Suleman *et al.LGU (IJECI)* 2018

LGU (IJECI)
ISSN: 2522-3429 (Print)
ISSN: 2616-6003 (Online)

LGU International Journal for
Electronic Crime Investigation

Research Article

Vol. 2 issue 3 July-September 2018

Classification of Website Phishing Data through Machine Learning Algorithms

Muhammad Taseer Suleman
Lahore Garrison University, Lahore, Pakistan
taseersuleman@lgu.edu.pk

Abstract:

Phishing is the dissemination of malicious web sites used to acquire passwords, credit card details or any sensitive personal information. Clients of web advancements deal with different security dangers and phishing is a standout amongst the most imperative dangers that should be addressed. Phishing sites have certain attributes and designs, in order to, distinguish those components that can help us to recognize phishing. In order to, recognize such elements information mining methods have been utilized. In this work, we depicted examination in arrangement of phishing sites utilizing diverse classification algorithms with genetic algorithms for enhancement, for example, as feature selection and generation. Keeping in mind the end goal to figure out which technique gives the prime outcomes in phishing sites characterization. Websites are characterized as “1” for "Legitimate”, “0” for "Suspicious" and “-1” for "Illegitimate". We have found that machine-learning algorithms along with feature selection algorithms were the best choice for detecting web phishing attacks.

Keywords: Phishing, Spamming, Features, Machine learning algorithms, Genetic algorithms.

1. Introduction

Web phishing is a mechanism of online fraud in which the victim is deceived by the attacker in gaining victim’s personal information like credit card number, financial accounts, address, phone numbers etc. The assailant makes a fake site page by replicating or rolling out a little improvement in the honest to goodness page. The fake sites are planned to look precisely like the bona fide site. The fast advancement of web applications give a ton of advantages to web clients to use these web application for making all their everyday exercises, for example, newspaper perusing, shopping, payment of many types of bills, ticket booking, and amusement and so forth. However, artisan create novel assaults that draws in more web client to be gotten in web of phishing. As per Gupta et al. [1] the entire number of specific phishing sites recognized in the primary quarter of 2014 alone was 125,215, delineating an expansion of more than 11% from the 2013 figures. While a greater part of the phishing effort utilizes malevolently enrolled areas and sub-spaces, they have made genuine money related harm

clients over the globe. Moreover, there had been a huge year on year increment in phishing assaults, which is appear in figure, expanding altogether from 203,983 of every 2013 to 448,126 of every 2017.

As indicated by the Anti-Phishing Working Group (APWG), the APWG Got reports of 630,494 extraordinary phishing locales recognized from the main quarter through the second from last quarter of 2017. The around the web phishing rate was 36.511% in the primary quarter, 32.211% in the second quarter, and 32.122% in the second from last quarter of 2017. Besides, ISP area observed to be the most under assault industry area from first to second from last quarter of 2017 as appeared in Fig 1 below.

Fig 1 Phishing detection rate in 2017

Web phishing attack is comprises for many stages. The choice of victim and the amount of benefit are important parameters in web phishing attack.
Phishing life cycle has following stages:

i. Planning and Setup

In first stage, the phisher determine the objective association, an individual or a country to be targeted for malicious purpose. They uncover the sensitive information with respect to their objective and its system. Normally phishing starts by sending spoofed emails to the victims [2]. Victims are supposed to send required information via replying to the email. However, most of the users do not reveal their information through email.

Another phishing technique can be adopted through creation of phishing websites. A combination of both aforementioned techniques can also be used for phishing as well [3] as shown in Fig 2.

Fig 2 Web-Phishing whole Plot

ii. Phishing

Assailants send mock messages to the dupe, utilizing gathered email tend to which request classified data from the dupe. Another special form of phishing is known as spear phishing. In spear phishing, target is generally a group of specific individuals. In addition, there are many other forms of phishing as depicted in Fig 3.

Fig 3 Different types of Phishing

iii. Break-in or Infiltration

In this stage, the victim taps on the pernicious connection and when he does that, a malware

naturally introduces on his gadget that enables the phisher to get to the system, irrupt and change its arrangements and get to rights to it.

iv. Information Accumulation

When the phisher access on victim's system, they remove the required information. On the off chance that the casualty gives classified record points of interest, the assailant would then be able to get to his record, which may, in the end, prompt budgetary misfortunes to the casualty. Once the attack is successful, the attacker does the information collection. Information may contain passwords, user identity number, contact lists, private images, and credit card information. The whole phishing life cycle is shown in Fig 4.

