VOL.5 NO.1 March 2016 - ISSN 2233 – 1859

Operations Research Society in Bosnia and Herzegovina

Southeast Europe Journal of Soft Computing

Available online: http://scjournal.ius.edu.ba

IUS Soft Computing Research Group

Malicious Web Sites Detection using C4.5 Decision Tree

Zerina Masetic, Abdulhamit Subasi, Jasmin Azemovic

International Burch University, Faculty of Engineering and Information Technologies,
71000, Sarajevo, Bosnia and Herzegovina
absubasi@effatuniversity.edu.sa
zerina.masetic@ibu.edu.ba
jasmin.azemovic@ibu.edu.ba

Article Info

Article history:
Received 10 Jan.2016
Received in revised form 16 Jan 2016

Keywords:
Malicious Web Sites, Blacklisting, URL, C4.5 Decision Tree

Abstract

The technology advancement poses the challenge to the cybercriminals for doing various online criminal acts, such as identity theft, extortion of money or simply, viruses and worms spreading. The common aim of the online criminals is to attract visitors to the Web site, which can be easily accessed by clicking on the URL. Blacklisting seems not to be the successful way of marking Web sites with the "bad" content, considering that many malicious Web sites are not blacklisted. The aim of this paper is to evaluate the ability of C4.5 decision tree classifier in detecting malicious Web sites, based on the features that characterize URLs. The classifier is evaluated through several performance evaluation criteria, namely accuracy, sensitivity, specificity and area under the ROC curve. C4.5 decision tree classifier achieved significant success in malicious Web sites detection, considering all four criteria (accuracy 96.5, sensitivity 96.4, specificity 96.5 and area under the curve 0.958).

1. INTRODUCTION

In recent years, with the advancements in technology and new e-commerce opportunities, World Wide Web has become fertile platform for wide range of internet criminal acts, starting from the online identity theft over some financial fraud to the hacking and spreading different types of viruses and worms. Internet crime refers to the any illegal activity committed on the Internet [1].

However, the specific commercial initiatives behind these acts may differ. Still, the common aim is to attract visitors to the Web site. These Web sites can be accessed via links through email, Web search results or links on the other Web sites by clicking on, so-called Uniform Resource Locator (URL). URL provides addressing scheme allowing the browser to request any Web page or document located on the Internet. Nonetheless, every time the user decides to click on specific URL, he needs to examine possible risk of accessing the specific Web site. Still, most common technique that marks the Web sites as "bad" is blacklists. In blacklists, URLs, hosts or networks are marked as containing malicious content. These lists are distributed to subscribers who use the information, to block any activity from or to the malicious Web site. Blacklisting can be done by Web sites blacklist entities (e.g. Google Blacklist, Norton Safe Browsing, etc.) or by AntiVirus products (e.g. AVG, AVAST, ESET, etc.) [2]. On the other hand, many malicious Web sites are not blacklisted; either being too new or being evaluated inaccurately [3]. Henry Harrsion, technical director at Detica, pointed out that anti-virus blacklisting "cannot detect things that are bad but not known" [4].

According to the security report done by Cisco [5], malicious actors continue innovating ways to exploit public trust for harmful consequences. Due to the malicious activity and Web site content, Cisco blocked 80 million web requests every day, which confirms enormous number of malicious activities done through the Internet and Web sites [5].

Web sites can be accessed by typing the domain name of the Web sites or its IP address. Domain names or IP addresses are used to identify any Web site. Domain name is contained in any URL, which defines the location of the required document, page or information on Web site. URL is composed of the protocol name, domain name and directory path. Considering following URL http://www.ibu.edu.ba/en/research/other/laboratories.html, we can notice protocol http, domain name is www.ibu.edu.ba together with Web server information and the rest of it is path to the page showing laboratories. The part ".ba" is so called URL country code, pointing out that domain name is registered in Bosnia and Herzegovina. URL cannot contain any space and it's case-sensitive. Moreover, URL can contain some special characters, such as "/", "?", "=", "-" and "_" [6].

The aim of this paper is to detect and classify the malicious Web sites, based on the content of URL, by applying C4.5 decision tree algorithm. When considering the content of URL, we will base on host-based and lexical-based features that characterize URLs.

The next section will briefly review recent work on malicious URLs detection. Host-based and lexical-based features, data collection and classification methodologies will be explain in Section 3. Results and performance evaluation will be given in Section 4. Moreover, we will discuss our findings and give summary of our work in Section 5.

