Asian Journal of Information Technology 15 (21): 4422-4425, 2016
ISSN: 1682-3915

Experimental Method to Improve the Accuracy of Phishing URL Detection Using Theil Classifier

R. Rakesh and A. Kannan
Department of Information Science and Technology, College of Engineering Guindy,
Anna University, 600025 Chennai, India

Abstract: Web security pertains with the proposal of efficient security measures to guard against attacks carried over the internet. Different attacks such as denial of service, cross site scripting, injection, authentication and session management, social engineering, etc., exist as a hindrance to web services and end users. Phishing is a kind of social engineering attack. Phishing is a malicious activity where personal and confidential information from the end user is obtained by luring them towards an illegitimate web page or Uniform Resource Locator (URL). In this study, a novel approach to anti-phishing using Theil decision tree classifier is proposed, where the proposed algorithm computes optimal node values, essential in identifying the splitting attribute for the constructed decision tree which is then used to classify malicious web pages or URL's.

Key words: Web security, anti-phishing, decision tree, classification, India

INTRODUCTION

Phishing is a type of malicious activity in which the attackers develop similar looking target web pages and lure unwary victims to reveal confidential information such as passwords, bank details, credit or debit card details, etc. Phishing has lead to huge economic losses over the years. In the year 2014, Phishing attacks cost organizations $4.5 billion in monetary losses as per the statistics released by RSA. In such cases, data mining techniques may prove useful in detecting and classifying such malicious activity. Decision trees are one of the familiar methods for feature classification. Decision tree classifiers require pre-classified dataset in order to learn patterns. In this study, we propose a decision tree that uses Theil index, as the splitting criterion for creating a knowledge base. This knowledge base helps our classifier to decide whether the suspected web page or URL is malign or not. From our test results obtained it is seen that our proposed classifier which uses the knowledge base obtained from the training data set, works efficiently when compared with other classifiers.

Literature review: In this study, we look at the different methodologies followed for identifying Phishing attacks. Many anti-phishing techniques have been proposed so far. McGrath and Gupta (2008) conducted a comparative analysis on phishing and legitimate web pages. Zhang *et al.* (2007), developed a tool called CANTINA which was used to classify phishing web pages by analyzing the keywords of a given web page using Term-Frequency metric. Later, they extended their work by adding 8 predominant features and proposed CANTINA+ (Xiang *et al.*, 2011) that depended upon third party services and Document Object Model (DOM) of the HTML page, in order to find phishing web pages. Whittaker *et al.* (2010) determined a given web page is phish or not by analyzing the given web page URL as well as the contents of the web page. Another familiar approach is blacklist approach. A blacklist contains a list of known phish web pages. PhishTank maintains a blacklist and updates it periodically as users identify and report it to PhishTank. These lists are implemented within commercial browsers such as Firefox, Internet Explorer and Chrome. Similarity based approach is another technique which employs visual features found in a web page such as layout, images, style, etc. (Chen *et al.*, 2010; Fu *et al.*, 2006). Sheen and Anitha (2012) proposed a new criterion for attribute node splitting using an inequality measure called Theil Index for the detection of malware samples. Ganapathy *et al.* (2013) surveyed the different classification techniques used in Intrusion Detection Systems (IDS). Sindhu *et al.* (2012) proposed a wrapper based feature selection algorithm to select an optimal feature set, by deploying a neural network procedure that increases the specificity and sensitivity of a malicious activity detector system.

Corresponding Author: R. Rakesh, Department of Information Science and Technology, College of Engineering Guindy, Anna University, 600025 Chennai, India

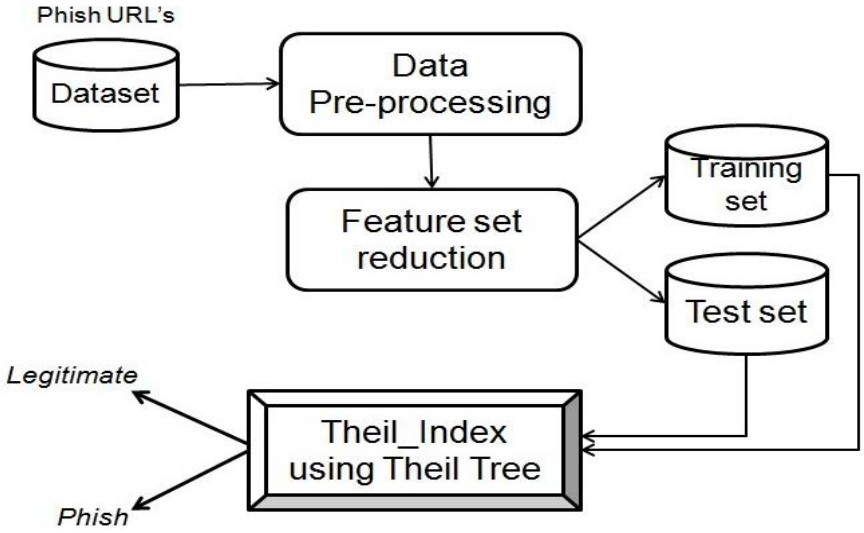

Fig. 1: Phishing web page identification using Theil Tree

Pradeepthi and Kannan (2015) proposed a classifier algorithm that classifies cloud based DOS attacks.

