PREDICTION OF PHISHING WEBSITE USING MACHINELEARNING

Kavitha R,
Student, Computer, Science Engineering, Prince Dr. K. Vasudevan College of Engineering and Technology, Chennai, India
Priyanka K,
Student, Computer, Science Engineering, Prince Dr. K. Vasudevan College of Engineering and Technology, Chennai, India
***Anitha M**,*
Professor, Computer Science Engineering, Prince Dr. K. Vasudevan College of Engineering and Technology, Chennai, India
Deepa R
Head of Department, Computer Science Engineering, Prince Dr. K. Vasudevan College of Engineering and Technology, Chennai, India

Abstract: —Phishing attack is a simplest way to obtain sensitive information from innocent users. The Internet has become an indispensable part of our life, However, It also has provided opportunities to anonymously perform malicious activities like Phishing. Phishers try to deceive their victims by social engineering or creating mockup websites to steal information such as account ID, username, password from individuals and organizations. Although many methods have been proposed to detect phishing websites, Phishers have evolved their methods to escape from these detection methods.Phishes use the websites which are visually and semantically similar to those real websites. One of the most successful methods for detecting these malicious activities is Machine Learning. 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.This is because most Phishing attacks have some common characteristics which can be identified by machine learning methods.Decision Tree, random forest and Support vector machine algorithms are used to detect phishing websites.we compared the results of multiple machine learning methods for predicting phishing websites.

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

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.Phishing is a kind of Cybercrime trying to obtain important or confidential information from users which is usually carried out by creating a counterfeit website that mimics a legitimate website. Phishing attacks employ a variety of techniques such as link manipulation, filter evasion, website forgery, covert redirect, and social engineering.In this paper, machine-learning algorithms have been used for modelling the prediction task.Training the features of phishing and legitimate websites creates the learning model.

Third party servicessuch as blacklist, search engine that contributes more for the accurate prediction of the phishing websitesare included as one of the features that are used to identify the phishing websites.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.

II LITERATURE SURVEY

1. Predictive Black listing to Detect Phishing Attacks

Author: P. Prakash, ManishKumar, M.Gupta ,R. Kompella

Phishing has been easy and effective way for trickery and deception on the Internet. While solutions such as URL blacklisting have been effective to some degree, their reliance on exact match with the black listed entries makes it easy for attackers to evade. We start with the observation that attackers often employ simple modifications (e.g., changing top level domain) to URLs. PhishNet, exploits this observation using two components. In the first component, we propose five heuristics to enumerate simple combinations of known phishing sites to discover new phishing URLs. The second component consists of an approximate matching algorithm that dissects a URL into multiple components that are matched individually against entries in the blacklist. We also show that our approximate matching algorithm leads to very few false positives (3%) andnegatives.

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

Author: Zhen Yang, YaochuJin, 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 multi objective 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.

3. 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.

4. Phishing Website Prediction A Machine Learning Approach

Author: Anjaneya Awasthi & Noopur Goel

In this paper, machine learning techniques are used for prediction. Data mining is used world wide by almost every face of the society viz. business organizations, govt. organizations, and other kind of data collectors to extract knowledge from the collected data. On the other hand, machine learning is a data mining technique that is used to analyze, classify the data, and efficiently predict the results for the estimation and planning by all of the organizations all around the globe. Classification algorithms, namely logistic regression, decision tree, and random forest classification, are used to predict the fake websites and presented their comparison of their predictions achieved. The results have been presented in numeric format as well as graphically with the help of chart. The data used is taken from UCI machine learning online repository.

5. 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.

III MACHINE LEARNING ALGORITHM

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

Random forest 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 boot strap method features and samples of data set are randomly selected with replacement to construct single tree.Among randomly selected features, random forest algorithmwillchoose bestsplitter for classification.

Decision tree are used as a well-known classification technique. A decision tree is a flowchart-like tree structure where an internal node represents a feature or attribute, the branch represents a decision rule, and each leaf node represents the outcome. The topmost node in a decision tree is known as the root node. It learns to partition based on the attribute value. It partitions the tree in a recursive manner called recursive partitioning. This particular feature gives the tree classifier a higher resolution to deal with a variety of data sets, whether numerical or categorical data.

