A Machine-Learning Framework for Supporting Intelligent Web-Phishing Detection and Analysis

Alfredo Cuzzocrea
University of Trieste
Trieste, Italy
alfredo.cuzzocrea@dia.units.it

Fabio Martinelli
Institute for Informatics and Telematics
Pisa, Italy
fabio.martinelli@iit.cnr.it

Francesco Mercaldo
Institute for Informatics and Telematics
Pisa, Italy
francesco.mercaldo@iit.cnr.it

ABSTRACT

This paper proposes a machine-learning framework for supporting intelligent web phishing detection and analysis, and provides its experimental evaluation. In particular we make use of state-of-the-art decision tree algorithms for detecting whether a Web site is able to perform phishing activities. If this is the case, the Web site is classified as a Web-phishing site. Our experimental evaluation confirms the benefits of applying machine learning methods to the well-known web-phishing detection problem.

CCS CONCEPTS

• **Security and privacy → Web application security**.

KEYWORDS

Web Phishing, Machine Learning for Supporting Web Phishing Detection, Web Phishing Analysis

ACM Reference Format:
Alfredo Cuzzocrea, Fabio Martinelli, and Francesco Mercaldo. 2019. A Machine-Learning Framework for Supporting Intelligent Web-Phishing Detection and Analysis. In *23rd International Database Engineering & Applications Symposium (IDEAS'19), June 10–12, 2019, Athens, Greece.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3331076.3331087

1 INTRODUCTION

Web security is an emerging trend, especially in the novel big data context (e.g., [1, 6, 15]). Traditionally, Web security has been addressed by exploiting several approaches, such as *privacy-preserving methodologies* (e.g., [1]), *hidden Markov models* (e.g., [21]), *logic-based approaches* (e.g., [9]), and so forth.

This traditional challenge, which involves in both academic and industrial research issues, is now emerging again due to its tight relation with novel *big data trends* (e.g., [7, 10, 19]), which has originated some very interesting approaches, among which [8, 18] are noticeable ones.

Among several problems, *Web phishing* (e.g., [5, 16, 17, 20]) is of relevant interest at now. Phishing is a method to imitating a official websites or genuine websites of any organization such as banks, institutes social networking websites, etc. Mainly phishing is attempted to theft private credentials of users such as username, passwords, PIN number or any credit card details etc. Phishing is attempted by trained hackers or attackers. Phishing is mostly attempted by phishy e-mails. This kind of Phishy e-mails may contains phishy or duplicate link of websites which is generated by attacker. By clicking these kinds of links, it is redirected on malicious website and it is easily to theft your personal credentials. Phishing Detection is a technique to detecting a phishing activity. There are various methods proposed by so many researchers. Among them Data Mining techniques are one of the most promising technique to detect phishing activity. Data mining is a new solution to detecting phishing issue. So data mining is a new research trend towards the detecting and preventing phishing website.

Starting from these considerations and in order to overcome the performances obtained, according by current literature, in this paper we propose a machine learning based method able to identify whether a web page is able to perform phishing activities.

Figure 1 shows the big picture of our framework. As shown in Figure 1, in our reference application scenario, several *Web Users* are interaction with *Web Phishing Sites* (still unknown, of course), and the goal of our framework is just to detect the Web phishing sites and notify the users on. To this end, the component *Feature Extraction* is in charge of extracting suitable features to drive the machine-learning-based detection phase. Features are extracted and an ad-hoc *Built-In Dataset* is populated this way. Finally, the *Decision Tree Algorithms* run over the latter dataset and the Web phishing event notification is finally reported to the *Web Users*.

This paper extends the previous short paper [4], where we introduced the main ideas of the proposed framework.

2 DECISION-TREE ALGORITHMS FOR WEB PHISHING DETECTION

In this Section we describe the method we propose for web phishing attacks detection.

Table 1 shows the features considered in the following study.

In order to collect data, we consider the PhishTank dataset [1]: PhishTank is a free community site where anyone can submit, verify, track and share phishing data. This dataset is in the form of .csv file format.

