IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

Impact Factor: 6.252

Volume 2, Issue 8, May 2022

Prediction of Phishing using Machine Learning

Sivalakshmi S[1], Vichitra Devi M[2], Kalpana R[3], Dr. M. Preetha[4]

Students, Department of Information Technology[1,2]
Professor, Department of Information Technology[3]
Head of Department, Department of Information Technology[4]
Prince Shri Venkateshwara Padmavathy Engineering College, Chennai, India

Abstract: *Phishing is a common attack on credulous people by making them to disclose their unique information using counterfeit websites. The objective of phishing website URLs is to purloin the personal information like user name, passwords and online banking transactions. Phishes use the websites which are visually and semantically similar to those real websites. As technology continues to grow, phishing techniques started to progress rapidly and this needs to be prevented by using anti-phishing mechanisms to detect phishing. Machine learning is a powerful tool used to strive against phishing attacks. Phishing is popular among attackers, since it is easier to trick someone into clicking a malicious link which seems legitimate than trying to break through a computer's defense systems. The malicious links within the body of the message are designed to make it appear that they go to the spoofed organization using that organization's logos and other legitimate contents. Here, we explain phishing domain (or Fraudulent Domain) characteristics, the features that distinguish them from legitimate domains, why it is important to detect these domains, and how they can be detected using machine learning and natural language processing techniques. In this paper, we compared the results of multiple machine learning methods for predicting phishing websites.*

Keywords: Phishing, Personal information, Machine Learning, Malicious links, Phishing domain characteristics

I. INTRODUCTION

Nowadays Phishing becomes a main area of concern for security researchers because it is not difficult to create the fake website which looks so close to legitimate website. Experts can identify fake websites but not all the users can identify the fake website and such users become the victim of phishing attack. Main aim of the attacker is to steal banks account credentials. Phishing attacks are becoming successful because lack of user awareness. Since phishing attack exploits the weaknesses found in users, it is very difficult to mitigate them but it is very important to enhance phishing detection techniques. Artificial intelligence (AI) refers to the simulation of human intelligence in machines that are programmed to think like humans and mimic their actions. The term may also be applied to any machine that exhibits traits associated with a human mind such as learning and problem-solving. The general method to detect phishing websites by updating

Blacklisted URLs, Internet Protocol (IP) to the antivirus database which is also known as "blacklist" method. To evade blacklists attackers uses creative techniques to fool users by modifying the URL to appear legitimate via obfuscation and many other simple techniques including: fast-flux, in which proxies are automatically generated to host the web-page; algorithmic generation of new URLs; etc. Major drawback of this method is that, it cannot detect zero-hour phishing attack. Four different Machine Learning Algorithms such as Random Forest technique is used to get accuracy of each method, Decision Tree Algorithm, Naïve Baeyer's Algorithm and Logistic Regression is used

DOI: 10.48175/IJARSCT-4456

II. LITERATURE SURVEY

A. A Bio-Inspired Self-learning Co evolutionary Dynamic Multiobjective Optimization Algorithm for Internet of Things Services

Author: Zhen Yang, Yaochu Jin, Fellow, and Kuangrong Hao, Member

The ultimate goal of the Internet of Things (IoT) is to provide ubiquitous services. To achieve this goal, many challenges remain to be addressed. Inspired from the cooperative mechanisms between multiple systems in the human being, this paper proposes a bio-inspired self-learning co evolutionary algorithm (BSCA) for dynamic multiobjective optimization of IoT services to reduce energy consumption and service time. BSCA consists of three layers. The first layer is composed of multiple subpopulations evolving cooperatively to obtain diverse Pareto fronts. Based on the solutions obtained by the first layer, the second layer aims to further increase the diversity of solutions. The simulation results demonstrate that the proposed algorithm is competitive in dynamic optimization of agricultural IoT services. In practice, IoT service system may select one of the extreme solutions or other Pareto optimal solutions on the front according to the service strategy specified by the decision-maker.

B. **Title:** A Prediction Model of DoS Attack's Distribution Discrete Probability

Author: Wentao Zhao, Jianping Yin, Jun Long

The process of prediction analysis is a process of using some method or technology to explore or stimulate some unknown, undiscovered or complicated intermediate processes based on previous and present states and then speculated the results [5]. In an early warning system, accurate prediction of DoS attacks is the prime aim in the network offence and defense task. Detection based on abnormity is effective to detect DoS attacks. This paper begins with the relation exists between network traffic data and the amount of DoS attack, and then proposes a clustering method based on the genetic optimization algorithm to implement the classification of DoS attack data. This method first gets the proper partition of the relation between the network traffic and the amount of DoS attack based on the optimized clustering and builds the prediction sub-models of DoS attack.

