛[1,a)]　馬場 謙介[2]　吉岡 克成[3]　山内 利宏[1,4]

**概要**：インターネットの発展に伴い，IoT 機器が増加している．IoT 機器は，Telnet サービスを動作させることにより外部からの遠隔操作が可能となっている場合があり，IoT 機器を対象としたマルウェアによる脅威にさらされている．本研究では，IoT 機器に Telnet で侵入された際のコマンドログを K-means 法を用いてクラスタリングし，どのような攻撃パターンがあるのか分析を行った．ハニーポットを用いて収集した Telnet ログをデータとして使用している．クラスタリングにあたり，コマンドを単語ごとに分割し，単語に対する n-gram を用いた．これは，複数のコマンドを同時に実行している場合やコマンド実行時にファイル等を指定している 場合を考慮するためである．本稿では，その分析結果を報告する．

**キーワード**：クラスタリング, Telnet ログ, 機械学習

# Analysis by Clustering Focusing on Telnet Connection Log Commands to IoT Devices

Takahiro Baba[1,a)]　Kensuke Baba[2]　Katsunari Yoshioka[3]　Toshihiro Yamauchi[1,4]

***Abstract:*** With the development of the Internet, the number of IoT devices is increasing. IoT devices may be able to be remotely controlled from the outside by operating Telnet, and are exposed to the threat of malware targeting IoT devices. In this study, we clustered the command log when Telnet was invaded into an IoT device using the K-means method, and analyzed what kind of attack patterns there are. The Telnet log collected using the honeypot is used as data. For clustering, the command was divided into words, and n-gram of the words was used. This is to consider the case where multiple commands are executed at the same time or the case where a file etc. is specified when executing a command. In this paper, we report the analysis results.

***Keywords:*** clustering, Telnet log, Machine learning

## 1. はじめに

インターネットの発展に伴い，IoT 機器が増加している．IoT 機器は，Telnet を動作させることにより，外部からの遠隔操作が可能となっている場合があり，IoT 機器を対象としたマルウェアによる脅威にさらされている．IoT 機器がマルウェアに感染すると，大規模な DoS 攻撃に使用されるなど，さらなる脅威となってしまう．例えば，2016 年 10 月「Mirai」[8] と呼ばれる IoT マルウェアがインターネットにつながっている家庭用ルーターやネットワークカ

[1] 岡山大学 学術研究院自然科学学域
Graduate School of Natural Science and Technology Okayama University
[2] 岡山大学 サイバーフィジカル情報応用研究コア
Cyber-Physical Engineering Informatics Research Core, Okayama University
[3] 横浜国立大学 先端科学高等研究院/大学院環境情報研究院
Institute of Advanced Sciences/Graduate School of Environment and Information Sciences, Yokohama National University
[4] JST さきがけ
JST PRESTO
[a)] baba-t@swlab.cs.okayama-u.ac.jp

メラ、ストレージ製品、デジタルビデオレコーダーといったIoT機器をターゲットにし、botをダウンロードさせ、攻撃者が操るサーバからの命令に従ってDDoS攻撃を実施している．そのため，IoT機器がマルウェアに感染しないように対策することは非常に重要な課題である．

IoTマルウェアの脅威を分析するため、IoT機器に対するサイバー攻撃を観察するためのハニーポットであるIoTPOT [4]が考案されている．このハニーポットを用いて収集されたデータを用いて実験を行った．ただし，収集されるデータが膨大であるため，本研究では，シェルコマンドのみを対象として，分析を行った．

本研究では，IoT機器にTelnetで侵入された際のコマンドログをK-means法 [3]を用いてクラスタリングし，分析を行う．クラスタリングにあたり，コマンドを単語ごとに分割し，単語に対するn-gramを用いている．n-gram [2]は特徴量にテキストの解析に有効である．n-gramを用いることで，複数のコマンドを実行している場合やコマンド実行時のファイル指定などを考慮するためである．