MATERIALS AND METHODS

Proposed system: In this study, a basic overview of how the structure of our proposed system works along with the function of the different components is discussed. The architecture of the proposed system is given in Fig. 1.

Theil tree classifier algorithm: Theil index which is used to measure the inequality or discrepancy between and within groups can be used to measure an attribute's impurity. Theil index for each and every attribute can be calculated for a given dataset so that an attribute that gives the least Theil index value can be identified. This attribute can be used as the splitting attribute while constructing the decision tree in order to output a tree with less height. The algorithm for the construction of Theil tree is given as. Theil Index value is calculated as follows:

$$N_r = \sum_{i=1}^{child_{root_{node}}} \frac{f_i}{f_p} \log\left(\frac{f_i / f_p}{m_i / m_p}\right) + \sum_{i=1}^{child_{root_{node}}} \frac{f_i}{f_p} N(r,i)$$

Where:

f_i = Denotes features that contributes much to the major output class

f_p = Denotes features that can classify itself as parent

m_i = Denotes the number of features that actually classifies at ith node

m_p = The total number of features in parent node

Algorithm Theil_tree (Dataset S, attribute list):

{
Input : Dataset S, a training feature set
An attribute list $\{(x_1,y_1)\ldots(x_n,y_n)\}$, where x denotes a specific value for a given y class label
Theil index as node splitting criterion
Output: A decision tree
create a node N
if attribute list is empty, then
return N as the leaf node with majority class in S
if the training feature set are all of the same class C, then
return N as the leaf node labeled with the class C
calculate Theil_index(x_i)
select Theil attribute, the attribute among attribute list with the least Theil index
label node N with with Theil attribute
attribute list = attribute list – N
for each known value, V of Theil index
grow a branch from node N for different test conditions
let S_v be the set of tuples in S satisfying V
repeat step 7 again
if S_v is empty, then
attach a leaf node labeled with the major class in S to node N
else, if no outcome is possible, attach the node returned by Theil_tree(S_v, attribute list) to node N
return N

Experimental setup: A dataset consisting of phishing URL's was used initially. In pre-processing, shortened URL's like URL's with bit.ly, tinyurl.com, tiny.cc, adf.ly, were removed using LongURL's url shortening service list. Also, non-english URL's were removed manually. Further, the dataset obtained after pre-processing was

Table 1: Frequency of top 8 features

Feature type	No. of URL's
URL link (the path of the phish or legitimate web page)	971
Is domain name used in URL	590
usage of HTTPS or SSL	903
URL length	887
usage of hyphen as domain separators	395
usage of IP address in URL's	354
URL's having @ symbol	272
Presence of multiple sub domains	598

Table 2: Performance analysis of Theil tree along with various classifiers

	Accuracy rate for the number of features used	
Classifier used	4	8
Random tree	96.43	95.71
ID3	93.40	94.87
CART	95.52	94.15
Naïve Baiyes	92.09	93.39
SVM	89.33	88.27
Theil classifier	96.68	96.04

analysed for the presence of heuristic features and the frequency of the top 8 features present in the dataset is given as in Table 1. Other features such as age of domain, website traffic, DNS record, Alexa reputation was not considered since the phish websites are short lived and hence these features won't be valid for a longer period of time. Hence, the top 8 heuristic features that are used in common while designing a phish web page is alone considered. It is followed by constructing the Theil Tree where Theil index values are calculated.

RESULTS AND DISCUSSION

This section briefs about the dataset used in the identification of phishing web pages and also discusses about the results obtained and analysis done as part of our proposed work. Also, the evaluation metrics used and the performance evaluation of our proposed system is discussed in brief. The initial dataset was created after obtaining phishing URL's. Table 1 shows the different phishing features that were considered for the experiment and their frequency found in the collected dataset.

