研究指導 中澤 真 教授

URL に基づく機械学習を用いたフィッシングサイト判別の精度向上 —差異係数に着目した特徴量選定—

中田 滋佑里

1. はじめに

インターネットの普及に伴い，サイバー犯罪のリスクが高まっている．特に，フィッシングサイトは正規のWebサイトを模倣し，ユーザの個人情報などを詐取する手法として深刻な問題となっている．フィッシングサイトは巧妙に偽装されており，ユーザが見分けるのは容易ではないためである．

従来の対策としてブラックリスト方式が用いられてきたが，未登録の新しいフィッシングサイトには対応できないという課題がある．この課題を解決するため，近年では機械学習を活用した判別手法が注目されている．機械学習を用いることで，フィッシングサイト特有のパターンを学習し，未知のサイトに対しても自動的に判別することが可能となる．

機械学習によるフィッシングサイトの判別では，多様な特徴量が利用される．中でも，URLの特徴量を用いた手法は，サイトの内容を解析せずに迅速に判別を行えるため，有効なアプローチの一つとされる．しかし，これまでの研究では，使用する特徴量は研究ごとに異なり，明確な基準が存在しなかった．そのため，特徴量の選定方法が課題となっていた．本研究では，この課題を解決するために，差異係数に基づいた特徴量の選定方法を提案し，新たな特徴量を選定する．差異係数を用いることで，フィッシングサイトと正規サイトの間で顕著な違いを示す特徴を数値的に評価し，判別に有効な特徴量を抽出することが可能となる．これにより，従来の手法では捉えきれなかったフィッシングサイトの特徴をより正確に抽出し，判別精度の向上を図ることを目的とする．

2. フィッシングサイトの判別の現状と課題

2.1 フィッシングサイトの現状

フィッシングとは，実在の組織になりすまし，パスワードやユーザ名などの認証情報，個人情報，またはクレジットカード番号などの金融情報を詐取する行為である[1]．IPA[1]によれば「情報セキュリティ10大脅威」の個人脅威として，2025年「フィッシングによる個人情報等の詐取」が7年連続ランクインしており，依然として重大な脅威と認識されている[2]．

次に，フィッシング詐欺とは，攻撃者が人間の心理の隙を突いてユーザの不安を煽り，メールやSMSなどからフィッシングサイトへ誘導し，認証情報，個人情報，金融情報などの重要な情報を騙し取る行為である．また，一部のフィッシングサイトは，情報の詐取に加え，閲覧するだけでマルウェア[2]に感染させる仕組みを持つものもある．

さらに，フィッシングサイトのURLには，いくつかの特徴がある．表 1にこれらのURLの例を示す[3]．①はユーザの打ち間違いを狙ったタイポスクワッティングを使用したもの，②はサブドメインに偽ドメイン名を含むもの，③は見た目が紛らわしいホモグラフ攻撃を利用したもの，④はサブドメインを過度に長くすることで偽装するものである．

表 1 フィッシングサイトの URL 例

	URLの例
①	https://**twItter**.com
②	https://google.com.**login.info**
③	https://**g0ogle**.com
④	https://**security.user...**facebook.com/dd...

2.2 フィッシングサイト判別システムの課題

フィッシングサイトは，表 1のURLのように巧妙に偽装されているため，ユーザが気づかずにアクセスしてしまう危険性がある．そのため，何らかのシステムでフィッシングサイトであることを判別して，これをユーザに通知する仕組みが必要になる．

フィッシングサイトの判別技術として，ブラックリスト方式が存在する．ブラックリスト方式とは，ユーザが特定のURLにアクセスを試みた際にリストを参照し，該当する場合には事前にアクセス遮断をする仕組みである．しかし，この手法ではリストに登録されていない新規のフィッシングサイトや未発見のサイトには対応できないという問題がある．

この課題に対して，近年では機械学習を用いたフィッシングサイトの判別手法が注目されている．この手法ではフィッシングサイト特有の特徴を抽出して数値化し，機械学習モデルに学習させることで，リストに登録されていないフィッシングサイトも判別できるようにしている．

機械学習によるフィッシングサイトの判別では，URLやサイトのHTML構造，使用されている画像の類似度，Whoisのドメイン登録情報など，多様な要素が判断材料となる．その中でもURLに基づく判別は，

[1] 独立行政法人 情報処理推進機構

[2] 情報の窃盗などを行う悪意を持ったソフトウェアの総称．

迅速かつ安全に実行できる．Jagdaleら[4]，Bireswarら[5]，田原ら[6]はいずれも，URLの特徴量を用いた機械学習モデルによるフィッシングサイト判別に取り組んだ．Jagdaleらの判別精度は86％以上であり，Bireswarらや田原らの判別精度は95％を超えている．

