Correlation Impact by Random Forest Towards Prediction of Phishing Website

Ainin Mardhiah Ishak¹, Aida Mustapha¹, Syed Zulkarnain Syed Idrus², Mohd Helmy Abd Wahab³, Salama A Mostafa¹

¹ Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, 86400 Parit Raja, Batu Pahat, Johor, Malaysia
² Faculty of Applied and Human Sciences, Universiti Malaysia Perlis, Jalan Wang Ulu Arau, Sg. Chuchuh, 01000 Kangar, Perlis
³ Faculty Electrical and Electronic Engineering, Universiti Tun Hussein Onn Malaysia, Parit Raja 86400 Batu Pahat, Johor, Malaysia

E-mail: aidam@uthm.edu.my

Abstract. Phishing is an online identity theft that lure unsuspecting victims into phishing website that asks for their personal information such as financial credentials or online access. The most obvious phishing attacks use spam emails to lure victims into visiting its phishing website. The less obvious attack is when a phisher spoofes legitimate sites to bring the victims to the phishing website by presenting a visually similar websites. To address this issue, this paper focuses on the effectiveness of the Random Forest algorithm in predicting a website whether a phishing or legitimate website. The prediction model is developed using the classification methodology and the results revealed that the prediction accuracy of Random Forest in only 66.66%. This is due to potential high correlation among the features when training the dataset since each Random Tree in the forest should protect each other from individual errors. In the future, this project is hoped to investigate features selection of the same dataset because the phishing website prediction model requires features that at least some predictive power.

1. Introduction

The Internet has long empowered us in many daily social and financial activities such that much of our personal information are scattered in different webpages that belong to different organizations. This makes us vulnerable to various types of possible web threats that can lead to identity theft, loss of personal information, loss of financial wealth or damage to brand reputation due to increasing fraud cases. In 2004 alone, more than 57 million users in the United States received phishing and phishing scams nearly 2 million of them have become victims of the attack [1]. Phishing is considered as a form web threats where a the web pages of a legitimate organization are attacked by phishers aimed at secret information such as user name, password and financial ID. The attack manipulate the users into believing the phishing website as a legitimate website [2]. Email may appears to come from a legitimate company to the victim who asked they update or verify their information when victim visit the link in e-mail is the starting of phishing attacks. Phisher is now using different one technique of creating websites to trick users and flirt they are, but all of them use a common features set to design phishing website.
Predicting a website whether it is a phishing website or a legitimate website is made possible via data mining approach. Data mining is one of the powerful techniques that can help researchers focus on the most important information in their data. Data collection is divided into several methods such as classification, anomaly detection, clustering, regression or summarization [3]. Each technique has its own advantages and methods in predicting large datasets. One approach to counter a phishing attack would be at the email level, since most current phishing attacks use spam emails to trick victims to a phishing website. The less obvious attack is when a phisher spoofes legitimate sites to bring the victims to the phishing website by presenting a visually similar websites. This is the main reason why detection of phishing website is necessary to protect the users from potential loss and harm.

The literature has shown a number of prediction works on phishing website using data mining approach such as the structural web mining (focusing on web links) [4], clustering approach [5, 6] or associative classification [7, 8]. This paper focuses on building a prediction model for phishing website using Random Forest that is able to classify a website into phishing or legitimate website based on various features. The remaining of this paper proceeds as follows. Section 2 presents the materials and methods of building the website phishing model including the algorithm and dataset along with the evaluation metrics. Section 3 presents the results and finally Section 4 concludes with some direction for future work.

2. Material and Method
According to [9], the most popular applied data mining technique is classification, whereby this method employs a set of pre-determined classes in building the classification model. This model, will in turn, be able to classify a large-sized dataset. The tasks of classification are organized frequently whether in binary or multiclass. The phishing website prediction model proposed in this paper is based on a classification methodology with Random Forest algorithm. The experiment is a binary prediction with two classes, which are legitimate vs. phishing. It will be performed by using hold-out validation method with 80% training and 20% testing data. Figure 1 shows the classification methodology in Azure Machine Learning Studio (https://studio.azureml.net/).