Phishing URL's were obtained from PhishTank repository. A total of 4970 random phishing URL's targeting popular brand names such as Paypal, Amazon, Facebook, Twitter, etc. were collected. For the training dataset, legitimate URL's were obtained from DMOZ repository and Yahoo Directory. An approximate number of 1000 legitimate URL's were collected. So in total, the training dataset contained nearly 6000 URL's. For our proposed work an open source data mining tool called WEKA was used in order to evaluate the performance of our proposed method with other classifiers. The tool contains many machine learning algorithms that can be used for data mining tasks such as classification,

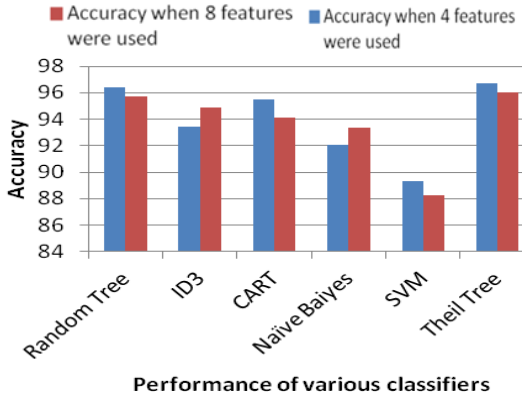

Fig. 2: Performance evaluation of Theil tree along with other classifiers

clustering, regression etc. Accuracy is defined as the ratio between the total number of samples that are correctly classified to that of the total samples present. It is given by:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

Where:

True Positive (TP) = No. of URL's correctly classified as phish

True Negative (TN) = No. of URL's correctly classified as legitimate

False Positive (FP) = No. of URL's incorrectly classified as phish

False Negative (TN) = No. of URL's incorrectly classified as legitimate

The dataset was given as input to various classifiers using WEKA tool and also to our proposed Theil classifier tree method. Classifiers that use different splitting criterion such as infromation gain, Gini Index were considered while evaluating the performance of our proposed method. Table 2 shows the accuracy rate of various classifiers along with our proposed method. Figure 2 shows the performance achieved by the proposed Theil classifier in comparison with other classifiers. It is seen that, the proposed Theil classifier scales up pretty well even when the number of features used is increased.

CONCLUSION

In this research, a new classification algorithm for classifying phishing URL's which uses Theil Index as splitting criterion has been proposed. A total of 8 heuristic features were considered based upon the

frequency presence of the features in the accumulated dataset. A decision tree learns patterns from a pre-classified feature set. Based upon the comparison and performance evaluation done along with various other commonly used classifiers, it is concluded that our method achieves improved accuracy.

ACKNOWLEDGEMENTS

The manuscript entitled "Experimental Method to Improve the Accuracy of Phishing URL Detection Using Theil Classifier", R. Rakesh and A. Kannan is submitted for the possible publication in Asian Journal of Information Technology. The manuscript has not been previously published is not currently submitted for review to any other journal and will not be submitted elsewhere before decision is made. This research was supported and funded in part by Anna University as part of Anna Centenary Research Fellowship.

REFERENCES

Chen, T.C., S. Dick and J. Miller, 2010. Detecting visually similar web pages: Application to phishing detection. ACM Trans. Internet Technol., 10: 1-38.

Fu, A.Y., W.Y. Liu and X. Deng, 2006. Detecting phishing web pages with visual similarity assessment based on Earth Mover's Distance (EMD). IEEE Trans. Dependable Secure Comput., 3: 301-311.

Ganapathy, S., K. Kulothungan, S. Muthurajkumar, M. Vijayalakshmi, P. Yogesh and A. Kannan, 2013. Intelligent feature selection and classification techniques for intrusion detection in networks: A survey. EURASIP J. Wireless Commun. Network. 10.1186/1687-1499-2013-271.

McGrath, D.K. and M. Gupta, 2008. Behind phishing: An examination of phisher modi operandi. Proceedings of the 1st Usenix Workshop on Large-Scale Exploits and Emergent Threats, April 15, 2008, San Francisco, CA, USA., pp: 1-8.

Pradeepthi, K.V. and A. Kannan, 2015. Cloud attack detection with intelligent rules. KSII. Trans. Internet Inf. Syst., 9: 4204-4222.

Sheen, S. and R. Anitha, 2012. A Novel Node Splitting Criteria for Decision Trees Based on Theil Index. In: Neural Information Processing, Tingwen, H., Z. Zeng, L. Chuandong and C.S. Leung (Eds.). Springer, Berlin, Germany, ISBN:978-3-642-34480-0, pp: 435-443.

Sindhu, S.S.S., S. Geetha and A. Kannan, 2012. Decision tree based light weight intrusion detection using a wrapper approach. Expert Syst. Applic., 39: 129-141.

Whittaker, C., B. Ryner and M. Nazif, 2010. Large-scale automatic classification of phishing. Proceedings of the Network and Distributed System Security Symposium, February 28-March 3, 2010, San Diego, California, USA -.

Xiang, G., J.I. Hong, C.P. Rose and L.F. Cranor, 2011. CANTINA+: A feature-rich machine learning framework for detecting phishing web sites. ACM Trans. Inform. Syst. Secur., 14: 21-48.

Zhang, Y., J. Hong and L. Cranor, 2007. CANTINA: A content-based approach to detecting phishing web sites. Proceedings of the 16th International Conference on World Wide Web, May 8-12, 2007, Banff, Alberta, Canada, pp: 639-648.