C. **Title:** Adversarial Examples: Attacks and Defenses for Deep Learning

Author: Xiaoyong Yuan, Pan He, Qile Zhu, and Xiaolin Li

With rapid progress and significant successes in a wide spectrum of applications, deep learning is being applied in many safety-critical environments. However, deep neural networks (DNNs) have been recently found vulnerable to well-designed input samples called adversarial examples. Adversarial perturbations are imperceptible to human but can easily fool DNNs in the testing/deploying stage. The vulnerability to adversarial examples becomes one of the major risks for applying DNNs in safety-critical environments. Therefore, attacks and defenses on adversarial examples draw great attention. We investigated the existing methods for generating adversarial examples.10 A taxonomy of adversarial examples was proposed. We also explored the applications and countermeasures for adversarial examples. This paper attempted to cover the state-of-the-art studies for adversarial examples in the DL domain. Compared with recent work on adversarial examples, we analyzed and discussed the current challenges and potential solutions in adversarial examples.

D. **Title:** Apriori Viterbi Model for Prior Detection of Socio-Technical Attacks in a Social Network

Author: Preetish Ranjan, Abhishek Vanish

Social network analysis is a basic mechanism to observe the behavior of a community in society. In the huge and complex social network formed using cyberspace or telecommunication technology, the identification or prediction of any kind of socio-technical attack is always difficult. This challenge creates an opportunity to explore different methodologies, concepts and algorithms used to identify these kinds of community on the basis of certain pattern, properties, structure and trend in their linkage. This paper tries to find the hidden information in huge social network by

DOI: 10.48175/IJARSCT-4456

compressing it in small networks through apriority algorithm and then diagnosed using viterbi algorithm to predict the most probable pattern of conversation to be followed in the network and if this pattern matches with the existing pattern of criminals, terrorists and hijackers then it may be helpful to generate some kind of alert before crime.

III. MACHINE LEARNING ALGORITHM

Four algorithms have been implemented to check whether a URL is legitimate or fraudulent.

Random forest algorithm creates the forest with number of decision trees. High number of tree gives high detection accuracy. Creation of trees is based on bootstrap method. In bootstrap method features and samples of dataset are randomly selected with replacement to construct single tree. Among randomly selected features, random forest algorithm will choose best splitter for classification.

Decision tree begins its work by choosing best splitter from the available attributes for classification which is considered as a root of the tree. Algorithm continues to build tree until it finds the leaf node. Decision tree creates training model which is used to predict target value or class in tree representation each internal node of the tree belongs to attribute and each leaf node of the tree belongs to class label.

Naïve Bayer's The Naive Bayes algorithm is an intuitive method that uses the probabilities of each attribute belonging to each class to make a prediction. It is the supervised learning approach you would come up with if you wanted to model a predictive modeling problem probabilistically. Naive bayes simplifies the calculation of probabilities by assuming that the probability of each attribute belonging to a given class value is independent of all other attributes. This is a strong assumption but results in a fast and effective method.

Logistic Regression is a statistical method for analyzing a data set in which there are one or more independent variable that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes).

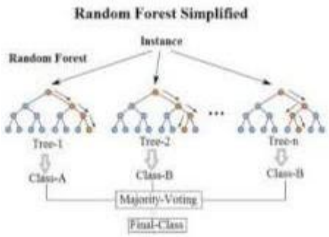


Fig: 1. Random Forest

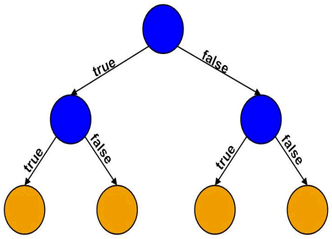


Fig: 2. Desicion Tree Algorithm

IV. PROJECT DESCRIPTION

We have developed our project using a website as a platformfor all the users. This is an interactive and responsive websitethat will be used to detect whether a website is legitimate or phishing. This website is made using different web designinglanguages which include HTML, CSS, JavaScript and Python. The basic structure of the website is made with the help of HTML. CSS is used to add effects to the website andmake it more attractive and user-friendly. It must be noted that the website is created for all users, hence it must be easyto operate with and no user should face any difficulty while making its use. Every naïve person must be able to use this website and avail maximum benefits from it. The website shows information regarding the services provided by us. It also contains information regarding ill practices occurring in today's technological world. The website is created with an opinion such that people are notonly able to distinguish between legitimate and fraudulentwebsite, but also become aware of the mal-practices occurring in current world. They can stay away from the people trying to exploit one's personal information, like email address, password, debit card numbers, credit card details, CVV, bank account numbers, and the list goes on.