2. LITERATURE OVERVIEW

In recent time, researchers examined various techniques for malicious Web site detection. However, one research to be highlighted is done by Ma et al. [3], where authors examined the ability of Naïve-Bayes, Support Vector Machine and Logistic Regression classifiers in detecting the malicious Web sites from suspicious URL. These authors achieved remarkable success in classifying suspicious URLs, by achieving 95-99% classification accuracy. Furthermore, Kazemian and Ahmed [7] compared three supervised machine learning techniques, namely k-Nearest Neighbor, Support Vector Machine and Naïve Bayes classifiers, and two unsupervised machine learning techniques, namely k-Means and Affinity Propagation in detecting malicious webpages. In detecting 100,000 webpages, supervised machine learning classifiers achieved 89-98% classification accuracy and unsupervised machine learning classifiers achieved 0.88 – 0.96 silhouette coefficient. Moreover, interesting research was done by Stevanovic, Vlajic and An [8]. They investigated ability of unsupervised neural network learning in detecting malicious and non-malicious Web site visitors. Based on the browsing behavior of Web site visitors, they investigated differences or similarities between malicious crawlers and non-malicious Web site visitors. Interesting finding is that 52% of malicious crawlers showed "human-like" browsing behavior. Spreading malicious URLs is quite popular on social networks. Therefore, Chen, Guan and Su [9] used Bayesian classifier to detect malicious URLs in social network environment, based on URL information and social behavior of users and obtained 95.7% classification accuracy.

In this paper, we investigated ability of C4.5 decision tree classifier in detecting malicious Web sites based on the host-based and lexical-based features that characterize URLs. Considering that, classification model presented on Fig. 1. is followed.

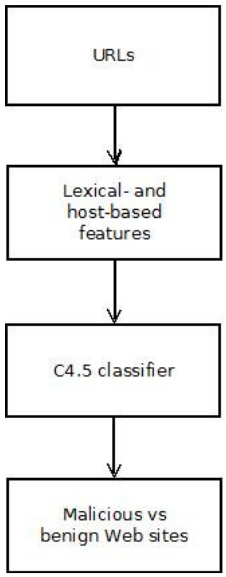

Fig. 1. – Block diagram of the classification process

3. METHODOLOGY

3.1 DATASET

Data set used for the purpose of this study is composed of 5000 different URLs, where 3324 URLs belong to the group of benign Web sites and 1676 URLs belong to the group of malicious Web sites. This data set is a part of larger study data set, created by Ma et al. [3] and freely available on http://sysnet.ucsd.edu/projects/url/. Originally, URLs are obtained from different resources. Benign URLs are obtained from DMOZ Open Directory Project [10] and random URL selector for Yahoo directory. Malicious URLs are obtained from PhishTank [11] and Spamscatter [12]. Data set used is composed of URLs from all four URL sources.

Features used to categorized URLs are either lexical or host – based. Lexical features are textual parts of the URLs which allow us to capture the attribute of malicious sites to look different in the eyes of the users. These features include length of the hostname, length of the whole URL and number of dots in URL – as real valued features. Moreover, binary features for each token in the hostname – delimited by "." and in the whole URL – delimited by "/", "?", "=", "-" and "_" are created [3,13,14].

Host – based features are used to see where the Web site is hosted and what the reputation of the hosting center is. These properties could be identified from the hostname part of the URL by checking whether the IP address is on the blacklist or not; what is the contact information of the owner or registrar of the domain, date of the registration, date of an update or expiration or any additional information related to the domain; what is time-to-live (TTL) value which determines how long it will take any change made to go into effect; what is the country code and what is the connection speed [3,13,14].

3.2 C4.5 DECISION TREE

Decision tree is a method that classifies instances by arranging them from the root node down the tree, to some leaf node, where each internal node represent test for some attribute of the tree and has no outgoing edges. The root is a node without incoming edges. The other nodes have exactly one incoming edge and are called leaves. The classification starts at the root node with testing the attribute specified by this node and continue down the tree branch according to the value of the attribute in the example. When a leaf node is reached, the instance is classified according to the class of the leaf [15].

Input values of the decision tree algorithm are [16]:

- Data partition D, which is actually training dataset together with targeted output;
- attribute_list, presenting the set of attributes used;
- Attribute_selection_method, a procedure used to determine the best splitting criterion, for partitioning dataset into individual classes.

An attribute selection measures is used for selecting the splitting criterion that best separates some dataset D. Popular attribute selection measure for C4.5 decision tree approach is gain ration [16]. C4.5 algorithm is based on ID3 algorithm, a very simple decision tree algorithm, presented by Quinlan [17]. This algorithm passes through decision tree, visits each node and select optimal split. It is achieved by using the gain ratio, represented by following formula [16]:

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo(A)}$$

Gain or information gain is attribute selection measure used in ID3 approach. In information gain, an attribute with the highest information gain is chosen as splitting attribute for the node N. So, this attribute minimizes the information needed to classify tuples *D* in a partition and returns minimum "impurity" in these partitions. Information gain is difference in entropy from before to after the set D is partitioned on attribute A. Moreover, it checks how much uncertainty in D is reduced after it is partitioned on attribute A. The uncertainty in the data set D is measured by entropy calculated as:

$$Entropy(D) = -\sum_{x \in X} p(x) \log_2 p(x)$$

where X is the set of classes in D and $p(x)$ is the proportion of number of elements in class x to the number of elements in set D. When entropy is 0, data set is perfectly classified [16].