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

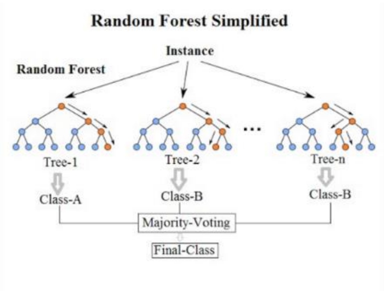
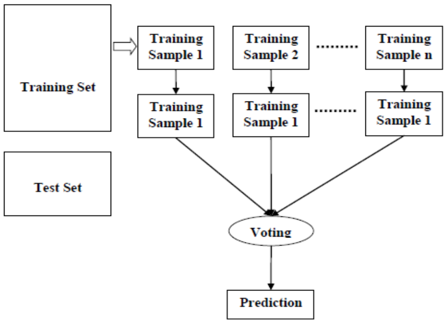


Fig 2 logistic regression



IV SYSTEM ARCHITECTURE

System Architecture is a generic discipline to handle objects (existing or to be created) called "systems" in a way that supports reasoning about the structural properties of these objects. The system architecture is a response to the conceptual and practical difficulties of the description and the design of complex systems

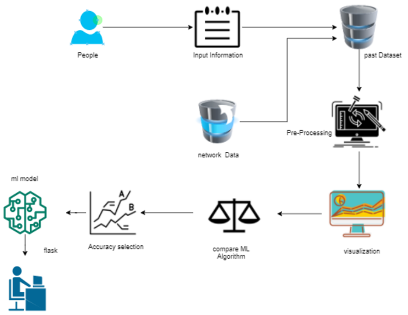


Fig 3 System architecture

V 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 orphishing.Random Forest Classifier and Support Vector Machine perform similar on the given dataset and have higher accuracy and lower FPR compared to other models. Apart from Logistic Regression and Categorical Naive Bayes, the performance of other models is also comparable.Support Vector Machine works well for linearly separable data. The data is not linearly separable directly, but after applying kernel, the data becomes separable and SVM is able to learn well from the data. The dataset which is for use for machine gaining knowledge of has to truly include these features. There is such a massive quantity of gadget learning algorithms and every set of rules has its own working mechanism whichuse are Hypertext Transport Protocol or HTTP (http), HTTP with Transport Layer Security (https), and File Transfer Protocol (ftp). This is the identifier for the net server on the web.

Sometimes it's a machine-readable Internet Protocol (IP) address, but more often especially from the user's perspective it is a humanreadable name.

V ACCURACY CALCULATION

General Formula:

F- Measure = 2TP / (2TP + FP + FN)

F1-Score Formula:

F1 Score = 2*(Recall * Precision) / (Recall + Precision)

VI CONCLUSION AND FUTURE SCOPE

This work, models the phishing website prediction as a classification task and demonstrates the machine learning approach for predicting whether the given website is legitimate website or phishing. Naïve Bayes classifier, Decision tree classifier,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 enhancement phishing website or not prediction to connect with cloud model. To optimize the work to implement in Artificial Intelligence environment.

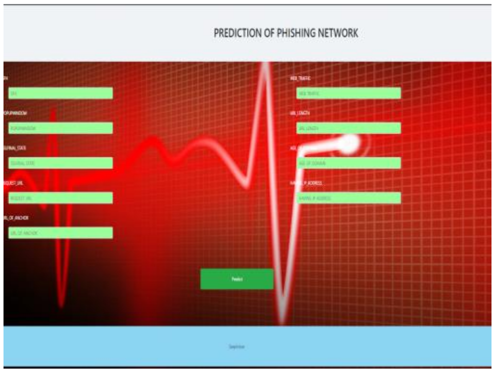


Fig 4 Prediction of Phishing

VII ACKNOWLEDGEMENT

We are using this opportunity to express our gratitude to everyone who supported us 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 paper.

VIII REFERENCES

1. 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.

2. X. Zhang, Y. Zeng, X. Jin, Z. Yan, and G. Geng, "Boosting the Phishing Detection Performance by Semantic Analysis,"2017.

3. L. Mac Hado and J. Gadge, "Phishing Sites Detection Based on C4.5 Decision Tree Algorithm," in 2017 International Conference on Computing, Communication, Control and Automation, ICCUBEA2017,2018,pp.1–5.

4. A. Desai, J. Jatakia, R. Naik, and N. Raul, "Malicious web contentdetection using machine leaning," RTEICT 2017 - 2nd IEEE Int.Conf. Recent Trends Electron. Inf. Commun. Technol. Proc., vol.2018–Janua,pp.1432–1436,2018.

5. Hossain M.A, Keshav Dahal, Maher Aburrous, "Modelling Intelligent Phishing Detection System for e-Banking using

6. Fuzzy Data Mining". Andrew H.Sung, Ram Basenet, Srinivas Mukkamala, "Detection of Phishing Attacks: A machine Learning Approach".

7. Ying Pan, Xuhus Ding “Anomaly Based Phishing page Detection”.