International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 07 | July 2020 www.irjet.net p-ISSN: 2395-0072

Detection of Phishing Websites based on Feature Extraction Using Machine Learning

Dr Anil GN[1], G Om Prakash[2], K Harsha Manoj[3], M Lokesh[4], Madhusudhan K M[5]

[1]Professor, [2,3,4,5]UG Student, Dept. of Computer Science & Engineering, BMS Institute of Technology & Management, Bengaluru, Karnataka, India

---

Abstract: *Phishing is one of the biggest threats in this era of the internet. Phishing is a smart mechanism where a legitimate website is cloned and victims are lured to a fake website to provide their personal as well as confidential information, sometimes it is expensive. Detection of the Phishing website is an intelligent and effective model based on the use of data mining algorithms for classification or association. However, the detection of phishing websites is a challenging task, as most of these techniques are unable to make a dynamically accurate decision as to whether the new website is phishing or legitimate. All rules and factors for grading phishing websites and the relationship between them were defined and characterized to detect by their efficiency, accuracy, number of rules produced and speed. They are the basis of these algorithms. We identify phishing websites using a combined approach by constructing resource description framework models and using group learning algorithms to classify websites. Our method is based on supervised learning techniques to train our program. This method has a very promising positive score, which is certainly appreciable. We also used a random forest classification to handle incomplete data sets as well as a software to eliminate features that allow quantifying the frequency of each task within the dataset. Our method of detecting by taking the URL's of the website as input helps us in better understanding of the features responsible for detection. The algorithm needs to be selected in such a way that it should result in better accuracy in most of the possible scenarios. As our system explores the strength of the Random Forest Algorithm and ensemble learning approaches, a promising accuracy is achieved.*

Keywords: **Data Preprocessing, Feature extraction, Machine learning, Phishing Detection, Random Forest, SVM, URL.**

1. Introduction

Phishing is a well-known act in which attackers capture users sensitive details by spoofing the websites or by tempting users to visit other false pages where their personal data is exposed available to the attackers, while they have been done inadvertently and innocently. Owing to the availability of various resources including online finance, entertainment, education, app uploading, and social networking, the Internet has explosively developed in recent years. In a web phishing attack, the attackers create phishing web sites which in order to gain their confidential financial and personal details, are identical to the legitimate web sites to confuse web users.

Analysis has found that attempts to divert phishing from 246.2 million attempts to 2017 to 2018 are up to 482.5 million in 2018. Such threats have incurred gross damage of $700 million. Initially, the phishing attack takes place by clicking on a link in emails. Victims should obtain an e-mail with an alert or confirmation connection. The Internet explorer should guide you to a page close to your initial one when you press this button on the intended victims. The attackers will instead capture the valuable network usage detail because personal information is requested to be entered on the phishing page. After phishing takes place, the attackers can ultimately execute financial theft. The effectiveness of the identification of phishing websites relies largely on the precision and timeliness of the acknowledgement of phishing websites. A variety of standard approaches for the identification of Phishing websites focused on black and whitelisting lists is proposed. Several sophisticated phishing approaches have been established to accurately forecast phishing websites, which are complementary alternatives to traditional Phishing website identification techniques.

In recent years, intelligent approaches for phishing websites based on supervised machine learning techniques have become widespread, smarter and more web- compared to traditional methods for phishing websites. Detection scheme for phishing pages with different features using support-vector computer data mining technology. Furthermore, the identification of Phishing Websites has used neural networks (NN), vector support (SVM), naïve bayes (NB), decision tree, random forest and other classification techniques. In general, the phishing websites used by two common approaches are the blacklist and whitelist, intelligent and heuristic approach. In these intelligent approaches, discriminatory features are selected manually or using statistical methods, which play an important part in increasing the classification efficiency.