## 2. 関連研究

本章では，Telnetサービスに対しての攻撃を分析している研究に関する文献を調査している．また，n-gramを特徴量として，マルウェアの分析を行っている文献の調査を行った．Telnetログをクラスタリングした研究 [7]では，3-gramの特徴量を用いて，純粋なLinuxコマンドで実行される新しいDoSサイバー攻撃を発見している．攻撃者の地理的分布を分析する研究 [6]では，ハニーポットで収集したログを用いて，攻撃者の地理的分布を世界地図により視覚化している．Telnetを用いたサイバー攻撃の分析 [10]では，Telnetログインの際に得られるID/パスワード情報とログイン後に使用されるシェルコマンド系列を分析することで，攻撃ホストの感 染マルウェアの推定を行い，さらに攻撃対象となっているIoT機器の増加を示している．IoT機器に対するパスワードリスト攻撃の収集と分析 [9]では，Telnetサービスへのログイン試行に用いられるログインIDとパスワードの組を分析し，安易なパスワードや製品のデフォルトのパスワードが狙われやすいことを示している．

本研究がTelnetログを利用した他の研究との違いは，シェルコマンドの分析にあたり，特徴量のn-gramの適切なnの値を検証しているところにある．n-gramはあまり大きな値を使用すると計算量が多くなるため，3-gramが使用される場合が多い．しかし，マルウェアのシェルコマンドは同時にいくつかの処理を行うため，複数のコマンドを一度に実行していることが多く，3-gramでは，連続する複数のコマンドを特徴として抽出できていないのではないかと考えた．また，(1,n)-gramとの比較を行い，n-gramと(1,n)-gramからシェルコマンドの分析に適切な特徴量を割り出した．

## 3. クラスタリング手法

### 3.1 目的

IoT機器は，外部からの遠隔操作が可能となっている場合があり，これらのIoT機器を対象としたマルウェアによる脅威にさらされている．そのため，マルウェアによりどのようなコマンドが行われているのかを分析することは重要である．そこで，本研究では，シェルコマンドに対して，クラスタリングを行いどのようなコマンドが出現するのか分析を行った．また，出現しているコマンドから考察を行った．

### 3.2 データセット

本研究では，IoTPOTを用いて収集したTelnetログをデータとして使用している．データは，2020年1月1日から2020年1月31日に収集した1,548,855件のログのうちランダムに抽出した100万件を使用した．

### 3.3 分析手法

本研究では，特徴量として，n-gramを用い，クラスタリング手法として，K-means法を用いている．

特徴量として，n-gramを用いているのは，連続する複数のコマンドを特徴として使用するためである．

K-means法は非階層クラスタリング手法であり，あらかじめ決められたクラスタに分けることにより，すべてのデータ同士の距離を計算する必要がないため，階層クラスタリング手法よりも計算量が少なく済むという利点がある．そのため，ビッグデータの分析に向いている．今回使用しているデータは，非常に膨大なため，非階層クラスタリング手法であるK-means法を使用することとした．

### 3.4 予備実験

本研究では，特徴量として，n-gramを用いているため，nをいくつにするべきか定量的に評価する必要がある．また，K-means法を用いるにあたり，クラスタ数を決定する必要がある．そのため，予備実験を行った．

まず，n-gram，(1,n)-gram，それぞれの特徴量を用いるにあたり，nをどの数値にするのが最も適切か検証を行った．(1,n)-gramは，1-gramからn-gramまでの和集合である.2-gramの例を**表 1**に，(1,2)-gramの例を**表 2**に示す．n-gram,(1,n)-gramにおいて，nの値が1から9までの範囲でクラスタ数が4の時の残差平方和を求めたものが**図 1**と**図 2**になる．その結果，5-gramの時が最も残差平方和が低く適切な値と言える．

また，本研究では，K-means法を用いているが，クラスタ数を手動で設定する必要がある．そこで，5-gramを特徴量とした場合の各クラスタ数の残差平方和を求め，**図 3**の

