2024 年度 卒業論文

ロジスティック回帰を用いた URL 特徴からのフィッシングサイト検出

龍谷大学 先端理工学部 数理・情報科学課程

Y200072 林 靖陽

指導教員 佐野 彰

概要

近年、正規ウェブサイトを装い、ユーザのログイン情報やクレジットカード情報を不正に取得する「フィッシングサイト」の被害が世界的に深刻化している。特に近年は、インターネット利用者の増加に伴い、フィッシング攻撃手法が多様化しており、サイト閲覧時にユーザが被害を察知することは容易ではない。こうした背景のもと、本研究では URL 文字列と SSL 証明書の特徴量に注目し、それらを用いたフィッシングサイト検出モデルを構築・評価することで、効率的かつ高精度な対策の可能性を探った。

本研究で採用したロジスティック回帰は、モデルがシンプルで計算コストが低いだけでなく、各特徴量の重要度を回帰係数として評価できるため、判定に寄与する要素を明確に把握できる利点がある。また、過学習を抑えつつ良好な精度を示し、結果の解釈も容易である点が、特徴量の寄与を分析する本研究に適している。非線形性の限界はあるものの、扱う特徴量は主に数値的なデータであり、ロジスティック回帰で十分対応可能と判断した。

まず、Mohammad らの先行研究を基に構築された UCI フィッシングサイトデータセット（約 11,000 件のサンプル、30 種類の特徴量）を活用し、URL 文字列から取得できるドメインの構造（サブドメイン数・ハイフンの有無など）や@記号の使用状況、URL の長さなどの特徴量を抽出するプログラムを自作した。具体的には、Python で URL を解析し、文字列操作や正規表現を用いて各種指標を数値化するアルゴリズムを開発した。続いて、UCI データセットには含まれていない SSL 証明書の情報を追加取得するため、ウェブサイトへの HTTPS 接続時にサーバーから返される証明書を解析するプログラムを別途作成した。

SSL 証明書の信頼性は以下の手順で評価した。HTTPS 接続時に取得するサーバー証明書から、発行者（Issuer）情報や証明書の有効期間（notBefore, notAfter）を取得する。次に、証明書の発行元が「正規認証局」であるかを Python の certifi パッケージが提供する信頼リストと照合する。いわゆる「正規認証局」とは、各 OS やブラウザ、プログラミング言語のライブラリで信用される認証局を指し、GlobalSign や DigiCert、GTS Root R1 などが該当する。こうした CA から発行された証明書であれば、一般的に信頼される。また、証明書の開始日と終了日の差分を計算し、「発行から 1 年以上経過しているか」「短期発行の格安証明書であるか」を判定基準とした。

さらに、新規に収集した 200 件の URL データセット（正規サイト 100 件＋フィッシングサイト 100 件）を用いて同様の手続きを実施した。フィッシングサイトは手動で稼働中の URL を確認し、正規サイトはランキングサイトから選定した。これらの特徴ベクトルをロジスティック回帰に入力し、フィッシング判定を行った結果、UCI データセットでは SSL 証明書の有無や発行元の信頼度が大きく精度向上に寄与する一方、新規 URL データセットでは短期間の証明書を使いながらも“正規認証局”から発行を受けているフィッシングサイトが多く、むしろハイフン多用やサブドメイン過多といった URL 構造上の特徴量が重要な判定要素になる場合があることがわかった。

以上の結果から、「URL 構造的特徴量＋ SSL 証明書情報」を合わせた多角的なアプローチが、フィッシングサイト検出の精度向上に寄与する可能性が示唆された。ただし、SSL 証明書に依存するのは危険であり、URL 内部の特徴量を含めた総合的な評価が必須である。今後は、機械学習法の拡充やデータセットの多様化を図り、より高精度で安定した検出プログラムの実現を目指す。