The evaluation consists of two different stages: (i) we provide hypotheses testing, to verify whether the features vector exhibit different distributions for attacks and normal messages populations; and (ii) decision-tree machine learning analysis in order to assess

[1] https://archive.ics.uci.edu/ml/datasets/Website+Phishing

IDEAS'19, June 10–12, 2019, Athens, Greece

ACM ISBN 978-1-4503-6249-8/19/06.
https://doi.org/10.1145/3331076.3331087

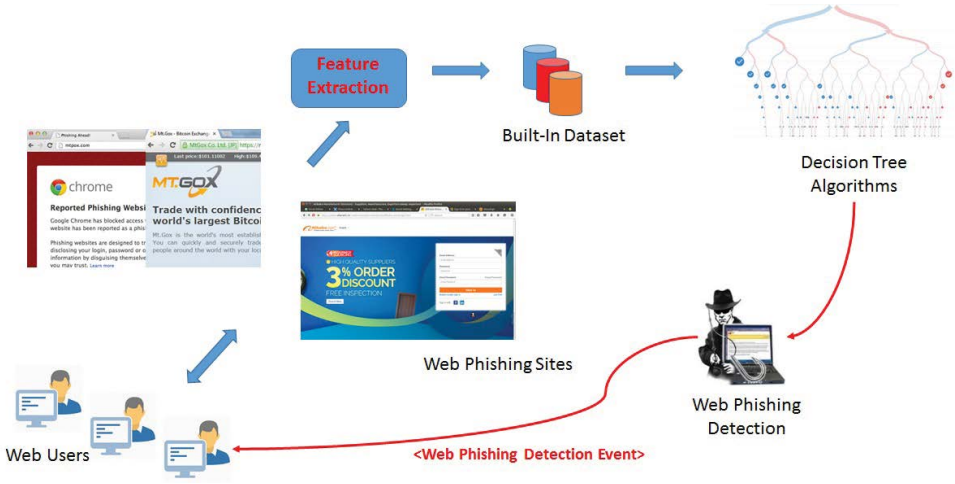

Figure 1: The Proposed Machine-Learning-Based Web Phishing Detection Framework

Variable	Feature
$F1$	URL Anchor
$F2$	Request URL
$F3$	Server Form Handler
$F4$	URL Length
$F5$	Having IP Address
$F6$	Prefix/Suffix
$F7$	IP
$F8$	Sub Domain
$F9$	Website Traffic
$F10$	Domain Age

Table 1: The Feature Set Involved in the Study

if the eight features are able to discriminate between attacks and normal messages.

Machine learning is a type of artificial intelligence able to provide computers with the ability to learn without being explicitly programmed [14].

Machine learning tasks are typically classified into two categories, depending on the nature of the learning available to a learning system:

- *Supervised learning*: the computer is presented with example inputs and their desired outputs, given by a "teacher", and the goal is to learn a general rule that maps inputs to outputs. It represents the classification: the process of building a model of classes from a set of records that contains class labels.
- *Unsupervised learning*: no labels are given to the learning algorithm, leaving it on its own to find structure in its input. Unsupervised learning can be a goal in itself (discovering hidden patterns in data) or a means towards an end (feature learning).

The algorithms considered are supervised decision tree-based i.e., they use a decision tree as a predictive model which maps observations about an item (represented in the branches) to conclusions about the target of the items value (represented in the leaves). These algorithms (i.e., *J48*, *HoeffdingTree*, *RandomForest*, *RetTree*, *LMT* and *DecisionStump*) are the most widespread to solve data mining problems [14] for instance, from malware detection [2, 3, 11, 13] to pathologies classification [12].

We consider in this work five different machine learning algorithms in order to enforce the conclusion validity. With regards to the hypotheses testing, the null hypothesis to be tested is:

H_0 : 'phishing and legitimate web pages exhibit similar values of the considered features'.

The null hypothesis was tested with Wald-Wolfowitz (with the p-level fixed to 0.05), Mann-Whitney (with the p-level fixed to 0.05) and with Kolmogorov-Smirnov Test (with the p-level fixed to 0.05). We chose to run three different tests in order to enforce the conclusion validity. The purpose of these tests is to determine the level of significance, i.e., the risk (the probability) that erroneous conclusions be drawn: in our case, we set the significance level equal to .05, which means that we accept to make mistakes 5 times out of 100. The analysis goal is to verify if the considered features are able to correctly discriminate between phishing and normal web pages. These algorithms were applied to the full feature vector. The classification analysis is performed using the Weka[2] tool, a suite of machine learning software, employed in data mining for scientific research.

