ログ解析のための Doc2Vec を用いた 区間特徴量の２次元表現の一考察

磯野　怜[1]　中野 心太[2]　関谷 信吾[2]　折田 彰[3]
岸本 頼紀[4]　早稲田 篤志[4]　花田 真樹[4]

概要：ログ解析において一定区間のイベント ID の特徴量を可視化するシステムが提案されている．この提案では，視認性を考慮して特徴量を１次元圧縮するため異常箇所の特徴が見え難くなる問題がある．別の研究では２次元圧縮の場合はある程度の特徴が表現できることが提案されている．そこで，２次元圧縮された特徴量を１次元で表現する手法について検討する．論文では２次元圧縮された特徴量に対して，中心からの距離，重心からの距離，２軸それぞれの値を時系列に表現した場合の表現方法について提案し，実際の例によるそれぞれの効果について報告する．

キーワード：セキュリティ，フォレンジック，ログ解析

A Consideration on Two-Dimensional Representation of Interval Features using Doc2Vec for Log Analysis

Rei Isono[1]　Shinta Nakano[2]　Shingo Sekiya[2]　Akira Orita[3]
Yorinori Kishimoto[4]　Atsushi Waseda[4]　Masaki Hanada[4]

Abstract:. A system has been proposed for visualizing the features of event IDs in sections of log analysis. In this proposal, features are compressed into one dimension for visibility reasons. This makes it difficult to see the features of abnormal logs. Another study has suggested that two-dimensional compression can express features to a certain extent. Therefore, we consider a method for expressing two-dimensionally compressed features in one dimension. In this paper, we propose a method for expressing the distance from the center, the distance from the center of gravity, and the values of each of the two axes as a time series for two-dimensionally compressed features, and report the effectiveness of each method using actual examples.

Keywords: Security,Forensic,Log Analysis

1. 序論

デジタルフォレンジックでは，攻撃を受けた端末に残されたログを解析することで，攻撃の内容や侵入経路を調査する．この際，対象とされるログにはアプリケーションログやセキュリティログなどが含まれるが，これらのログは平常時から継続的に記録されているため膨大であり，その中から攻撃の痕跡を発見するには多大な時間を要する[1].

企業などで使用される PC は平常時に定常業務を行っているため，記録されるログには規則性が見られる場合が多い．そのため，ログを一定期間ごとに分割した際，特定の区間でのみ出現するログには，攻撃の痕跡が含まれている可能性が高いと考えられる．これに基づき，イベントログを一定期間で区切り，各区間のイベント ID に対して Doc2Vec で得られる特徴量を可視化する手法が提案されている．これは得られた特徴量ベクトルを次元圧縮し，結果を時系列でプロットした折れ線グラフとして描画することで，低頻度で出現するログを含む区間を特定し可視化する手法である[2]．しかし，この手法では折れ線グラフに変換する際に次元を 1 次元に圧縮する必要があり，情報の多くが失われてしまう．その結果，攻撃の痕跡が十分に可視化できないという問題が指摘されている．一方で，同様に Doc2Vec によるベクトル化を行った後，2 次元に圧縮して散布図としてマッピングした場合，異常な区間を外れ値として識別できることが報告されている[3]．しかしながら，分割区間が多い場合には散布図に点が密集してしまい，視認性が低下するため，調査支援には不向きである．本研究では，2 次元に圧縮した特徴量に対して，中心からの距離，重心からの距離，xy 各軸の値を時系列で折れ線グラフに反映させる手法を提案する．この手法により，より精度の高い可視化が可能となり，調査支援の一助になると考えられる．

1 東京情報大学大学院 総合情報学研究科
Graduate School of Informatics. Tokyo University of Information Sciences.
2 株式会社日立システムズ セキュリティ技術 R&T センタ
Hitachi Systems, Ltd. Security Technology R&T Center.
3 株式会社日立システムズ セキュリティリスクマネジメント本部
Hitachi Systems, Ltd. Security Risk Management Division.
4 東京情報大学 総合情報学部
Faculty of Informatics, Tokyo University of Information Sciences
*g24001ir@edu.tuis.ac.jp
【 論文原稿：上記*の文字書式「隠し文字」 】