Fig 4 Scenario of web phishing

Detection of phishing website is a problem of major concern. Various techniques such as fuzzy, neural systems and data mining methods applied, in order to, counter web-phishing attacks [4]. Several machine-learning methods also applied for detection of fake websites. Machine learning approach is based on both supervised as well as unsupervised. We have tested many machine-learning algorithms on the given data downloaded from UCI machine learning dataset. These algorithms include Naïve Bayes (NB), Support Vector Machine (SVM), Neural Net (NN), Random Forest (RF), IBK lazy classifier and Decision Tree (ID3). However, we have observed that phishing detection results can be enhanced by applying feature selection algorithms like Generating Genetic Algorithms (GGA), Another Genetic Algorithm (AGA) etc. In the end, we have shown the difference of accuracy between the results of those machine-learning algorithms applied to the data to those machine-learning algorithms used with feature selection algorithms.

The rest of research article is organized as follows: section II contains the previous related research work. Section III describes methodology of our work. In Section IV we have shown the results. In section V we have concluded our work.

2. Related Work

In [5] Tahir et al. have proposed a hybrid model, in order to, overcome phishing issue. Their proposed hybrid model show beats as far as high precision and less mistake rate. They completed tests in two stages. In the first stage, they separately performed classification algorithm and select the best three models on criteria of execution and high precision. In the second stage, they additionally consolidated each model with their best "Three" singular models.

In [6] authors have proposed the classification algorithm named as PAC (Phishing Associative Classification). They observed the execution of proposed calculation in term of precision measurements with four well-known calculations that are C4.5, PRISM, CBA, and MCAR.

In [7], Authors have portrayed examination in arrangement of phishing sites utilizing

distinctive Machine learning calculations. They have applied various machine learning techniques including Random Forest (RF), C4.5, REP Tree, Decision Stump, Hoeffding Tree and Rotation Forest. From the outcomes, it has been discovered that the Rotation Forest calculation with REP Tree as a classifier and MLP plays out the best on a full preparing and on diminished set, separately.

In [8] authors have utilized information-mining approach like supervised characterization, which enhances the frameworks precision and distinguishes more measure of spam and harmful URLs.

Google in [9] gives a support of safe perusing that enables the applications to check the URLs utilizing a file of suspicious areas, which is consistently refreshed by Google. It is a trial Programming interface, however, is utilized with Google Chrome and Mozilla Firefox, and it is anything but difficult to utilize.

Authors in [10] connected distinctive sorts of machine learning based arrangement calculations, including Naive Bayes (NB), Support Vector Machine (SVM), Neural Net (NN), Random Forest (RF), IBK relaxed classifier and Decision Tree (J48) and broaden Pradeep and Ravendra's work by presenting new order calculation named Neural Net in their test. In the end, they Shield clients from nasty or unstructured connections in Site pages and Texts.

Measured and looked at the execution of the classifier as far as precision. Neural Net demonstrated a decent order exactness contrast with others.

Authors in [11] proposed another calculation Linkguard calculation to give up from phishing assaults. This calculation utilizes attributes of hyperlinks to deduct the attacks. Linkguard algorithm examines the contrast between the visual connection

and genuine connection. Link Guard is valuable for recognizing phishing assaults, as well as can shield clients from nasty or unstructured connections in Site pages and Texts.

3. Methodology

In our work, we utilized dataset for the examination is "Phishing Websites Dataset" ("UCI Machine Learning Vault: Phishing Sites Informational collection," 2016) [12]. This dataset was accumulated fundamentally from Phish Tank archive, Miller Smiles archive, and Google’s seeking administrators.

The dataset is separate into training (70%) as well as testing (30%) datasets. Dataset includes total 11054 instances. All occasions sorted as “1” for "Real”, “0” for "Suspicious" and “-1” for "Phishy". We have used Python 3.6 for data analysis.

The creators illuminate the key features that have been turned out to be strong and effective in foreseeing phishing sites while proposing some new features, tentatively allocating new standards to some outstanding features and refreshing some different features.