2. Existing Techniques/ Systems

Many of the techniques for the detection of phishing websites have been introduced which prevented indeed many attacks but attackers have developed different techniques for collecting confidential information about the

victims. It can be related to the analogy of applying different forms of encryption techniques only because most people assume that you can decrypt it. Some of the techniques available include:

2.1 Black Listing Technique:

Blacklisting is the way to block access to certain suspicious websites. In this procedure, we list and prevent access to these websites by blocking suspicious websites.

2.2 White Listing Technique:

This is the procedure that may be related to the blacklist approach, but the difference is that we are making a list of legitimate websites and only those websites have been given access rights. Compared to the blacklist approach, this technique is preferable because there are chances of different phishing websites coming up and the blacklist approach becoming vulnerable.

2.3 Heuristics approach:

A website has many features which were responsible for phishing detection. So we select particular features that were considered main in this procedure and detect the phishing web site by training the dataset.

In this way, an approach based on heuristics is more efficient when compared with the approach to white and black listing. Here, we basically use the techniques of machine learning to classify a particular website as phishing, suspect or legitimate. There are many algorithms which are used to detect phishing. Of which algorithms based on decision trees are efficient and accurate. New features are detected, and algorithms are modified based on various approaches. The techniques of phishing detection suffer from low detection precision and high false alarm, particularly when new phishing methods are implemented.

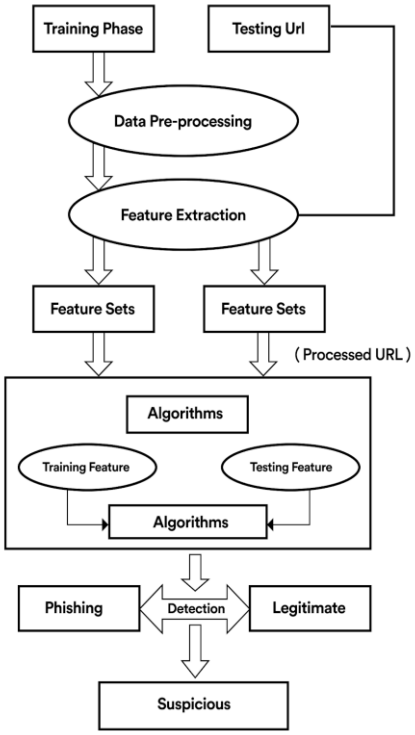
Blacklist method is not good enough for recognizing phishing attacks as registering the new domain is very easy and no blacklist will ensure the perfect up-to-date database. If we see further to oversee false-negative problems, page content inspection is used by different strategies. The algorithms of page content inspection have a variety of approaches the phishing website detections with different degrees of accuracy.

The problem with detecting a phishing website is that attackers constantly look for new ways to make the users believe that each of them is on a legitimate website. Phishers have been constantly improved to become a legitimate user by copying the original logos and being exactly as the original website. Phishers have also started to develop psychology behind their emails that playoff urgency, greed or trust. There are many algorithms which are used to detect phishing. Of which algorithms based on decision trees are efficient and accurate. New features are detected, and algorithms are modified based on various approaches. Many algorithms which were written are using different approaches and tried to achieve maximum accuracy using machine learning. Features of website detection play an important role and detection based on feature extraction has been challenging to maintain the maximum accuracy of detection.

3. Proposed Methodology

We seek to take the user input of the URL or consider a file containing the URL's and find whether it is a phishing or legitimate website by using random forest algorithm and other tools. The tools used will make all the detection process faster. We also seek to maintain the maximum possible accuracy of the algorithm used for better phishing detection. This study of phishing detection helps us understand the features and their importance behind phishing detection. The main reason behind taking the url as input is to understand the features better and compare the features between phishing and legitimate websites.

3.1 Architecture:



Machine learning techniques are used to detect phishing websites. The above representation is the architectural design representing the whole process. In the training phase, we pre-process the information to evacuate unneeded data, where it mainly focuses on removing the complex structures with attributes. The following is the feature extraction and it is of high importance for obtaining a better accuracy and better understanding of the website features that which are responsible for phishing websites detection.