SplitInfo is the term which describes how equally the attribute splits the data and is calculated by formula [16]:

$$SplitInfo(A) = -\sum_{j=1}^{n} \frac{|D_j|}{|D|} \times \log_2 \left(\frac{|D_j|}{|D|} \right)$$

The term $\frac{|D_j|}{|D|}$ represents the weight of *jth* partition.

Moreover, tree pruning is important step in decision tree algorithm. It is the method that is used to solve the problems of overfitting the data in trees. This technique removes sections of the tree that provide less power in instances classification. Pruned trees tend to be smaller and less complex. They are faster and better in correctly classifying the data than unpruned trees [18].

4. RESULTS

In this section, we evaluate the ability of C4.5 decision tree classifier in detecting malicious Web sites, based on the features of URLs. The ability of classifier is evaluated using different performance evaluation criteria, such as accuracy, sensitivity, specificity and ROC curve. Initially, 10-fold cross validation is applied to the whole data set, which is composed of 5000 different URLs. In 10-fold cross validation [19], the whole data set is divided into 10 random partitions, or folds of equal size, in which classes are represented in approximately same proportion in each partition. One tenth is used for testing and remaining nine tenth are used for training the classifier. The procedure is repeated 10 times, where each fold is taken as test and remaining nine as training. Finally, the 10 error estimates are averaged and overall error estimate is found [20].

Classification accuracy is one of the most significant criteria in classification process. It evaluates how accurately the classifier will classify future data in dataset. However, classifier might be trained to classify only "positive" or "negative" sets of data. Therefore, two new terms are introduced: sensitivity and specificity [15].

Sensitivity specifies how good the classifier can recognize positive samples and is defined by [15]:

$$Sensitivity = \frac{TP}{TP + FN} \times 100$$

where TP is number of true positive samples and FN is the number of false negative samples. In our case, sensitivity

measures proportion of malicious Web sites that are correctly classified as such.

Specificity specifies how good the classifier can recognize negative samples and is defined by [15]:

$$Specificity = \frac{TN}{TN + FP} \times 100$$

where TN is number of true negative samples and FP is the number of false positive samples. In our case, specificity measures proportion of benign Web sites that are correctly classified as such.

Based on the definitions above, the accuracy can be defined as a function of sensitivity and specificity [15]:

$$Accuracy = \frac{Sensitivity + Specificity}{2}$$

Receiver operating characteristic (ROC) curve is the term used to describe the classifier performance without regard to the class distribution or error rate. The ROC curve is the graph of sensitivity on y – axis, which represents percentage of the total number of positive samples, versus 1- specificity on x – axis, representing percentage of the total number of negative samples. The ideal point on the ROC curve would be (0, 1), meaning that all positive instances are classified as positive and none of negative instances are misclassified as positive. The classification performance is measured by mean area under the curve (AUC). The bigger area it is, the better classifier model is [15].

Moreover, classification model with C4.5 decision tree classifier is created and results are obtained. Results are shown through above mentioned performance criteria in Table 1.

Table 1 - Performance evaluation of C4.5 decision tree

Perf. crit. / Web types	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC
Benign	98.3	96.4	96.5	0.958
Malicious	92.9			0.958
Average	96.5			0.958

According to the Table 1, C4.5 decision tree model is 96.5% accurate in detecting malicious Web sites. However, important for classifier is to be able to classify both, positive and negative class samples of the data set. Ability of the classification model to classify both, positive and negative samples is presented through sensitivity and specificity values. Sensitivity value of the model is 96.4%, which shows proportion of malicious Web sites that are detected and correctly classified. Moreover, specificity value of the model is 96. 5% which shows proportion of benign Web sites that are correctly classified. Area under the ROC curve is quite high, with the average value of 0.958, showing that the rate of true positive samples is high. ROC curve is shown in the Fig. 2.

