Research Article

A Rule Based Prediction of Phishing Websites Using Data Mining Classification Techniques

M.H F. Fazliya[1*], H.M.M Naleer[2]

**fathifazly@gmail.com*

[1, 2] Department of Mathematical Sciences, Faculty of Applied Sciences, South Eastern University of Sri Lanka

Abstract

Phishing is a major cybercrime which challenges the information security by mimicking legitimate websites to steal personal identity data and financial account credentials. Victims suffer from financial aspect and other private data theft. As there is a rapid advance in phishing activities via fake websites many anti phishing techniques are emerging but phishers use different and efficient techniques to trick victims so an intelligent and resilient model is required to detect the phishing sites. Data mining techniques can be used to develop an efficient model with the nontrivial and implicit information that can be mined from large datasets using classification algorithms to classify a websites legitimacy. Four different classification algorithms were implemented to classify the data set and comparatively studied for their performances, accuracy and number of rules created. The rules generated from the classification algorithms showed 96.76% accuracy whereas the prevailing phishing detection tools promises 92% as the highest accuracy which may effectively contribute for real time phishing website detection.

Keywords: *Phishing, Data mining*

1. Introduction

One of the primary problems in web security nowadays is phishing. According to APWG trend report of 2016, Phishing is a criminal mechanism employing both social engineering and technical subterfuge to steal consumer's personal identity data and financial account credential. Phishing is a major cybercrime where the personal identity data such as username, password and the financial credentials such as credit card credentials are stolen during an electronic transaction by third party mimicking them as authentic entity. Phishers create fake websites, malicious links and spoofed emails and send to users (Khan *et al.,* 2018). When the unsuspected user tries the link it will redirect to the fake site which lookalike the

original website and soon after the submission of personal or financial information the user is vulnerable for phishing attack. APWG (Anti Phishing Working Group) started monitoring the phishing activity from 2004 and it reported a highest rate in the first quarter of 2018. Phishing attacks recorded in the beginning of 2018 showed a 46 percent increase when compared to last quarter of 2017 (Anti-Phishing Working Group, 2018). In order to save the online users from phishing it is important to have a system that is practical, adaptive and low in false detection.

In general, two approaches are employed in identifying the phishing websites. Conventionally a blacklist based approach is employed as an anti-phishing mechanism in various phishing detecting tools where the requested URL is compared with the previously blacklisted ones but this is highly ineffective as the blacklist usually cannot cover all phishing websites since within seconds a new fraudulent website is expected to be launched. Second approach is called white list method where several features are gathered from the website in the real time to classify as legitimate or not. Browser extensions are used to perform client-side verification of the URL for its legitimacy.

Spoofguard and Netcraft are popular extensions used by web surfers which guarantees only 85% of accuracy (Krishnan *et al.*,2015) . An anti-phishing mechanism which promises a higher accuracy can be implemented to provide a safer experience in online transactions. In order to attain this goal, classification data mining techniques were used to solve the difficulty in detecting the phishing websites by the nontrivial extraction of implicit, previously unknown, and potentially useful information from large data set obtained from reliable repository. Typical classification algorithms are Naïve Bayes, Decision tree, rule induction and neural networks. For this context rule-based decision tree algorithms were implemented. Rule based approach is simple, user friendly, intuitive and follows separate and conquer method in which the rules are produced recursively for the subsets of the training data set and all the other subsets that produce same rules will be discarded (Tayel *et al.*, 2013)as well as It is not difficult to extend the rule set with the emerging phishing detection techniques when new rules are encountered. Different rule-based classification algorithms were chosen as they deal with different mechanisms in pruning and error reduction to classify.

2. Related Work

Phishing websites are a major and a recent issue that is viewed as an obstacle in online trading and e-commerce and since preventing such attacks is an important step towards defending against website phishing attacks, there are several promising approaches to this problem and a comprehensive collection of related works. Different work had been conducted from 2007 to 2015 to track phishing e-mails, URLs and webpages. Blacklist method has been the popular method and

most of the browsers nowadays have inbuilt functionalities to identify blacklisted websites.

