JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Predicting Hazard Messages Using Machine Learning

Mr. Byalla Brahma Sekhar (M.C.A). Rajeev Gandhi Memorial college Of Engineering and Technology, Nandyal

*Mr. O. SAMPATH MTech, Ph.D. Rajeev Gandhi Memorial college Of Engineering and Technology, Nandyal

Abstract

Internet usage has integrated into our daily lives. Therefore, to capture users' attention, several browser suppliers compete to build up new functionality and complex capabilities that serve as a target for intrusion attempts and put websites at risk. However, the current methods are insufficient to protect web users, who need a quick and accurate model that can tell apart between safe and dangerous websites. Using machine learning classifiers like random forest, support vector machine, naive bayes, logistic regression, and some special URL (Uniform Resource Locator) based on extricated features, the classifiers are designed in this research article to create a new classification system to analyze and detect malicious web pages. educated to anticipate harmful web pages. According on the experimental findings, the random forest classifier performs better than other machine learning classifiers, achieving an accuracy of 95%.

1. INTRODUCTION

1.1 Introduction

With the web's rapid expansion, users now have access to an increasing number of services, including online banking, e-commerce, social networking, shopping, bill payment, e-learning, etc. Users browse the internet using browsers or web applications. As more sophisticated features and functionalities are added to browsers, users run the danger of losing their sensitive and personal data. Simply simply clicking on one of the malicious websites, which allows the invaders to detect the vulnerabilities on the web page and inject the payloads to get remote access to the victim's web page, naive people who are unaware of the various infections are easily captured by the invader. Consequently, accurate web page identification in an The constantly expanding web environment is crucial. To address these issues, blacklisting services were built into browsers, although they have a number of drawbacks, such as inaccurate listing. In this paper, we investigate a self-learning method for web page classification using a limited feature set. Four machine learning classifiers are used to divide the website into benign and harmful web page categories.

2. Literature Survey

- Researchers recommend three different strategies, namely blacklisting, static analysis, and dynamic analysis, to identify harmful web pages. Each strategy has a goal to achieve, and we've talked about some of these tactics in order. By employing supervised machine learning techniques, Tao et al. [1] established a unique framework for automatically determining whether a web page is harmful or benign. Based on features, the web pages were classified as malicious or not. Unfavorable web sites were gathered from the dataset.

Adware et al.'s [3] novel lightweight self-learning method for recognizing harmful online pages using categorized attributes The Genetic Algorithm (GA) was employed in the MALURL framework to train classifiers that can recognize malicious online pages. Consideration was given to the data sets Phis Tank for harmful websites and Alexa for benign websites. It was discovered that the typical system precision was 87%. "Adaptive SVM(SVM) machine learning technique" is used by Hwang et al. Because of its flexibility to adapt, the SVM can handle new training data. SVM aims to lower the possibility of new web pages being incorrectly classified. Using the machine learning algorithms K-NN and SVM, Yue et al. [6] suggested a method for categorizing fraudulent web pages using 30 features. K-NN produced results that were superior to SVM. For the purpose of identifying malicious web pages and certain threat kinds, two categorization models were employed. To identify known and undiscovered harmful web pages, Yoo et al. [4] developed two types of detection methods: abuse detection and anomaly detection. Though the false positive rate was significant at 30.5%, the detection rate was comparatively high at 98.9%. They used the RafaBot dataset and the WEKA tool to carry out their experiment.

3. OVERVIEW OF THESYSTEM

3.1 Existing System

Malicious URLs from a single attack type are often found using existing approaches. In this research, we present a method using machine learning to determine the type of attack a malicious URL undertakes to detect malicious URLs of all the common attack types.

3.1.1 Disadvantages of Existing System

- Less feature compatibility
- Low accuracy.