V. FEATURE EXTRACTION

We have implemented python program to extract features from URL. Below are the features that we have extracted for detection of phishingURLs

1. **Server from Handler (SFH):** Request URL examines whether the external objects SFHs that contain an empty string or about: blank are considered doubtful because an action should be taken upon the submitted information.
2. **Pop up Window:** It is unusual to find a legitimate website asking users to submit their personal information through a pop-up window. On the other hand, this feature has been used in some legitimate websites and its main goal is towarn users about fraudulent activities or broadcast a welcome announcement, though no personal information was asked to be filled inthrough these pop-up windows. Rule: IF {Pop-up Window Contains Text Fields → Phishing Otherwise → Legitimate}
3. **SSL Final Certificate**: SSL is an acronym of secure socket layer. It creates an encryptedconnection between the web server and the user's web browser allowing for private information to betransmitted without the problems of eavesdropping. All legitimate websites will have SSL certificate. Butphishing websites do not have SSL certificate. The SSL certificate of a website is extracted by providing thepage address.
4. **Request URL:** contained within a webpage such as images, videos andsounds are loaded from another domain.
5. **URL of Anchor:** An anchor is an element defined by the <a> tag. Thisfeature is treated exactly as Request URL.
6. **Website Traffic:** This feature measures the popularity of the website bydetermining the number of visitors and the number ofpages they visit.
7. **URL Length**: Phishes can use a long URL to hide the doubtful part in theaddress bar.
8. **Age of Domain:** This feature can be extracted from WHOIS database. Most phishing websites live for a short period of time. By reviewing our dataset, we find that the minimum ageof the legitimate domain is 6 months.
9. **Presence of IP address in URL**: If IP address present in URL then the feature is set to 1 else set to 0. Mostof the benign sites do not use IP address as an URL to download a webpage. Use of IP address in URL indicates that attacker is trying to steal sensitive information.

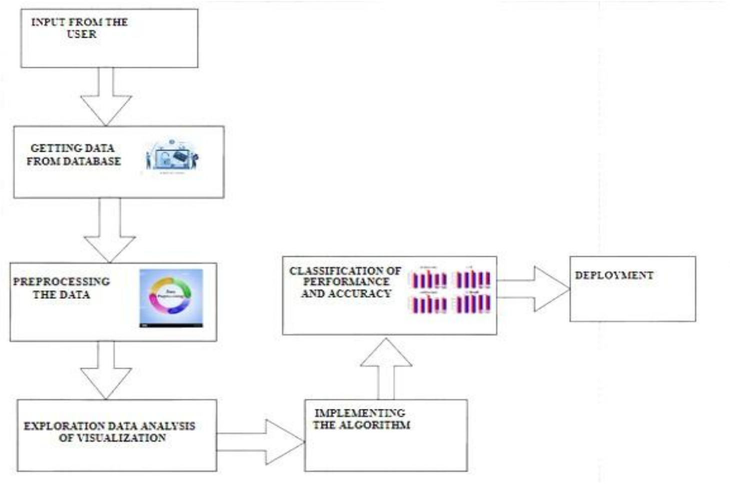
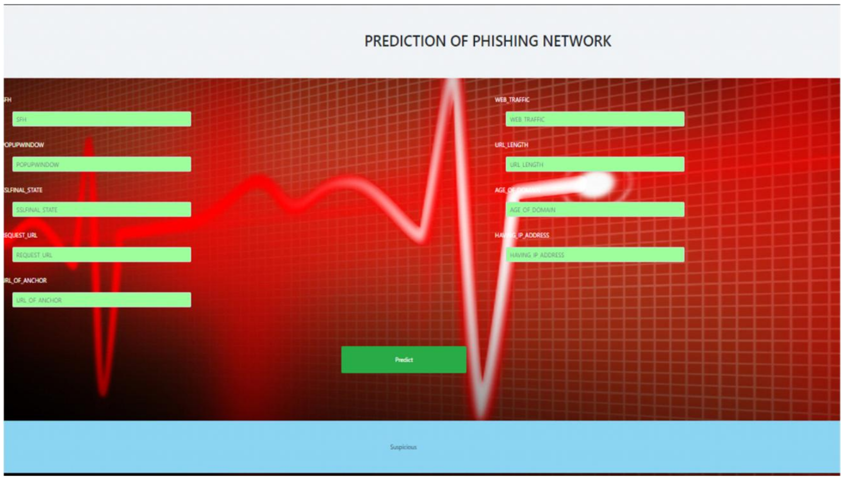