In 2007 phishing web page detection has been performed by extracting phishing website URLs from various known phishing attacks. In 2008 phishing detection is done (Basnet *et al.*,2008) considering 16 features with as Support Vector Machine (SVM), Biased Support Vector Machine (BSVM), Neural Network (NN) and Self Organizing Map (SOMs). Later URLs are classified based on the lexical features and host based features using Naive Bayes (NB), Support Vector Machine (SVM) and Logistic Regression (LR).

In 2010 phishing detection was done with reduced features of CANTINA (Xiang *et al.,* 2011) Anti-Phishing and Network analysis tool by using machine learning techniques such as Naive Bayes (NB), Neural Network (NN), Support Vector Machine (SVM), Random Forest (RF), J48 Decision tree and Adaboost. In 2011 in order to detect phishing websites, obfuscation in URLs domain name, features related to webpage source code were considered. Later on in (Santhana *et al.*, 2011) Multi-Layer Perceptron (MLP), Decision tree induction and Naive Bayes were used to evaluate seventeen website features updated with Meta title, meta description, content attributes and "href" attributes of tag <a> for detecting phishing attacks. The correlation based and wrapper based feature selection techniques were studied to improve the classification accuracy. In 2012 domain based features along with the lexical features were proposed in the classification framework (Mohammad *et al.*, 2012). In 2015 thirty effective minimal set of features that shows improved detection were extracted. Different techniques and algorithms were implemented for improved accuracy in detection.

3. Experiments

Deciding a websites legitimacy is a classification problem in which datamining techniques can be implemented to extract the hidden and nontrivial knowledge with the availability of a big data. When the data set was chosen the effective minimal set of features that is considered as attributes was selected from Machine Learning Repository (Center for machine learning and intelligent systems**.** Dataset comprises 11055 instances of 3793 phishing and 7262 legitimate websites with 30 attributes falls under four major categories (Mohammad *et al.*,2015) Attributes are the effective minimal set of phishing website features. Attributes can have mostly two values -1, 1 values that represent phishing and legitimate features and some features include 0 value that represents suspicious features. The data feature values are created based on rules proposed in the literature Phishing Websites Features (Mohammad *et al.*,2015). Features are divided into four main categories as below.

Table 1: Categories of website features

Category	Website Features
Address Bar based Features	Using the IP Address
	Long URL to Hide the Suspicious Part
	Using URL Shortening Services “TinyURL”
	URL’s having “@” Symbol
	Redirecting using “//”
	Adding Prefix or Suffix to the Domain
	Sub Domain and Multi Sub Domains
	HTTPS (Hyper Text Transfer Protocol with Secure Sockets Layer)
	Domain Registration Length
	Favicon
	Using Non-Standard Port
	The Existence of “HTTPS” Token in the Domain of the URL
Abnormal Based Features	Request URL
	URL of Anchor
	Links in <Meta>,<Script>and<Links>tags
	Server Form Handler (SFH)
	Submitting Information to Email
	Abnormal URL
HTML and JavaScript based Features	Website Forwarding
	Status Bar Customization
	Disabling Right Click
	Using Pop-up Window
	IFrame Redirection
Domain based Features	Age of Domain
	DNS Record
	Website Traffic
	PageRank
	Google Index
	Number of Links Pointing to Page
	Statistical-Reports Based Feature

In classification problems, a classifier tries to learn several feature attributes as inputs to predict an output. In phishing website classification, a classifier rule tries to classify a website as phishing or legitimate by learning certain features and patterns and the correlation among them in the website.

Four different kind of classification algorithms are implemented on the phishing website data set namely Decision tree J48, JRip rules, PART, Decision table. These algorithms were chosen because of the different strategies they use on datasets.

J48 Algorithm: J48 is an implementation of the classic C4.5 decision tree algorithm. The C4.5 algorithm employs a divide-and-conquer approach. The J48 algorithm is used to extract a decision tree that can classify web pages as legitimate or phishing.

