Detection of URL based phishing attacks using machine learning: A Survey

[1]Ms. Sophiya. Shikalgar, [2]Dr. S. D. Sawarkar, [3]Mrs. Swati Narwane
[1]Research assistant, [2]Principal, [3]Assistant Professor
Datta Meghe College Of Engineering

Abstract - A fraud effort to get sensitive and personal information like password, username, and bank details like credit / debit card details by masking as a reliable organization in electronic communication. It most of the time redirects the users to similar looking website as legitimate website. The phishing website will appear same as the legitimate website and directs the user to a page to enter personal details of the user on the fake website. The system administration is very important these days as any failure can be detected and solved instantly. The system administration also need to define rules and set firewall settings to avoid phishing attacks through URL. Researchers have been studying various machine learning algorithm in lines to predict and avoid phishing attacks. Through machine learning algorithms one can improve the accuracy of the prediction. The machine learning, no one algorithm works best for every problem, and it's especially relevant for supervised learning. Using a single machine learning algorithm will give us good accuracy to predict the phishing attacks but to get better accuracy we need something more. The proposed system predicts the URL based phishing attacks with maximum accuracy. We shall talk about various machine learning, the algorithm which can help in decision making and prediction. We shall use more than one algorithm to get better accuracy of prediction. The algorithms namely the Naive Bayes and Random forest are used in the proposed system to detect URL based phishing attacks. The hybrid algorithm approach by combining two of the mentioned algorithms will increase accuracy.

Keywords - Phishing, legitimate, URL, feature extraction, machine learning, applications, classification, approach, algorithm.

I. INTRODUCTION

Phishing is a configuration of fraud that happens when a hostile website pretends to be a legitimate website. This is done to gather crucial information of a user like the password, OTP, credit/ debit card numbers CVV etc. Currently there are many techniques and software available through which one can detect phishing attacks, also there are many anti phishing software which can block phishing websites. But still the attackers produce new and hybrid approaches to bypass these anti phishing software. Phishing is a deceitful approach that uses a fusion of technology with the social engineering and gathers all personal / sensitive and important information of the user as mentioned above. The phishers use these details to misuse the password, credit card number of a user in an electronic communication. [6]

Phishing imitates the characteristics and features of emails and makes it look the same as the original one. It appears similar to that of the legitimate source. The user thinks that this email has come from a genuine company or an organisation. This makes the user to forcefully visit the phishing website through the links given in the phishing email. These phishing websites are made to mock the appearance of an original organisation website. The phishers force user to fill up the personal information by giving alarming messages or validate account messages etc so that they fill up the required information which can be used by them to misuse it. They make the situation such that the user is not left with any other option but to visit their spoofed website. [8]

1.1. Reason to tackle Phishing

Phishing is a cyber crime, the reason behind the phishers doing this crime is that it is very easy to do this, it does not cost anything and it effective. The phishing can easily access the email id of any person it is very easy to find the email id now a day and you can sending an email to anyone is freely available across the world. These attakers put very less cost and effort to get valuable data quickly and easily. The phishing frauds leads to malware infections, loss of data, identity theft etc. The data in which these cyber criminals are interested is the crucial information of a user like the password, OTP, credit/ debit card numbers CVV, sensitive data related to business, medical data, confidential data etc. Sometimes these criminals also gather information which can give them direct access to the social media account their emails. [3]

1.2. Features Used for Phishing Domain Detection

A lot of software / approaches and algorithms are used for phishing detection. These are used at academic and commercial organisation levels. A phishing URL and the parallel page have many features which are different from the malignant URL. Let us take an example to hide the original domain name the phishing attacker can select very long and confusing name of the domain. This is very easily visible. Sometimes they use the IP address instead of using the domain name. On the other hand they can also use a shorter domain name which will not be relevant to the original legitimate website. Apart from the URL based feature of phishing detection there are many different features which can also be used for the detection of Phishing websites namely the Domain-Based Features, Page-Based Features and Content-Based Features. [16]

II. Types of Approaches in machine learning

To make predictions or decisions, machine learning creates a mathematical model using a sample data which can also be called as training data. These algorithms do not explicitly perform a task or programmed to make predictions. Machine learning provides systems the capability to learn and improve from experience automatically without any explicit program execution because it is an application of artificial intelligence. Machine learning pays attention on the development of the computer programs which accesses data and uses it for learning.