しかし，Webサイトの仕様は定期的に変更されることがあり，それに伴ってURL構造なども変化する．このような変化により，従来の特徴量が最新のWebサイトでは取得できなくなり，判別精度が低下する可能性がある．特に，フィッシングサイトはこうした変化を悪用し，より検出されにくい手法を取り入れることが考えられる．したがって，既存の特徴量が最新のフィッシングサイトに対しても有効であるかどうかを検証することが重要である．

ここで，フィッシングサイトの判別精度を向上させる重要な要素として，URLからの特徴量の抽出と選択が挙げられる．そこで，Bireswarらによって抽出された特徴量を表 2に示す．この表において，前述の3つの先行研究で共通の特徴量ついては※で示した．また，これらの特徴量は，URLを悪用する際にしばしばみられる重要な指標とされ，多くの研究で有効性が示されている．

表 2 Bireswar[5]が使用した特徴量

番号	特徴量
※1	URL の長さ
2	全体の文字の数に対する特殊文字の割合
3	特殊文字の数 （スキーム部分を除く）
4	全体の文字数に対するパス長の割合
5	疑わしいキーワードの数 （自身の観察などによる選定）
6	使用されているプロトコル
7	ハイフン(-)の数
8	最後の文字として特殊文字が使用されているか
※9	リダイレクトが発生しているか
※10	アットマーク(@)の数
※11	スラッシュ(/)の数 （スキーム部分を除く）
※12	ドメイン部分の中の IP アドレスの有無
13	クエスチョンマーク(?)の数
14	サブドメインの数
15	ドメイン前に www の有無
※16	URL の中に「http」が含まれているか
17	ポート番号の有無
※18	Unicode 文字 （%エンコーディングされた特殊文字の数）

本研究では，Bireswarらが提案した特徴量（表 2）を基盤として，ここに新たな特徴量を加えることで判別精度向上を図る．Bireswarらの手法を採用した理由は，他の先行研究と比べてより多くかつ詳細な特徴量を提示しており，追加的な特徴量の検討や拡張を行う上で有用性が高いと考えられるためである．

3. フィッシングサイト判別システムの構成

3.1 使用するツールとアルゴリズム

新たな特徴量の抽出には，フィッシングサイトと正規サイトのURLを比較・分析し，両者をより正確に区別できる性質を見出すことが重要である．特に，誤判別を減らすためには，一方のサイトにのみ顕著に現れる特徴量を抽出する必要がある．本研究では，Python[3]を用いた分析プログラムでこれを明らかにする．特徴量抽出プログラムでは，URLから特徴量を抽出するために次のPythonモジュールを使用している．まず，urllib.parseを用いてURLスキームやプロトコルなどの情報を取得し，reによって正規表現を扱う．また，CSVファイルの読み書きにはcsvを利用し，オペレーティングシステムとのやり取りのためにosを使用している．

機械学習のプログラムでは，機械学習ライブラリscikit-learn[4]を使用する．これはPythonを用いて多くの機械学習アルゴリズムを提供するオープンソースの機械学習ライブラリである．そして，抽出した特徴量に基づく正規サイトとフィッシングサイトを判別するアルゴリズムとして，ロジスティック回帰を用いる．機械学習以外の判別手法としてExcel統計の判別分析も使用する．

3.2 判別システムの構成

システムの構成を図 1に示す．まずフィッシングサイトと正規サイトを含むデータセットを用意する（①）．正規サイトを20,000件，フィッシングサイトを20,000件，計40,000件のURLを用意する．正規サイトのURLはGitHub[5]およびDMOZ Directory[6]（現在:Curlie）から収集する．フィッシングサイトのURLは，PhishTank[7]から収集したものを使用する．正規サイトのURLは2014年までのデータを使用する．ただし，正規サイトの構造や特徴は一般に大きく変わらないため，過去のデータであっても有用である．一方，フィッシングサイトのURLは，2024年12月までのものを含んでいる．Githubからのデータは，エンジニアや研究者などが投稿するものであり，一般的なウェブサイトなどに比べ，技術知識を持つ人々によって提供されるため，高品質なデータが多い．PhishTankに提供されているURLはコミュニティによる確認プロセスを経てフィッシ

[3] https://www.python.org/about/
[4] https://scikit-learn.org/stable/index.html
[5] https://github.com/gr33ndata/dmoz-urlclassifier
[6] https://curlie.org/
[7] https://phishtank.org/

ングサイトと認定されているため，誤検知のリスクが低い．

次に，これらのデータセットから特徴量を抽出し，CSVファイルとして保存する(②)．ここで，抽出する特徴量は表 2の特徴量に新たな特徴量を加えたものとする．その後，Pythonのプログラムにより機械学習モデルを作成する(③)．このプログラムでは，まず②で抽出したCSVファイルを読み込む．これらをランダムサンプリングにより，学習データとテストデータの比率が等しくなるように分割する．こうして分けられた学習データを用いて機械学習モデルを作成した．次に，残りのテストデータを用いてフィッシングサイトの判別実験による検証をする(④)．最終的にその判別から予測値，混合マトリックス，正解率，再現率，F-1スコア等の評価結果を算出する(⑤)．

