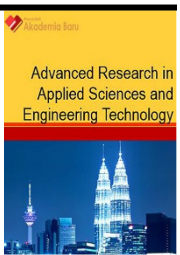
Penerbit Akademia Baru

Journal of Advanced Research in Applied Sciences and Engineering Technology

Journal homepage: www.akademiabaru.com/araset.html
ISSN: 2462-1943


Akademia Baru
Advanced Research in Applied Sciences and Engineering Technology

Classification of phishing websites using machine learning techniques

Open Access

Hadi Zamani [1], Muhamad Kamal bin Mohammed Amin [1,*]

[1] Bio Cognition Laboratory, Bio-Inspired System and Technology iKohza (Research Group), Malaysia – Japan International Institute of Technology (MJIIT), Universiti Teknologi Malaysia (UTM) 54100 Kuala Lumpur, Malaysia

ARTICLE INFO

Article history:
Received 31 May 2016
Received in revised form 30 June 2016
Accepted 15 August 2016
Available online 19 December 2016

ABSTRACT

Phishing detection is a momentous problem which can be deliberated by many researchers with numerous advanced approaches. Current anti-phishing mechanisms such as blacklist-base anti-phishing, Heuristic-based anti-phishing does suffer low detection accuracy and high false alarm. There is need for efficient mechanism to protect users from phishing websites. The purpose of this study is to investigate the capability of 6 machine learning algorithms i.e. Multi-Layer Perceptron (MLP), Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT), Logistic Regression (LR) and Naïve Bayes (NB) to classify phishing and non-phishing websites. These algorithms were trained with two different groups of training in WEKA environment and then were tested in terms of accuracy, precision, TP rate, and FP rate on a 3 different sets of dataset which contains dissimilar portion of phishing and non-phishing instances. Results presented that Naïve Bayes classifier has better detection accuracy between other classifiers for predicting phishing websites while Multi-Layer Perceptron gave worst result in terms of detection accuracy. The result also showed that Support Vector machine has better FP rate between other classifier. In addition, Random Forest, Decision Tree, and Naïve Bayes can classify all phishing websites as phishing correctly. It means that TP rate is 100% for these classifiers. In conclusion this paper suggests using NB as the best classifier for predicting phishing and non-phishing websites.

Keywords:
Classification, Machine learning techniques, Phishing detection

1. Introduction

The World Wide Web services are employed daily by huge numbers of people to communicate around the world and are some sort of mission-critical application for a lot of businesses. People have benefited from new services such as virtual marketplace, online shopping and online banking, while realizing these benefits, we have also opened ourselves and our system to a number of threats. Computer system security attacks might be categorized into several types: physical attacks, syntactic

[*] *Corresponding author.*
E-mail address: mkamalma@utm.my (Muhamad Kamal bin Mohammed Amin)

attacks, in addition to semantic attacks. Physical infrastructure for instance computers and wires are classified as the example of physical attacks. The syntactic attacks mark the running logic of desktops and networks, for instance, software vulnerabilities, cryptographic algorithms vulnerabilities, and so on, while semantic attacks are geared towards people vulnerabilities. Phishing is certainly one of semantic attacks [1].

The word Phishing originates from the similarity of "fishing", wherever net criminals used fake emails or other different social engineering techniques to "fish" for user name, passwords and monetary information from an outsized ocean of net users, the employment of "ph" may be a common hacker replacement that controlled hacking of phone to phone systems. The term was initial employed by hackers in throughout 1996 UN agency were stealing America on-line (AOL) accounts by then and initial mention of the term "phishing" on the web was created in 2600 hacker newsgroup in January 1996. The major phishing attack, in its present variety, toward financial institutions has been recounted in July 2003. The attacks primarily targeted E-loan, E-gold, Wells city, and Citibank [2].

The rest of the paper is organized as follows: in Section 2 this paper discusses some related work. In Section 3 this paper reviews six machine learning algorithms used in the study. In Section 4 This paper discusses dataset collection and division, evaluation metrics, and WEKA environment. In Section 5 the results are presented and discussed. This paper draws conclusions and motivate for future work in Section 6.

2. Related work

Phishing attacks are growing very fast each year with new techniques. Although the web users are conscious of these kinds of phishing attacks, many of users become prey to these attacks [3]. There are many different techniques to fighting against phishing attacks that reported earlier.

Chen, et al. [4] proposed a client approach based on five key features. This technique come to 96% of detection rate. The main advantage of their method is that learning phase to the classifier is not necessary. However, if one of five main phishing features change, the formula utilized in detection phishing could fail. Another disadvantage of their proposed method is the possible absence of a test database with legitimate messages; consequently, it isn't possible to measure the FP (False Positive) rate. Moreover, the features are viewed as isolated and no mix of them was studied.

