International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 06 Issue: 06 | June - 2022 Impact Factor: 7.185 ISSN: 2582-3930

Prediction Of Phishing Website Using Machine Learning

AHAMED KURAISHI S M

AMEENUDEEN B

MOHAMED HUZAIFA J

DHAANISH AHMED COLLEGE OF ENGINEERING

Abstract: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. One of the most successful methods for detecting these malicious activities is Machine Learning. This is because most Phishing attacks have some common characteristics which can be identified by machine learning methods. In this paper, we compared the results of multiple machine learning methods for predicting phishing websites.

INTRODUCTION

Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data, and apply knowledge and actionable insights from data across a broad range of application domains.The term "data science" has been traced back to 1974, when Peter Naur proposed it as an alternative name for computer science. In 1996, the International Federation of Classification Societies became the first conference to specifically feature data science as a topic. However, the definition was still in flux.The term “data science” was first coined in 2008 by D.J. Patil, and Jeff Hammerbacher, the pioneer leads of data and analytics efforts at LinkedIn and Facebook. In less than a decade, it has become one of the hottest and most trending professions in the market.Data science is the field of study that combines domain expertise, programming skills, and knowledge of mathematics and statistics to extract meaningful insights from data.Data science can be defined as a blend of mathematics, business acumen, tools, algorithms and machine learning techniques, all of which help us in finding out the hidden insights or

patterns from raw data which can be of major use in the formation of big business decisions.

Objectives

The goal is to develop a machine learning model for phishing website or not Prediction, to potentially replace the updatable supervised machine learning classification models by predicting results in the form of best accuracy by comparing supervised algorithm

Literature Survey

Title: A Bio-Inspired Self-learning Coevolutionary Dynamic Multiobjective Optimization Algorithm for Internet of Things Services

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

Year : 2018

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 coevolutionary 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 third layer refines the solutions found in the second layer by adopting an adaptive gradient refinement search strategy and dynamic optimization method to cope with changing concurrent multiple service requests, thereby effectively improving the accuracy of solutions.

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

Author: Wentao Zhao, Jianping Yin, Jun Long

Year : 2008

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. A various studies focused on DoS attacks from different respects [2][6][10]. However, these methods required a priori knowledge being a necessity and were difficult to discriminate between normal burst traffics and flux of DoS attacks. Moreover, they also required a large number of history records and can not make the prediction for such attacks efficiently. Based on data from flux inspecting and intrusion detection, we propose a prediction model of DOS attack's distribution discrete probability based on clustering method of genetic algorithm and Bayesian method. Due to various interference factors, the frequency of the DoS attack is considered to be a random variable. And probability is an effective way to describe randomness

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. In this paper, we review recent findings on adversarial examples for DNNs, summarize the methods for generating adversarial examples, and propose a taxonomy of these methods. Under the taxonomy, applications for adversarial examples are investigated. We further elaborate on countermeasures for adversarial examples. In addition, three major challenges in adversarial examples and the potential solutions are discussed In this paper, we reviewed the recent findings of adversarial examples in DNNs.

Title: Apriori Viterbi Model for Prior Detection of Socio-Technical Attacks in a Social Network
Author: Preetish Ranjan, Abhishek Vaish
Year: 2014

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 compressing it in small networks through apriori 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.

Title: New Attack Scenario Prediction Methodology
Author: seraj Fayyad, cristoph meinel
Year: 2013

Intrusion detection system generates significant data about malicious activities run against network. Generated data by IDS are stored in IDS database. This data represent attacks scenarios history against network. Main goal of IDS system is to enhance network defense technologies. Other techniques are also used to enhance the defense of network such as Attack graph. Network attack graph are used for many goals such as attacker next attack step prediction. In this paper we propose a real time prediction methodology for predicting most possible attack steps and attack scenarios. Proposed methodology benefits from attacks history against network and from attack graph source data. it comes without considerable computation overload such as

checking of attack plans library. It provides parallel prediction for parallel attack scenarios.

Existing System

Existing CTI for phishing website detection methods can be divided into three types: lookup systems, fraud cuebased methods, and deep representation-based methods. The lookup system detects a phishing website by "looking up" the website URL against a blacklist of phishing URLs and an alarm is raised when the website's URL appears in the list. The blacklists are classifiers (e.g., SVM, decision tree) and novel machine learning methods (e.g., statistical learning theory based methods, genre tree kernel methods and recursive trust labeling algorithm) have been devised to detect phishing websites. Similarly, website traffic based fraud cues requires to analyze the website traffic within a period of time, making them hard to meet the real-time detection requirement.