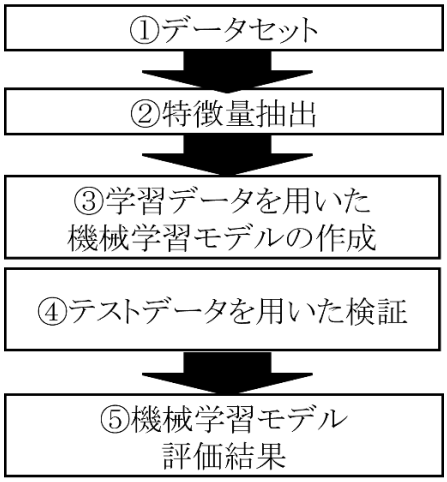

図 1 判別システムの構成図

4. 新たな特徴量選定と評価実験

4.1 実験方法

本研究の実験は以下の通りである．

1. 予備実験による既存の特徴量の有効性の検証
2. 判別分析結果の検証
3. データセット1，2を用いた組み合わせ実験
   - ① データセット1から差異係数の算出
   - ② ロジスティック回帰による判別実験
     - ➢ 既存の特徴量
     - ➢ 既存の特徴＋新しい特徴量
2. マクネマー検定による有意差の検証

4.2 予備実験

予備実験では，最新のフィッシングサイトのURLに対応しているかを検証するため，判別分析を用いて簡易的な実験を実施した．Bireswarら[5]の分析条件と揃え，正規サイトのURLを19,853件，フィッシングサイトのURL12,799件をデータセットとして使用する．その結果，判別分析では97.99％の精度が見られた．

これらの結果から，既存の特徴量でも最新のデータセットに対して有効であることが判明した．

4.3 新たな特徴量候補

予備実験の結果から既存の特徴量に新たな特徴量を追加するための分析をする．まず，判別分析で誤判別されたURLの傾向を確認し，正規サイトをフィッシングサイトと誤ってしまうケースよりも，フィッシングサイトを見逃す誤りが多いことを確認した．具体的には，フィッシングサイトのURL12,799件のうち，638件が正規サイトと誤判別されていた．この誤判別されたURLに対して，語句の出現傾向を確認し，「webmail」という語が正規サイトには低頻度でしか出現せず，フィッシングサイトでは一つのURL中に2回以上出現するケースが多いことが判明した．このため，本研究ではこの語を新たな特徴量の候補とした．

次に，表 2では，URL内に含まれるスラッシュやハイフンといった一部の特殊文字の出現頻度が特徴量として採用されていた．そこで本研究では，既存の特徴量として使用されていない特殊文字にも着目した．特殊文字は「!#$&’()*+,/:;=?@[],-._~」と定義した[7]．

ここで，どの特殊文字を使用するかは，差異係数に基づいて選定した．差異係数は，2つのデータにおける単語の使用頻度を比較することで，どちらのデータにより使用されているのかを判断する指標である[8]．差異係数を特徴量選定に使用することで特徴量のばらつきを考慮し，判別に役立つ特徴量を選択できるのではないかと考えた．

各特殊文字の出現有無を基に出現頻度と差異係数を求め，差異係数の中でも絶対値が1に近いものを特徴量の候補とした．表 3は差異係数の上位順(絶対値が1に近い順)に特徴量候補を並べている．特殊文字の上位4つは，差異係数の絶対値が1であるが，出現頻度が100未満と低い．出現頻度が低い特徴量は，判別に寄与しづらいと考え，差異係数の絶対値が1に近いものの中でも，どちらかのサイトで出現頻度が100以上であることを基準とし，上位5つの特殊文字の出現有無を特徴量の候補として選定した(~&#;=)．

表 3 差異係数に基づく特徴量候補一覧

特徴量名	差異係数	絶対値	出現頻度(正規)	出現頻度(フィッシング)
webmail	-1.00	1.00	0	101
,	-1.00	1.00	0	17
!	-1.00	1.00	0	12
[	-1.00	1.00	0	2
]	-1.00	1.00	0	2
~	**0.93**	**0.93**	**541**	**21**
&	**-0.79**	**0.79**	**248**	**2120**
#	**-0.70**	**0.70**	**21**	**120**
;	**-0.70**	**0.70**	**254**	**1447**
=	**-0.59**	**0.59**	**703**	**2759**

4.4 ロジスティック回帰の結果

差異係数の絶対値が1に近く，出現頻度が 100 以上であった語「webmail」と，特殊文字の上位 5 つを新しい特徴量として加え，ロジスティック回帰を用いた実

験を行った. その特徴量の組み合わせと精度の関係を**表 4** に示す. この結果から特徴量の組み合わせとして最適なものは, 既存の特徴量に「~」, 「&」, 「#」, 「;」, 「=」, 「webmail」を加えたものという結果になった.