ようにエルボー図を作成した．これにより，クラスタ数 4 以下でほとんど変化が見られないことから，クラスタ数を 4 と設定している．これらの残差平方和を求めるにあたり，クラスタの中心は K-means++ [1] を用いて計算している．

**表 1** 2-gram の例

| Telnet コマンド | /mnt/.ptmx && cd /mnt/ |
|---|---|
| 2-gram | “mnt .ptmx”, “.ptmx &&”, “&& cd”, “cd mnt” |

**表 2** (1,2)-gram の例

| Telnet コマンド | /mnt/.ptmx && cd /mnt/ |
|---|---|
| (1,2)-gram | “mnt”, “.ptmx”, “&&”, “cd”, “mnt .ptmx”, “.ptmx &&”, “&& cd”, “cd mnt” |

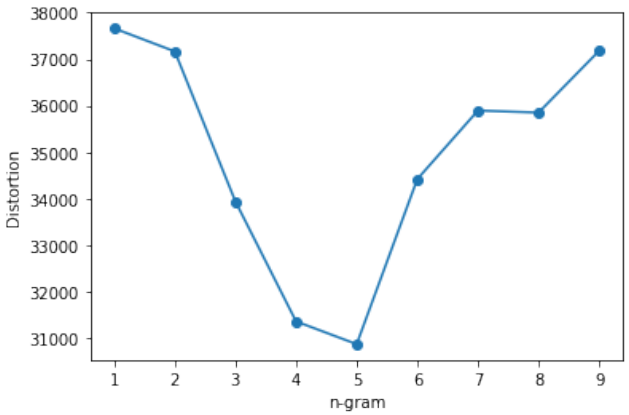


**図 1** n-gram の残差平方和

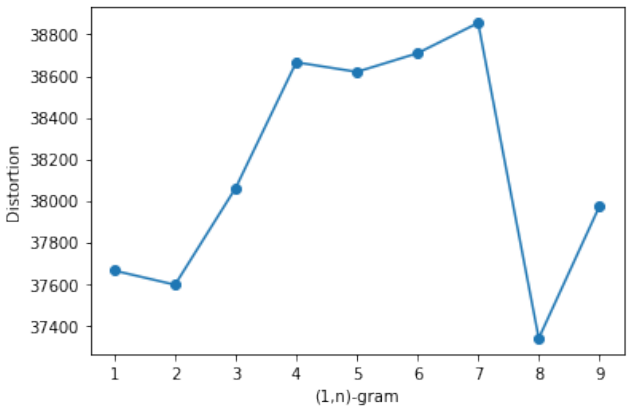


**図 2** (1,n)-gram の残差平方和

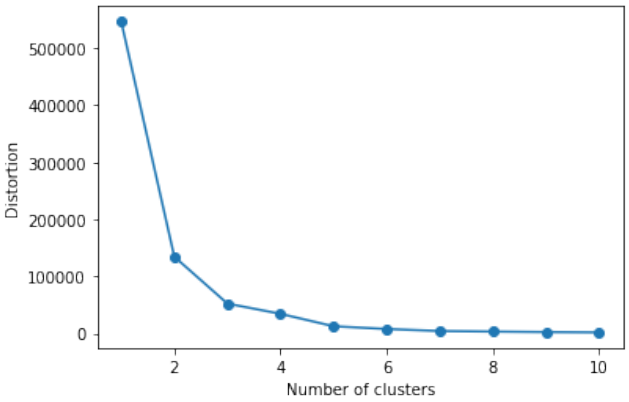


**図 3** 各クラスタ数の残差平方和

### 3.5 K-means 法を用いたクラスタリング

本研究では，K-means 法を用いて，Telnet ログのコマンドのクラスタリングを行った．今回使用したデータは 100 万件とデータ数が多いため，あらかじめ決められたクラスタ数に要素を分けていく K-means 法を用いることで，計算量を少なくすることができる．クラスタリングの特徴量として，5-gram の tf-idf を用いた．まず，コマンドをスペース，バックスラッシュ，改行，スラッシュで分割し，分割した単語ごとの 5-gram を作成した．それぞれに対して，tf-idf を求めることで，特徴量としている．その際に，全 telnet ログの中で 10%〜90%に出現するもののみを使用している．また，コマンドが絶対パスや相対パスで指定してある場合を考慮し，パスが指定されている部分を除去することで，同じコマンドであれば，同じ特徴量となるようにしている．