3.2 Feature Extraction

Considering the software tools for feature extraction saves time as we are automating the manual process and hence improving the quality of phishing detection. All the features are categorized as by the effects on a website. Firstly, we examine if a webpage contains any text fields because a phishing webpage asks users to disclose their personal information through these fields. If the webpage has at least one text field we will continue to extract other features. Otherwise, the extraction process is terminated. For measuring the significance of features, we've collected datasets from the websites and using the tool we computed each feature frequency within the dataset in order to reflect the feature importance. By the ratio of the feature in the dataset, some weights will be given to the features. These frequencies will give us an initial indication of how influential is the feature of a website.

The featured sets are extracted as .csv files from the software tools and the model training will be done using the algorithms SVM and the random forest. The best-suited algorithm is taken into consideration and is used for the phishing website detection. Here, the user enters the URL of the suspicious website and results are interpreted. The feature sets are extracted after processing the dataset in the software tool that selects the attributes that are required to detect the phishing websites. We can select the attributes required to push to the dataset in prior using the tool.

Tool: A group of functionality features that can be extracted using our own software tool is distinct from previous researches. In predicting phishing websites, these features are evaluated using rules derived from different algorithms to reduce the false-negative rate by classifying phishing websites as legitimate. We have shown also that extraction functions are automatically faster than manual extraction, increasing the size of the dataset and allowing further experiments; improving the accuracy of the estimate.

Here based upon the efficient machine learning algorithm selected to detect the phishing website, the detection phase of the project is done and hence thereby resulting in the model accuracy as well as the output when the input URL is passed. We get the output in the form of a set of featured attributes with their respective values and hence detected based upon the output. Finally, based upon the featured attributes we tend to classify the websites into Legitimate, suspicious and phishing.

3.3 Algorithms

There are different algorithms used for classification such as random forest, SVM, KNN. But we have used the Random Forest algorithm.

Random forest can classify and give a 90% and above accuracy. This algorithm works by randomly generating a number of classification trees. Such trees are created by using different samples from the same dataset, and each time they build trees they can use different types of characteristics. Therefore, the different subsets of the same dataset generate the arbours at random, and the characteristics are often used to generate any tree alone. Instead, like the decision trees, Random Forest guarantees that the numbers do not over fit. After the trees are formed, we can classify each tree by finding the results and then assigning it to the class defined by the largest number of trees. For larger datasets, Random Forest is way better than SVM because if the dataset rows exceed 20,000 rows it tends to be unusable.

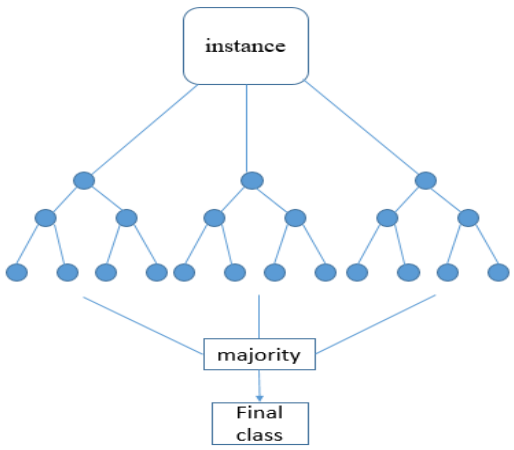
3.3.1 Random Forest

Random Forest can be easily said as a collection of decision trees. Where each decision tree which gives a yes or no value and finally by sending the data through different decision trees we will get the desired output. There are 2 ways for combining outputs of decision trees. They are:-

1. Bagging
2. Boosting

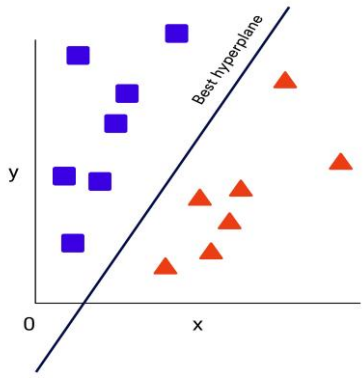
Bagging:- Different training data subsets are randomly drawn with replacement from the entire training dataset.