3.2 Proposed System

The textual qualities, link structures, webpage contents, DNS data, and network traffic are only a few of the discriminative features that our system makes use of. Many of these features are innovative and quite powerful. Our experimental experiments using 40,000 benign URLs and 32,000 harmful URLs gathered from actual Internet sources demonstrate that our method performs better than others: the accuracy was over 98% in recognizing attack types and over 98% in detecting malicious URLs. The effectiveness of each collection of discriminative traits is also reported, and their readability is discussed.

3.3 Methodology

User

1.System:

1.1 Train data set:

System can give training to the data set

1.2 model performance:

The three main metrics used to evaluate a classification **model** are accuracy, precision, and recall. Accuracy is defined as the percentage of correct predictions for the test data. It can be calculated easily by dividing the number of correct predictions by the number of total predictions.

1.3 predictions:

Using the machine leaning algorithms, we can predict the result

2.User:

2.1 upload dataset

The user uploads the dataset.

2.2 view dataset

The uploaded dataset is viewed by the user.

2.2 viewing graphs

Graphs can be generated by the system and the user can be view that graphs

Algorithms:

XGBoost:

XGBoost is an algorithm that has recently been dominating applied machine learning and Kaggle competitions for structured or tabular data. XGBoost is an implementation of gradient boosted decision trees designed for speed and performance.

XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. In prediction problems involving unstructured data (images, text, etc.) artificial neural networks tend to outperform all other algorithms or frameworks. However, when it comes to small-to-medium structured/tabular data, decision tree-based algorithms are considered best-in-class right now.

Bagging: Now imagine instead of a single interviewer, now there is an interview panel where each interviewer has a vote. Bagging or bootstrap aggregating involves combining inputs from all interviewers for the final decision through a democratic voting process.

XGBoost and Gradient Boosting Machines (GBMs) are both ensemble tree methods that apply the principle of boosting weak learners (CARTs generally) using the gradient descent architecture. However, XGBoost improves upon the base GBM framework through systems optimization and algorithmic enhancements.

4. Architecture

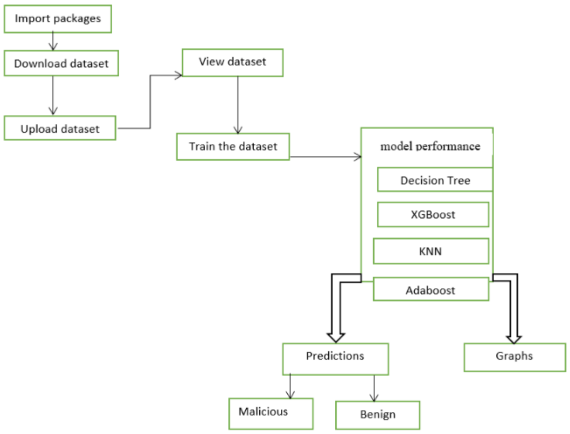


Fig 1: Frame work of proposed method

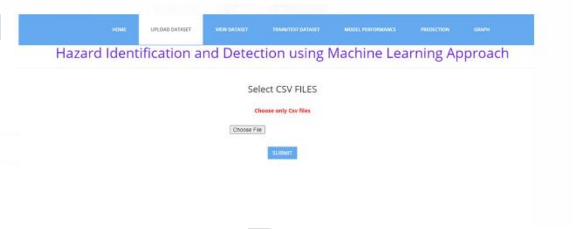
Above architecture diagram shows three stages of data flow form one module to another module. Data collection, preprocessing, and algorithm training.

4.RESULTS SCREEN SHOTS

Home Page:

Hazard Identification and Detection using Machine Learning Approach

Upload Data:

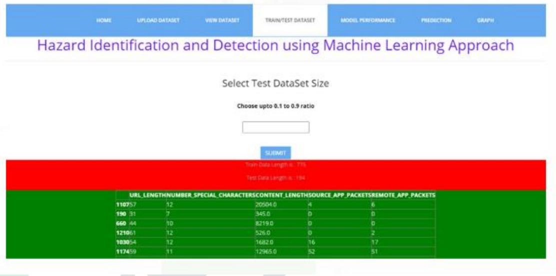


Hazard Identification and Detection using Machine Learning Approach

Select CSV FILES

Choose only Csv files

Choose options:



