IJITCE

ISSN 2347- 3657

International Journal of
Information Technology & Computer Engineering

Email : ijitce.editor@gmail.com or editor@ijitce.com

International Journal of
Information Technology & Computer Engineering

ISSN 2347–3657
Volume 12, Issue 1, 2024

DETECTION OF PHISHING WEBSITE USING MACHINE LEARNING

Dr. N C Sendhil Kumar, Professor, Department Of ECE SICET ,Hyderabad
R.Vani, M.Nikhitha, S.Shiva Nayak, P.Kusuma
UG Student, Department Of ECE, SICET ,Hyderabad

ABSTRACT

Phishing attacks are the easiest way to obtain sensitive information from innocent users. Phishers aim to capture sensitive information such as usernames, passwords and banking information. Cyber warfare In this article, various legitimate Machine learning techniques for identifying phishing URLs by capturing and analyzing URLs and phishing URLs are examined. Decision trees, random forests, and support vector machine algorithms are used to detect phishing websites. The purpose of this article is to identify phishing URLs and narrow down the best machine learning algorithms by comparing the accuracy, vulnerability, and vulnerability of each algorithm.

Keywords:
Phishing attack, Machine learning

1. INTRODUCTION

Phishing has become a major focus of security researchers today because it is not easy to create fake websites that appear close to being legitimate. Professionals can detect fake websites, but not all users can and these users fall victim to phishing attacks. The killer's main goal is to steal bank credentials. In the US economy, $2 billion is lost every year due to lost customers who fall victim to phishing [1]. The third Microsoft Compute Security Index, published in February 2014, estimated that the annual global impact of phishing could be as high as $5 billion[2]. Phishing attacks have been successful due to lack of awareness among users. Phishing attacks are difficult to mitigate because they exploit user vulnerabilities, but it is important to improve phishing detection techniques.A way to identify phishing websites by updating blacklisted URLs into Internet Protocol (IP) antivirus databases, also known as the "blacklist" method. Attackers use effective strategies to trick users to evade blacklists, changing URLs to be clear through obfuscation, and many other simple strategies: fast streaming, where proxies are created to host web pages; algorithms new URLs etc. creates. The biggest disadvantage of this method is that it cannot detect zerohour phishing attacksHeuristicbased detection, which involves detecting signatures of reallife phishing attacks, can detect zeroday phishing attacks, but there is no guarantee that the attack will occur and the heuristic is false. a lot [3].To overcome the shortcomings of blacklists and heuristics, many security researchers are now focusing on machine learning. Machine learning has many algorithms that require historical data to improveDecisions or predictions about future information. Using this technology, algorithms will analyze various blacklisted and legitimate URLs and their characteristics to accurately identify phishing websites, including zero-day phishing websites

2. DATASET

URLs of harmless websites were collected from www.alexa.com, and URLs of phishing web sites were collected from www.phishtank.com. This database contains a total of 36,711 URLs, including 17058 harmless URLs and 19653 phishing URLs. Benign URLs are marked "0" and phishing URLs are marked "1".

3. FEATURE EXTRACTION

We used a python program to extract attributes from URLs. Below are the features we extracted to detect phishing URLs.

Presence of IP address in URL:

This property is set to 1 if the URL contains an IP address, and to 0 otherwise. Most malicious websites do not use IP addresses as URLs to download web pages. The use of an IP address in a URL indicates that an attacker is trying to steal sensitive information.

Presence of @ symbol in URL:

This property is set to 1 if the URL contains the @ symbol, and to 0 otherwise. Phishers adding the special symbol @ to the URL causes the browser to ignore everything before the - @ - symbol; The real address is usually after the - @â symbol [4].

Number of dots in Hostname

Phishing URLs

URL redirection

If there is – // – in the URL path, the function is set to 1, otherwise it is set to 0. The presence of - // - in the URL path indicates that the user will be sent to another website [4].

HTTPS token in URL

This property is set to 1 if the URL contains the HTTPS token, and to 0 otherwise. Phishers may add the "HTTPS" token to the domain name of the URL to mislead users. For example, http://https-www-paypal-it-mpp-home.soft-hair.com [4].

