5J-07

機械学習を用いたクロスサイトスクリプティング攻撃の検知実験と特徴抽出法に関する考察

梅原 章宏[1] 松田 健[2] 園田 道夫[1] 水野 信也[3] 趙 晋輝[4]
中央大学大学院理工学研究科情報工学専攻[1]
長崎県立大学情報システム学部情報セキュリティ学科[2]
静岡理工科大学情報学部コンピュータシステム学科[3]
中央大学理工学部情報工学科[4]

1. はじめに

近年のインターネット普及に伴い,個人情報の窃取や成りすましといった被害が増加している.そのような被害を引き起こす攻撃の一つにクロスサイトスクリプティング(XSS)攻撃[1]がある.従来の対策としては構文解析によるフィルタ[2]が提案されているが, XSS 攻撃に用いられる攻撃入力は正常に用いられるスクリプトとの区別が難しく,機械的な攻撃検知が容易ではないと考えられる.

私たちは先行研究[3][4][5]として入力に対する特徴抽出法を提案し,機械学習アルゴリズムを用いた検知手法について考察を行ってきた.本文では先行研究での特徴抽出法と実験結果をまとめ,学習方式による検知結果の比較や,予測ラベルを用いた場合における SCW の検知結果の差異について考察する.

2. XSS 攻撃[1]

XSS 攻撃は HTML の入力値を参照する部分などに脆弱性が存在する場合に,不正なスクリプト文を入力することで不正な動作を誘発させる攻撃である.

既存の対策の一つとして,あらかじめ登録されたシグネチャに基づいて入力に対する処理を行うブラックリスト・ホワイトリスト方式や HTML の特殊文字を別の記号に置換するエスケープ処理がある.しかし,シグネチャに基づく方式では事前にシグネチャを登録する必要があるため,登録されていない未知の攻撃に対しては対応が難しいと考えられる.また,エスケープ処理は漏れなく行うことが可能であれば根本的な対策として非常に有効であるが,漏れが存在した場合脆弱性が生じる恐れがある.

3. 機械学習アルゴリズム

3.1 SVM

SVM はバッチ式の教師あり学習を行う機械学習アルゴリズムの一つであり,データ分類は以下の式で定義される[6].

$$y(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}) + b = \sum_{n=1}^{N} a_n t_n k(\mathbf{x}, \mathbf{x}_n) + b$$

ここで $\mathbf{x}$ は入力ベクトル, $\mathbf{w}$ は重みベクトル, ϕ は特徴空間変換関数, a_n はラグランジュ乗数, b はバイアスパラメータである.またt_nは目標値, $k(\mathbf{x}, \mathbf{x}_n)$はカーネル関数である.

SVM は分類境界に対して最も近い入力(サポートベクトル)との距離をあらわすマージンを最大化することで,汎化誤差が最小になるような分類境界を求める.また,ペナルティ項を導入することで分類にどの程度の汎化性を与えるか制御を行う.

3.2 SCW[7]

SCW はオンライン式の教師あり学習を行う機械学習アルゴリズムの一つであり,データ分類は以下の式で定義される.

$$\hat{y}(x) = \mathrm{sgn}(\boldsymbol{\mu}_{t-1}\mathbf{x}_t)$$
$$\text{if } \boldsymbol{\mu}\mathbf{x} \geq 0 : \hat{y}(\mathbf{x}) = 1$$
$$\text{else} \quad : \hat{y}(\mathbf{x}) = -1$$

ここで $\mathbf{x}$ は入力ベクトル, $\hat{y}(\mathbf{x})$ は予測ラベル, $\boldsymbol{\mu}$ は重みをあらわす平均ベクトルである.

$\boldsymbol{\mu}, \boldsymbol{\Sigma}$の更新式は以下の式であらわす.

