L-006

Web アクセス履歴に着目した VAE による異常検出
Anomaly detection using VAE focused on web access log

川上 颯太 *　青木 茂樹 *　宮本 貴朗 *
Souta Kawakami　Shigeki Aoki　Takao Miyamoto

1　はじめに

近年, マルウェアによるサイバー攻撃が増加しており, サイバー攻撃に対する社会の関心が高まっている．従来多くの組織ではサイバー攻撃への対策として, 組織外ネットワークと組織内ネットワークの境界に Firewall や IDS などの侵入防御装置を設置し, 組織内への侵入を試みる不審な通信の遮断などを行ってきた．しかしながら, 現在問題となっている標的型攻撃等は従来の対策だけでは検出が難しい．そこで侵入後の不正な動きを検出するシステムが必要となっている．

マルウェア感染後の不正な通信を検知する手法として文献 [1] が挙げられる．文献 [1] ではマルウェア感染後のプロキシログから, マルウェアによる悪性サイトへのアクセスを特定する手法を提案している．この手法では, 既知の悪性サイトのドメインと一般サイトのドメインの特徴を機械学習手法で学習し, 識別している．しかし, 一般的な機械学習手法をネットワークの異常検知に応用する場合, 教師データとして必要な多岐にわたる異常データの収集が難しいことが課題となっている．そこで, 教師つきデータを必要としない教師なし学習による異常検知手法が研究されている．文献 [2] では, 組織内の端末の通信にはユーザ固有の規則性が現れると仮定し, 確率モデルによりユーザの通信パターンを学習して異常検知する手法を提案している．この手法では, プロキシログから特徴量を抽出し, 通信パターンの遷移の確率を求め, 普段現れないような遷移を異常として検出している．この手法では, セッション単位で異常を検出しているが, 短い間隔で頻繁にアクセスが発生している場合にセッションを分割できず, 正しく異常検出できない可能性がある．

本稿では, 教師なし深層学習モデルの一つである VAE(Variational AutoEncoder) を用いてプロキシログから抽出した端末ごとの Web アクセス履歴に注目して異常通信を検知する手法を提案する．文献 [2] ではセッション単位で異常を検出していたのに対し, 本稿では Web アクセス履歴を一定の行数で分割し異常を検出することで, 文献 [2] の問題を解決している．実験では, 大阪府立大学の職員用端末のログと CTU データセット [3] を用いて有効性を確認した．

* 大阪公立大学大学院情報学研究科 Graduate School of Informatics, Osaka Metropolitan University

2　提案手法

まず, プロキシログから端末ごとの日常の Web アクセス履歴を抽出し, 特徴ベクトルに変換する．変換した特徴ベクトルを, VAE で学習する．次に学習とは別の期間の Web アクセス履歴を学習時と同様に特徴ベクトルに変換し, テスト用データとする．その後, 学習済みの VAE によりテスト用データの正常通信と異常通信を識別する．

2.1　Web アクセス履歴の特徴抽出

Web アクセス履歴から得られる特徴量を VAE へ入力する特徴ベクトルへと変換する．まず端末ごとの Web アクセス履歴から 6 次元の特徴量を抽出する．使用する特徴量は送信データサイズ, 受信データサイズ, ドメインの文字列の全長, ドメインに含まれる数字の数, ドメインに含まれるピリオドの数, ステータスコードである．抽出した値を 0~255 に正規化する．この処理を Web アクセス履歴 8 行分に対して行い, 8×8 のベクトルへと変換する．ここで横方向に足りない 2 要素分については 0 で補完する．

2.2　VAE による学習

学習期間の Web アクセス履歴の特徴ベクトルを, 端末ごとに VAE で学習する．VAE の特徴として潜在変数の表現に確率分布を用いることが挙げられる．VAE は Encoder と Decoder の 2 つで構成される．Encoder では入力データ $\boldsymbol{x}$ の平均と分散を求め, 潜在変数 $\boldsymbol{z}$ をその確率分布からサンプリングする．Decoder では潜在変数 $\boldsymbol{z}$ から元の次元のデータ $\boldsymbol{x}'$ へと再構成する．その際, 入力データ $\boldsymbol{x}$ と再構成データ $\boldsymbol{x}'$ が同じになるように VAE を学習する．

2.3　異常検出

2.2 節で学習した VAE を用いて異常を検出する．学習された VAE は学習期間の Web アクセスの確率分布に従って入力データを再構成する．したがって, 学習後の VAE にテスト用データ $\boldsymbol{x}$ を入力した際に, $\boldsymbol{x}$ が学習期間に存在する Web アクセスの特徴に類似する場合, $\boldsymbol{x}$ と $\boldsymbol{x}'$ の再構成誤差は小さくなる．一方で, 学習期間に存在する Web アクセスとは異なる場合は, 学習期間の Web アクセスの確率分布には従わないため, $\boldsymbol{x}$ に似たベクトル $\boldsymbol{x}'$ を再構成することが出来ず, $\boldsymbol{x}$ と $\boldsymbol{x}'$ の再構成誤差は大きくなる．そこで, 式 (2) を用いて入力したベクトル $\boldsymbol{x}$ と再構成したベクトル $\boldsymbol{x}'$ の再構成誤差を異常度 $A(\boldsymbol{x})$ として算出する．$A(\boldsymbol{x})$ が閾値未満ならば正常, 閾値以上ならば異常とし, 正常な通信と異常な通信を識