2.1. Supervised learning algorithms and Semi-supervised algorithms

Supervised learning algorithms uses a mathematical model of a set of data which have both the inputs and the desired outputs. This set of data is called as the training data, which have of examples. These examples usually have one or more inputs and the desired output, this is also mentioned as a supervisory signal. In **semi-supervised learning algorithms**, the desired output is missed by few of the training examples In the mathematical model, we represent the training example by an array and in matrix form is the training data. To predict the output related to new inputs supervised algorithms pursue a function through iterative optimisation. [20]

To flawlessly know the output for inputs which are not included in the training data, an optimal function which allows the algorithm is used. This improves the correctness of the output for a particular task. This algorithm involve regression and classification. These classifications are used only when we want the output to restricted value whereas the regression algorithm is used for numerical output which is in the range.[10]

2.2. Unsupervised learning algorithms

Unsupervised learning algorithms uses a dataset which has only input values, these inputs are in the form of structure for example it could be in a cluster of data points or a group of data.The unsupervised algorithm recognizes similarities in the data and behave based on the existence and non-existence of such simlarities in the dataset. The cluster is a set of observations of the dataset, these clusters are analysed in such a manner that the similar observations which are pre-defined data or criteria of a data form one cluster and remaining data form a different cluster. These are then evaluated and metric to find the outputs[15]

2.3. Reinforcement Learning is a type of **Machine Learning**, and thereby also a branch of **Artificial Intelligence**. It allows **machines** and software agents to automatically determine the ideal behaviour within a specific context, in order to maximize its performance.[20]

III. Classification Algorithms

Classification is the procedure followed to predict the class of the given dataset. These classes are defined or also known as targets/ labels or categories. Classification is a type of predictive modelling which maps the function from input to desired output. Let us take an example of spam detection in email to understand further the concept of classification. In spam detection in email service providers can be pointed as a classification problem. This will be a binary type of classification as only two classes namely, as spam and not spam are defined. A classifier deploys some data to learn how given input variables relate to the class. In this example, spam and non-spam emails have to be given as the training data. When the classifier is trained correctly, it can be used to detect a spam email. In classification the targets are also given along with the input data, hence this comes under supervised learning category.[19]

3.1. Classification algorithms

Depending on the application and nature of the dataset used we can use any classification algorithms mentioned below. As there are different applications, we can not differentiate which of the algorithms are superior or not. Each of classifiers have its own way of working and classification. Let us discuss each of them in detail. [5]

3.1.1. Naive Bayes Classifier

This classifier can also be called as Generative Learning Model. The classification here is based on Baye's Theorem, it assumes independent predictors. In simple words this classifier will assume that the existence of a specific feature in a class is not related to the existence of any other feature. If there is a dependency among the features of each other or on the presence of other features, all of these will be considered as an independent contribution to the probability of the output. This classification algorithm is very much useful to large datasets and is very easy to use. [14]

3.1.2. Logistic Regression

Logic regression is used for Predictive Learning Model. To determine output in this classifier, we use a statistical method to analyse the dataset. These data set can have one or more than one independent values. The output is calculated with a data in which there could be two outputs. The aim of this classification algorithm is to find the relationship between the dichotomous category and predictor variables.[6][14]

3.1.3. Decision Trees

This classification algorithm builds the regression models. These models are built in form of structure which is similar to tree - a tree like structure is created by this classifier. It keeps on dividing the data set into subsets and smaller subsets which develops an associated tree, incrementally. The decision tree is finally created which has decision nodes and leaf nodes. In this tree the leaf node will have details about the classification or the decision taken for classification whereas the decision will have branches. The highest decision node which will be at the top of the tree will correspond to the root node. This will be the best predictor. [3][14]

3.1.4. Random Forest

This classification algorithm are similar to ensemble learning method of classification. The regression and other tasks, work by building a group of decision trees at training data level and during the output of the class, which could be the mode of classification or prediction regression for individual trees. This classifier accuracy for decision trees practice of overfitting the training data set[8][14]

3.1.5. Neural Network

As the name suggests this classifier has units known as neurons, which are arranged in layers that convert the input vector to relevant output. Each single neuron takes an input, this is most often a non-linear input, this is given to a function which is them passed to the next layer to get the output. The input given to the first layer will act as an output for the next layer and so on, thus this classification algorithm follows a feed-forward method. But in this method there is no feedback to the previous layer, so weighting are also given to the signals passing through the neurons and the layers, these signals then are turned into a training phase this eventually then become a network to handle any particular problem.[2][14]

3.1.6. Nearest Neighbor

As the name suggests the nearest neighbour algorithm is based on the nearest neighbour and this classification algorithm is supervised. It is also called as k-nearest neighbour classification algorithm. A cluster of labeled points are used to understand how the other points should be labelled. For labelling a new point it checks the already labelled points which could be closest to the point to be labelled, i.e closest to the neighbour. In this way depending on the votes of the neighbour the new point is labelled the same label which most of neighbours have. In algorithm 'k' is the number of neighbours which are checked.[5][14]

3.1.7 Support vector machine (SVM)

This is also one of the classification algorithms which is supervised and is easy to use. It can used for both classification and regression applications, but it is more famous to be used in classification applications. In this algorithm each point which is a data item is plotted in a dimensional space, this space is also known as n dimensional plane, where the 'n' represents the number of features of the data. The classification is done based on the differentiation in the classes, these classes are data set points present in different planes. [5]

