Detecting of Phishing Websites Using Various Data Mining Classification Techniques

Team: Kofi Agyeman, Josiah Anderson, Nathan Gootee, Poorvaja Sundar, Cindy Wei

CS235 Data Mining Techniques - Spring 2020

Abstract

The internet today is irreplaceable. The amount of life, industry and commerce is immense. There are various applications that make life easier for people in many ways, from banking to online shopping. But alongside the many advantages, one of the primary disadvantages can be identified as phishing. There are many phishing websites whose aim is to trap users and extract sensitive information from them. The main aim of this paper is to identify legitimate websites versus phishing websites and obtain the accuracy measures of these predictions. This is done using five well-known data mining algorithms: Support Vector Machine, Logistic Regression, Naive Bayes, AdaBoost Technique and Decision Trees. We determine which algorithms best classify the dataset considered and give good accuracy measurements.

Problem statement

Given the online phishing website dataset from *Kaggle*, we are investigating and comparing the performance of various data mining classification algorithms implementations (Decision Tree, Naive Bayes, Logistic Regression, Support Vector Machines and Adaptive Boosting) to detect phishing websites.

Introduction

Phishing is defined as the social engineering process of luring users into fraudulent websites to obtain their personal or sensitive information such as the usernames, passwords, addresses, credit card details, etc… According to the Anti-Phishing Working Group (APWG) phishing trend report, in the first quarter of 2020, a number of cyber criminals have launched a variety of COVID-19 themed phishing and malware attacks against workers, healthcare facilities and the recently unemployed. There was a significant increase in the number of the phishing sites detected in the first quarter of 2020 when compared to the last quarter of 2019, approximately an increase of 3000 phishing sites.

A major challenge faced while detecting phishing sites is the discovery of the techniques utilized by the cyber criminals. They continuously enhance their techniques and develop web pages that are able to withstand themselves against various forms of detection. Due to this, there is also a need to actively predict and classify the phishing sites that are out there on the internet.

The aim of this paper is to present a study of existing methods that are used to classify the phishing sites as either safe or unsafe for people to use.

Methods & Approach

Overview of Project Classification Techniques - After careful review of literature we have narrowed down the implementation of phishing detection to the following classification techniques: Decision Tree, Naïve Bayes, Logistic Regression, Support Vector Machines and Adaptive Boosting. General overview and specific implementation approach of each technique is given below. We will implement and evaluate performance of these classification schemes with the same training and test dataset.

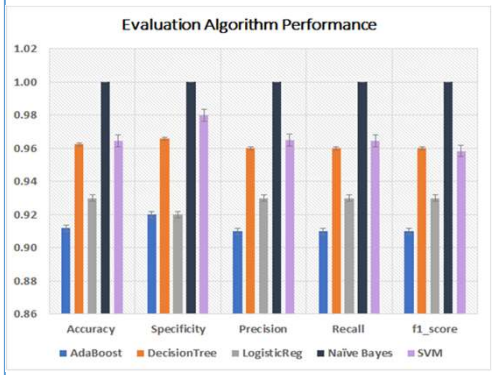
- Support Vector Machine
- Naïve Bayes
- Logistic Regression
- Adaptive Boosting
- Decision Tree

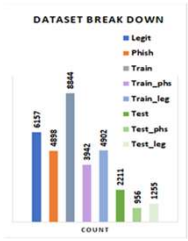
Implementation steps and evaluation scheme:

Each team member, following the steps outlined below, investigated, and implemented one of the proposed data mining techniques using the same dataset. Performance of each algorithm will be evaluated, compared, and results presented appropriately.

1. Preprocessing: Data sourcing, cleaning, integration, and transformation,
   - After extensive search and evaluation, we used dataset from Kaggle repository (https://www.kaggle.com/akashkr/phishing-website-dataset) because it is the cleanest, has extensive descriptions of features, and has no missing values. This is a dynamic search, so the dataset could be changed or added to in the future.
2. Data visualization
   - Precision-Recall Curve
   - Correlation matrices
   - Scatter plots
3. Mining approach and algorithms implementation
4. Validation and testing of mining algorithms
   - For testing purposes, we intend to set aside ~20% of the dataset for testing.
   - Remaining 80% of the dataset will be used for training and predictive model generation with careful attention to overfitting this set
   - We intend to explore the effects of algorithm performance using different k values for k-fold cross-validation processes, potentially will investigate k = 2, 5 and 10.
5. Performance evaluation of implemented algorithms based on accuracy of detection
   - Confusion matrix and Kappa statistic measures
   - Sensitivity (true positive rate), Specificity (true negative rate)
   - F1-score computation. To avoid the class imbalance problem, a weighted average of Precision and Recall will be considered.