2. 先行研究

本節では，先行研究について概要を示す．

先行研究では，分割されたログファイルのベクトル化において，ログの情報の中でも特に日付とイベント ID に着目している．時系列順に並ぶログを一定期間で分割し，各区間ごとにイベント ID の並びを生成する．この際，イベント ID を単語，イベント ID の並びを文章として解釈することで，Doc2Vec を適用可能となり，時系列順に並んだ文書の比較を特徴量ベクトルの値で実現している．

次に，可視化手法について示す．Doc2Vec によって得られるベクトル値は高次元であるため，折れ線グラフや散布図などの可視化を行うには，PCA（Principal Component Analysis）などを用いた次元圧縮が必要である．この次元圧縮により，ログに含まれる特徴を視覚的に表現することが可能となる．

3. 考え方

3.1 特徴量

先行研究の結果から，特徴量として 2 次元データは 1 次元データよりも精度が高いと考えられる．このため，本研究でも先行研究と同様に 2 次元データを対象とする．具体的には，イベントログを一定期間で区切り，各区間に対して Doc2Vec を適用した後，次元圧縮を行い 2 次元データを生成する．この特徴量を対象とする．

しかし，これらを可視化するにあたり日時を考慮するならば 1 次元の特徴量の方が望ましい．そこで，２次元データから得られる値として，特徴量そのものである中心からの距離，集合全体における重心からの距離，２次元マッピングされた場合の x 軸 y 軸それぞれの 0 軸からの距離が考えられる．これらの値であれば，横軸を日付時刻，縦軸をこれらの値として２次元で表現できる．

3.2 折れ線グラフでの可視化

先行研究では，2 次元マッピングを散布図で可視化する手法が提案されている．しかし，散布図では区間ごとの外れ値がどこに存在するのかが直感的に把握しづらいという課題がある．そこで本研究では，折れ線グラフを用いた時系列可視化を提案する．折れ線グラフの横軸を時系列とすることで，折れ線の値に変化が生じている箇所を特定することで，攻撃が発生した可能性の高い日時を直感的に把握できると考えられる．

縦軸に用いる値については，2 次元データを 1 次元データに変換する必要があるため，次の 3 種類の指標を適用する．

・中心からの距離
・重心からの距離
・重心からの x の距離，重心からの y の距離

これらの指標に対して攻撃ログを適用することで，それぞれの手法の効果について確認する．

4. 攻撃ログの作成と環境

特徴量の２次元表現を行う上で使用する攻撃ログについて概要を示す．ログ作成はホスト OS を Windiows11 上の VirtualBox7.2.0 で仮想環境として構築した．

4.1 攻撃の種類

4.1.1 PassTheTicket の環境

AD 環境: 構築済みの Active Directory 環境
被攻撃端末:
・ドメインサーバ
OS: Windows Server 2019
・ドメインに所属するクライアント端末
OS: Windows 10
攻撃端末:
OS: Kali Linux
使用ツール: Metasploit，Mimikatz

4.1.2 PassTheHash の環境

AD 環境: 構築済みの Active Directory 環境
被攻撃端末:
・ドメインサーバ
OS: Windows Server 2019
・ドメインに所属するクライアント端末（2 台）
OS: Windows 10
攻撃端末:
・OS: Kali Linux
使用ツール: Metasploit，PsExec，Mimikatz

4.2 攻撃手順

4.2.1 PassTheTicket 攻撃の手順

1．Kali Linux 上で，攻撃用ペイロードを内包した実行ファイル（bayload.exe）および認証情報抽出ツール（mimikatz.exe）を準備する．

- bayload.exe は，Metasploit Framework の msfvenom コマンドを使用して生成される攻撃用ファイルであり，ターゲットシステム上で Meterpreter セッションを確立するために使用される．