PART Algorithm: The choice of PART algorithm is based on the fact that it adapts separate and conquer to generate a set of rules and uses divide-and-conquer to build partial decision trees. The way PART builds and prunes a partial decision tree is similar to the C4.5 implementations with a difference which can be explained as follows: C4.5 generates one decision tree and uses pruning techniques to simplify it; each path from the root node to one of the leaves in the tree represents a rule. On the other hand, PART avoids the simplification process by building up partial decision trees and choosing only one path in each one of them to derive a rule. Once the rule is generated, all instances are associated with it, and the partial tree is discarded.

Decision Table: Decision Table algorithm summarizes the dataset to a decision table which contains the same number of attributes as the original dataset. The classifier categories a new data item into the decision table by finding the matches. Wrapper method is used to find the effective subset of features for building the table.

JRip: JRip contains two major stages a building stage and an optimization stage. Building process is split into two phases a grow phase and a prune phase. Grow phase grows one rule by greedily adding antecedents (or conditions) to the rule until the rule is perfect. Overlarge rule set is then repeatedly simplified by applying pruning techniques. Simplification ends when applying any pruning would increase error.

For the whole classification process data mining tool WEKA was used in this research. Waikato Environment for Knowledge Analysis (WEKA) is a suite of machine learning software written in Java, developed at the University of Waikato, New Zealand. Weka is an open source, portable workbench which is a collection of machine learning algorithms for data mining tasks (Aksenova *et al.*,2004).

Classification algorithms were implemented one by one for the data set. Four different classification algorithms were utilized to classify the phishing website data set with 10 folds cross validation test option in order to produce rules that effectively identify phishing websites. In order to study their performance accuracy, error rates, number of rules produced and the time taken for classification are considered in this study.

4. Experimental Result Analysis

11055 instances of phishing website relations are classified with four classifiers. Following chart shows the correct and incorrect classification of instances and their percentages.

Table 2: Correct and incorrect classification instances of algorithms

Algorithm	Correctly Classified Instances	Incorrectly Classified Instances
PART	1069	358
JRip	1054	551
Trees.J48	10599	456
Decision Table	10308	747

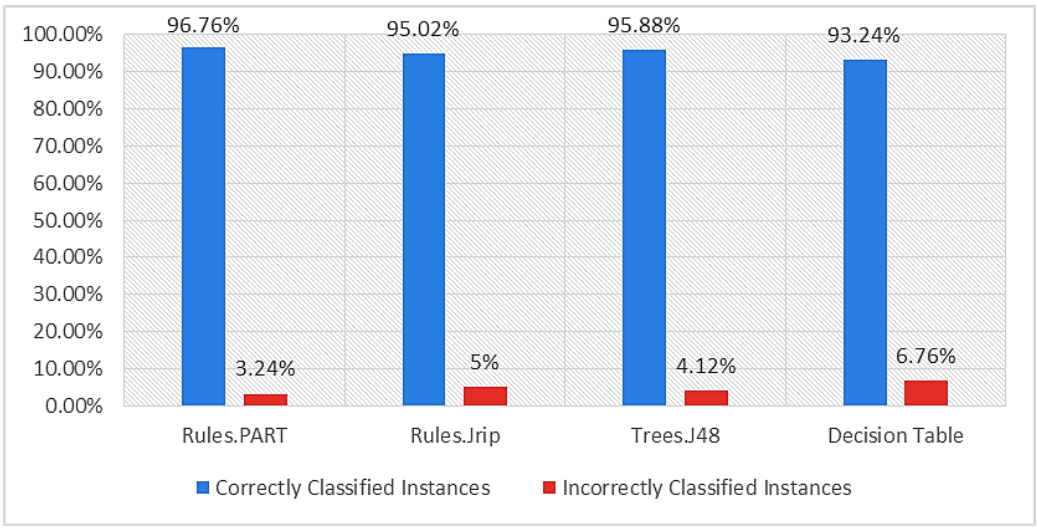

Figure 1: Algorithms and its classification of correct and incorrect instances

All the algorithms taken for the study show a higher prediction rate and among all of them PART algorithm shows highest accuracy by attaining 96.76%. Algorithms are further evaluated with the error rate parameters such as MAE, RMSE, RAE and RRSE.