損失関数: $l^{\phi} = \max(0, \phi\sqrt{\mathbf{x}_t^{\mathrm{T}}\Sigma\mathbf{x}_t} - y_t\boldsymbol{\mu}\mathbf{x}_t)$

if $l^{\phi} > 0$:

$$(\boldsymbol{\mu}_{t+1}, \boldsymbol{\Sigma}_{t+1}) = \arg\min_{\mu,\Sigma} D_{\mathrm{KL}}(\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \| \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)) + C l^{\phi}(\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}); (\mathbf{x}_t, y_t))$$

損失関数において y はデータ $\mathbf{x}$ がどのクラスに所属するかの正解ラベル, $\boldsymbol{\Sigma}$ は共分散行列, $\phi = \Phi^{-1}(\eta)$である.Φ は正規分布の累積密度関数, η は誤差を許容するパラメータをあらわす.更新式において D_{KL} はカルバック情報量, $\mathcal{N}$ は $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ の多変量正規分布, C は重みの更新を制御するパラメータである.

4. 実験

本章ではそれぞれの先行研究[3][4][5]における実験と結果についてまとめる.なお SVM の実行プログラムには Python の機械学習ライブラリである scikit-learn0.15.2 の関数 SVC を用い,カーネル関数は rbf カーネルとした. SCW の実行プログラムは論文[7]を参考に Python で実装した.

4.1 特徴文字の変更による結果の差異

先行研究[3]では特徴文字(特徴次元)の変更による結果の変化を調査した.入力として用いたデータは収集した攻撃と正常の入力からそれぞれ 500 個を抽出し,要素数 1000 個のデータセットを生成した.

特徴ベクトルは,正常に多い記号(x_0)と攻撃に多い記号(x_1)を考慮して各要素に当てはまる記号の頻度

An Analysis of Detection Experiment and Feature Extraction of Cross-Site Scripting Attack with machine learning
1 Akihiro Umehara, Graduate School of Chuo University
2 Takeshi Matsuda, University of Nagasaki
3 Michio Sonoda, Graduate School of Chuo University
4 Shinnya Mizuno, Shizuoka Institute of Science and Techonology
5 Jinhui Chao, Chuo University

を数えた 5 次元ベクトルと,標準 ASCII 記号全ての頻度を単純に数え上げた 128 次元ベクトルの 2 種類のベクトルを生成した. 5 次元ベクトルの例を表 1 に, それぞれの特徴ベクトルを用いた際の SVM, SCW による検知実験結果を表 2 に示す.

4.2 学習方式の違いによる結果の差異

先行研究[4]では学習方式の異なる SVM と SCW において,学習データと異なる形式のテストデータを与えた際の挙動について調査を行った.

使用したデータは URL 形式の攻撃入力と正常入力,スクリプト形式の攻撃入力, TeX の数式形式の正常入力の 4 パターンを収集・生成した.これらの入力から 200 個ずつ抽出し,攻撃と正常を組み合わせることで 4 つのデータセットを生成した.特徴抽出法は 4.1 節における 128 次元のものを使用した.

実験は学習データセット 4 パターンに対しテストデータセット 4 パターンの合計 16 パターンの実験を行った.その中から正常 URL-攻撃 URL で学習した際にテストデータとして正常 URL-攻撃 URL(1)と正常 URL-攻撃スクリプト(2)の 2 パターンを与えた場合の結果を表 3 に示す.

4.3 特徴抽出法の変更による結果の差異

先行研究[5]では 4.1 節から特徴抽出法を変更し,結果に対する考察を行った.データは 4.2 節の 4 種類の入力からからそれぞれ 250 個ずつを用いた. 0~600 個目(a)のデータは 4 つの入力集合から 150 個ずつ,601~800 個目(b)のデータは 2 種類の攻撃入力をそれぞれ 100 個ずつ,801~1000 個目(c)までのデータは 2 種類の正常入力をそれぞれ 100 個ずつ, 合計 1000 個のデータをテストデータとし,学習データにはテストデータの 0~200 個目までを用いた.

また特徴ベクトルの生成において,4.1 節の方法で得た 128 次元ベクトルの各要素に定数値を足し,その後 1 に正規化を行った.結果を表 4 に示す.