2．Windows 端末でセキュリティ設定を変更し，攻撃環境を整備する．

3．Metasploit を使用してセッションを確立し，権限昇格を実施する．

4．Mimikatz でデバッグ特権を有効化し，Kerberos チケットを取得する．

5．Kerberos チケットを使用して PassTheTicket 攻撃を実行する．

6．ドメインリソースへのアクセスを確認する．

4.2.2 PassTheHash 攻撃の手順

1．ネットワーク内の端末を nmap でスキャンし，ターゲッ

トを探索する（nmblookup を使用して NetBIOS 名やドメイン情報を確認）.
2. Metasploit を使用してターゲット端末 A に接続する.
3. Mimikatz を端末 A に転送し，実行して NTLM ハッシュを取得する.
4. 取得したハッシュを用いて端末 A に再侵入し，PassTheHash 攻撃を実施する.
5.（試行結果）端末 B へのアクセスは成功しなかった.

作成したログの中からアプリケーションログ，セキュリティログ，システムログの中から攻撃ログの攻撃実行時間のみをトリミングし，長時間起動していた別端末から取得したログを通常時のログとして両者を統合し，攻撃痕跡が含まれているログとして作成した.

5. 適用例

4 で示した手順に基づいて作成したログファイルに対し，3 種類の手法を用いて折れ線グラフを作成した．これにより，攻撃を実行した痕跡のある日時において，縦軸の値がどのように変化するかを確認した.

各ログファイルに対する 3 手法の結果を次に示す．xy の距離の図では赤が x,青が y の値を示す.