Table 3: Error rate evaluation of algorithms.

Algorithm	**Mean Absolute Error (MAE)**	**Root Mean Squared Error (RMSE)**	**Relative Absolute Error (RAE)**	**Root Relative Squared Error (RRSE)**
PART	0.0411	0.1646	8.3355	33.1447
JRip	0.0823	0.2121	16.67	42.70
Trees.J48	0.0567	0.1853	11.49	37.30
Decision Table	0.1228	0.2268	24.8852	45.6664

PART exhibits reduced error rates and J48 shows second least error rates for all the four parameters.

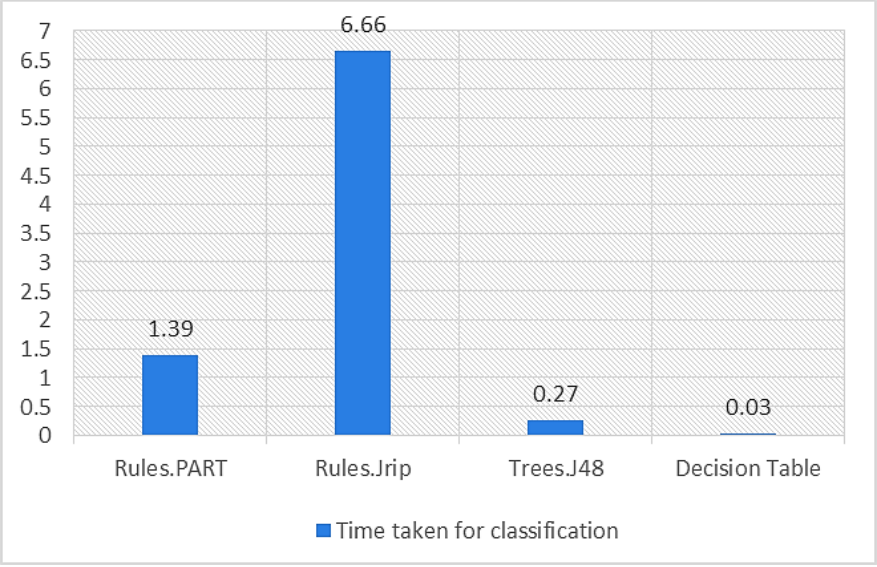

Figure 2: Algorithms and their classification time

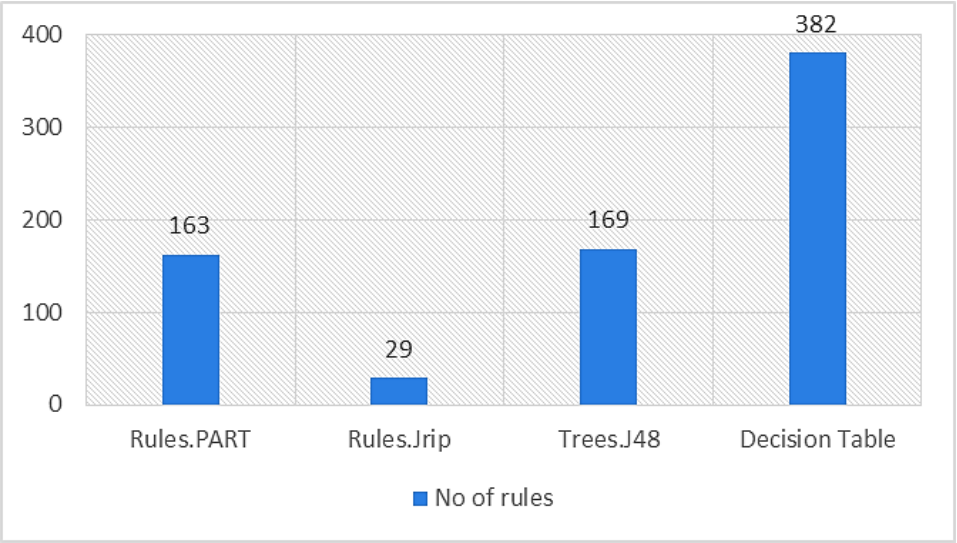

Figure 3: Algorithms and the number of rules produced

Decision table takes the least time for classification whereas JRip records the highest according to Figure 2. Rule pruning is very effective in JRip and it generates less number of rules to classify. Among all the four algorithms used, PART achieves 96.76% of accuracy with least error rates in a feasible time duration. These experiment results shows that these set of rules can be used as a model to build real applications involving large databases. The rules generated from the classification model show the correlation and relationships between all phishing features and patterns. It showed that there is a significant relationship between phishing website criteria's URL & Domain Identity and Security & Encryption for identifying phishing website.