表 4 ロジスティック回帰の結果

使用する特徴量(有無)								精度	
								データセット	
既存	webmail	~	&	#	;	=	残り すべて	1	2
○								0. 9775	0. 9759
○	○							0. 9793	0. 9777
○	○	○						0. 9791	0. 9778
○		○	○	○				0. 9776	0. 9761
○	○	○	○	○				0. 9790	0. 9777
○		○	○	○	○			0. 9777	0. 9761
○	○	○	○	○	○			0. 9789	0. 9779
○		○	○	○	○	○		0. 9779	0. 9766
○	○	○	○	○	○	○		**0. 9795**	**0. 9786**
○		○	○	○	○	○	○	0. 9779	0. 9764

4.5 マクネマー検定の結果

既存の特徴量のみを用いた場合の精度と, 新しい特徴量を加えた最も精度の高い組み合わせの精度に差異があるかを検証するため, マクネマー検定を実施した[9]. マクネマー検定の結果を表 5に示す. この結果から, 既存の特徴量のみの場合と1回目のデータセット, および既存の特徴量のみの場合と2回目のデータセットの比較において, P値が0.001未満であったため有意差が認められた. したがって, 既存の特徴量に「~」, 「&」, 「#」, 「;」, 「=」, 「webmail」を新たに加えることは, 精度の向上に有効であることが示された.

表 5 マクネマー検定の結果

データセット	カイ 2 乗値	自由度	P 値
1	26.25.45	1	P<0.001
2	20.41.67	1	P<0.001

4.6 考察

予備実験の結果, 最新のフィッシングサイトのURLに対して既存の特徴量のみを用いても高い判別精度が得られた. これは従来の手法が一定の有効性を持つことを示唆しているが, 実際には誤判別も見られたため, 新たな特徴量の導入による精度向上の余地は残っている. 本実験では, 差異係数の絶対値が高く, 出現頻度も一定水準以上である語や特殊文字を新たな特徴量として追加した結果, 判別精度が改善し, 統計的にも有意であることが確認された. したがって, 差異係数を基にした特徴量の選定方法は有効であると判断できる.

5. むすび

本研究では, 機械学習を用いたフィッシングサイトの判別精度向上を目的とし, 既存の特徴量の有効性を検証するとともに, 新たな特徴量の選定を試みた. また, 差異係数に基づく特徴量の選定を行い, その有効性を評価した結果, 判別精度の向上が確認された. これにより, 従来の手法では判別が困難であったフィッシングサイトに対して, より高精度な判別が可能となることが示された.

今後の課題として, さらなる新しい特徴量の選定や, 他のアルゴリズムとの比較が挙げられる. これらの課題に取り組むことで, より実用的なフィッシングサイト判別システムの構築を目指す.

参考文献

[1] フィッシング対策協議会, フィッシングとは, https://www.antiphishing.jp/consumer/abt_phishing.html, (参照 2025-01-31).

[2] IPA, 情報セキュリティ 10 大脅威 2025, https://www.ipa.go.jp/security/10threats/10threats2025.html, (参照 2024-12-14).

[3] NEC, その URL はそっくりさん？~ミスリード URL に注意~, https://jpn.nec.com/cybersecurity/blog/220107/index.html, (参照 2024-12-14).

[4] Nikita Jagdale, Pallavi Chavan, "Hybrid Ensemble Machine Learning Approach for URL Phishing Detection", 2022 2nd Asian Conference on Innovation in Technology (ASIANCON) Pune, India, 2022.

[5] Bireswar Banik and Abhijit Sarma, "Phishing URL detection system based on URL features using SVM", International Journal of Electronics and Applied Research (IJEAR) vol.5, issue 2, pp40-55, 2018.

[6] 田原 舜大, 秦野 亮, 西山 裕之, "URL 文字列の特徴量からユーザーへの攻撃を判別する機械学習モデル", 情報処理学会第 85 回全国大会講演論文集, 2023 巻, 1 号, pp461‐462, 2023.

[7] Cyberlibrarian, 図書館員のコンピュータ基礎講座 URI で使用できる文字, 2014-10-17, https://www.asahi-net.or.jp/~AX2S-KMTN/ref/uric.html, (参照 2025-02-05).

[8] 小林 雄一郎, ことばのデータサイエンス, 朝倉書店, 2019.

[9] 上田拓治, 44 の例題で学ぶ統計的検定と推定の解き方, オーム社, 2009.