IV. Machine learning in detection of phishing attacks

In the training phase, we should use the labeled data in which there are samples such as phish area and legitimate area. If we do this then classification will not be a problem for detecting the phishing domain. To do a working detection model it is very crucial to use data set in the training phase. We should use samples whose classes are known to us, which means the samples whom we label as phishing should be detected only as phishing. Similarly the samples which are labeled as legitimate will be detected as legitimate URL. The dataset to be used for machine learning must actually consist these features.There so many machine learning algorithms and each algorithm has its own working mechanism which we have already seen in the previous chapter. The existing system uses any one of the suitable machine learning algorithms for the detection of phishing URL and predicts its accuracy. The existing system has good accuracy but it is still not the best as phishing attack is a very crucial, we have to find a best solution to eliminate this. In the currently existing system, only one machine learning algorithm is used to predict the accuracy, using only one algorithm is not a good approach to improve the prediction accuracy. Each of the algorithms which explain in the earlier chapter has some disadvantages hence it is not recommended to use one machine learning algorithm to further improve the accuracy. [10]

Once the model is trained it is very important to evaluate the classifier which we shall use and validate its capability. Now in the above section we have seen all the advantages and disadvantages of all the available classifier. Hence we propose to use more than one classifier that is we can use a combination of two classifiers to improve the accuracy further of prediction. We shall evaluate each of the classifiers and use Naive Bayes and Random forest, by using the combination mentioned in this section we shall improve the accuracy and make it better. After applying the classification the results are generated and the URLs are classified into phishing and legitimate URLs. The Phishing URLs are blacklisted in the database and the legitimate are white list in the database. [12]

V. Conclusion

According to the survey we have found out that phishing attacks is very crucial and it is important for us to get a mechanism to detect it. As very important and personal information of the user can be leaked through phishing websites, it becomes more critical to take care of this issue. This problem can be easily solved by using any of the machine learning algorithm with the classifier. We already have classifiers which gives good prediction rate of the phishing beside, but after our survey we conclude that it will be better to use a hybrid approach for the prediction and further improve the accuracy prediction rate of phishing websites. We recommend the use combination of the Naive Baye's and Random Forest classifiers with any of the above explained algorithm to be used to improve the accuracy prediction rate to detect a phishing URL. In future we can also use a combination of any other two or more classifier to give us the best results of prediction.

VI. References

[1] [Antonio J. Talln-Ballesteros, Simon James Fong, and Raymond Kwok-Kay Wong], An Empirical Study on Performance Server Analysis and URL Phishing Prevention to Improve System Management Through Machine Learning, 2019

[2] [Routhu Srinivasa Rao, Alwyn Roshan Pais], Jail-Phish: An improved search engine based phishing detection system, 2019

[3] [Yan Ding, Nurbol, Keqin, Wushour Slamu], A Keyword-based Combination Approach for Detecting Phishing Webpages, 2019

[4] [Samuel Marchal, Kalle Saari, Nidhi Singh and N. Asokan], Know Your Phish: Novel Techniques for Detecting Phishing Sites and their Targets, 2012.

[5] [Narendra. M. Shekokar, Chaitali Shah, Mrunal Mahajan, Shruti Rachh], An Ideal Approach for detection and Prevention of Phishing Attacks, 2015

[6] [Jigar Rathod, Prof. Debalina Nandy], Anti-Phishing Technique to Detect URL Obfuscation, 2014

[7] [Adnan Hodi, Jasmin Kevri, Adem Karadag], Comparison of machine learning techniques in Phishing website Classification, 2016

[8] [Purvi Pujara, M. B.Chaudhari], Phishing Website Detection using Machine Learning : A Review, 2018

[9] [Anand Desai,Janvi Jatakia, Rohit Naik, Nataasha Raul], Malicious Web Content Detection Using Machine Learning, 2017

[10] [Santhana Lakshmi V, Vijaya MS], Efficient prediction of phishing websites using supervised learning algorithms, 2012

[11] [Ankit Kumar Jain and B. B. Gupta], PHISH-SAFE: URL Features-Based Phishing Detection System Using Machine Learning, 2018

[12] [H.B. Kazemian, S. Ahmed], Comparisons of machine learning techniques for detecting malicious web pages, 2014

[13] [Jian Mao, Jingdong Bian, Wenqian Tian, Shishi Zhu, Tao Wei, Aili Li and Zhenkai Liang], Phishing page detection via learning classifiers from page layout feature, 2019

[14] [Rami M. Mohammad, Fadi Thabtah, Lee McCluskey], An Assessment of Features Related to Phishing Websites using an Automated Technique, 2012

[15] https://www.researchgate.net/publication/226420039-Detection-of-Phishing-Attacks

[16] https://www.proofpoint.com/us/threat-reference/phishing

[17] https://towardsdatascience.com/phishing-domain-detection-with-ml-5be9c99293e5

[18] https://en.wikipedia.org/wiki/Phishing

[19] https://www.techrepublic.com/article/how-to-tackle-phishing-with-machine-learning

[20] https://www.irjet.net/archives/V5/i3/IRJET-V5I3580.pdf