Figure 1. Classification Methodology
2.1. Dataset

The dataset used in this project is sourced from the UCI website at https://www.kaggle.com/akashkr/phishing-website-dataset/version/2. The data is intended to be used for experimental verification throughout the project. Figure 2 shows the excerpt of the dataset.

![Figure 2. Excerpt of Dataset](image)

There are 32 attributes and 11,055 instances included with various important features such as having IP Address, Request URL and Abnormal URL. Due to the limited space in this paper, only five features are described in detail as follows.

2.1.1. Using IP Address: IP address is used as an alternative of the domain name in the URL, such as http://125.98.3.123/fake.html, users can be sure that someone is trying to steal their personal information. Sometimes, the IP address is transformed into hexadecimal code as shown http://0x58.0xCC.0xCA.0x62/2/paypal.ca/index.html. The rule is as follows:

```
if Domain Part has an IP Address then
    phishing
else
    legitimate
end if
```

2.1.2. Long URL: Long URL can be used by phishers in hiding the doubtful part in the address bar. It is calculated by the length of URLs in dataset and produced an average URL length. If the length of the URL is greater than or equal 54 characters then the URL classified as phishing were shown in the results. The rule is as follows:

```
if URL length < 54 then
    legitimate
else if URL length >= 54 and <= 75 then
    suspicious
else
    phishing
end if
```

2.1.3. Having “@” Symbol: This symbol in the URL leads the browser to ignore everything. The rule is as follows:
if URL Having “@” Symbol then
    phishing
else
    legitimate
end if

2.1.4. Domain Registration Length: The rule of thumb is that a phishing website lives for a short period of time, which can be formalized as follows:

if Domain expires <= 1 year then
    phishing
else
    legitimate
end if

2.1.5. Favicon: If the favicon is load from a domain other than that shown in the address bar, then the webpage is likely to be considered a phishing attempt. The rule is as follows:

if Favicon loaded From external domain then
    phishing
else
    legitimate
end if

Other than the five features described, more features are available in the dataset such as HTTPS Token, Popup Window, DNS Record, Page Rank, and Google Index. A phishing website usually add the “HTTPS” token in domain as part of the URL or has no DNS records for the domain or has a Page Rank < 0.2 or it is not indexed in Google due to its short lifespan.

2.2. Algorithms

Random forest is chosen for the data mining algorithm of this study. This algorithm is an ensemble learning technique that combines the decision trees rather than using the prediction of a single decision tree of classification [10]. Classification of all trees in random forest will provide a class to an unknown samples (test samples) and the class having maximum votes will be assigned to the unknown sample. The Random Forest algorithm is defined as an ensemble of $n$ classifiers $h_1(x), h_2(x), \ldots, h_n(x)$ that consisted of $x$ attributes. The pseudocode for Random Forest is shown in Algorithm 1 [11].

Algorithm 1 Pseudocode for Random Forest

Set $k = 1$
for each sample $i$ in bootstrapped dataset do
    Grow an unpruned classification tree $h_i(x)$ for sample $i$
    for each node in the classification tree $h_i(x)$ do
        Randomly sample $k$ of the predictor variables
        Choose the best split from among those variables
    end for
end for
Predict new data by combining the predictions of the trees
Calculate summary statistics and variable importance

One thing to note is that Random Forest is a classification algorithm that is aggregated by a number of decisions trees with the idea that prediction by a group of trees is better than
prediction of a single tree. The key to the success of Random Forest algorithm is the low correlation between the individual decision tree model because uncorrelated trees protect each other from their errors. This means all the individual decision tree models are trained on different testing set and well as different features when building the model.

2.3. Evaluation Metrics

The evaluation metrics used in measuring the performance of Random Forest algorithm in predicting a phishing website are accuracy, precision, and recall, which are defined as follows.

- Positive (P): A phishing website is detected.
- Negative (N): A legitimate website is detected.
- True Positive (TP): Website is a phishing web, and is predicted to be phishing web.
- False Negative (FN): Website is a phishing web, but is predicted to be legitimate web.
- True Negative (TN): Website is a legitimate, and is predicted to be legitimate web.
- False Positive (FP): Website is a legitimate web, but is predicted to be phishing web.