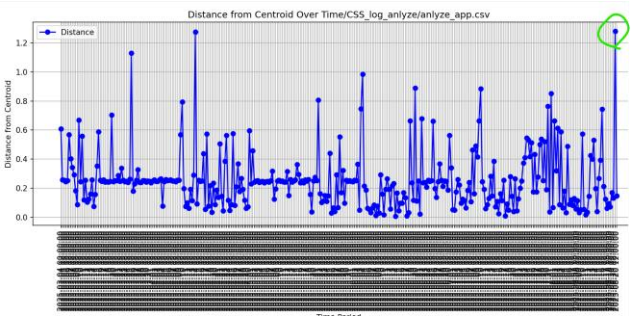

図 1 アプリケーションログ/重心からの距離/HashTheTicket

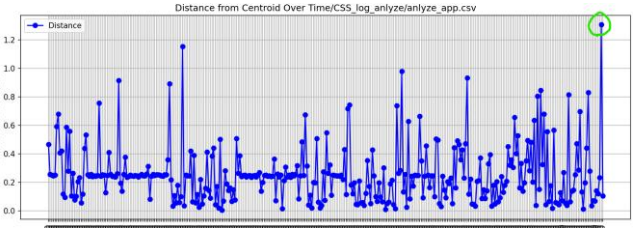

図 2 アプリケーションログ/中心からの距離/HashTheTicket

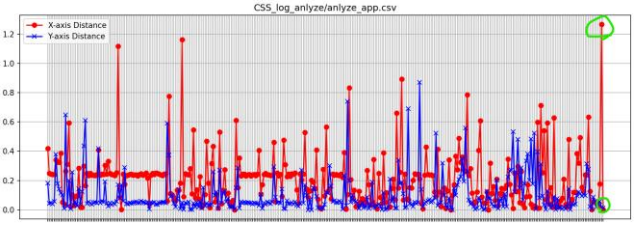

図 3 アプリケーションログ/x,y 軸の距離/HashTheTicket

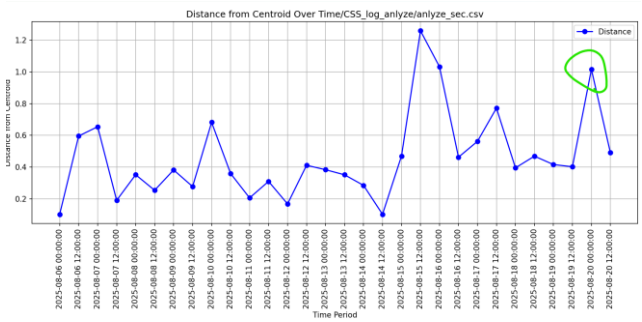

図 4 セキュリティログ/重心からの距離/HashTheTicket

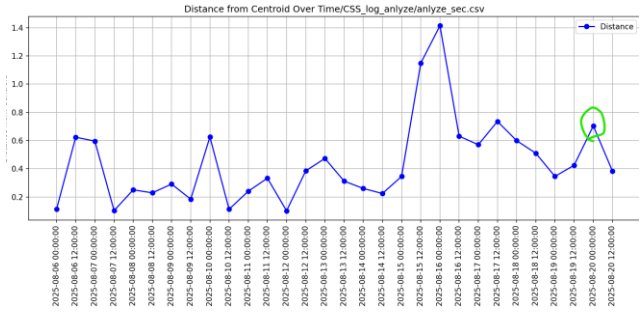

図 5 セキュリティログ/中心からの距離/HashTheTicket

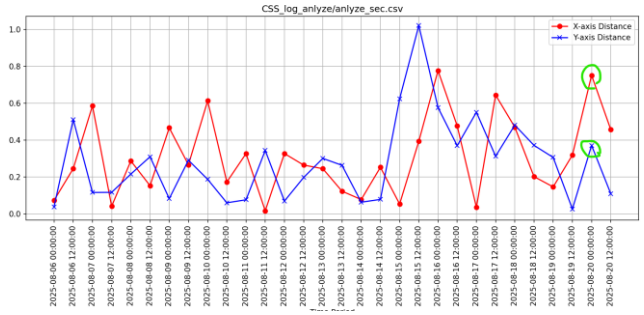

図 6 セキュリティログ/x,y 軸の距離/HashTheTicket

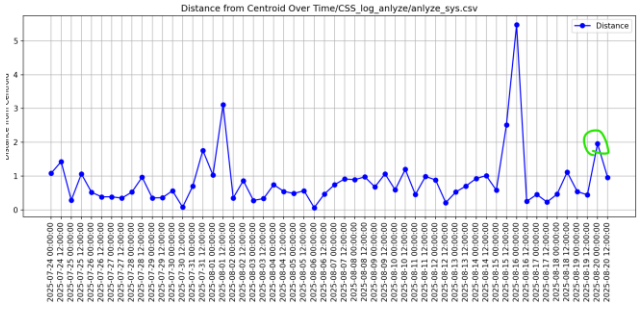

図 7 システムログ/重心からの距離/HashTheTicket

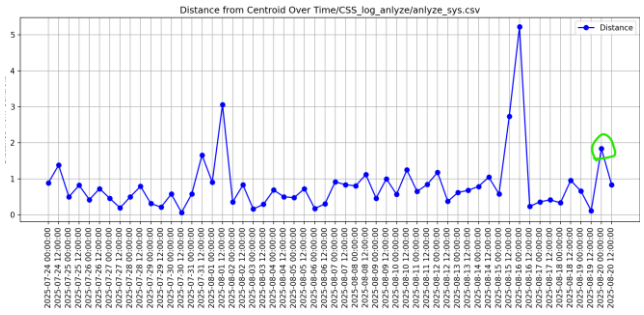

図 8 システムログ/中心からの距離/HashTheTicket

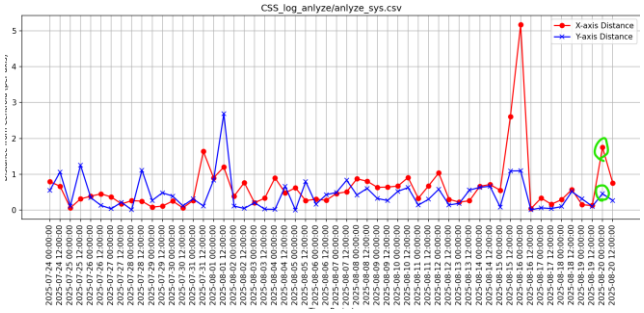

図 9 システムログ/x,y 軸の距離/HashTheTicket

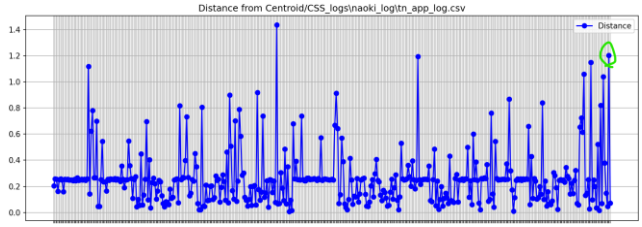

図 10 アプリケーションログ/重心からの距離/HashThePass

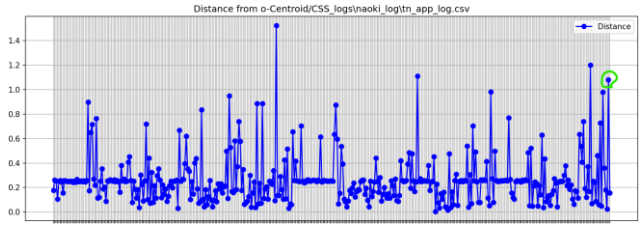