5.考察

先行研究[3]の特徴抽出法については,特徴次元を増加させることで良好な結果を得ることができた. 一方で制御文字などほとんど出現しない記号も特徴として見ているため,特徴次元をさらに減らすことは可能であると考えられる.また,先行研究[3]では SVM 優位な結果となったが先行研究[5]で用いた特徴抽出法では SCW が優位になる傾向がみられた. 以上のことから特徴抽出法と機械学習の組み合わせは検知結果に大きな影響を及ぼすことが予測される.

先行研究[4]ではオンライン学習が可能な SCW ではテストデータの傾向が学習データと異なる際にもある程度の対応が可能であることが考えられた.ただし SCW はデータにより結果がばらつく傾向があるため,対策をとる必要がある.また本来の SCW アルゴリズムは正解ラベルを知っている前提での再学習を行うため, SVM とフェアな比較をするためには予測ラベルを用いた学習に変更することが必要となる.先行研究[5]の SCW 検知実験では正解ラベルと予測ラベルそれぞれを用いた学習を行った.その際自身の実験環境では検知結果に大差があらわれることはなかったが,影響を検討する必要がある.

表 1　5 次元特徴ベクトル(x_0, x_1, x_2, x_3, x_4)

x_0	/	.	_	?	%
x_1	<	>	=	;	SP
x_2	文字				
x_3	数字				
x_4	その他記号・制御文字				

表 2　先行研究[3]の実験結果 (単位:%)

	SVM		SCW	
特徴次元	5-dim	128-dim	5-dim	128-dim
正解率	97.0	98.6	62.9	89.3
攻撃精度	96.5	97.8	62.1	91.7
正常精度	97.6	99.4	63.8	86.5
攻撃 F 値	97.0	98.6	62.1	89.5
正常 F 値	97.0	98.6	59.7	89.0

表 3　先行研究[4]の実験結果 (単位:%)

	SVM		SCW	
パターン	(1)	(2)	(1)	(2)
正解率	97.5	82.5	85.3	85.0
攻撃精度	96.5	97.2	95.6	91.2
正常精度	98.4	75.0	75.1	79.0
攻撃 F 値	97.5	79.4	86.5	85.9
正常 F 値	97.5	84.8	83.5	83.8

表 4　先行研究[5]の実験結果 (単位:%)

	SVM			SCW		
フェーズ	(a)	(b)	(c)	(a)	(b)	(c)
正解率	82.3	98.5	63.5	94.3	94.5	93.0
攻撃精度	100.0	98.5	---	94.6	97.5	---
正常精度	64.5	---	63.5	94.0	---	93.0
攻撃 F 値	85.0	99.2	---	94.3	97.2	---
正常 F 値	78.3	---	77.7	94.4	---	97.2

参考文献

[1]"安全なウェブサイトの作り方 改訂第 7 版", 独立行政法人情報処理推進機構セキュリティセンター, "https://www.ipa.go.jp/security/vuln/websecurity.html", (最終閲覧日:2017/1/7)
[2]"IE8 Security Part Ⅳ: The XSS Filter - IEBlog - SiteHome - MSDN Blogs", "http://blogs.msdn.com/b/ie/archive/2008/07/02/ie8-security-part-ⅳ-the-xss-filter.aspx, (最終閲覧日:2015/1/7)
[3] 梅原 章宏, 松田 健, 園田 道夫, 水野 信也, 趙 晋輝, "機械学習を用いたクロスサイトスクリプティング(XSS)攻撃の検知に関する考察", IPSJ第71回CSEC研究会研究報告
[4] 梅原章宏, 松田健, 園田道夫, 水野信也, 趙晋輝, "クロスサイトスクリプティング(XSS) 攻撃の検知におけるバッチ学習とオンライン学習の比較実験", IPSJ第78 回全国大会
[5] 梅原章宏, 松田健, 園田道夫, 水野信也, 趙晋輝, "機械学習によるクロスサイトスクリプティング攻撃検知と誤検知要因の分析", IPSJ第111回MPS研究会研究報告
[6]Christopher M.Bishop, "PATTERN RECOGNITION AND MACHINE LEARNING", Springer(2006)
[7]Jialei Wang, Peilin Zhao, Steven C.H. Hoi, "Exact Soft Confidence-Weighted Learning", ICML(2012)