Based on this definition, the specific formula for accuracy, precision, and recall are defined in Equation 1 to Equation 4.

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}
\]

\[
\text{Precision} = \frac{TP}{TP + FP} \tag{2}
\]

\[
\text{Recall} = \frac{TP}{TP + FN} \tag{3}
\]

\[
\text{Recall} = \frac{2TP}{2TP + FP + FN} \tag{4}
\]

3. Results and Discussion

Table 1 shows the evaluation metrics produced after modeling the phishing detection website.

| Table 1. Confusion Matrix |
|---------------------------|
|               | Phishing | Legitimate |
| Phishing       | 802      | 38         |
| Legitimate     | 357      | 670        |

Figure 3 shows the Receiver Operating Characteristic (ROC) curve that illustrates the diagnostic ability of Random Forest. An ROC space is defined by False Positive Rate at the \(x\) axis vs. the True Positive Rate at the \(y\) axis. The space shows a relative trade-off between true positive (a phishing website successfully detected as a phishing website) and false positive (a legitimate website but detected as a phishing website). In this figure, the diagonal divides the ROC space. Points above the diagonal line represent classification results that are better than random, while points below the line represent results that are worse than random. From this curve, we can expect that the closer the curve result is to the upper left corner, the better it predicts. However, the distance from the diagonal line (random result) already shows the predictive power of Random Forests. The details of the prediction results are shown in Table 2. Based on the results, we can see that the performance of Random Forests is only 65.6% accuracy.
Figure 3. ROC Curve

Table 2. Prediction Results of Web Phishing

| Algorithm      | Accuracy | Precision | Recall | F1 Score |
|----------------|----------|-----------|--------|----------|
| Random Forest  | 0.666    | 0.692     | 0.677  | 0.685    |

4. Conclusion
Prediction of phishing websites is essential in order to protect Internet users from becoming phishing victims as well organizations from losing their brand and customers. This paper presented a website phishing model using the Random Forest, a data mining classification algorithm that is able to automatically classify phishing pages. This project has proposed and evaluated a phishing detection model to detect in terms of accuracy. The performance of 67% accuracy indicates potential areas to be explored in the feature space such that they have more predictive power in the future.

References
[1] Dhamija R, Tygar J D and Hearst M 2006 Why phishing works Proceedings of the SIGCHI conference on Human Factors in computing systems pp 581–590
[2] Mohammad R M, Thabtah F and McCluskey L 2012 An assessment of features related to phishing websites using an automated technique 2012 International Conference for Internet Technology and Secured Transactions (IEEE) pp 492–497
[3] Hussain S, Atallah R, Kamsin A and Hazarika J 2018 Classification, clustering and association rule mining in educational datasets using data mining tools: A case study Computer Science On-line Conference (Springer) pp 196–211
[4] Jain A K and Gupta B B 2018 International Journal of Information and Computer Security 10 398–417
[5] El-Alfy E S M 2017 The Computer Journal 60 1745–1759 ISSN 0010-4620 (Preprint https://academic.oup.com/comjnl/article-pdf/60/12/1745/22297461/bxx035.pdf) URL https://doi.org/10.1093/comjnl/bxx035
[6] Hamid I R A, Samsudin N A, Mustapha A and Arbaiy N 2017 Dynamic trackback strategy for email-born phishing using maximum dependency algorithm (mda) Recent Advances on Soft Computing and Data Mining ed Herawan T, Ghazali R, Nawi N M and Deris M M (Cham: Springer International Publishing) pp 263–273 ISBN 978-3-319-51281-5

[7] Hadi W, Aburub F and Alhawari S 2016 Applied Soft Computing 48 729–734

[8] Abdelhamid N, Ayesh A and Thabtah F 2014 Expert Systems with Applications 41 5948–5959

[9] Ahmed A B E D and Elaraby I S 2014 Data mining: A prediction for student’s performance using classification method

[10] Breiman L 2001 Machine learning 45 5–32

[11] Yin H, Yao X, Tino P, Corchado E and Byrne W 2007 Intelligent Data Engineering and Automated Learning-IDEAL 2007: 8th International Conference, Birmingham, UK, December 16-19, 2007, Proceedings vol 4881 (Springer)