図 11 アプリケーションログ/中心からの距離/HashThePass

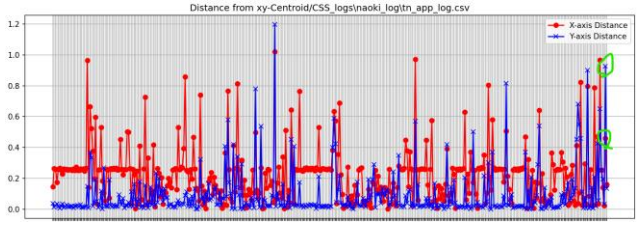

図 12 アプリケーションログ/x,y 軸の距離/HashThePass

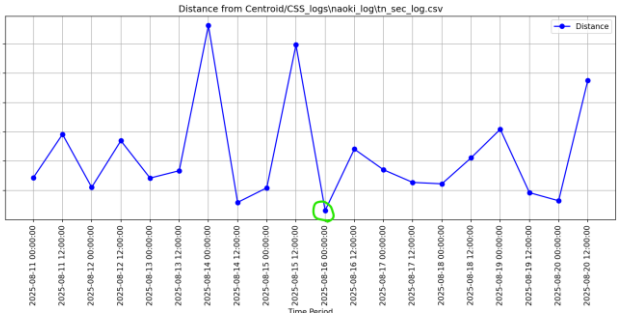

図 13 セキュリティログ/重心からの距離/HashThePass

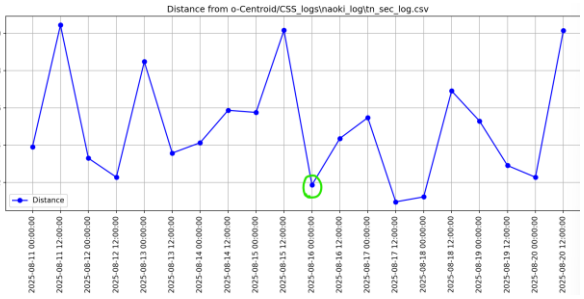

図 14 セキュリティログ/中心からの距離/HashThePass

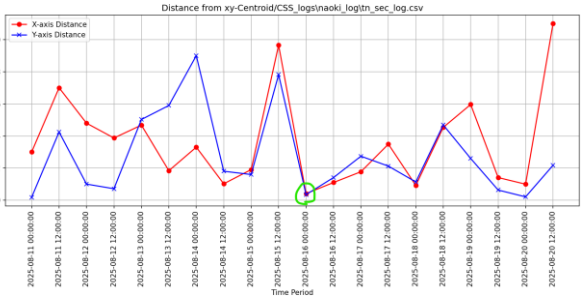

図 15 セキュリティログ/x,y 軸の距離/HashThePass

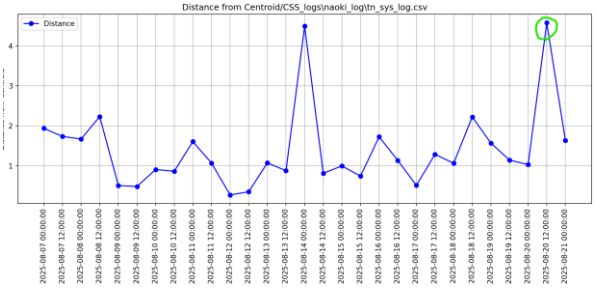

図 16 システムログ/重心からの距離/HashThePass

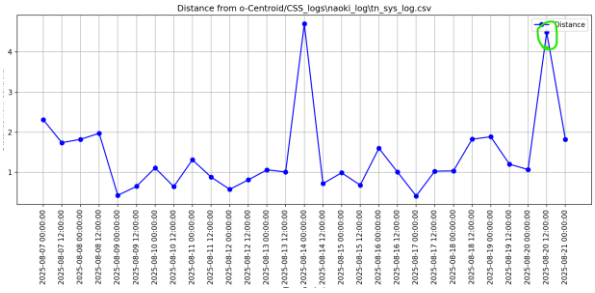

図 17 システムログ/中心からの距離/HashThePass

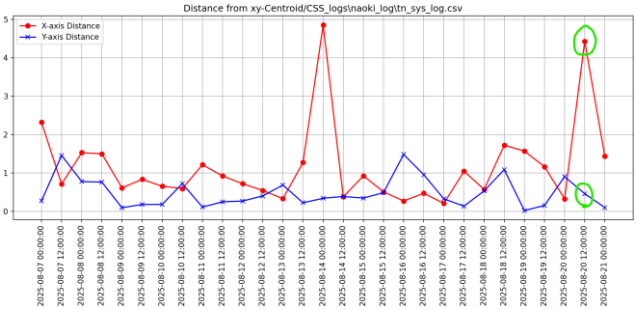

図 18 システムログ/x,y 軸の距離/HashThePass

PassTheTicket 攻撃では，アプリケーションログとセキュリティログにおいて，他の区間と異なる値を示すケースが多く見られた．一方で，システムログは他の区間との差異が小さく，目立った変化は確認されなかった．

PassTheHash 攻撃では，アプリケーションログとシステムログで値の変化が確認された．一方，セキュリティログは重心付近に位置しており，大きな変化は見られなかった．

6. 考察