Features have been grouped into following categories:

- Address Bar-based features
- Abnormal based features
- HTML and Javascript Based features
- Domain Based Features

The address bar based features is a heuristic approach towards web phishing detection [13].

Abnormal based feature includes abnormal URL, anchor URL, abnormal DNS etc. [13]. We have conducted few experiments on our data in terms of its covariance, variance.

In Fig 5, we have shown various characteristics of our data.

These characteristics include total count, mean, standard deviation (std) and data range (min & max). Therefore, analysis between different features is easy enough.

	having_IP_Address	URL_Length	Shortining_Service	having_At_Symbol	double_slash_redirecting	Prefix_Suffix	having_Sub_Domain	SSLfinal_State
count	11054.000000	11054.000000	11054.000000	11054.000000	11054.000000	11054.000000	11054.000000	11054.000000
mean	0.313914	-0.633345	0.738737	0.700561	0.741632	-0.734938	0.064049	0.251040
std	0.949495	0.765973	0.674024	0.713625	0.670837	0.678165	0.817492	0.911856
min	-1.000000	-1.000000	-1.000000	-1.000000	-1.000000	-1.000000	-1.000000	-1.000000
25%	-1.000000	-1.000000	1.000000	1.000000	1.000000	-1.000000	-1.000000	-1.000000
50%	1.000000	-1.000000	1.000000	1.000000	1.000000	-1.000000	0.000000	1.000000
75%	1.000000	-1.000000	1.000000	1.000000	1.000000	-1.000000	1.000000	1.000000
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

	Domain_registeration_length	Favicon	...	popUpWidnow	Iframe	age_of_domain	DNSRecord	web_traffic	Page_Rank	Google_Index
count	11054.000000	11054.000000	...	11054.000000	11054.000000	11054.000000	11054.000000	11054.000000	11054.000000	11054.000000
mean	-0.336711	0.628551	...	0.613353	0.816899	0.061335	0.377239	0.287407	-0.483626	0.721549
std	0.941651	0.777804	...	0.789845	0.576807	0.998162	0.926158	0.827680	0.875314	0.692395
min	-1.000000	-1.000000	...	-1.000000	-1.000000	-1.000000	-1.000000	-1.000000	-1.000000	-1.000000
25%	-1.000000	1.000000	...	1.000000	1.000000	-1.000000	-1.000000	0.000000	-1.000000	1.000000
50%	-1.000000	1.000000	...	1.000000	1.000000	1.000000	1.000000	1.000000	-1.000000	1.000000
75%	1.000000	1.000000	...	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
max	1.000000	1.000000	...	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

Fig 5: Count, Mean and Standard Deviation of each dataset feature

In the next step, we have applied different machine learning techniques on our dataset. The dataset, as earlier said, was used to foresee the exactness of the acknowledgment using assorted classifier. For the earlier examination, the component assurance is not used and just classifiers are used to get the required accuracy for each of the classifiers. The data is obviously portrayed however this time particular segment decision methodologies are used for the update of the results or to check for any possible upgrades. The usage of feature decision procedures furthermore help in dimensionality diminish as feature reducing.

In Fig 6, we have correlated each feature name with a feature ID. In this study for feature selection algorithm, we used Generating Genetic Algorithm (GGA), Another Genetic Algorithm (AGA), Yet Another Generating Genetic (YAGGA) and Yet Another Generating Genetic Algorithm-2 (YAGGA2). The classifiers utilized were Naïve Bayes, ID3, KNN, Decision Tree, and Random Forest. The Characteristics of highlight choice calculations are that they select the best components on the premise of properties weights.

Feature Name	Feature ID
IP_Address	A
URL_Length	B
Shortining_Service	C
having_At_Symbol	D
double_slash_redirecting	E
Prefix_Suffix	F
having_Sub_Domain	G
SSLfinal_State	H
Domain_registeration_length	I
Favicon	J
Port	K
HTTPS_token	L
Request_URL	M
URL_of_Anchor	N
Links_in_tags	O
SFH	P
Submitting_to_email	Q
Abnormal_URL	R
Redirect	S
on_mouseover	T
RightClick	U
popUpWidnow	V
I frame	W
age_of_domain	X
DNSRecord	Y
web traffic	Z
PageRank	AA
Google Index	AB
Links_pointing_to_page	AC
Statistical_report	AD
Result	AE

Fig 6: Each Feature get associated with a unique ID

Based on the above features in our dataset, we have conducted series of experiments that involved two streams.

