COMPARISON OF MACHINE LEARNING TECHNIQUES IN PHISHING WEBSITE CLASSIFICATION

Adnan Hodžić
International Burch University
Bosnia and Herzegovina
adnan.hodzic@ibu.edu.ba

Jasmin Kevrić
International Burch University
Bosnia and Herzegovina
jasmin.kevric@ibu.edu.ba

Adem Karadag
Turkey
nuhadem@gmail.com

Abstract: *Phishing is one among the luring strategies utilized by phishing artist in the aim of abusing the personal details of unsuspected clients. Phishing website is a counterfeit website with similar appearance, but changed destination. The unsuspected client post their information thinking that these websites originate from trusted financial institutions. New antiphishing techniques rise continuously, yet phishers come with new strategy by breaking all the antiphishing mechanisms. Hence there is a need for productive mechanism for the prediction of phishing website. This paper described comparison in classification of phishing websites using different Machinelearning algorithms. Random Forest (RF), C4.5, REP Tree, Decision Stump, Hoeffding Tree, Rotation Forest and MLP were used to determine which method provides the best results in phishing websites classification. All instances are categorized as 1 for "Legitimate", 0 for "Suspicious" and 1 for "Phishy". Results show that RF with REP Tree show the best performance on this dataset for classification of phishing websites.*

Keywords: *Machine Learning, Phishing Websites*

Introduction

Internet is not only significant for individual users but also for online business organizations. These organizations usually offer online trading(Liu & Ye, 2003). Nevertheless, Internet-users can be prone to different types of webthreats that can make financial damages, identity theft, loss of private information, brand reputation damage and loss of client's trust in ecommerce and online banking. Therefore, Internet appropriateness for commercial sales becomes doubtful.

Phishing websites is a semantic intrusion which targets the user instead of computer. It is a fairly new Internet crime when compared to other forms, such as virus and hacking. The phishing problem is a tough problem due to the fact that it is extremely easy for an attacker to make a replica of a good website, which looks very authentic to users.

Phishing attacks usually aim to acquire confidential information like usernames, passwords and financial IDs by tricking users. Phishing attacks typically start by sending an email that appears to come from authentic company to victims requesting them to update or validate their information by visiting a link within the email.

The idea is that bait is dropped out hoping that a user will take it and bite into it just like the fish. Usually, bait is an instant messaging website or an email, which will take the user to hostile phishing websites(James, 2005).

The motivation behind this study is to make a strong and effective technique which uses Data Mining algorithms and mechanisms to detect phishing websites. Associative and classification algorithms can be very helpful in identifying Phishing websites. It can give us answers about what the most important phishing website features and indicators are and how they link to each other. Comparing between various Data Mining classification and association systems and techniques is also a goal of this study since there are only few investigations that compares different data mining methods in predicting phishing websites.

Literature Review

Numerous methodologies are being implemented at present to classify phishing websites.(Aburrous, Alamgir, Keshav, & Fadi, 2009) suggests a method for intelligent phishing detection using fuzzy data mining. In this study, ebanking phishing website detection degree is achieved based on six attributes: URL & Domain Identity, Security and Encryption, Source Code and Java script, Page Style and Contents, Web Address Bar, and Social Human Factor. Fuzzy logic and data mining algorithms are applied to classify ebanking phishing websites.

(Basnet, Ram, Srinivas, & Sung, 2008) adopts machine learning way for identifying phishing attacks. Support vector machine, biased support vector machine and neural network are used for the effective prediction of phishing emails. The objective of this study is to classify phishing emails by combining basic features in phishing emails and utilizing several machine learning algorithms for the classification process.

(Mohammad, Fadi, & Lee, 2013) suggested an intelligent prototype for predicting phishing attacks based on Artificial Neural Network. Same authors shed light on the key features that classify phishing websites from real ones and evaluate how good rulebased data mining classification methods are in detecting phishing websites and which classification approach is proven to be more reliable (Mohammad, Lee, & Fadi, 2014).