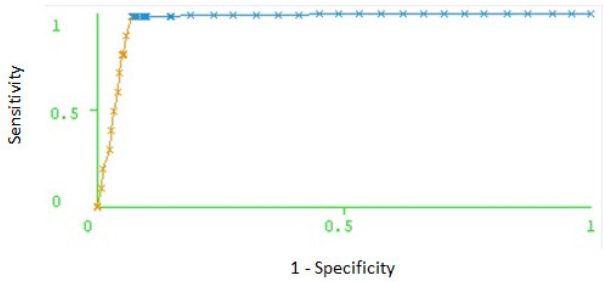

Fig. 2. ROC curve for benign and malicious classes

5. CONCLUSIONS

The results in previous section confirmed the ability of the proposed C4.5 decision tree algorithm in distinguishing malicious and benign Web sites, based on the URL features. When considering all statistical indices in Table 1, it is obvious that C4.5 decision tree algorithm achieved significant success in detecting malicious Web sites. Characteristics of classifier, together with features selected determine classifier's performance.

Popular techniques in previous works were Naïve Bayes and Support Vector Machine. Ma et al. [3] achieved remarkable success in identifying malicious Web sites using both of the mentioned techniques, with 95-99% classification accuracy. Moreover, both techniques are used by Kazemian and Ahmed [7], who achieved 89-98% classification accuracy. Our proposed technique is C4.5 decision tree, which is simple, but powerful algorithm, similar to human decision process. However, performance of classifier depends on several factors, such as the size and complexity of the tree. Optimal classifier's performance was achieved by pruning the tree, which reduced the tree size and its complexity, using the classification accuracy as fitting function. The proposed technique achieved remarkable success and is capable of detecting malicious Web sites, based on the URL features.

6. REFERENCES

[1] B. Henson, B. Reyns, and B. Fisher, "Internet crime," *Key Issues in Crime and Punishment: Crime and criminal behavior*, pp. 155-168, 2011.

[2] My Website is Blacklisted – This Site May be Hacked. [Online]. https://sucuri.net/my-website-is-blacklisted-this-site-may-be-hacked

[3] Justin Ma, Lawrence K. Saul, Stefan Savage , and Geoffrey M. Voelker, "Beyond Blacklists: Learning to Detect Malicious Web Sites from Suspicious URLs," in *KDD '09 Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, 2009, pp. 1245-1254.

[4] Kevin Townsend. (2011, August) Security centric issues,

news and rants - and other things. [Online]. https://kevtownsend.wordpress.com/2011/08/24/whitelisting-vs-blacklisting/

[5] "Cisco 2014 Annual Security Report," Cisco Systems, Report 2014.

[6] Dave Taylor, *Creating Cool Web Sites with HTML, XHTML, and CSS*. USA: Wiley Publishing, Inc., 2004.

[7] H. B. Kazemian and S. Ahmed, "Comparisons of machine learning techniques for detecting malicious webpages," *Expert Systems with Applications*, pp. 1166-1177, 2015.

[8] Dusan Stevanovic, Natalija Vlajic, and Aijun An, "Detection of malicious and non-malicious website visitors using unsupervised neural network learning.," *Applied Soft Computing*, vol. 13, pp. 698-708, 2013.

[9] Chia-Mei Chen, D. J. Guan, and Qun-Kai Su, "Feature set identification for detecting suspicious URLs using Bayesian classification in social networks," *Information Sciences*, vol. 289, pp. 133-147, 2014.

[10] Netscape. DMOZ Open Directory Project. [Online]. http://www.dmoz.org

[11] OpenDNS. PhishTank. [Online]. http://www.phishtank.com

[12] D. S. Anderson, C. Fleizach, S. Savage, and G. M. Voelker, "Spamscatter: Characterizing Internet Scam Hosting Infrastructure. ," in *Proc. of the USENIX Security Symposium*, Boston, 2007.

[13] D. K. McGrath and M. Gupta, "Behind Phishing: An Examinaton of Phisher Modi Operandi," in *Proc. of the USENIX Workshop on Large-Scale Exploits and Emergent Threats (LEET)*, San Francisco, 2008.

[14] P. Kolari, T. Finin, and A. Joshi, "SVMs for the Blogosphere: Blog Identification and Splog Detection," in *Proc. of the AAAI Spring Symposium on Computational Approaches to Analysing Weblogs*, Stanford, 2006.

[15] O. Maimon and L. Rokach, *Data Mining and Knowledge Discovery Handbook*. New York: Springer Science+Business Media, 2005.

[16] Jiawei Han, Micheline Kamber, and Jian Pei, *Data Mining Concepts and Techniques*.: Elsevier Inc., 2012.

[17] J. R. Quinlan, *C4.5: programs for machine learning*.: Morgan Kaufmann Publishers, Inc., 1993.

[18] Rezaul Begg, Daniel T. H. Lai, and Marimuthu Palaniswami, *Computational intelligence in biomedical engineering*.: Taylor & Francis Group, LLC , 2008.

[19] S. Salzberg, "On comparing classifiers: pitfalls to avoid and a recommended approach," *Data Mining and Knowledge Discovery*, pp. 317-328, 2007.

[20] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed.: Elsevier Inc., 2005.