## 4. 実験結果

実験の結果，各クラスタが効果的に分割されていることを確認するため，特異値分解による次元削減を行い**図 4** のように可視化を行った．4 クラスタに分割できている．**表 3** では，クラスタ毎の出現頻度の高い単語上位 10 個を抽出した．

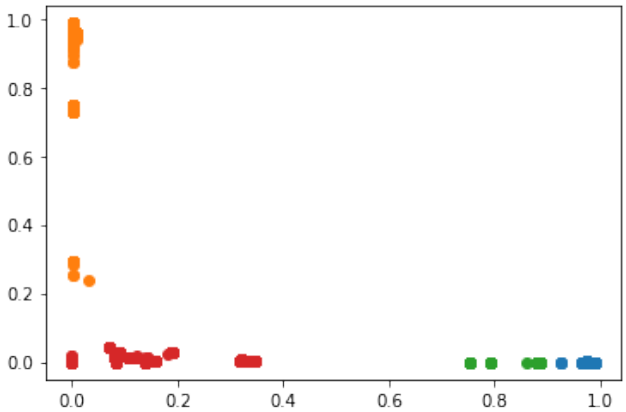


**図 4** 特異値分解によるクラスタの可視化

表 3　各クラスタの出現頻度の高い特徴量

| | |
|---|---|
| クラスタ 1 | .ptmx cd netslink shm .ptmx<br>.ptmx cd run .ptmx cd<br>.ptmx cd run tmp .ptmx<br>.ptmx cd shm .ptmx cd<br>.ptmx cd tmp .ptmx cd<br>busybox cat busybox while read<br>busybox wget; busybox tftp; busybox<br>busybox while read i; do<br>cat busybox while read i;<br>cd .ptmx cd busybox rm |
| クラスタ 2 | echo x6b x61 x6d x69<br>.nippon busybox echo x6b x61<br>busybox rm .nippon busybox echo<br>rm .nippon busybox echo x6b<br>.t; rm .sh; rm .human<br>rm .t; rm .sh; rm<br>.human cd busybox cp echo<br>.human rm .t; rm .sh;<br>.nippon busybox BOT rm .t;<br>.sh; rm .human cd busybox |
| クラスタ 3 | $i; done busybox busybox UNSTABLE<br>.ptmx cd .ptmx cd busybox<br>.ptmx cd .ptmx cd run<br>.ptmx cd busybox rm .xxx666<br>.ptmx cd netslink .ptmx cd<br>.ptmx cd netslink shm .ptmx<br>.ptmx cd run .ptmx cd<br>.ptmx cd run tmp .ptmx<br>.ptmx cd shm .ptmx cd<br>.ptmx cd tmp .ptmx cd<br>.xxx666 s2m busybox cp busybox |
| クラスタ 4 | cat busybox while read i;<br>do echo $i; done busybox<br>echo $i; done busybox busybox<br>enable system linuxshell shell sh<br>i; do echo $i; done<br>read i; do echo $i;<br>while read i; do echo<br>$i; done busybox busybox AKASHICY<br>$i; done busybox busybox Exploit1<br>$i; done busybox busybox KowaiSlump |