Methodology

- **Dataset**

Dataset used for the research is "Phishing Websites Data Set" ("UCI Machine Learning Repository: Phishing Websites Data Set," 2016). This dataset was gathered mainly from: PhishTank archive, MillerSmiles archive, Google's searching operators.

The authors shed light on the key features that have been proven to be solid and efficient in predicting phishing websites while proposing some new features, experimentally assigning new rules to some wellknown features and updating some other features.

The dataset is divided into 3 parts, training set and 2 test sets. The training set has 11055 and test sets have 2456 and 2670 instances. All instances are categorized as 1 for "Legitimate", 0 for "Suspicious" and 1 for "Phishy".

Dataset phishing criteria is divided into 4 sections (Address Bar based Features, Abnormal Based Features, HTML and JavaScript based Features and Domain based Features) and it has 30 attributes.

Table 1: Phishing features

Features group	Features Factor Indicator
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 Separated by (-) to the Domain
	Sub Domain and Multi Sub Domains
	HTTPS (Hyper Text Transfer Protocol with Secure Sockets Layer)
	Domain Registration Length
	F avicon
	Using NonStandard Port
	The Existence of "HTTPS" Token in the Domain Part of the URL
Abnormal Based Features	Request URL
	URL of Anchor
	Links in <Meta>, <Script> and <Link> 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 Popup Window
	IFrame Redirection
Domain based Features	Age of Domain
	DNS Record
	Website Traffic
	PageRank
	Google Index
	Number of Links Pointing to Page
	StatisticalReports Based Feature

- **Algorithms**

Several different machine learning algorithms were used for experiments.

1. Multilayer Perceptron (MLP)

Multilayer Perceptron is the most frequently used neural network classifier. MLP is a neural network and a neural network can be described as an artificial neural network which consists of a huge number of interconnected processing components known as neurons that act as a microprocessor. It is a mathematical model for classification of nonlinear data into distinct classes. Multilayer Perceptron is the most popular and frequently used neural network design (Bishop, 1995). The MLP is feedforward network architecture which involves two layers with one or more than one hidden layers; the layers are named as the input layer, hidden layer, the output layer.

2. Random Forest

Random forests are a mixture of tree predictors where each tree depends on the values of an arbitrary vector sampled individually and with the same allocation for all trees in the forest. The generalization error for forests converges a.s. to a limit as the amount of trees in the forest becomes great. The generalization error of a forest of tree classifiers hangs on the strength of the individual trees in the forest and the relationship between them (Breiman, 2001).

3. Decision Trees

Decision Tree Classification produces the output as a binary tree like construction called a decision tree. A Decision Tree model includes rules to predict the target variable. This algorithm scales well, even where there are changing numbers of training examples and significant numbers of attributes in big databases.

a) J48

J48 algorithm is an implementation of the C4.5 decision tree algorithm. J48 uses the greedy technique to induce decision trees for classification (Chen, Zheng, Lloyd, Jordan, & Brewer,
2004). A decisiontree model is built by examining training data and the model is used to classify hidden data

b) ReducedError Pruning (REPTree)

REPTree is a quick decision tree learner. Constructs a decision/regression tree utilizing data gain/variance and prunes it adopting reducederror pruning (with backfitting). REPTree only sorts values for numeric features once. Missing values are dealt with by splitting the related instances into pieces (i.e. as in C4.5).

c) Decision Stump

Decision stump is an algorithm for building and using a decision stump. It is typically used in combination with a boosting algorithm. Decision stump algorithm does regression (meansquared error) or classification (entropy). Missing is handled as a separate value ("DecisionStump", 2016).

d) Hoeffding Tree

A Hoeffding tree (VFDT) is an incremental, anytime decision tree induction algorithm that can learn from great data streams, supposing that the distribution generating examples does not vary over time. Hoeffding trees uses the fact that a small sample can often be adequate to choose a best splitting attribute. This idea is supported by the Hoeffding bound, which quantifies the number of observations (Hulten, Geoff, Laurie, & Pedro, 2001).