Boosting:- Every new subset contains the elements that were misclassified by previous models.



3.3.2 One Class SVM

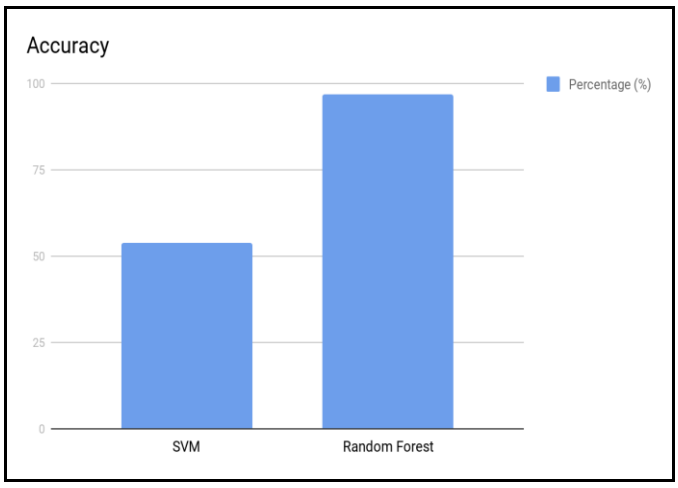
SVM is support vector machines which mainly classifies the given data into 2- group classification problems using algorithms. SVM comes under supervised learning model. Let's imagine we have two tags: *red* and *blue*, and our data has two features: x and y.



An SVM takes these data points and outputs the hyper plane (which is a line) that best separates the tags. This line is the decision boundary one side will be blue and the other side is red. Once the URL is fed to the system, the system extracts features such as number of visitors, number of pages visited by them. All the relevant features of the URLs are extracted which are used to differentiate between phishing URLs and legitimate URLs. The important consideration here is we ask the users to enter the URL as input. Once the URL is fed to the system, the system extracts features such as number of visitors, number of pages visited by them. All the relevant features of the URLs are extracted which are used to differentiate between phishing URLs and legitimate URLs. The important consideration here is we ask the users to enter the URL as input.

4. Observations

Our interpretation of the detection of phishing websites represents the idea of considering the features of the websites and of classifying them as phishing and legitimate. The most preferred algorithms to detect phishing sites are SVM and Random Forest. Here, before we give the url as input, we run the algorithms to get the accuracy value. The accuracy of the random forest turned out to be around 97.10 % and the accuracy of one SVM class turned out to be around 50.16 %.



When we give the URL as input. Our implementation is in quite a way that the features that are eligible to be extracted from various servers are also taken into account. We get the output as the accuracy score and the attribute value of the features. On the basis of these standards, we identify websites as phishing or legitimate. The following is a list of attribute values for various websites that have been taken into account.

4.1 Table:

The table below shows few of the most prominent features to be considered in the role of identification of phishing websites.

Features	Legitimate 1	Legitimate 2	Phishing 1	Phishing 2
URL_Length	1	1	1	1
having_Sub_Domain	-1	-1	0	0
SSLfinal_State	-1	-1	-1	-1
URL_of_Anchor	-1	-1	-1	-1
Abnormal_URL	-1	-1	1	1
Iframe	1	1	-1	-1
webb_traffic	1	1	1	1
Links_pointing_to page	-1	-1	1	1
Results	**-1**	**-1**	**1**	**1**

The results column in the table shows whether the website is a phishing website (1) or a legitimate website (-1).

We could find drastic differences for a few essential aspects when checked on phishing and on legitimate websites where some of them are briefly discussed below and how we should recognise the particular feature as a sign of phishing.

