7ZC-01

ハニーポットで収集したサイバー攻撃数の時系列分析による予測

西田 圭佑 †　　蓑原 隆 ‡
拓殖大学工学部情報工学科 †　　拓殖大学工学部情報工学科 ‡

1　はじめに

近年，情報システムから不正に情報を入手したり，運用を妨害したりするサイバー攻撃が増加している．このようなサイバー攻撃に対抗するためにはシステムへの攻撃数をあらかじめ予測し，事前に対策をとることが重要である．サイバー攻撃の予測には，離散モデル，連続モデル，機械学習などの様々な方法が提案されている [1] が，大域的な攻撃の強さを予測する方法として，攻撃数の変化について時系列分析を行う方法の有益さが指摘されている [2][3].

本研究では，大域的な攻撃の強さと局所的な攻撃の関係に着目し，故意に攻撃を受けやすいように設定した多数のハニーポット [4] を分散配置して攻撃情報を収集している DSheild プロジェクト [5] のデータを利用して，ローカルに設置したハニーポットのデータに対して時系列分散を行うことで，局所的な攻撃の強さの予測を行うことを目的とする．

2　攻撃予測の概要

本研究の攻撃予測システムの構成を図 1に示す．ハニーポットとして DShield[5] をベースに Raspberry Pi に実装した擬似サーバ (以下，ローカルハニーポットと呼ぶ) を大学の DMZ に設置しインターネット上に公開する．擬似サーバに対するアクセスを攻撃とみなして，DShield プロジェクトのサイトである Internet Storm Center(以下 ISC) に登録するとともに，ローカルのデータベースに登録する．また，DShield プロジェクトに参加している世界中のハニーポットから収集された 1 日ごとの攻撃総数を ISC から取得し，ローカルのデータベースに登録する．

攻撃予測は，ローカルハニーポットで収集した 1 日ごとの攻撃数 Y_t に対して，ISC から取得した 1 日ごとの攻撃数 X_t を説明変数として時系列分析を行う．このとき両者のデータの変動を抑えるために対数変換を前処理として行い $y_t = \log Y_t$ と

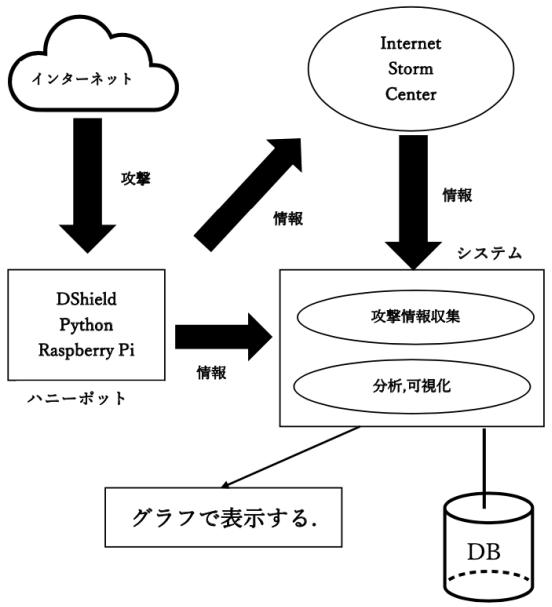

図 1　攻撃予測システムの構成

$x_t = \log X_t$ を用いる．時系列分析のモデルには，式 (1) に示す ARIMAX(p, d, q) を使用する．

$$\Phi(L)(1-L)^d y_t = \beta x_t + \Theta(L) w_t \qquad (1)$$

ここで　L : ラグ演算子
　　w_t : ホワイトノイズ
$$\Phi(L) = 1 - \phi_1 L - \cdots - \phi_p L^p$$
$$\Theta(L) = 1 + \theta_1 L + \cdots + \theta_q L^q$$

モデルのパラメータの決定は，実測値と計算値の残差が小さくなるように行うが，モデルの次数 p, q によってパラメータ数が増加するので，パラメータ数を評価に加えた AIC を使用し，AIC が最小になるようにパラメータを決める．

パラメータを決定したモデルに過去の実測値を代入することで 1 日先の攻撃数の予測ができる．さらに予測値をモデルに加えることを繰り返すことで数日先までの予測を行う．ただし，ARIMAX モデルを使用するためには，ISC から取得した説明変数のデータ x_t として未来の値が必要になる．そこで，x_t について ARIMA(p, d, q) モデルを使った時系列分析を行い，説明変数の予測値を算出してか

Cyber Attacks Forecasting by using Time Series Analysis of Honeypot Data
† Keisuke NISHIDA, Takushoku University
‡ Takashi MINOHARA, Takushoku University