4. Rotation Forest

Rotation Forest is an ensemble technique which trains L decision trees separately, using a different set of obtained features for each tree. Rotation Forest (Rodriguez, Kuncheva, & Alonso, 2006) draws upon the Random Forest idea. The base classifiers are also separately built decision trees, but in Rotation Forest every tree is trained on the whole data set in a rotated feature space. While the tree learning algorithm constructs the classification regions using hyperplanes parallel to the feature axes, a small rotation of the axes may guide to a very different tree.

- **Feature Ranking**

Feature ranking was applied through WEKA software using Correlation Attribute Evaluation("CorrelationAttributeEval," 2016). It evaluates the value of an attribute by measuring the correlation (Pearson's) between it and the class. Nominal attributes are measured on a value by value basis by regarding each value as an indicator. A general correlation for a nominal attribute is reached at via a weighted average.

We selected all attributes whose weight is above 0.1. Those are:

- ❏ HTTPS
- ❏ URL of Anchor
- ❏ Adding Prefix or Suffix Separated by () to the Domain
- ❏ DNS Record
- ❏ Sub Domain and Multi Sub Domains
- ❏ Request URL
- ❏ Domain Registration Length
- ❏ Server Form Handler (SFH)
- ❏ Links in <Meta>, <Script> and <Link> tags
- ❏ Google Index
- ❏ Age of Domain
- ❏ PageRank

Experiments and Results

All experiments were conducted in WEKA tool ("Weka 3 Data Mining with Open Source Machine Learning Software in Java," 2016) which is an open source data mining application created in JAVA at Waikato University.

Table 2: Full training set results

Classifier	Test 1	Test 2
MLP	85.5%	85%
Random Forest	85.7%	84.5%
C4.5	74.6%	73%
REPTree	88.4%	88%
Decision Stump	86.1%	87%
Hoeffding Tree	87.3%	88.4%
Rotation Forest (REP Tree)	**89.1%**	**88.5%**
Rotation Forest (Hoeffding Tree)	88%	84.6%

The results show that Rotation Forest algorithm with REP Tree as a classifier give the best results for both test sets with 89.1% and 88.5% accuracy respectively. Other classifiers were not far behind, except C4.5 with 74.6% and 73% for two test sets.

After doing the ranking features with Correlation Attribute Evaluation, we applied the same classifiers. The results are very close to the ones with full training set. Surprisingly, MLP results improved for both test sets to 89% and 86.4%. MLP is also the best classifier for first test set with just 0.1% drop in comparison to Rotation Forest with REP Tree results with the full training set. REP Tree was the best classifier for test set 2 with 87.6% correct classification.

The drop in correct classification after feature reduction is applied is 1.17%.

Table 3: Reduced training set results

Classifier	Test 1	Test 2
MLP	**89%**	86.4%
Random Forest	81.8%	80.2%
C4.5	73.9%	73%
REPTree	87.1%	**87.6%**
Decision Stump	86.1%	87%
Hoeffding Tree	82.1%	83.4%
Rotation Forest (REP Tree)	88.9%	87%
Rotation Forest (Hoeffding Tree)	87.5%	84%

If we compare two result tables, we can see that Rotation Forest with REP Tree as a classifier gives the overall best results with 88.37% correct classification, while MLP outshines all other classifiers when feature reduction is applied.

Discussion

(Mohammad et al., 2014) conducted the similar feature selection where they selected nine features (Request URL, Age of Domain, HTTPS and SSL, Website Traffic, Long URL, Sub Domain and Multi Sub Domain, Adding prefix or Suffix Separated by (−) to Domain, URL of Anchor and Using the IP Address). If we compare their selected attributes with ours, we can see that we share 6 same features: Request URL, Age of Domain, HTTPS and SSL, Sub Domain and Multi Sub Domain, Adding prefix or Suffix Separated by (−) to Domain and URL of Anchor).