1. Machine learning algorithms along with Feature Selection
2. Machine learning algorithms without Feature Selection.

These both streams involved series of stages involved in them with difference of one or more stages. The details of all stages involved are as follows.

V. Read CSV: A simple method involved is the reading of CSV. A comma-delimited Phishing.CSV is given as an input to the system. The data contains as many as 11054 instances.

VI. Cross-Validation: It is a statistical method, which involves evaluation and comparison of learning algorithms through dataset division [14]. The division of dataset brought two segments: Train dataset and Test dataset. K-fold cross-validation is the basic form of cross-validation. In our case, we have also cross over our data through cross-validation also known as X-Validation. The division between train dataset and Test Dataset also came into practice.

VII. Testing: After Cross validation data is passed through the Testing stage. Testing stage involves the application of multiple machine learning algorithms on the specific data [15].

VIII. Classifier: Classifier used to perform classification on the given data. Classification is actually the task of mapping function from input features to finite output values [16]. In our case, our task is to classify the given data according to Phishy, non-phishy (normal) instance based on the previous data learning. In our case, the classifiers are Naïve Bayes, ID3, KNN, Decision Tree and Random Forest.

IX. Model Application: After then we apply different models to our dataset based on the aforementioned various models.

X. Feature Selection: As we have mentioned before, our testing, analysis and result generation based on the comparison of two streams. One with Feature selection and other without feature selection. This is an important stage in the data analysis. Feature selection aims to choose a subset of feature from the available features [17][18]. We have used feature selection algorithm like GGA, AGA, YAGGA, YAGGA-2. We have used these feature selection algorithms along with classification algorithms. In result section, we will show the effect of performance with and without using feature selection algorithm.

XI. Performance: Performance is measured against all the stages that we have mentioned above. This would be the last stage of each stream. Results had been collected and compared, in order to, find best.

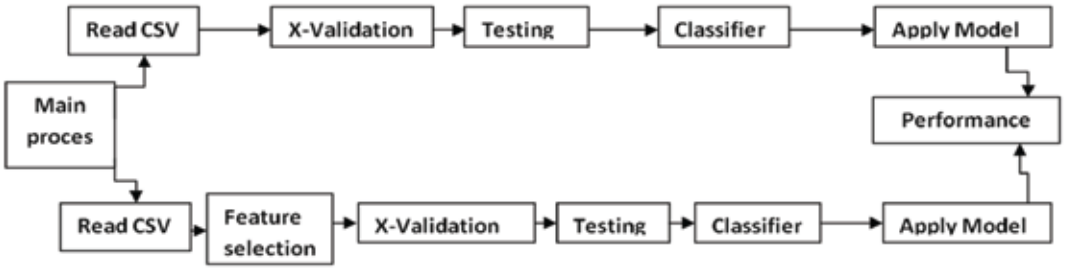

Fig 7: Model for applying Machine learning algorithms with & without Feature Selection Algorithms

4. Results and Discussions

In stage 1 classifiers i.e.; Naive Bayes, ID3, KNN, Decision Tree and Random Forest and are utilized to get the required exactness for each of the classifiers. In stage 2 the information is, on the other hand, characterized yet this time distinctive features strategies i.e.; GGA, AGA, YAGGA, and YAGGA2 are utilized for the upgrade of the outcomes or to check for any believable changes. Results demonstrate that ID3 with YAGGA with 15 features chosen, lessened from 30 highlights, demonstrate the best execution on this dataset for order of phishing sites. The results are shown in Fig 8 with "YAGGA + ID3" shows the maximum accuracy up to 95%.