INPUT FROM THE USER
GETTING DATA FROM DATABASE
PREPROCESSING THE DATA
EXPLORATION DATA ANALYSIS OF VISUALIZATION
IMPLEMENTING THE ALGORITHM
CLASSIFICATION OF PERFORMANCE AND ACCURACY
DEPLOYMENT

Fig: 3: System Architecture

VI. CONCLUSION AND FUTURE SCOPE

The analytical process started from data cleaning and processing, missing value, exploratory analysis and finally model building and evaluation. The best accuracy on public test set is higher accuracy score will be finding out. This application can help to find the Prediction of phishing website or not. In future hybrid technology will be implemented to detect phishing websites more accurately, for which random forest algorithm of machine learning technology and blacklist method will be used. We have detected phishing websites using Random Forest algorithm with and accuracy of 97.31%.

For future enhancements, we intend to build the phishing detection system as a scalable web service which will incorporate online learning so that new phishing attack patterns can easily be learned and improve the accuracy of our models with better feature extraction.



Suspicious

Fig: 4: Prediction of Phishing Network

VII. ACKNOWLEDGEMENT

I am using this opportunity to express my gratitude to everyone who supported me throughout the course of research project report. We Would like to thank all our faculties, friends and family members who have directly and indirectly helped in easy completion of our project. Special thanks to Mrs. S. Kalpana, Associate Professor, Department of Information Technology, Prince Shri Venkateshwara Padmavathy Engineering College, Chennai, Tamil Nadu, India for guiding us in completing this project successful.

REFERENCES

[1]. Desai, J. Jatakia, R. Naik, and N. Raul, “Malicious web content detection using machine leaning,” RTEICT 2017 - 2nd IEEE Int. Conf. Recent Trends Electron. Inf. Commun. Technol. Proc., vol. 2018–Janua, pp. 1432–1436, 2018.

[2]. P. Prakash, M. Kumar, R. R. Kompella, and M. Gupta, “Phish Net: Predictive Blacklisting to Detect Phishing Attacks,” 2010.

[3]. Bradley Barth, “SOC teams spend nearly a quarter of their day handling suspicious emails,”https://www.scmagazine.com/home/email-security/soc-teams-spend-nearly-a-quarter-of-their-day-handling-suspicious-emails. 2021.

[4]. Crane Hassold, “Employee-Reported Phishing Attacks Climb 65%, Clobbering SOC Teams,” https://www.agari.com/email-security-blog/employee-reported-phishing-attacks-soc/. 2020.

DOI: 10.48175/IJARSCT-4456

[5]. A. Y. Fu, W. Liu, and X. Deng, “Detecting phishing web pages with visual similarity assessment based on Earth Mover’s Distance (EMD),” IEEE Trans. Dependable Secur. Comput, vol. 3, no.4, pp. 301–311, 2006, doi: 10.1109/TDSC.2006.50.

[6]. Neupane, N. Saxena, J. O. Maximo, and R. Kana, “Neural Markers of Cyber security: An fMRI Study of Phishing and Malware Warnings,” IEEE Trans. Inf. Forensics Secur., vol. 11, no.9,pp. 1970–1983, 2016, doi: 10.1109/TIFS.2016.2566265.

[7]. X. Zhang, Y. Zeng, X. Jin, Z. Yan, and G. Geng, “Boosting the Phishing Detection Performance by Semantic Analysis,” 2017.

[8]. L. MacHado and J. Gadge, “Phishing Sites Detection Based on C4.5 Decision Tree Algorithm,” in 2017 International Conference on Computing, Communication, Control and Automation, ICCUBEA 2017, 2018, pp. 1–5.

[9]. A. Desai, J. Jatakia, R. Naik, and N. Raul, “Malicious web content detection using machine leaning,” RTEICT 2017 - 2nd IEEE Int. Conf. Recent Trends Electron. Inf. Commun. Technol. Proc., vol. 2018–Janua, pp. 1432–1436, 2018.

[10]. https://www.researchgate.net/publication/355263255_Detecting_phishing_websites_using_machine_learning_technique

DOI: 10.48175/IJARSCT-4456