3 EXPERIMENTAL ASSESSMENT AND ANALYSIS

We used five metrics in order to evaluate the results of the classification: Precision, Recall, F-Measure, MCC and RocArea. The results that we obtained with this procedure are shown in table 2.

[2]http://www.cs.waikato.ac.nz/ml/weka/

Algorithm	Precision	Recall	F-Measure	MCC	Roc Area	Class
J48	0,904	0,892	0,898	0,829	0,958	legitimate
	0,923	0,916	0,919	0,833	0,958	phishing
HoeffdingTree	0,840	0,892	0,865	0,770	0,948	legitimate
	0,882	0,916	0,899	0,786	0,953	phishing
RandomForest	0,891	0,892	0,892	0,818	0,968	legitimate
	0,917	0,912	0,914	0,822	0,966	phishing
RepTree	0,856	0,911	0,882	0,799	0,964	legitimate
	0,933	0,872	0,901	0,804	0,961	phishing
LMT	0,876	0,892	0,884	0,804	0,970	legitimate
	0,922	0,905	0,913	0,821	0,972	phishing
DecisionStump	0,794	0,849	0,820	0,692	0,835	legitimate
	0,836	0,913	0,873	0,726	0,845	phishing

Table 2: Classification results.

As shown in Table 1 the proposed method is able to obtain a precision equal to 0,923 and a recall equal to 0,916 in phishing attack detection using the J48 algorithm. The classification algorithms obtaining the best precision are J48 and RepTree, but considering also the recall metric, we highlight that the RepTree recall is lower if compared with the one obtained by the J48 classification algorithm: this is the reason why we confirm the J48 algorithm as the one obtaining the best performances in terms of precision and recall in order to detect web phishing attacks. As a matter of fact, the remaining algorithms (i.e., HoeffdingTree, RandomForest, LMT and DecisionStump) exhibit lower performances than J48 and RepTree in terms of precision and recall.

4 CONCLUSIONS AND FUTURE WORK

Recently, a more effective approach to fight phishing that relies on machine learning techniques has emerged. In this approach, models extracted by a ML technique are used to classify websites either as legitimate or phishy, based on certain features. In this paper we propose a method machine learning-based able to detect whether a web page exhibits phishing attacks. As future work, we plan to extend the proposed features in order to increase the detection accuracy, in addition we plan to apply formal methods with the aim to detect the code in which the malicious action happens.

ACKNOWLEDGMENTS

This work has been partially supported by H2020 EU-funded projects SPARTA contract 830892 and C3ISP and EIT-Digital Project HII.

REFERENCES

[1] Saad A. Abdelhameed, Sherin M. Moussa, and Mohamed E. Khalifa. 2018. Privacy-preserving tabular data publishing: A comprehensive evaluation from web to cloud. *Computers & Security* 72 (2018), 74–95.

[2] Pasquale Battista, Francesco Mercaldo, Vittoria Nardone, Antonella Santone, and Corrado Aaron Visaggio. 2016. Identification of Android Malware Families with Model Checking. In *Proceedings of the 2nd International Conference on Information Systems Security and Privacy, ICISSP 2016, Rome, Italy, February 19-21, 2016.* SciTePress, 542–547.

[3] Gerardo Canfora, Francesco Mercaldo, Corrado Aaron Visaggio, and Paolo Di Notte. 2014. Metamorphic malware detection using code metrics. *Information Security Journal: A Global Perspective* 23, 3 (2014), 57–67.

[4] Alfredo Cuzzocrea, Fabio Martinelli, and Francesco Mercaldo. 2018. Applying Machine Learning Techniques to Detect and Analyze Web Phishing Attacks. In *iiWAS*. ACM, 355–359.

[5] Serge Egelman, Lorrie Faith Cranor, and Jason I. Hong. 2008. You've been warned: an empirical study of the effectiveness of web browser phishing warnings. In *Proceedings of the 2008 Conference on Human Factors in Computing Systems, CHI 2008, 2008, Florence, Italy, April 5-10, 2008.* 1065–1074.

