SCIS 2022 2022 Symposium on
Cryptography and Information Security
Osaka, Japan & Online, Jan. 18 – 21, 2022
The Institute of Electronics,
Information and Communication Engineers

フィッシングを識別するための機械学習におけるデータセットの影響
Effects of datasets on machine learning to identify Phishing

中本　雄一 *　　加藤　雅彦 *
Yuichi Nakamoto　　Masahiko Katoh

キーワード フィッシング　機械学習　データセット

あらまし

インターネットの利活用が進む一方で，フィッシングはインターネットにおける重大な脅威の一つとなっている．フィッシングは，正規のサイトを装って ID やクレジットカード番号などの個人情報をだまし取ろうとする行為である．そこで，機械学習を用いてフィッシングをリアルタイムに検出するフィッシング識別器を作成し，フィッシングによる被害を未然に防止するための研究が行われている[1] [2]．機械学習を利用したフィッシング対策は，リアルタイムにフィッシングサイトを検出することができるが，誤検出の可能性がある．誤検出を減らして，フィッシング識別器の精度を向上させることが機械学習を用いたフィッシング対策の課題の一つである．

本論文では，フィッシング識別器の精度を向上させる手法として，データセットに着目する．先行研究において機械学習に用いるデータセットを作成する際に，「PhishTank」がデータソースとして利用されている[3]．しかし，PhishTank には多くの報告者が参加しており，明らかにフィッシングサイトではないと思われるサイトも報告されている．PhishTank をデータソースとして用いる場合，PhishTank に報告されたサイトがフィッシングサイトの定義に当てはまるかどうか照らし合わせた上で用いる必要がある．そうしなければ，作成したデータセットの中に正解ラベルの誤りが含まれることになる．

そこで、フィッシングの識別において，データセットが検出精度にどのような影響を与えるか評価検証を行う．まず，先行研究の手法と同様の方法で，オリジナルのデータセットを作成する．フィッシングの定義を明確にした上で，ソースに含まれているフィッシングの定義に当てはまらないサイトについて，オリジナルのデータセットをもとに正解ラベルを訂正し，修正版のデータセットを作成する．それぞれのデータセットを分割し，一方のデータセットを用いてフィッシング識別器に学習させ，他方のデータセットを用いてフィッシング識別器を評価する．

表１　データセットごとの正解率

	(評価 1) 学習用：オリジナル 評価用：修正版	(評価 2) 学習用：修正版 評価用：修正版
手法 A[1]	91.8%	93.0%
手法 B[2]	90.5%	91.8%

検証の結果，オリジナルの学習用データセットを修正版のデータセットを用いて評価する評価１よりも，正解ラベルを訂正した修正版のデータセットを用いた評価 2 の方が良い結果が得られ，機械学習によるフィッシングサイトの検出において，データセットを見直すことによりさらに検出精度を向上させることが出来ることを確認した．

参考文献

[1] Che-Yu Wu, Cheng-Chung Kuo and Chu-Sing Yang, "A Phishing Detection System Based on Machine Learning", International conference on Intelligent Computing and its Emerging Applications (ICEA), 2019.
[2] Mohammad, Rami, McCluskey, T.L. and Thabtah, Fadi (2012) An Assessment of Features Related to Phishing Websites using an Automated Technique. In: International Conferece For Internet Technology And Secured Transactions. ICITST 2012 . IEEE, London, UK, pp. 492-497. ISBN 978-1-4673-5325-0
[3] "PhishTank," October 2006. [Online]. Available. httpl/www.phishtank.com

* 長崎県立大学，長崎県西彼杵郡長与町まなび野 1−1−1