Performance of Several Machine Learning Algorithms on Predicting Phishing Websites.

Sean M Kennedy[1], Ryan Pavsek[1]

1 Department of Electrical Engineering and Computer Science, University of Cincinnati
812 Rhodes Hall, 2851 Woodside Drive, Cincinnati, OH 45219

Abstract. The goal of this project is to determine the performance and practicality of using machine learning algorithms to detect spoofed and malicious websites used in phishing attacks. To this end we analyzed the performance of four machine learning classification algorithms - Gaussian Naïve Bayes, Support Vector Machine, Random Forest, and Multi-Layer Perceptron on the Website Phishing Data Set (found here https://archive.ics.uci.edu/ml/datasets/phishing+websites). In this project, we show which of the algorithms tested performed the best and comment on the practicality of using machine learning algorithms in detecting phishing websites.

Introduction and Motivation

Some of the most pervasive web-based attacks in recent years have been due to malicious URL links included in emails. These emails can appear to be sent from a legitimate source such as a bank or online store. Once the victim clicks on a link they are directed to a site which also appears legitimate that requests sensitive information, or infects the victim's device with malware. This attack is a form of phishing and according to the United States Federal Bureau of Investigations, costs American businesses upwards of 500 million dollars every year [1]. Some recent high-profile examples include the Yahoo.com data breach that occurred in late 2014 that led to over 500 million affected users. According to the FBI this massive breach began with a single phishing email sent to a Yahoo employee in early 2014 [2]. Other notable victims of targeted email phishing include the Australian Aerospace Company FACC, US-based retailer Target and the security firm RSA Security LLC.

Our motivation for choosing this topic for our project is partly due to the impact email phishing has on a breadth of victims ranging from individuals to multinational corporations to government entities. It is an attack that everyone is subjected to on a regular basis. The cost to an individual can equate to identity theft, which could lead to devastating financial harm. Corporations affected by these attacks can also suffer not only the embarrassment of a breach but financial fallout as well. Importantly, governments can lose sensitive information impacting national security. Website phishing attacks are rising in number and lead to more monetary theft every year [1]. Obviously, better solutions to protect against these attacks are sorely needed.

Some challenges in this project include a lack of email phishing datasets publicly available. After some searching, we found an appropriate dataset on the Machine Learning Repository maintained by the Center for Machine Learning and Intelligent Systems at the University of California, Irvine. Also, there is a challenge in determining what features of a website can be used to effectively predict if it is a phishing website or not. This is the subject of ongoing research and falls outside the scope of this project.

Basic Approach and Setup

The approach we took was to apply four machine learning classification algorithms, Gaussian Naïve Bayes, Support Vector Machine, Random Forest, and Multi-Layer Perceptron[1] on a dataset of websites containing legitimate and phishing websites with the goal of creating classifier models to categorize a website as a phishing website or a legitimate website. A Random Forest classifier fits a number of decision tree-based classifiers to subsets of the training data and uses averaging to mitigate over-fitting and increase the accuracy of prediction [3].

The dataset we used in this project was obtained from previously mentioned Machine Learning Repository website [4] and was provided by Mohammad et. al [5]. This dataset contains 2456 instances with 30 attributes. The data was collected from the PhishTank archive, the MillerSmiles archive and Google's searching operators. Attributes include URL properties, whether HTTPS is used, domain registration length, using a non-standard port, redirects, disabling right clicks, etc. A complete description of the dataset can be found in Mohammad et. al [5]. The features were categorized as -1 indicating 'phishy', a 0 for 'suspicious', and a 1 for

1 We have covered Naïve Bayes, Support Vector Machine and Perceptron in class, so we will not explain the details of these algorithms.

‘legitimate’. The websites known to be phishing websites are categorized with a -1 and legitimate websites are categorized with a 1. The file type of the dataset is Attribute-Relation File Format (.arff). We used Python with the machine learning library scikit-learn (which is built on the NumPy and SciPy libraries) to implement the algorithms, the matplotlib library for generating figures, and the liac-arff library to read the .arff file into Python. These libraries can be installed using pythons built in package installer *pip* using the command ‘pip install 'library_name''.