Moreover, all of the 30 features fall within 4 different feature groups: Address Bar based Features, Abnormal Based Features, HTML and JavaScript based Features, and Domain based Features. However, none of the 12 selected feature falls within "HTML and JavaScript" based Features. This raises the question whether this group of features is relevant in classification of phishing websites.

Conclusion

Phishing websites detection has gotten a colossal consideration by greater part of the individuals as it serves to recognize the undesirable data and dangers. Hence, the greater part of the analysts focuses in discovering the best classifier for recognizing phishing websites.

This work models the phishing website prediction as a classification task and presents the machine learning approach for predicting whether the given website is legitimate website or phishing. Multilayer perceptron, Decision tree classifiers, and Rotation Forest have been applied for training the prediction model. Training set of 11055 and two test sets of 2456 and 2670 instances with 30 attributes have been

prepared in order to facilitate training and implementation.

From the results it has been found that the Rotation Forest algorithm with REP Tree as a classifier and MLP performs the best on a full training and on reduced set, respectively. When training set was reduced from 30 attributes to 12, the overall results for all classifiers dropped for 1.17%. In the meantime, MLP's overall results increased from 85.5% to 87.7%.

It is hoped that more interesting results will follow on further exploration of data.

References

- Liu, Jiming, and Yiming Ye. *Ecommerce Agents: Marketplace Solutions, Security Issues, and Supply and Demand*. Berlin: Springer, 2001. Print.
- Aburrous, M. R., Alamgir, H., Keshav, D., & Fadi, T. (2009). Modelling Intelligent Phishing Detection System for Ebanking Using Fuzzy Data Mining. In *2009 International Conference on CyberWorlds*. http://doi.org/10.1109/cw.2009.43
- Basnet, R., Ram, B., Srinivas, M., & Sung, A. H. (n.d.). Detection of Phishing Attacks: A Machine Learning Approach. In *Studies in Fuzziness and Soft Computing* (pp. 373–383).
- Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford University Press. Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32.
- Chen, M., Zheng, A. X., Lloyd, J., Jordan, M. I., & Brewer, E. (n.d.). Failure diagnosis usingdecision trees. In *International Conference on Autonomic Computing, 2004. Proceedings.* http://doi.org/10.1109/icac.2004.1301345
- CorrelationAttributeEval. (n.d.). Retrieved May 9, 2016, from http://weka.sourceforge.net/doc.dev/weka/attributeSelection/CorrelationAttributeEval.html
- DecisionStump. (n.d.). Retrieved May 9, 2016, from http://weka.sourceforge.net/doc.dev/weka/classifiers/trees/DecisionStump.html
- Hulten, G., Geoff, H., Laurie, S., & Pedro, D. (2001). Mining timechanging data streams. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining KDD '01*. http://doi.org/10.1145/502512.502529
- James, L. (2005). *Phishing Exposed*. Syngress.
- Liu, J., & Ye, Y. (2003). *ECommerce Agents: Marketplace Solutions, Security Issues, and Supply and Demand*. Springer.
- Mohammad, R. M., Fadi, T., & Lee, M. (2013). Predicting phishing websites based on selfstructuring neural network. *Neural Computing & Applications*, *25*(2), 443–458.
- Mohammad, R. M., Lee, M., & Fadi, T. (2014). Intelligent rulebased phishing websites classification. *IET Information Security*, *8*(3), 153–160.
- Rodriguez, J. J., Kuncheva, L. I., & Alonso, C. J. (2006). Rotation Forest: A New Classifier
- Ensemble Method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(10), 1619–1630.
- UCI Machine Learning Repository: Phishing Websites Data Set. (n.d.). Retrieved May 9, 2016, from https://archive.ics.uci.edu/ml/datasets/Phishing+Websites
- Weka 3 Data Mining with Open Source Machine Learning Software in Java. (n.d.). Retrieved: May 9, 2016, from http://www.cs.waikato.ac.nz/ml/weka/