ら，ARIMAX モデルによる攻撃予測を行う．

3 攻撃予測の実験

ローカルハニーポットによるデータの取得は 2021 年 10 月 4 日から開始している．2021 年 10 月 20 日から 2022 年 7 月 20 日まで，各月の 20 日を起点として 55 日分の実測攻撃数でモデルのパラメータを決定し，その後の 1 週間の攻撃数を予測する実験を行った．

予測された攻撃数と実際の攻撃数の MAPE を，説明変数の有無で 3 日先の予測まで比較した結果を表 1に示す．この表から ISC で観測された攻撃数を説明変数として使用した場合の方が平均して予測精度が高いことがわかる．

表 1 説明変数の有無による予測精度の比較

モデル	1 日後	2 日後	3 日後
説明変数あり	50.6 %	80.9 %	35.8 %
説明変数なし	55.3 %	85.4 %	37.6 %

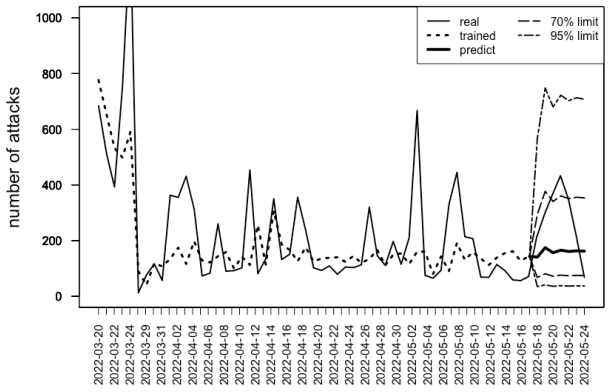

図 2 攻撃数の変化の予測 (説明変数あり)

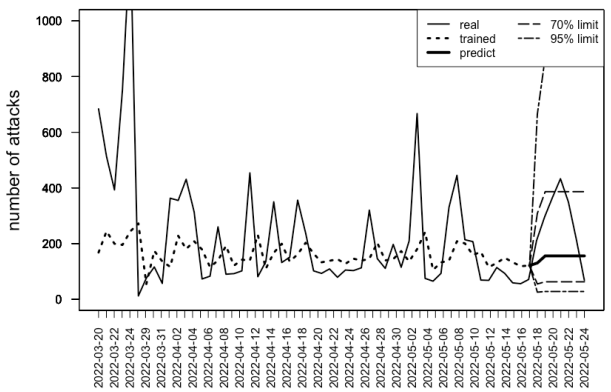

図 3 攻撃数の変化の予測 (説明変数なし)

また，2022 年 3 月 20 日からのローカルハニーポットへの攻撃数について，実数，訓練ずみモデル，予測値，70% 信頼区間，95% 信頼区間でプロットした結果を図 2と，図 3に示す．両者を比較すると，説明変数を用いた方が実際の攻撃の変化に追従している．

一方，約 1 年のデータ収集期間において，2022 年 3 月 17 日に平均的な攻撃数の約 64 倍の攻撃が観測されている．ISC への同日の攻撃数も平均の 12.5 倍になっているが，前日までは傾向がみられていないため予測値が大きく異なる結果になった．このような突発的な増加については大域的なハニーポットへの攻撃の情報を使っても対応が難しいことがわかった．

4 まとめ

本研究では，ローカルに設置したハニーポットで観測された攻撃を使って，今後の攻撃の強さを予測するために，世界各地のハニーポットから集められたグローバルな攻撃数を説明変数として時系列解析を行った．約 1 年間のデータ収集期間について実験を行なった結果，説明変数を用いる方が予測精度を上げられることを確認した．

10 通りの区間で時系列モデルの評価を行なったが，区間によって最適なモデルの次数 (p, d, q) が変化しており，モデルのパラメータを決定する期間などの設定について，さらに検討を行うことが必要である．

参考文献

[1] Husák, M., Komaárková, J., Bou-Harb, E. and Čeleda, P.: Survey of Attack Projection, Prediction, and Forecasting in Cyber Security, *IEEE Communications Surveys & Tutorials*, Vol. 21, No. 1, pp. 640–660 (2019).

[2] Werner, G., Yang, S. and McConky, K.: Time Series Forecasting of Cyber Attack Intensity, *Proceedings of the 12th Annual Conference on Cyber and Information Security Research*, CISRC '17 (2017).

[3] Zuzčák, M. and Bujok, P.: Using Honeynet Data and a Time Series to Predict the Number of Cyber Attacks, *Computer Science and Information Systems*, Vol. 18, No. 4, pp. 1197–1217 (2021).

[4] Sanders, C.: *Intrusion Detection Honeypots: Detection Through Deception*, Applied Network Defense (2020).

[5] SANS Internet Storm Center `https://www.dshield.org/` (accessed on 21, Dec. 2022).