Results





	Accuracy	Specificity	Precision	Recall	f1_score
AdaBoost	0.91	0.92	0.91	0.91	0.91
DecisionTree	0.96	0.97	0.96	0.96	0.96
LogisticReg	0.93	0.92	0.93	0.93	0.93
Naïve Bayes	1.00	1.00	1.00	1.00	1.00
SVM	0.96	0.98	0.97	0.96	0.96

Utilizing all five machine learning models we trained phishing predictive classification models using the designate test dataset. Dataset courtesy of Kaggle data repository was loaded, inspected, and partitioned into training and testing (20%) predictors array and output series. See **Fig. 1** for details of the data break down. There were no missing values and all predictor feature values were either [-1, 0, or 1], thus mitigating the need for min/max or Gaussian normalization of the feature values. **Figures 2 and 3** present a table and bar plots of the summary of classification performance of the phishing detection algorithms implemented. Average accuracy, specificity, recall, precision and f-scale measures are given for predictive tests on the 20% test data set.

Conclusions

Comparing the obtained accuracy measures in each algorithm, it is observed that Naive Bayes is by far the best performing algorithm for the chosen dataset. The Naive Bayes algorithm gives an accuracy measure of 99%, which is the best accuracy measure when compared to the others. Figure A depicts the comparison of accuracy, precision, recall and f1-score of all the five algorithms in the form of a bar chart. This chart greatly helps in the observation of the best algorithm with ease

References

1. Abu-Nimeh, S., Nappa, D., Wang, X., & Nair, S. (2007). A comparison of machine learning techniques for phishing detection. *Proceedings of the anti-phishing working groups 2nd annual eCrime researchers summit on - eCrime '07*. doi:10.1145/1299015.1299021
2. Aburrous, M., Hossain, M. A., Dahal, K., & Thabtah, F. (2010). Predicting phishing websites using classification mining techniques with experimental case studies. *2010 Seventh International Conference on Information Technology: New Generations*. doi:10.1109/itng.2010.117
3. Almseidin, Mohammad & Abuzuraiq, Almaha & Al-kasassbeh, Mouhammd & Alnidami, Nidal. (2019). Phishing Detection Based on Machine Learning and Feature Selection Methods. International Journal of Interactive Mobile Technologies (iJIM). 13. 171. 10.3991/ijim.v13i12.11411
4. Altaher, A. (2017). Phishing websites classification using hybrid SVM and KNN approach. *International Journal of Advanced Computer Science and Applications*, *8*(6). doi:10.14569/ijacsa.2017.080611
5. Anandkumar, Vanitha. (2019). Malicious-URL Detection using Logistic Regression Technique. International Journal of Engineering Business Management. 9. 108-113. 10.31033/ijemr.
6. James, Joby & L., Sandhya & Thomas, Ciza. (2013). Detection of phishing URLs using machine learning techniques. 304-309. 10.1109/ICCC.2013.6731669.
7. M. A. Adebowale, K. T. Lwin, E. Sánchez and M. A. Hossain, "Intelligent web-phishing detection and protection scheme using integrated features of Images frames and text", *Expert Syst. Appl.*, vol. 115, no. December 2017, pp. 300-313, 2018.
8. Marchal, S., Asokan, N. (2018). On Designing and Evaluating Phishing Webpage Detection Techniques for the Real World. *CSET '18: 11th USENIX Workshop on Cyber Security Experimentation and Test*.
9. Moghimi, Mahmood & Varjani, A.. (2016). New Rule-Based Phishing Detection Method. Expert Systems with Applications. 53. 10.1016/j.eswa.2016.01.028.
10. Mohammad, R.M., Thabtah, F., McCluskey, L., (2013). Intelligent Rule-Based Phishing Websites Classification. *The Institution of Engineering and Technology Information Security, Vol. 8, Iss. 3, pp. 153–160*. doi:10.1049/iet-ifs.2013.0202.