5. Conclusion

Experiments are conducted using four different rule based algorithms to discover the hidden knowledge from the large dataset to predict the phishing websites. Classified outputs are compared for their performances in terms of accuracy, error rate, time duration and the number of rules produced. From the results it is obvious that all the selected algorithms achieve higher prediction rate. The PART decision rules predict phishing websites with the accuracy rate 96.76% and records a less error rate. The rules generated showed the correlation and relationship between website features which can help us in building phishing website detection frameworks. This phishing detection model is able to protect users from being phished by performing verification during a confidential data submission.

6. Reference

Abdelhamid, N., Ayesh , A. & Thabtah, F.(2014). Phishing detection based associative classification data mining. *Expert Systems with Applications* ,41(13), 5948-5959.

Aksenova , S.S. (2004) Machine Learning with WEKA,WEKA Explorer Tutorial for WEKA Version 3.4., School of Engineering and Computer Science,Department of Computer Science,California State University, Sacramento,California,

Anti-Phishing Working Group (2018). *Phishing Activity Trends Report (1st Quarter 2018)*.
Available at: https://docs.apwg.org/reports/apwg_trends_report_q1_2018.pdf

Basnet, R., Mukkamala, & S., Sung, A. (2008). Detection of Phishing Attacks: A Machine Learning Approach, 373-383.
https://doi.org/10.1007/978-3-540-77465-5_19

Khade, A. & Shinde, S.K. (2013). Detection of Phishing Websites Using Data Mining Techniques. *International Journal of Engineering Research & Technology (IJERT)*. 2(12), 3725-3729.

Khan, A. & Sharma, R. (2018) A Survey Paper on Detection of Phishing Website by URL Technique, *International Journal of Computer Science and Mobile Applications.* 6, 33–37,

Krishnan, D.M. & Subramaniyaswamy, V., (2015) Phishing website detection system based on enhanced itree classifier, *ARPN J. Eng. Appl. Sci.,* vol. 10, no. 14, 5688–5699.

Available at: https://archive.ics.uci.edu/ml/machine-learning-databases/00327/

Mohammad R., Thabtah F & McCluskey L (2012). *An Assessment of Features Related to Phishing Websites using an Automated Technique*. In The 7th International Conference for Internet Technology and Secured Transactions (ICITST-2012); London: ICITST

Mohammad R, Thabtah F & McCluskey L., (2015). *Phishing websites dataset*,

Available through: UCI machine learning Repository
https://archive.ics.uci.edu/ml/datasets/phishing+websites

Mohammad, R., McCluskey, T.L. & Thabtah, F. A. (2014*) Intelligent Rule based Phishing Websites Classification*. IET Information Security, 8 (3). 153-160. ISSN 1751-8709

Mohammad, R.M., Thabtah, F. & McCluskey, L., (2015). *Phishing Websites Features*. UCI machine learning Repository.

Lakshmi, V.S. & Vijaya, M.S.,(2011). *Efficient prediction of phishing websites using supervised learning algorithms*. International Conference on Communication Technology and System Design 798-805.

Tayel, S., Reif, M. & Dengel, A., (2013). *Rule-based Complaint Detection using RapidMiner*, Conference: RCOMM , At Porto, Portugal, 141 – 149

Xiang, G., Hong, J., Rose, C.P. & Cranor, L., (2011) *CANTINA+: A feature-rich machine learning framework for detecting phishing Web site*. ACM Trans. Inf. Syst. Secur. Vol.14, No.2, 1-21.