Hazard Identification and Detection using Machine Learning Approach

Select Test DataSet Size

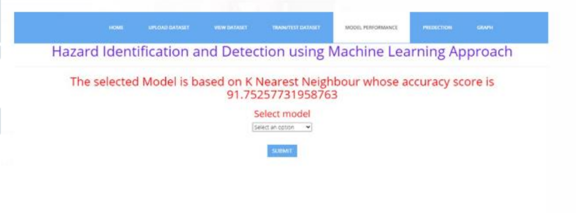
Choose upto 0.1 to 0.9 ratio

Train Data Length is : 775

Test Data Length is : 194

	URL_LENGTH	NUMBER_SPECIAL_CHARACTERS	CONTENT_LENGTH	SOURCE_APP_PACKETS	REMOTE_APP_PACKETS
1107	57	12	20504.0	4	6
190	31	7	345.0	0	0
660	44	10	8219.0	0	0
1210	61	12	526.0	0	2
1030	54	12	1682.0	16	17
1174	59	11	12965.0	52	51

Predict Result:



Hazard Identification and Detection using Machine Learning Approach

The selected Model is based on K Nearest Neighbour whose accuracy score is 91.75257731958763

Select model

6. CONCLUSION

- ✓ Identification of malicious web pages is a developing area of cybersecurity. Despite the fact that numerous research studies on the subject of malicious web page identification have been carried out, these are quite expensive because they require more time and resources. In this study, we used machine learning techniques to predict whether online sites were harmful or benign using a novel web site classification system based on URL attributes.

Random Forest (RF), a machine learning classifier, achieves a greater accuracy of 95%. The experimental results have demonstrated the effectiveness of our technology in identifying malicious web pages.

Future Enhancement

✓ In order to improve the performance of the classifier, it has been proposed to expand the feature sets and conduct analysis utilizing a variety of data sources.

7. References

[1] Tao, Wang, Yu Shunzheng, and Xie Bailin. "A novel framework for learning to detect malicious web pages." In 2010 International Forum on Information Technology and Applications, vol. 2, pp. 353-357. Ieee, 2010.

[2] Eshete, Birhanu, Adolfo Villafiorita, and Communist Weldemariam. "Malicious website detection: Effectiveness and efficiency issues." In 2011 First SysSec Workshop, pp. 123-126. IEEE, 2011.

[3] Aldwairi, Monther, and Rami Alsalman. "Malurls: A lightweight malicious website classification based on URL features." Journal of Emerging Technologies in Web Intelligence 4, no. 2 (2012): 128-133.

[4] Yoo, Suyeon, Sehun Kim, Anil Choudhary, O.P. Roy, and T. Tuithung. "Two-phase malicious web page detection scheme using misuse and anomaly detection." International Journal of Reliable Information and Assurance 2, no. 1 (2014): 1-9.

[5] Hwang, Young Sup, Jin Baek Kwon, Jae Chan Moon, and Seong Je Cho. "Classifying malicious web pages by using an adaptive support vector machine. "Journal of Information Processing Systems 9, no. 3 (2013): 395-404.

[6] Yue, Tao, Jianhua Sun, and Hao Chen. "Fine-grained mining and classification of malicious Webpages." In 2013 Fourth International Conference on Digital Manufacturing & Automation, pp. 616-619. IEEE, 2013.

[7] Krishnaveni, S., and K. Sathiyakumari. "Spider Net: An interaction tool for predicting malicious web pages." In International Conference on Information Communication and Embedded Systems (ICICES2014), pp. 1-6. IEEE, 2014.

[8] Sun, Bo, Mitsuaki Akiyama, Takeshi Yagi, Mitsuhiro Hatada, and Tatsuya Mori. "Automating URL blacklist generation with similarity search approach." IEICE TRANSACTIONS on Information and Systems 99, no. 4 (2016): 873-882..

.