4.1.1 Sub Domain and Multi Sub Domains

If the points are larger than two, it is classified as "Phishing," as multiple subdomains are provided. If not, we must add "Legitimate" to the feature if the URL has no subdomains. Here, the attribute having sub-domains has resulted in (-1) for legitimate websites since the URL has no sub-domains

but (0) for phishing websites indicating more than 2 sub-domains.

4.1.2 Frame Redirection

IFrame is an HTML tag used to display an additional webpage into one that is currently shown. Phishers can make use of the "iframe" tag and make it invisible i.e. without frame borders. In this regard, phishers make use of the "frameBorder" attribute which causes the browser to render a visual delineation. As shown above, the iframe attribute displays (1) on a legitimate website, which means that no Iframe is available, whereas the same attribute for a legitimate website shows (-1) as a legitimate website does not contain iframe.

4.1.3 Number of Links Pointing to Page

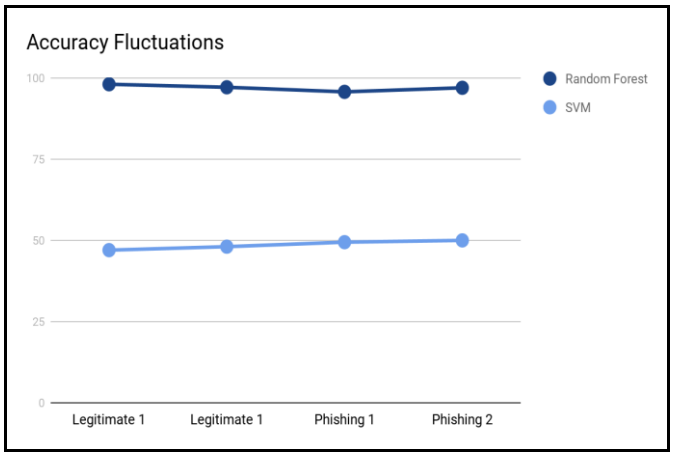
The number of links pointing to the webpage indicates its legitimacy level, even if some links are of the same domain. In our datasets and due to its short life span, we find that 98% of phishing dataset items have no links pointing to them. On the other hand, legitimate websites have at least 2 external links pointing to them. The "links pointing to a page" attribute shows (-1) for a legitimate website that shows there are no links, but for websites that indicate (1) that there are links to the website are for phishing websites.

4.1.4 Abnormal URL

This feature can be extracted from WHOIS database. For a legitimate website, identity is typically part of its URL. The abnormal url attribute indicates (-1) when the url is not present in the WHOIS database and (1) when it is present. Finding the accuracy of the random forest and the SVM after each iteration. We've seen these changes in accuracy. The remaining features in the table tend to have no specific differences when comparing both phishing and legal website outputs.

4.2 Graph showing the accuracy fluctuations of algorithms:

The graph below shows the accuracy fluctuations we came across when the programs are reiterated to extract the attribute values. The whole point in showing the accuracy variations is that random forest and svm algorithms when added to data sets with different data, do not lose their accuracy on a large scale. As a result, Random Forest accuracy will be around 96% to 97% in phishing detection. When considered for different websites, the accuracy of Random Forest for legitimate websites is 97.07% and 97.14%, while for phishing websites it is 96.73% and 96.98%. The accuracy of one SVM class for legitimate websites is 48.37% and 48.07%, while for phishing websites it is 48.45% and 49.19%.



We can therefore deduce that, in any case, the random forest accuracy is maintained and that it is better to detect phishing when compared to other algorithms. We've considered a few features and run the algorithm, which resulted in a decrease in accuracy. When considered to be the main features responsible, there are cases of high accuracy for some websites and also cases of low accuracy which have dropped to 86.24 percent using random forest. It is therefore better to consider maximum features of the website for phishing detection. The main reason is that hackers have unique attributes to attack users for different websites, so most features for detection of websites are always better considered.