Ying, et al. [5] Ding used discrepancies in their study which exist in the website’s identity, structural features and HTTP transactions to be able to distinguish the fake websites. It needs user skill and former knowledge of the websites. They have utilized Support vector machine (SVM) as page classifier. Okanovic, et al. [6] applied three supervised learning algorithms include support vector machine, Naïve Bayes, and K-Nearest Neighbor, and two unsupervised learning techniques like K-Means and Affinity Propagation, and compares the results. The main advantage of their work is, to use unsupervised learning techniques which have not been applied to detect fake pages by any other researchers in the past. In addition, the research employs more complex structures for better accuracies. Marchal et al. [7] used Markov model to make possible phishing URLs unlike regular monitoring of URLs. The main disadvantage of their scheme for generating possible URLs is intensive computational that result of complex operations involved.

According to study of Maher, et al. [8] their proposed model is relying on FL (Fuzzy Logic) functions that is used to characterize the website phishing elements and in addition to indicators as fuzzy variables and generate six measures and criteria’s [URL & Domain Identity, Security & Encryption, Source Code & Java script, Page Style & Contents, Web Address Bar and Social Human Factor] associated with website phishing attack aspects with a layer structure.

Mohammad et al. [9] proposed self-structuring neural network for classification on top of the 17 features for phishing detection. One drawback of their work is, self-structuring neural network makes only server side implantation. Singh et al. [10] used 15 features of phishing websites and applied back propagation training on top of SVM to classify phishing website in order to achieve more effective and accurate classification.

3. Machine learning techniques

One important branch of artificial intelligence is Machine learning. Machine learning is known as the construction and study of systems that capable to learn from data. For instance, a machine learning system could be trained on websites to learn to distinguish between legitimate websites and fake websites. After learning it can then be used to classify new webpages into legitimate websites or fake websites.

Machine learning focuses on prediction, based on known properties learned from the training data. Machine learning algorithms can be organized into a taxonomy based on the desired outcome of the algorithm. It can classify into supervised algorithm, Unsupervised learning, Semi-supervised learning, Reinforcement learning, Transduction, and Learning to learn.

3.1. Support vector machine

Among machine learning techniques, Support Vector Machines (SVM) are considered as supervised learning models. This technique used for classification and regression that analyse data and recognize pattern. The basic SVM require a group of input data and predicts, for every given input, which of two possible classes forms the output, which makes it a non-probabilistic binary linear classifier. Given a group of training samples, each labelled as belonging to one of two categories, an SVM training algorithm creates a model that labelled new examples into one category or the other. For example for this paper the new webpages can be labelled as legitimate website or phishing websites [11, 6].

3.2. Multi-layer perceptron

Multi-Layer Perceptron (MLP) is a feedforward neural network algorithm for supervised classification that assign group of input data into a group of possible outputs. It has linear function, for example a classification algorithm that makes its predictions based on a linear predictor function combining a set of weights with the feature vector describing a given input. In this technique the appropriate weights are applied to the inputs and the resulting weighted sum passed to a function that produces the output [11].

3.3. Random forest

Another learning algorithm that combine several tree predictors is Random Forest (RF), where each tree relies on the weight of a random vector tested autonomously. Additionally, all trees in RF have similar distribution.

Random forest is able to handle huge numbers of irregular data in a dataset. In addition, throughout the forest creating process they make an inner unbiased guesstimate of the generalization error. Additionally, they can appraisal missing data effectively. A main disadvantage of random forests is not able to be reproductive, during the process of creating the forest is random.

Furthermore, understanding the final model and subsequent results is very hard, as it comprises many self-governing decisions trees [11].

3.4. Decision tree

Decision Tree (DT) produces the end result like a binary tree, for that reason is named as decision tree. Model of decision tree comprises instructions to identify the target variable. Also, this technique measures properly, even where the numbers of training examples are too large and the numbers of attributes in outsized databases are huge. J48 algorithm is a kind of implementing of the C4.5 decision tree learner. This algorithm applies the greedy method to make decision trees for classification. Analysing training data are able to make a decision-tree model which can be applied to classify unseen data [3].

3.5. Logistic regression

One of the most favoured statistical algorithm in variety area for binary data is Logistic Regression (LR). Logistic regression calculates the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function, which is the cumulative logistic distribution. It has been extensively used because of its simplicity and fantastic interpretability. As a part of comprehensive linear models it typically utilizes the logic operation. [11].