Proposed System

The proposed model is to build a machine learning model for anomaly detection. Anomaly detection is an important technique for recognizing fraud activities, suspicious activities, network intrusion, and other abnormal events that may have great significance but are difficult to detect. The machine learning model is built by applying proper data science techniques like variable identification that is the dependent and independent variables. Then the visualisation of the data is done to insights of the data .The model is build based on the previous dataset where the algorithm learn data and get trained different algorithms are used for better comparisons. The performance metrics are calculated and compared.

Algorithm Explanation

In machine learning and statistics, classification is a supervised learning approach in which the computer program learns from the data input given to it and then uses this learning to classify new observation. This data set may simply be bi-class (like identifying whether the person is male or female or that the mail is spam or non-spam) or it may be multi-class too. Some examples of classification problems are: speech recognition, handwriting recognition, bio metric identification, document classification etc. In Supervised Learning, algorithms learn from labeled data. After understanding the data, the algorithm determines which label should be given to new data based on pattern and associating the patterns to the unlabeled new data.

SYSTEM SPECIFICATION

4.1 Environmental Requirements:

1. Software Requirements:

Operating System : Windows

Tool : Anaconda with Jupyter Notebook

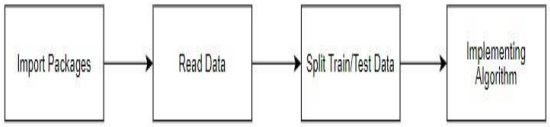
2. Hardware requirements:

Processor : Pentium IV/III

Hard disk : minimum 80 GB

RAM : minimum 2 GB

MODULE DIAGRAM



GIVEN INPUT EXPECTED OUTPUT

input : data

output : getting accuracy

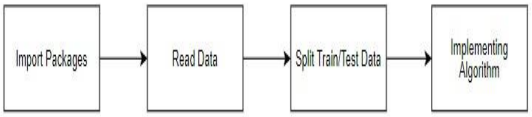
Random Forest Classifier

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of over fitting to their training set. Random forest is a type of supervised machine learning algorithm based on ensemble learning. Ensemble learning is a type of learning where you join different types of algorithms or same algorithm multiple times to form a more powerful prediction model. The random forest algorithm combines multiple algorithm of the same type i.e. multiple decision tree , resulting in a forest of trees, hence the name "Random Forest". The random forest algorithm can be used for both regression and classification tasks.

The following are the basic steps involved in performing the random forest algorithm:

- Pick N random records from the dataset.
- Build a decision tree based on these N records.
- Choose the number of trees you want in your algorithm and repeat steps 1 and 2.

In case of a regression problem, for a new record, each tree in the forest predicts a value for Y (output). The final value can be calculated by taking the average of all the values predicted by all the trees in forest. Or, in case of a classification problem, each tree in the forest predicts the category to which the new record belongs. Finally, the new record is assigned to the category that wins the majority vote.



GIVEN INPUT EXPECTED OUTPUT

input : data

output : getting accuracy

Decision Tree Classifier

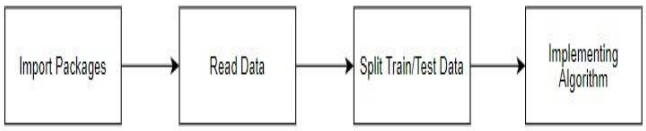
It is one of the most powerful and popular algorithm. Decision-tree algorithm falls under the category of supervised learning algorithms. It works for both continuous as well as categorical output variables. Assumptions of Decision tree:

- At the beginning, we consider the whole training set as the root.
- Attributes are assumed to be categorical for information gain, attributes are assumed to be continuous.

- On the basis of attribute values records are distributed recursively.
- We use statistical methods for ordering attributes as root or internal node.

Decision tree builds classification or regression models in the form of a tree structure. It breaks down a data set into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. A decision node has two or more branches and a leaf node represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data. Decision tree builds classification or regression models in the form of a tree structure. It utilizes an if-then rule set which is mutually exclusive and exhaustive for classification. The rules are learned sequentially using the training data one at a time. Each time a rule is learned, the tuples covered by the rules are removed.