3 種類の手法について，中心からの距離と，重心からの距離および重心からの xy の距離の比較では，図 4,5,6 より重心をもとにした 2 手法の方が，攻撃の区間が他より顕著に可視化できていることが確認できた．これより，2 次元の特徴量を折れ線グラフで可視化する際は，重心からの距離をもとにした方が有効だと考えられる．また，図 3,9,18 より，重心からの x の距離と重心からの y の距離をプロットした場合はどちらか片方の値だけ大きく特徴が出ることが多いことが確認できる．ここから，1 次元表現する際には，どちらか一方ではなく，同時に x の距離と y の距離をプロットする方が効果的であると考えられる．

7. 結論

本論文では，デジタルフォレンジックのために，Windows ログのイベント ID に着目した Doc2vec を用いた特徴量の可視化手法として，2 次元マッピング時の中心からの距離，重心からの距離，重心からの xy の値を折れ線グラフで可視化する手法について提案した．

本手法を pass the hash および pass the ticket 攻撃の例に適用した結果，本例においては，中心よりも重心からの距離の方が，攻撃痕跡が顕著に現れた．また，重心からの x 軸 y 軸それぞれ距離では，どちらか片方に顕著な特徴が見られる傾向が確認できた．

8. 参考文献

[1] 企業における情報システムのログ管理に関する実態調査，https://warp.da.ndl.go.jp/info:ndljp/pid/11440710/www.ipa.go.jp/files/000052999.pdf

[2] 磯野怜，機械学習を用いた異常ログ可視化のための誤検知された正常ログ対策の検討，コンピュータセキュリティシンポジウム 2024 論文,p. 1884-1887，発行日 2024-10-15

[3] 岸本頼紀，デジタルフォレンジックのための Doc2vec を用いたインシデント日時可視化システムの検討，東京情報大学研究論集,巻 28，号 1, p. 29-38，発行日 2024-09-30