3.6. Naïve bayes

Naive Bayes (NB) is a simple method for creating classifiers. This classifier considered all features independent from one to another inside each class, but it seems can work properly in practice even if that independence hypothesis is invalid. It categorizes data in 2 ways: 1) while using the training instances, this technique appraisal the factors of a probability distribution, supposing features are provisionally independent given the class. 2) For new test instances, the technique calculates the probability of those instances while they are belonging to each class. And then this technique classifies the test instances based on the largest probability [6, 8].

4. Experimental design

This section starts with explanatory of data collection and dataset in this paper, and describes dataset division that used for training and testing classifiers, and then explain evaluation metrics which are used to evaluate the result of each algorithm.

4.1. Data collection

Dataset in this paper include 5249 instants which contains 3611 phishing 1638 and non-phishing. Phishing websites are collected from PhishTank website (www.phishtank.com) and non-phishing websites are created manually by Google search engine. The dataset consists of 10 columns, 9 columns are features of phishing and non-phishing websites and 1 column is nominal which indicates that website is phishing=1 or non-phishing=0. Table 1 shows the features of database and their brief description.

Table 1
Database Features

Features	**Description**
IP_Address	1 if the URL has IP address or -1 if not
SSL_Connection	1 if the SSL connection is provided and -1 if not
Long_URL	the length of the URL
Dots	Shows how many dots in the URL which can reflect how many subdomains are used
At_Symbol	1 if there is @ symbol embedded in the URL or -1 if not
Hexadecimal	1 if the URL have hexadecimal codes and -1 If not
Frame	1 if the webpage has a frame and -1 If not
Redirect	1 if the webpage has a code to redirect to another destination and -1 If not
Submit	1 if the webpage has a form to send data and -1 if
Label	The classification of each webpage, 1 mean phishing and 0 mean non-phishing

The features represent the frequency of the most frequent terms that appear in phishing and non-phishing websites. In addition, these website features can be investigated to detect phishing websites and also capable to distinguish between phishing and non-phishing websites.

4.2. Data distribution

Dataset in this paper is divided into three sets to be used for training and testing of machine learning algorithm for predicting of phishing and non-phishing websites. There are two steps were taken to divide the dataset. The first step is, divide the dataset into three dissimilar sets with different proportion of phishing and non-phishing websites for each set. First set has 30% of phishing and 70% of non-phishing websites, second set has 50% of phishing and 50% of non-phishing websites, and third set has 70% of phishing and 30% of non-phishing websites. Percentage and number of phishing and non-phishing websites for each set is shown in Table 2.

Table 2
Division of Dataset

Type of Sets (Phishing Raito, None-Phishing Raito)	Number of Phishing	Number of Non-Phishing
Set 1 (70%,30%)	70% phishing = 1147	30% non-phishing = 491
Set 2 (50%,%50)	50% phishing = 819	50% non-phishing = 819
Set 3 (30%,%70)	30% phishing = 491	70% non-phishing = 1147

Table 3
Number of rows and percentage of training and testing process

Different Percentage of Training and Testing for three Different sets	Group 1 (60:40)	Group 2 (70:30)
	60% Training Process	70% Training Process
	40% Testing Process	30% Testing Process
Set 1, Set 2, Set 3	983 Rows for Training Process	1147 Rows for Training Process
	655 Rows for Testing Process	491 Rows for Testing Process

Second step is, employ three above mentioned sets with two different group of training and testing include different percentages of dataset rows. In first group 60% of Set1, Set2, and Set3 are used for training which includes 983 rows, and 40% of Set1, Set2, and Set3 are used for testing which includes 655 rows. In second group 70% of Set1, Set2, and Set3 are used for training which includes

1147 rows, and 30% of Set1, Set2, and Set3 are used for testing which include 491 rows. Table 3 shows with two different percentages of training and testing process for all different sets.

4.3. Evaluation metrics

Performance concerning to any classifier requires to be evaluated with some metric, to evaluate the result and therefore the superiority of the algorithm. In this paper, to compute the results of the experiments of six machine learning algorithms, four frequently used metrics are hired, that are Classification Accuracy (CA), Precision (P), True Positive rate (TP), and False Positive rate (FP).