This process is continued on the training set until meeting a termination condition. It is constructed in a top-down recursive divide-and-conquer manner. All the attributes should be categorical. Otherwise, they should be discretized in advance. Attributes in the top of the tree have more impact towards in the classification and they are identified using the information gain concept.A decision tree can be easily over-fitted generating too many branches and may reflect anomalies due to noise or outliers.



GIVEN INPUT EXPECTED OUTPUT

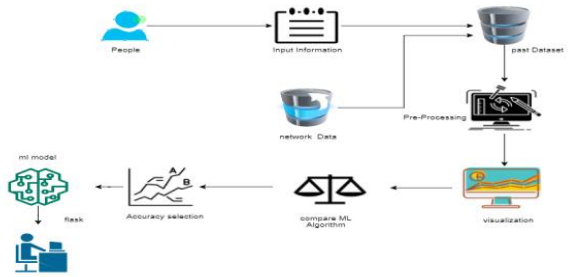
input : data

output : getting accuracy

Naive Bayes algorithm:

- 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.
- The probability of a class value given a value of an attribute is called the conditional probability. By multiplying the conditional probabilities togeth er for each attribute for a given class value, we have a probability of a data instance belonging to that class. To make a prediction we can calculate probabilities of the instance belonging to each class and select the class value with the highest probability.

- Naive Bayes is a statistical classification technique based on Bayes Theorem. It is one of the simplest supervised learning algorithms.Naive Bayes classifier is the fast, accurate and reliable algorithm. Naive Bayes classifiers have high accuracy and speed on large datasets.
- Naive Bayes classifier assumes that the effect of



a particular feature in a class is independent of other features. For example, a loan applicant is desirable or not depending on his/her income, previous loan and transaction history, age, and location.

- Even if these features are interdependent, these features are still considered independently. This assumption simplifies computation, and that's why it is considered as naive. This assumption is called class conditional independence.

Architecture Diagram

SCREEN SHOTS

Conclusion

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 find out. This application can help to find the Prediction of phishing website or not

REFERENCES

1. P. Prakash, M. Kumar, R. R. Kompella, and M. Gupta, "PhishNet : Predictive Blacklisting to Detect Phishing Attacks," 2010

2. R. M. Mohammad, "An Assessment of Features Related to Phishing Websites using an Automated Technique," pp. 492–497, 2012.

3. W. Hou, X. Gao, D. Tao, and X. Li, "Blind image quality assessment for stereoscopic images via deep learning," *IEEE Trans. NEURAL NETWORKS Learn. Syst.*, vol. 26, no. 6, pp. 1275–1286, 2015

4. P. Liu, X. Qiu, and X. Huang, "Syntax-based Attention Model for Natural Language Inference," 2016

5. C. Hori *et al.*, "Attention-Based Multimodal Fusion for Video Description," *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2017-Octob, pp. 4203–4212, 2017

6. P. Yi, Y. Guan, F. Zou, Y. Yao, W. Wang, and T. Zhu, "Web Phishing Detection Using a Deep Learning Framework," *Wirel. Commun. Mob. Comput.*, vol. 2018, 2018.

7. M. Chatterjee and A. S. Namin, "Detecting Phishing Websites through Deep Reinforcement Learning," in *IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC)*, 2019, no. 1, doi: 10.1109/COMPSAC.2019

8. J. Li, K. Lu, Z. Huang, L. Zhu, and H. T. Shen, "Heterogeneous Domain Adaptation Through Progressive Alignment," *IEEE Trans. NEURAL NETWORKS Learn. Syst.*, vol. 30, no. 5, pp. 1381– 1391, 2019.

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

10. European Union Agency for Network and Information Security, "ENISA Threat Landscape," 2020.

11. S. Sountharrajan, M. Nivashini, S. K. Shandilya, E. Suganya, A. B. Banu, and M. Karthiga, "Dynamic Recognition of Phishing URLs Using Deep Learning Techniques," in *Advances in Cyber Security Analytics and Decision Systems*, 2020

12. F. Feng, X. He, J. Tang and T. -S. Chua, "Graph Adversarial Training: Dynamically Regularizing Based on Graph Structure," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 6, pp. 2493-2504, 1 June 2021