Vol.2016-MPS-108 No.8
Vol.2016-BIO-46 No.8
2016/7/4

記号の順序によるウェブアプリケーション攻撃の特徴抽出

松田 健[1,a)]　大谷 康介[2]

概要：インターネット上では大量のデータのやりとりが行われており，それに紛れるかのようにサイバー攻撃が日常的に行われている．本稿では，サイバー攻撃の観測を支援することを目的とした攻撃特徴抽出のための数理モデルを提案する．

Feature Extraction of Web Application Attack Based on Order of Symbols

Matsuda Takeshi[1,a)]　Ootani Kousuke[2]

1. はじめに

近年の急速なICT技術の利活用に伴い，様々なデータがインターネット上で扱われるようになり，利便性が向上している一方で，サイバー攻撃の脅威もますます増加していることを忘れてはならない [1]．インターネット上で扱うデータが膨大になればなるほど，サイバー攻撃の痕跡をログデータから見つけ出すことはますます困難となる．このようなサイバー攻撃を例からの学習のような統計的推測に基づいて検出する手法に関する研究は，様々な方法を用いて提案されている [2][3]．また，脆弱性を作り込まないようにアプリケーションを開発する効果的な手法も考案され，実用化されているにもかかわらず，完全に攻撃からアプリケーションやデータを防御することは難しいのが実情である．このような中，統計的推測に基づいたサイバー攻撃の自動検出方法の確率は，上に述べた通り，大量のログデータを解析するためにも重要な技術であるといえる．そのためには，攻撃検出の精度を向上させることは言うまでもなく重要であるが，このような自動検出法を用いたとしても，実際に攻撃があったかどうか確認するのは専門知識を持った人間であるため，このような作業をサポートする技術として攻撃検出技術を開発することが実用上ももっとも重要であるといえる．本研究では，ウェブアプリケーションのデーターベースを攻撃対象とするSQLインジェクション攻撃の特徴を単項式として表現し，そこから攻撃特徴を抽出する確率モデルを提案する．提案モデルでは，攻撃特徴とそれ以外の特徴（以下，正常特徴）を単項式の指数部分を用いて表現するため，データに基づいて，明らかな攻撃特徴や正常特徴には重み付けを行わず，攻撃やそれ以外のものに共通して含まれる特徴量の関係性を重みとして数値化できることであり，特徴空間がどのようになっているか代数的に表現できる特徴をもっている．

[1] 長崎県立大学 1-1-1 Manabino, Nagayo-cho, Nishi-Sonogi-gun, Nagasaki 851–2195, JAPAN
[2] 合同会社 binary lab, mail : kousuke3346@gmail.com
[a)] tmatsuda@sun.ac.jp

2. 記号順序に基づくSQLインジェクション特徴抽出モデル

まず，使用する記号について定義していく．i, j を非負整数とし，以下の記号を用いて攻撃と正常の特徴づけを行う．検査対象の文字列に含まれる記号を調べ，記号 s_i の次に記号 s_j が観測されたとき，初めから用意されている1に p_{ij} を掛け算していくことで単項式を生成する．例えば，文字列

10';DROP members –

はSQLインジェクション攻撃の一例であり，対応する単項式は $p_{02}p_{11}p_{18}p_{21}$ となる．また，文字列

(^ − ^);

はSQLインジェクション攻撃とならない一例であり，対応する単項式は $p_{36}p_{62}$ となる．

番号	記号
s_0	シングルクォート
s_1	スペース
s_2	セミコロン
s_3	左丸括弧
s_4	右丸括弧
s_5	カンマ
s_6	マイナス
s_7	ピリオド
s_8	コメント (/* #)
s_9	クエスチョンマーク

以上の準備の元で，L_A 個の攻撃サンプルと L_N 個の正常サンプルを収集することができたとき，そのサンプルにおける攻撃特徴量を単項式の指数部分 a_{ij} に以下のようにして与える．

$$F(A) = \prod_{l_a=1}^{L_A} \left(\prod_{ij} p_{ij}^{a_{ij}} \right) \prod_{l_n=1}^{L_N} \left(\prod_{ij} p_{ij}^{a_{ij}} \right)$$

本提案モデルは本質的にはベルヌーイ分布を考えることと同値であり，目標は $F(A)$ を最大にする p_{ij} を求めることである．

3. 攻撃検出モデルによる特徴抽出

典型的な攻撃サンプル $L_A = 5$ 個, 攻撃サンプルに含まれる記号を用いて構成される顔文字を含む文字列 $L_N = 5$ 個を用意して計算する．サンプル数を少なくしたのは，提案モデルの特徴を紹介するためである．$f(a_i), f(n_j)$ はそれぞれ，攻撃，正常サンプルから生成される単項式である．

$$\begin{aligned}
f(a_1) &= p_{01}p_{11}^5 \\
f(a_2) &= p_{01}p_{11}^4p_{13}p_{15}^4p_{17}p_{33}p_{35}p_{44}p_{45}p_{51}^5p_{54}p_{55} \\
f(a_3) &= p_{00}p_{01}p_{03}p_{11}^2p_{15}p_{30}p_{33}p_{35}^2p_{53}^2p_{55}^4 \\
f(a_4) &= p_{02}p_{11}p_{18}p_{21} \\
f(a_5) &= p_{08} \\
f(n_1) &= p_{36}p_{46}p_{64} \\
f(n_2) &= p_{36}p_{62} \\
f(n_3) &= p_{04}p_{30}p_{34}p_{43}
\end{aligned}$$