Table 3
Number of rows and percentage of training and testing process

Different Percentage of Training and Testing for three Different sets	Group 1 (60:40) 60% Training Process 40% Testing Process	Group 2 (70:30) 70% Training Process 30% Testing Process
Set 1, Set 2, Set 3	983 Rows for Training Process 655 Rows for Testing Process	1147 Rows for Training Process 491 Rows for Testing Process

To evaluate the effectiveness and the efficiency of machine learning algorithms in this paper for classifying Phishing websites, this paper will be conducted simulation using WEKA simulator. WEKA (Waikato Environment for Knowledge Analysis) is a popular suite of machine learning software written in Java. The WEKA Workplace contains a collection of visualization tools and algorithms for data analysis and predictive modelling, together with graphical user interfaces for easy access to this functionality. These six machine learning algorithms will be trained through WEKA with two different groups to create proper classifier model for testing.

5. Result and discussion

The phishing website classification model is generated by implementing machine learning algorithms. As a testing method, this paper executed the experiments in ten-fold cross-validation approach. Cross validation is a statistical technique of appraising and comparing learning algorithms by dividing data into two portions of data for training and validating or testing the model [9]. Cross validation can also be used to understand the generalization power of a classifier. Cross validation estimates the error rate efficiently and in unbiased way. Cross validation divides data set into K sub-sample (in this study k=10). One of k sub-sample is selected as testing data, and the remaining k-1 sub-sample are used as training data. This procedure is repeated k times, in which each of the k sub-samples is used exactly once as the testing data. All results are averaged and single answer is executed [12, 13].

The goal is to compare the performance of different classifiers and find out the best approach for classification phishing and non-phishing websites. Table 4 shows the classification accuracy, precision, True positive(TF), and False Positive (FP) rate of Multi-Layer Perceptron (MLP), Support Vector Machine (SVM), Random Forest (RF), Decision tree C4.5 (DT), Logistic Regression (LR) and Naive Bayes (NB) on divided dataset with different percentage of training and testing.

Table 4
Comparison among Six Machine Learning Algorithms

Percentage of Sets		Set 1 (70%,30%)		Set 2 (50%,50%)		Set 3(30%,70%)		
Percentage of Groups		Group1 (60:40)	Group2 (70:30)	Group1 (60:40)	Group2 (70:30)	Group1 (60:40)	Group2 (70:30)	**Average**
Algorithm	Metrics							
MLP	Accuracy	98.93%	98.98%	99.54%	99.38%	98.01%	97.75%	98.76%
	Precision	98.9%	98.8%	99.1%	98.8%	93.5%	92.1%	96.86%
	True Positive	99.5%	99.7%	100%	100%	100%	100%	99.86%
	False Positive	2.3%	2.5%	0.9%	1.2%	2.8%	3.0%	2.11%
SVM	Accuracy	99.54%	99.38%	99.08%	98.98%	99.08%	99.18%	99.20%
	Precision	100%	100%	98.1%	98.0%	97.1%	97.2%	98.4%
	True Positive	99.3%	99.1%	100%	100%	100%	100%	99.7%
	False Positive	0%	0%	1.8%	2.0%	1.3%	1.1%	1.03%
RF	Accuracy	99.54%	99.59%	99.39%	99.18%	98.62%	98.37%	99.11
	Precision	99.4%	99.4%	98.7%	98.3%	95.7%	94.5%	97.66%
	True Positive	100%	100%	100%	100%	100%	100%	**100%**
	False Positive	1.5%	1.3%	1.2%	1.6%	2.0%	2.3%	1.65%
DT	Accuracy	99.69%	99.59%	98.77%	98.77%	98.93%	98.77%	99.08%
	Precision	99.6%	99.4%	97.6%	97.7%	96.7%	96.1%	97.85%
	True Positive	100%	100%	100%	100%	100%	100%	**100%**
	False Positive	1.0%	1.4%	2.5%	2.6%	1.5%	1.8%	1.8%
LR	Accuracy	99.23%	99.18%	99.38%	99.18%	99.08%	98.98%	99.17%
	Precision	99.4%	99.2%	98.9%	98.5%	97.3%	96.9%	98.36%
	True Positive	99.6%	99.7%	100%	100%	100%	100%	99.88%
	False Positive	1. %7	2.2%	1.3%	1.7%	1.4%	1.5%	1.63%
NB	Accuracy	99.84%	99.79%	99.69%	99.59%	98.62%	98.37%	**99.31%**
	Precision	99.8%	99.7%	99.4%	99.2%	95.7%	94.5%	98.05%
	True Positive	100%	100%	100%	100%	100%	100%	**100%**
	False Positive	0.5%	0.7%	0.6%	0.8%	2.0%	2.3%	1.15%

The result of dataset with different ratio of phishing and non-phishing websites with two different portions of training and testing dataset indicates that Naïve Bayes classifier has better classification accuracy between other classifiers. Random Forest, Decision Tree, and Naïve Bayes can classify all phishing websites as phishing correctly. It means TP rate is 100% for these classifiers. In terms of classification precision SVM has highest value. Support Vector machine has better FP rate among other classifiers.