別する．

$$A(\boldsymbol{x}) = \sum |\boldsymbol{x} - \boldsymbol{x}'| \quad (1)$$

3 実験

3.1 実験環境

実験では，正常データのみで構成された大阪府立大学のプロキシログから抽出した職員用端末 10 台の Web アクセス履歴を 6 対 4 の比率で学習用とテスト用に分割し，テスト用 Web アクセス履歴のそれぞれに CTU が公開しているデータセット [3] から抽出した異常ログを時系列情報を維持したまま挿入した．CTU データセットはチェコ工科大学が研究用に収集，作成したデータセットである．本実験ではマルウェアによるボットネットトラフィックをキャプチャし，端末の挙動を観測したログ形式のラベル付きデータ (CTU-Malware-Capture-Botnet-13/2013-10-10_capture-win14.weblogng.labeled) を使用した．また，Web アクセス履歴から生成したベクトルの内，4 行以上が異常ログで構成されているものを異常ベクトルとしてラベル付けを行った．ラベル付け後の実験データの内訳を表 1 の 1 列目 ~4 列目に示す．

3.2 実験結果，考察

実験結果を表 1 の 5 列目に示す．実験の結果，AUC 値は 0.85~0.99 となり，大半の端末で 0.95 以上を記録し，本手法の有効性を確認できた．しかし，端末 A では AUC 値が 0.85 と他の端末と比較して低い検知精度となっている．図 1, 2 に端末 A と端末 B の異常度のヒストグラムを示す．図中，異常ラベルの異常度の分布が確認できる拡大図を示している．図 1, 2 の異常ラベルの異常度の分布を比較すると，端末 A の異常度が全体として低くなっていることが確認できる．また，端末 A の正常ベクトルに対する異常度のばらつきも端末 B と比較して大きいことを確認できる．端末 A と端末 B のアクセス先サイトの種類数を比較したところ，端末 A が 4,028 種類に対し，端末 B が 1,778 種類であり，学習データはほぼ同数だった．このことから，学習時に多様なサイトへの少量のアクセスが発生するような場合では，安定した学習が困難であった可能性が考えられる．そのため，正常ベクトルと異常ベクトルの異常度に違いが現れにくかったと推察される．

4 まとめ

本稿では，Web アクセス履歴に着目し，VAE を用いて異常な Web アクセスを検知する手法を提案した．実験では大阪府立大学の職員用端末の Web アクセス履歴と CTU の公開データセットを用いて本手法の有効性を確認した．今後の課題として，検知精度向上のために新たな特徴量を検討することなどが挙げられる．

表 1. 実験結果の AUC 値

端末	学習データ	正常ラベル	異常ラベル	AUC
A	85,225	21,419	136	0.85
B	85,823	21,585	119	0.94
C	77,359	19459	129	0.93
D	105,059	26,391	122	0.98
E	43,312	10,986	91	0.91
F	58,745	14,772	163	0.95
G	74,721	18,810	119	0.94
H	69,310	17,438	138	0.99
I	55,203	13,921	128	0.95
J	100,403	25,229	120	0.96

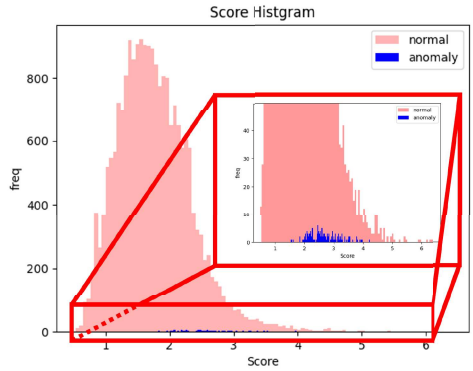

図 1. 端末 A の異常度のヒストグラム

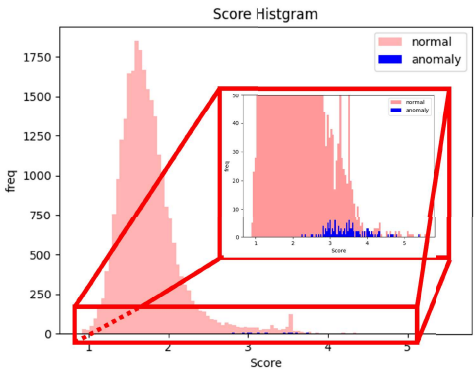

図 2. 端末 B の異常度のヒストグラム

参考文献

[1] 松岡 裕和, 佐々木 良一：インシデント後におけるログ解析での機械学習を用いた悪性ドメインの抽出手法の提案, マルチメディア, 分散協調とモバイルシンポジウム 2019 論文集,2019,478-486 (2019-06-26),

[2] 名倉 悠, 青木 茂樹, 宮本 貴朗：組織内端末の通信の規則性に着目したプロキシログの異常検知, コンピュータセキュリティシンポジウム 2022 論文集,1301-1308 (2022-10-17)

[3] Stratosphere,(2015),StratosphereLaboratoryDatasets, from https://www.stratosphereips.org/datasets-overview(参照 2021-10-19)