以上の単項式より，

$$F(A) = \prod_{i=1}^{5} f(a_i) \prod_{j=1}^{3} (1 - f(n_j)) \qquad (1)$$

を最大にするパラメータ空間を計算する．$F(A)$ において，$p_{00}, p_{01}, \cdots, p_{99}$ の全てにおいて偏微分を施してできる，連立多項式の解空間が求めるパラメータ空間となる．$F(A)$ の $\prod_{i=1}^{5} f(a_i)$ の部分と $\prod_{j=1}^{5}(1 - f(n_j))$ の部分に現れる変数について考察すると，$\prod_{i=1}^{5} f(a_i)$ のみに含まれる変数は，このサンプルの場合は $p_{00}, p_{01}, p_{03}, p_{05}, p_{08}, p_{11}, p_{13}, p_{15}, p_{17}, p_{33}, p_{35}, p_{44}, p_{45}, p_{51}, p_{53}, p_{54}, p_{55}$ であり，これらの変数で偏微分を施した場合，$\frac{\partial}{\partial p_{ij}} F(A) = 0$ は，$p_{ij} \neq 0$ とすると

$$\prod_{j=1}^{5} (1 - f(n_j)) = 0$$

となる．同様に，$\prod_{j=1}^{5}(1 - f(n_j))$ の部分のみに含まれる変数は $p_{04}, p_{34}, p_{36}, p_{43}, p_{46}, p_{62}, p_{64}$ であるから，$\frac{\partial}{\partial p_{ij}} F(A) = 0$ は，$p_{ij} \neq 0$ とすると

$$\frac{\partial}{\partial p_{ij}} \prod_{j=1}^{5} (1 - f(n_j)) = 0$$

となる．さらに，$\prod_{i=1}^{5} f(a_i), \prod_{j=1}^{5}(1 - f(n_j))$ に共通して現れる変数はこのサンプルの場合は，p_{30} であり，$\frac{\partial}{\partial p_{30}} F(A) = 0$ は

$$\prod_{j=1}^{5} (1 - f(n_j)) + p_{30} \frac{\partial}{\partial p_{30}} \prod_{j=1}^{5} (1 - f(n_j)) = 0$$

となる．これらの連立多項式の解空間は，グレブナー基底を計算することで，以下の多項式の実代数多様体 $V(f_1, f_2, f_3, f_4, f_5, f_6)$ となることがわかる．

$$\begin{aligned}
f_1 &: p_{36}^2p_{46}p_{62}p_{64} - (p_{46}p_{64} + p_{62})p_{36} + 1 \\
f_2 &: p_{36}p_{46}p_{62}p_{64} + (p_{04}p_{30}p_{34}p_{43} - 1)(p_{46}p_{64} + p_{62}) \\
f_3 &: p_{36}p_{64}(1 - p_{04}p_{30}p_{34}p_{43})(p_{36}p_{62} - 1) \\
f_4 &: p_{36}(1 - p_{04}p_{30}p_{34}p_{43})(p_{36}p_{46}p_{64} - 1) \\
f_5 &: p_{36}p_{46}(1 - p_{04}p_{30}p_{34}p_{43})(p_{62} - 1) \\
f_6 &: 1 - 2p_{04}p_{30}p_{34}p_{43}
\end{aligned}$$

実代数多様体 $V(f_1, f_2, f_3, f_4, f_5, f_6)$ はパラメータ空間を一意に定めることはしないが，パラメータ同市の関係を表現しており，使用したサンプルデータの特徴空間を表すものと考えることができる．

4. おわりに

本研究では，SQL インジェクション攻撃の特徴を記号の順序から抽出するモデルを提案した．この特徴を用いて攻撃検出を行うには，機械学習のアルゴリズムなどを適用すれば可能であるため，今後は実験を行って提案モデルの精度について検証を行う．

参考文献

[1] OWASP : https://www.owasp.org/index.php/Main_Page (2016.6.02).
[2] Yi Wang.; Zhoujun Li.; *SQL Injection Detection with Composite Kernel in Support Vector Machine*, International Journal of Security & Its Applications, Vol. 6 Issue 2, pp. 191-196 (2012).
[3] 松田健; **非線形な潜在曲線モデルを応用した *SQL* インジェクション攻撃の特徴抽出**，情報処理学会論文誌数理モデル化と応用（TOM）, Vol. 8-3 pp. 1-9 (2015).