[6] Hsiu-Chuan Huang, Zhi-Kai Zhang, Hao-Wen Cheng, and Shiuhpyng Winston Shieh. 2017. Web Application Security: Threats, Countermeasures, and Pitfalls. *IEEE Computer* 50, 6 (2017), 81–85.

[7] IBM, Paul Zikopoulos, and Chris Eaton. 2011. *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data* (1st ed.). McGraw-Hill Osborne Media.

[8] Igor V. Kotenko, Igor Saenko, and Alexander Branitskiy. 2018. Applying Big Data Processing and Machine Learning Methods for Mobile Internet of Things Security Monitoring. *J. Internet Serv. Inf. Secur.* 8, 3 (2018), 54–63.

[9] Rafal Kozik, Michal Choras, and Witold Holubowicz. 2017. Packets tokenization methods for web layer cyber security. *Logic Journal of the IGPL* 25, 1 (2017), 103–113.

[10] Kuan-Ching Li, Hai Jiang, Laurence T. Yang, and Alfredo Cuzzocrea (Eds.). 2015. *Big Data - Algorithms, Analytics, and Applications.* Chapman and Hall/CRC.

[11] Fabio Martinelli, Fiammetta Marulli, and Francesco Mercaldo. 2017. Evaluating Convolutional Neural Network for Effective Mobile Malware Detection. *Procedia Computer Science* 112 (2017), 2372–2381.

[12] Francesco Mercaldo, Vittoria Nardone, and Antonella Santone. 2017. Diabetes Mellitus Affected Patients Classification and Diagnosis through Machine Learning Techniques. *Procedia Computer Science* 112, C (2017), 2519–2528.

[13] Francesco Mercaldo, Corrado Aaron Visaggio, Gerardo Canfora, and Aniello Cimitile. 2016. Mobile malware detection in the real world. In *Software Engineering Companion (ICSE-C), IEEE/ACM International Conference on.* IEEE, 744–746.

[14] Tom M Mitchell. 1999. Machine learning and data mining. *Commun. ACM* 42, 11 (1999), 30–36.

[15] Paulo Jorge Costa Nunes, Iberia Medeiros, José Fonseca, Nuno Neves, Miguel Correia, and Marco Vieira. 2018. Benchmarking Static Analysis Tools for Web Security. *IEEE Trans. Reliability* 67, 3 (2018), 1159–1175.

[16] Nuttapong Sanglerdsinlapachai and Arnon Rungsawang. 2010. Using Domain Top-page Similarity Feature in Machine Learning-Based Web Phishing Detection. In *WKDD*. IEEE Computer Society, 187–190.

[17] Guang Xiang, Jason I. Hong, Carolyn Penstein Rosé, and Lorrie Faith Cranor. 2011. CANTINA+: A Feature-Rich Machine Learning Framework for Detecting Phishing Web Sites. *ACM Trans. Inf. Syst. Secur.* 14, 2 (2011), 21:1–21:28.

[18] Zheng Xu, Zhiguo Yan, Lin Mei, and Hui Zhang. 2015. The Big Data Analysis of the Next Generation Video Surveillance System for Public Security. In *WEB (Lecture Notes in Business Information Processing)*, Vol. 258. Springer, 171–175.

[19] Chao-Tung Yang, Jung-Chun Liu, Ching-Hsien Hsu, and Wei-Li Chou. 2014. On improvement of cloud virtual machine availability with virtualization fault tolerance mechanism. *The Journal of Supercomputing* 69, 3 (2014), 1103–1122.

[20] Yue Zhang, Jason I. Hong, and Lorrie Faith Cranor. 2007. Cantina: a content-based approach to detecting phishing web sites. In *Proceedings of the 16th International Conference on World Wide Web, WWW 2007, Banff, Alberta, Canada, May 8-12, 2007.* 639–648.

[21] Zhongliu Zhuo, Yang Zhang, Zhi-Li Zhang, Xiaosong Zhang, and Jingzhong Zhang. 2018. Website Fingerprinting Attack on Anonymity Networks Based on Profile Hidden Markov Model. *IEEE Trans. Information Forensics and Security* 13, 5 (2018), 1081–1095.