Information submission to Email:

Phishers can use the "mail()" or "mailto:" function to forward the user's information to their own email[4]. The property will be set to 1 if the URL contains such a function, otherwise it will be set to 0.

URL Shortening Services “TinyURL”:

The TinyURL service allows phishers to disguise long phishing URLs by shortening them. The aim is to redirect users to phishing websites. The property is set to 1 if the URL was created using a shortening service (such as bit.ly), 0 otherwise

Length of Host name:

The average length of benign URLs found is 25, this is set to 1 if the length of the URL is greater than

25, otherwise it is set to 0

Presence of sensitive words in URL:

Phishing websites use sensitive words in their URLs to trick users into thinking they are facing a legitimate web page. Words found in most of the phishing URLs are:- "confirm", "account", "bank", "security", "ebyisapi", "webscr", "login", "email", "setup", " Toolbar", "Backup", " Paypal", "Password", "Username" etc.

Number of slash in URL:

The number of slashes in a benign URL is considered 5; If the number of slashes in the URL is more than 5, the property is set to 1, otherwise it is set to 0.

Unicode in URL:

Phishers may use Unicode characters in the URL to trick users into clicking on the URL. For example, the domain –xn--80ak6aa92e.com – is equivalent to “Ð°ÑÑÓÐµ.com”. The URL shown to the user is "Ð°ÑÑÓÐµ.com", but clicking on this URL redirects the user to "xn--80ak6aa92e.com", a phishing website.

Age of SSL Certificate:

The presence of HTTPS is important to leave the impression that the website is legitimate [4]. However, for harmless websites, the minimum age of an SSL certificate is between 1 and 2 years.

URL of Anchor:

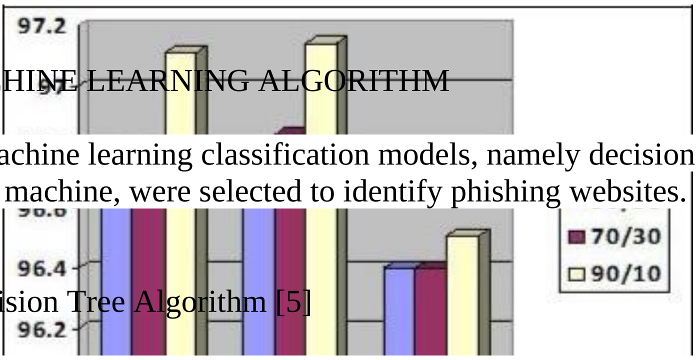
We extract this feature by accessing the source code and URL. The URL of the link is identified by the tag. This property is set to 1 if the tag has the maximum number of hyperlinks from other sources, and to 0 otherwise.

IFRAME:

We extract this feature by accessing the source code of the URL. This form is used to add another page to an existing home page. Phishers can use the "iframe" icon and make it invisible, i.e. no border [4]. Since the edges of the home page are not visible, users may think that the home page is also part of the home page and access sensitive information.

Website Rank:

We extract the website ranking and compare it to the top hundred thousand websites in the Alexa database. This property is set to 1 if the website rank is higher than 10.0000, otherwise it is set to 0.



4. MACHINE LEARNING ALGORITHM

Three machine learning classification models, namely decision tree, random forest, and support vector machine, were selected to identify phishing websites.

4.1 Decision Tree Algorithm [5]

It is one of the most used algorithms in machine learning. The decision tree algorithm is easy to understand and use. A decision tree starts by selecting the best discriminator among the available features, which is considered the root of the tree. The algorithm continues building the tree until it finds a leaf. Decision trees build training models that predict meaningful targets or clusters in tree representations, where each root of the tree is an attribute and each leaf of the tree is a member with a set of letters. Gini index and data climbing method are used to calculate nodes in the decision tree algorithm.