Finally, we conclude that the manual detection of phishing websites will measure the degree to which new websites with the same features cannot be identified. This method of extracting features by giving input as URL allows us to understand the significance of the features for classification. In our research, we notice that random forest algorithm is better suited to detecting phishing websites than a single class SVM. The accuracy of the random forest is approximately 96-97%. Based on the attribute values, we have been able to detect features that play a key role in the detection of websites.

5. Conclusions

In this paper, we have tried to explain the significance and benefits of random forests over other algorithms. Therefore, random forests are compared to one class of SVM, which is far more precise than a single type of SVM. For the given URL as input we obtained the attribute values as output based on which we were able to compare both the phishing as well as the legitimate websites and identified some of the important features responsible for the phishing detection. This allows us to understand the combination of important features and how machine learning is used to detect them. The detailed use of random forests by introducing new details to the dataset allows us to understand that precision is maintained constantly and hence improved detection efficiency.

6. Future Work

Our work is based on the random forest machine learning algorithm. The further implementation can be done to improve the accuracy of the phishing detection by making certain modifications in the algorithm and also the important part is the feature selection. The features selection is the core of phishing detection as it makes the process more accurate. The further implementation can also be done by reducing the time taken to detect the phishing website by making necessary changes in the algorithm. Although there are certain procedures for selecting the features and improving the accuracy, the attackers always come up with the new undetectable features and certain modifications need to be done based upon the current scenarios of attack. Hence, the updates in improving the algorithm and feature selection should be done periodically to make sure that there would be no loss of important credentials.

REFERENCES

[1] Immadisetti Naga Venkata Durga Naveen, Manamohana K, Rohit Verma. (2019). Detection of Malicious URLs using Machine Learning Techniques.

[2] Ebubekir Buber,Önder Demir Ozgur Koray Sahingoz (2017)Feature Selections for the Machine Learning based Detection of Phishing Websites .

[3] Abdulhamit Subasi, Esraa Molah, Fatin Almkallawi, Touseef J. Chaudhery (2017) Intelligent Phishing Website Detection using Random Forest Classifier.

[4] Xun Dong, John A. Clark, Jeremy L. Jacob (2008). User Behaviour Based Phishing Websites Detection.

[5] Amani Alswailem, Bashayr Alabdullah, Norah Alrumayh, Dr.Aram Alsedrani. (2016). Detecting Phishing Websites Using Machine Learning.

[6] Hemali Sampat, Manisha Saharkar, Ajay Pandey, Hezal Lopes Detection of Phishing Website Using Machine Learning.

[7] Yasin Sönmez, Türker Tuncer(2017). Phishing Web Sites Features Classification Based on Extreme Learning Machine.

BIOGRAPHIES

Dr Anil G.N is Professor and HOD in the Department of Computer Science and Engineering at BMS Institute of Technology, affiliated to Visvesvaraya Technological University, Belgaum, KARNATAKA, INDIA. He has around 20 years of experience in teaching.

M Lokesh is currently pursuing a Bachelor of Engineering degree majoring in Computer Science and is in his final year at BMS Institute of Technology. His research interests are in the field of Cyber Security and Entrepreneurship.

K Harsha Manoj is currently pursuing a Bachelor of Engineering degree majoring in Computer Science and is in his final year at BMS Institute of Technology. His research interests are in the field of Cyber Security, Machine Learning and Cloud Computing.

G Om Prakash is currently pursuing a Bachelor of Engineering degree majoring in Computer Science and is in his final year at BMS Institute of Technology. His interest lies in the fields of Human Computer Interaction, Product designing, and Web development. He is looking forward to working in projects on Human Computer Interaction.

Madhusudhan KM is currently pursuing a Bachelor of Engineering degree in computer science and is in his final year at BMS institute of technology. His research interest are in field of Machine Learning and web development.