## 5. 考察

K-means 法によるクラスタリングの結果，4 クラスタに分割することができた．クラスタ 1 では，「ptmx」というコマンドが多く出現している．これは，疑似ターミナルを作成するために使用される．このコマンドは，Mirai と Bashlite [5] にみられる特徴である．ほかに「while read」というコマンドも出現しており，入力を受け取り処理を行おうとしていることが分かる．クラスタ 2 では，「nippon」や「human」というコマンドが多く出現している．クラスタ 3 ではクラスタ 1 と同様に「ptmx」というコマンドが出現しているため，Mirai や Bashlite の特徴が表れている．また，クラスタ 3 の特徴として，「UNSTABLE」というコマンドが出現している．これは，安定版ではないパッケージやライブラリを探し脆弱性を利用するためではないかと推測される．クラスタ 4 では，「AKASHICY」，「Exploit1」，「KowaiSlump」というコマンドが出現していることが分かる．

## 6. おわりに

本研究では，K-means 法を用いてクラスタリングを行うにあたり，適切な特徴量の検証を行った．その結果,n-gram と (1,n)-gram の中でも 5-gram が適切であると結論付けた．この特徴量を用いて，シェルコマンドを 4 クラスタに分割し，どのようなコマンドがあるか検証した．

今後の課題として，本研究では，2020 年 1 月に収集されたデータを使用しているが，データを 2020 年全体へ拡張し，シェルコマンドの分析を行う必要がある．また，各クラスタに出現したコマンドから具体的にどのような攻撃を行っているのかを検証する必要がある．最近では，シェルスクリプトファイルの形で攻撃ペイロードが送られ、実行される場合も多いため，これらの分析をシェルコマンドから行うことを検討している．

**謝辞**　本研究の一部は，JST さきがけ JPMJPR1938，および JSPS 科研費 JP19H04111 の助成を受けたものです．

### 参考文献

[1] Arthur, D. and Vassilvitskii, S.: k-means++: The advantages of careful seeding, Technical report, Stanford (2006).

[2] Cavnar, W. B., Trenkle, J. M. et al.: N-gram-based text categorization, *Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval*, Vol. 161175, Citeseer (1994).

[3] Hartigan, J. A. and Wong, M. A.: Algorithm AS 136: A k-means clustering algorithm, *Journal of the royal statistical society. series c (applied statistics)*, Vol. 28, No. 1, pp. 100–108 (1979).

[4] Pa, Y. M. P., Suzuki, S., Yoshioka, K., Matsumoto, T., Kasama, T. and Rossow, C.: IoTPOT: A Novel Honeypot for Revealing Current IoT Threats, *Journal of Information Processing*, Vol. 24, No. 3, pp. 522–533 (online), DOI: 10.2197/ipsjjip.24.522 (2016).

[5] Tom, S.: BASHLITE Family Of Malware Infects 1 Million IoT Devices, https://threatpost.com/bashlite-family-of-malware-infects-1-million-iot-devices/120230/.

[6] Visoottiviseth, V., Jaralrungroj, U., Phoomrungraungsuk, E. and Kultanon, P.: Distributed honeypot log management and visualization of attacker geographical distribution, *2011 Eighth International Joint Conference on Computer Science and Software Engineering (JC-*

*SSE)*, IEEE, pp. 23–28 (2011).
[7] Wu, C.-J., Huang, S.-Y., Yoshioka, K. and Matsumoto, T.: IoT Malware Analysis and New Pattern Discovery Through Sequence Analysis Using Meta-Feature Information, *IEICE Transactions on Communications*, Vol. E103.B, No. 1, pp. 32–42 (online), DOI: 10.1587/transcom.2019CPP0009 (2020).
[8] 高橋睦美：史上最悪規模の DDoS 攻撃「Mirai」まん延、なぜ？, https://www.itmedia.co.jp/news/articles/1802/21/news034.html.
[9] 清松天樹，池部　実，吉田和幸：ハニーポットを用いた IoT 機器に対するパスワードリスト攻撃の収集と分析，情報処理学会研究報告，Vol. 2018-IOT-42, No. 8, pp. 1–6 (2018).
[10] 中山　颯，鉄　穎，楊　笛，田宮和樹，吉岡克成，松本　勉：IoT 機器への Telnet を用いたサイバー攻撃の分析，情報処理学会論文誌，Vol. 58, No. 9, pp. 1399–1409 (2017).