4.2 Random Forest Algorithm [6]

The random forest algorithm is one of the most important algorithms in machine learning. The random forest algorithm creates a forest consisting of many decision trees. The larger the tree, the more accurate the diagnosis.The creation of the tree is based on the bootstrapping method. In the bootstrapping method, a tree is created by randomly selecting and changing the features and structure of the data set. The random forest algorithm, just like the decision tree algorithm, selects the best discriminator of randomly selected features for classification; The random forest algorithm also uses the Gini index and data augmentation method to find the best distribution. This process will continue until the random forest produces n trees.Each tree in the forest predicts a target value and the algorithm then calculates the votes for each target prediction. Finally, the random forest algorithm determines the target prediction with the highest votes as the final prediction.

4.3 Support Vector Machine Algorithm [7]

Support vector machine is another powerful algorithm in machine learning. In the support vector machine algorithm, each data object is drawn as a location in n-dimensional space, and the support vector machine algorithm creates a dividing line, called a hyperplane, for the two groups.The support vector machine looks for the closest points, called support vectors, and when it finds the closest points, it draws a line connecting them. The support vector machine then creates a line bisecting the connecting line and perpendicular to it. The margin must be maximum for the product to be distributed properly. The margin here is the distance between the hyperplane and the support vectors. In real cases, it is not possible to separate data. To solve this problem, support vector machine uses kernel operation to convert low point to high point.

Fig. 1 Detection accuracy comparison

5. IMPLEMENTATION AND RESULT

Scikit-learn tool has been used to import Machine learning algorithms. Dataset is divided into training set and testing set in 50:50, 70:30 and 90:10 ratios respectively. Each classifier is trained using training set and testing set is used to evaluate performance of classifiers. Performance of classifiers has been evaluated by calculating classifier's accuracy score, false negative rate and false positive rate.

Table 1: Classifier's performance

Dataset Split ratio	Classifiers	Accuracy Score	False Negative Rate	False Positive Rate
50:50	Decision Tree	96.71	3.69	2.93
	Random Forest	96.72	3.69	2.91
	Support vector machine	96.40	5.26	2.08
70:30	Decision Tree	96.80	3.43	2.99
	Random Forest	96.84	3.35	2.98
	Support vector machine	96.40	5.13	2.17
90:10	Decision Tree	97.11	3.18	2.66
	Random Forest	97.14	3.14	2.61
	Support vector machine	96.51	4.73	2.34

The results show that compared with the decision tree and support vector machine algorithms , the random forest algorithm has a higher detection accuracy of 1 m (97.14) and the lowest er ror rate.The results also show that the detection accuracy of phishing websites increases as m ore data is used as training data. All classes performed well when 90% of the data was used a s training data.Fig. Figure 1 shows the accuracy of each classification when 50%, 70%, and 9 0% of the data are used as reference data, and the figure clearly shows that the detection accu racy increases when 90% of the data is used for training. The dataset and random forest searc h accuracy is better than the other two classifiers.

6. CONCLUSION

This article is designed to improve the detection of phishing websites using machine learning techniques. Using the random forest algorithm, we achieved a detection accuracy of 97.14% with the lowest false positive rate. The results also show that the classifier provides better per formance when we use more data than the training data.In the future, hybrid technology will be used where machine learning's random forest algorithm and blacklisting process will be us ed to more accurately identify phishing websites.

7. REFERENCES

[1] Gunter Ollmann, “The Phishing Guide to Understanding and Preventing Phishing Attacks ,” IBM Internet Security
Systems, 2007.

[2] https://resources.infosecinstitute.com/category/ enterprise / phishing/phishing-environment/phishing-data-attack-stats/#gref

[3] Mahmoud Khonji, Youssef Irak, “Phishing Detection: IEEE Literature Survey and Andre w Jones, 2013
><br< b="" style="margin: 0px; padding: 0px;"></br<> >[4] Mohammad R., Thabtah F. McCluskey L., (2015)

Phishing Website Dataset Muaj: https://archive.ics.uci.edu/ml/datasets / Phishing+ Websites Visited in January 2016

[5] http://dataaspirant .com/2017/01/30/how-decision -tree-algorithm-works/

[6] http://dataaspirant.com /2017/05/22/random-forest-algorithm-machine- Learning/

[7] https http://www.kdnuggets.com/2016/07/ support -vector-machines -simple-explanation .html

[8] www.alexa.com

[9] www.phishtank .com