Classification Algorithm	Unprocessed Data	Feature Selection Algorithms			
		GGA	AGA	YAGGA	YAGGA2
Naïve Bayes	91.08%	93.31 %	75.53 %	92.94%	73.53%
ID3	87.16%	94.63 %	75.53 %	94.99%	74.5%
KNN	89.51%	92.55 %	73.53 %	94.72%	73.53%
Decision Tree	91.65%	94.00 %	59.51 %	93.18%	86.04%
Random Forest	76.33%	89.27 %	57.88 %	92.46%	85.92%

Fig 8: Results of accuracy

5. Conclusion and Future work

Web Phishing attack is of serious concern. This work models the phishing site expectation as a characterization undertaking and exhibits the machine learning approach for foreseeing whether the given site is genuine site or phishing. The phishing dataset was taken from UCI learning website. The dataset contains as many as 11054 instances. In this research, several machine learning algorithms. The results were compared with the application of same machine learning algorithm along with feature selection algorithm. It has been noted that YAGGA along with ID3 has given the best results with approximately 95% accuracy of website phishing detection. In future, we will extend this work for other famous website attacks with the help of machine learning algorithms.

6. References

[1] W. Zhuang, Q. Jiang and T. Xiong,"An Intelligent Anti-Phishing Strategy Model for Phishing Website Detection," 32nd IEEE International Conference on Distributed Computing Systems Workshops, 2012, China.

[2] B. B Gupta, A. Tewari, A.K. Jain, D.P Agrawal," Fighting against phishing attacks: state of the art and Future challenges," Neural Computing and Applications, Vol 28 Issue 12, December 2017.

[3] C.E Drake, J.J. Oliver, E.J. Koontz," Anatomy of a Phishing Email," CEAS, 2004.

[4] G. Varshney, R.C. Joshi, A. Sardana,

”Personal Secret Information Based Authentication towards Preventing Phishing Attacks,” In: Meghanathan N., Nagamalai D., Chaki N. (eds) Advances in Computing and Information Technology. Advances in Intelligent Systems and Computing, vol 176. Springer, Berlin, Heidelberg

[5] M. Islam, N.K. Chowdhury,” Phishing Websites Detection Using Machine Learning Based Classification Techniques.

[6] M.A.U.H Tahr, S. Asghar, A. Zafar, S. Gillani,” A Hybrid Model to Detect Phishing-Sites using Supervised Learning Algorithms,” International Conference on Computational Science and Computational Intelligence (CSCI), 15 Dec 2016, Las Vegas, NV, USA.

[7] S.Wedyan, F. Wedyan,” An Associative Classification Data Mining Approach for Detecting Phishing Websites,” Journal of Emerging Trends in Computing and Information Sciences, Vol. 4 No.12, Dec 2013.

[8] A. Hodzic, J. Kevric, A. Karadeg,” Comparison of Machine Learning Techniques in Phishing Website Classification,” International Conference on Economic and Social Studies, April 2016.

[9] S.B Rathod, T.M. Pattewar,” A comparative performance evaluation of content based spam and malicious URL detection in E-mail,” IEEE International Conference on Computer Graphics, Vision and Information Security (CGVIS), 2015.

[10] Google safe browsing API Available. Retrieved 1st June 2018 from https://developers.google.com/safe-browsing/.

[11] J. James, L. Sandhya, C. Thomas,”Detection of Phishing using machine learning techniques,” International Conference on Control Communication and Computing (ICCC), December 2013.

[12] G. Varshney, M. Misra, P.K Atrey,”A survey and classification of web phishing detection schemes,” Security and Communication Networks, Wiley Online Library, October 2016.

[13] G. Varshney, M. Misra, P.K Atrey,”A survey and classification of web phishing detection schemes,” Security and Communication Networks, Wiley Online Library, October 2016.

[14] Phishing Dataset. Retrieved on 25th April 2018 from https://archive.ics.uci.edu/ml/datasets/phishing+websites

[15] M.G. Alkhozae, O.A. Batarfi,” Phishing Websites Detection based on Phishing Characteristics in the Webpage Source Code,” International Journal of Information and Communication Technology Research, Vol 1No.06, October 2011.

[16] S. Arlot, A. Celisse,” A survey of cross-validation procedures for model selection,” Statistics Surveys, Vol 4, 2010.

[17] How to evaluate Machine Learning Algorithms. Retrieved 2nd July from https://machinelearningmastery.com/how-to-evaluate-machine-learning-algorithms/

[18] F.Y. Osisanwo, J.E.T. Akinsola, x O. Akinsola , J. O. Hinmikaiye , O. Olakanmi , J. Akinjobi,” Supervised Machine Learning Algorithms: Classification and Comparison,” International Journal of Computer Trends and Technology (IJCTT), Volume 48 No. 3, June 2017.