6. Conclusion

In recent years phishing websites have been raised and lots of the existing anti-phishing tools are applying the black list approach that is not effectual. Phishing websites are applying new techniques which often can allow them to help misbehave successfully. Therefore, the white list along with black list will not be effectual any more specifically new phishing web sites. Machine learning algorithm have been developed and used to detect these phishing websites. There are a few existing tools making use of machine learning tactic by examine the information of each webpage as a way to detect phishing web sites. This paper compared six machine learning algorithm which they are Multi-Layer Perceptron (MLP), Support Vector Machines (SVM), Random Forest (RF), Decision Trees (DT), Logistic Regression (LR), and Naïve Bayes (NB) in order to find best machine learning algorithm to classify phishing and non-phishing websites. Base on experimental results of this study, the highest classification accuracy among these six machine learning algorithm is related to Naïve Bayes that is equal 99.31% and the worst classification accuracy is belongs to Multi-Layer Perceptron that is equal 98.76. This result also indicates that FP rate of Support Vector machine is equal 0% which means none of non-phishing websites are not classified as phishing. This study did not comprehensively investigate all the possible parameters for machine learning algorithm.

References

[1] He, Mingxing, Shi-Jinn Horng, Pingzhi Fan, Muhammad Khurram Khan, Ray-Shine Run, Jui-Lin Lai, Rong-Jian Chen, and Adi Sutanto. "An efficient phishing webpage detector." *Expert Systems with Applications* 38, no. 10 (2011): 12018-12027.

[2] Dunham, Ken. *Mobile malware attacks and defense*. Syngress, 2008.

[3] Lakshmi, V. Santhana, and M. S. Vijaya. "Efficient prediction of phishing websites using supervised learning algorithms." *Procedia Engineering* 30 (2012): 798-805.

[4] Chen, Juan, and Chuanxiong Guo. "Online detection and prevention of phishing attacks." In *2006 First International Conference on Communications and Networking in China*, pp. 1-7. IEEE, 2006.

[5] Pan, Ying, and Xuhua Ding. "Anomaly Based Web Phishing Page Detection." In *Acsac*, vol. 6, pp. 381-392. 2006.

[6] Kazemian, H. B., and S. Ahmed. "Comparisons of machine learning techniques for detecting malicious webpages." *Expert Systems with Applications* 42, no. 3 (2015): 1166-1177.

[7] Marchal, Samuel, Jérôme François, and Thomas Engel. "Proactive discovery of phishing related domain names." In *International Workshop on Recent Advances in Intrusion Detection*, pp. 190-209. Springer Berlin Heidelberg, 2012.

[8] Aburrous, Maher Ragheb, Alamgir Hossain, Keshav Dahal, and Fadi Thabatah. "Modelling intelligent phishing detection system for e-banking using fuzzy data mining." In *CyberWorlds, 2009. CW'09. International Conference on*, pp. 265-272. IEEE, 2009.

[9] Mohammad, Rami M., Fadi Thabtah, and Lee McCluskey. "Predicting phishing websites based on self-structuring neural network." *Neural Computing and Applications* 25, no. 2 (2014): 443-458.

[10] Singh, Priyanka, Yogendra PS Maravi, and Sanjeev Sharma. "Phishing websites detection through supervised learning networks." In *Computing and Communications Technologies (ICCCT), 2015 International Conference on*, pp. 61-65. IEEE, 2015.

[11] Abu-Nimeh, Saeed, Dario Nappa, Xinlei Wang, and Suku Nair. "A comparison of machine learning techniques for phishing detection." In *Proceedings of the anti-phishing working groups 2nd annual eCrime researchers summit*, pp. 60-69. ACM, 2007.

[12] Wang, Suge, Deyu Li, Xiaolei Song, Yingjie Wei, and Hongxia Li. "A feature selection method based on improved fisher's discriminant ratio for text sentiment classification." *Expert Systems with Applications* 38, no. 7 (2011): 8696-8702.

[13] Tjahyanto, Aris, Yoyon K. Suprapto, and Diah P. Wulandari. "Spectral-based features ranking for gamelan instruments identification using filter techniques." *TELKOMNIKA (Telecommunication Computing Electronics and Control)* 11, no. 1 (2013): 95-106.