Results

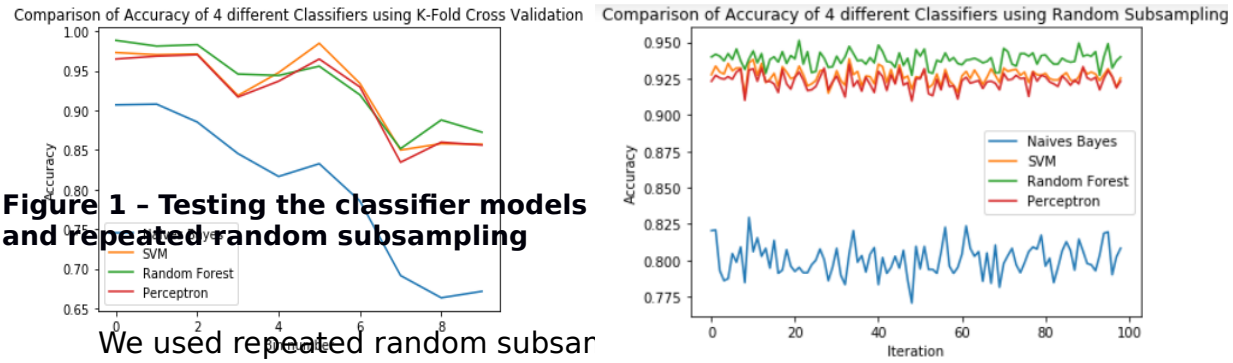

Figure 1 – Testing the classifier models and repeated random subsampling

We used repeated random subsar

training and 20 percent testing) over 100 iterations as well as k-fold cross validation with a k of 10 to train and test the classifiers. Both generated similar models, with repeated random subsampling producing the least standard deviation during testing. The algorithm that performs the best on accurately classifying phishing websites is the Random Forests Classifier. The Multi-Layer Perceptron and the SVM are not far behind and the Naïve Bayes Classifier performed the worst at 80 percent prediction accuracy. Training and testing the models generated from the 10-fold cross validation took 51.4 seconds and training and testing using the models generated from repeated random sub-sampling took 410.8 seconds.

Algorithm	**Mean Accuracy**	**Standard Deviation**

Naïve Bayes	80 %	8.9 %
Multi-Layer Perceptron	92 %	4.9 %
SVM	92.6 %	5 %
Random Forest	93.2 %	4.6 %

Table 1- Performance of 4 the classifiers using 10-fold cross validation

Algorithm	**Mean Accuracy**	**Standard Deviation**
Naïve Bayes	80 %	1 %
Multi-Layer Perceptron	92.3 %	0.5 %
SVM	92.7 %	0.5 %
Random Forest	93.8 %	0.5 %

Table 2- Performance of the 4 classifiers using repeated randomized subsampling

Future Directions and Conclusions

Another promising method used in detection of phishing websites is a 'layered' approach. This method combines blacklisting known phishing websites based on prior knowledge and classifying unknown phishing websites based on their features into one unified framework. By combining these two techniques, such a framework would lower the percentage of false positives generated by classification models while supplementing the amount of blacklisted phishing websites. One implantation of this approach is CANTINA+ by Zhang et. al [6].

With the recent uptick in website phishing attacks and their impact, a resolution is necessary to mitigate their effect. Fortunately, it seems that email filtering techniques are improving their prediction of spoofed emails and malicious links. However, these methods are not doing enough to stop the 500-million-dollar theft yearly from businesses, identity theft of individuals, and the negative consequences to governments. A more complete solution appears to be necessary to solve this problem.

References

[1] "PSA FBI Phishign.pdf." [Online]. Available:

https://www.ic3.gov/media/2017/170504.aspx#fn3.

[2] U.S. District Court Northern District of California, “United States of America v Dimitry Dokuchaev,” 2017.

[3] “SKLearn Doc.” [Online]. Available: http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html.

[4] “Phishing Dataset.” [Online]. Available: https://archive.ics.uci.edu/ml/datasets/phishing+websites.

[5] Mohammad, R. M., Thabtah, F., & McCluskey, L. (2014). Intelligent rule-based phishing websites classification. IET Information Security, 8(3), 153-160. Retrieved from https://search-proquest-com.proxy.libraries.uc.edu/docview/1558846544?accountid=2909

[6] G. Xiang, J. Hong, C. P. Rose, and L. Cranor, “Cantina+,” *ACM Trans. Inf. Syst. Secur.*, vol. 14, no. 2, pp. 1–28, 2011.