View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Electronic Archive of Kyiv Polytechnic Institute

UDC 004.9

Malicious and benign websites classification using machine learning methods

M. Lavreniuk[1], O. Novikov[1]

[1] *National Technical University of Ukraine «Igor Sikorsky Kyiv Polytechnic Institute», Institute of Physics and Technology*

Abstract

Nowadays web surfing is an integral part of the life of the average person and everyone would like to protect his own data from thieves and malicious web pages. Therefore, this paper proposes a solution to the discrimination of malicious and benign websites problem with desirable accuracy. We propose to utilize machine learning methods for classification malicious and benign websites based on URL and other host-based features. State-of-the-art gradient-boosted decision trees are proposed to use for this task and they have been compared with well-known machine learning methods as random forest and multilayer perceptron. It was shown that all machine learning methods provided desirable accuracy which is higher than 95% for solving this problem and proposed gradient-boosted decision trees outperforms random forest and neural network approach in this case in terms of both overall accuracy and f1-score.

Keywords: cybersecurity, malicious websites, machine learning, gradient-boosted decision trees, neural networks.

Introduction

The popularity of the Internet grows every year and every day. In 2010, the population of Internet users was about two billion [1] and at the end of June 2019, the population of Internet users reached more than 4.5 billion [2]. Moreover, within the technical progress in the nearest future, this number will rapidly increase. According to the popularity of the Web, it attracts the attention of hackers and people who have bad intentions. One of the possible threats for the average Web user is malware distribution when user visit malicious websites. Such malware is designed to conduct various cybercrimes, such as gaining control of the victim system, stealing private information, launching denial-of-service attacks, and spamming [1]. Another form of attack is to make a machine or network resource unavailable or make it so slow that it is practically impossible to use it [3]. Therefore, the problem of detection of malicious websites is vital nowadays. For example, 24,000 malicious mobile applications are blocked every day and information that most applications release is 63% mobile phone numbers and 37% device location [4]. From 2016 to 2017 the number of new mobile malware variants increased by 54 per cent and the percentage of cybersecurity costs increased by 22.7%, with malware attack costing companies 2.4 million dollars on average [4].

The standard approach for solving the problem of detection of malicious websites consists of blacklist databases exploitation either through extensive analysis or crowdsourcing. However, these standard approaches have issues in case of observing new attacks due to the flexibility of malicious websites [5]. To overcome these issues, in the last decade, researchers have applied machine learning techniques for malicious Uniform Resource Locator (URL) detection [5], [6]. For example, in [6], [7] the authors used only URL information for features extraction by machine learning approaches. In this paper, we propose machine learning methods for classification websites on malicious and benign based on not only URL itself but also utilizing additional information such as host-based features that could be obtained from the website.

Data description

In this experiment, we utilized dataset with 1781 samples (1565 samples of benign websites and 216 samples of malicious websites) [8]. This dataset consists of 18 features [9]:

- URL: it is the anonymous identification of the URL analyzed in the study.
- URL_LENGTH: it is the number of characters in the URL.
- NUMBER_SPECIAL_CHARACTERS: it is a number of special characters identified in the URL, such as, “/”, “%”, “#”, “&”, “. “, “=”.
- CHARSET: it is a categorical value and its meaning is the character encoding standard (also called character set).
- SERVER: it is a categorical value and its meaning is the operative system of the server got from the packet response.
- CONTENT_LENGTH: it represents the content size of the HTTP header.
- WHOIS_COUNTRY: it is a categorical variable, its values are the countries we got from the server response (specifically, our script used the API of Whois).
- WHOIS_STATEPRO: it is a categorical variable, its values are the states we got from the server

response (specifically, our script used the API of Whois).
- WHOIS_REGDATE: Whois provides the server registration date, so, this variable has date values with format DD/MM/YYY HH:MM.
- WHOIS_UPDATED_DATE: Through the Whois, we got the last update date from the server analyzed.
- TCP_CONVERSATION_EXCHANGE: This variable is the number of TCP packets exchanged between the server and our honeypot client.
- DIST_REMOTE_TCP_PORT: it is the number of the ports detected and different to TCP.
- REMOTE_IPS: this variable has the total number of IPs connected to the honeypot.
- APP_BYTES: this is the number of bytes transferred.
- SOURCE_APP_PACKETS: packets sent from the honeypot to the server.
- REMOTE_APP_PACKETS: packets received from the server.
- APP_PACKETS: this is the total number of IP packets generated during the communication between the honeypot and the server.
- DNS_QUERY_TIMES: this is the number of DNS packets generated during the communication between the honeypot and the server.

In addition, each sample had attribute TYPE: this is a categorical variable, its value represents the type of web page analyzed, specifically, 1 is for malicious websites and 0 is for benign websites. The dataset has been randomly split on training and test set in 70:30 proportion.

Methodology

In this paper, a random forest approach has been chosen as a baseline, because it is one of the traditional machine learning approaches that provides good results in any cases. Combination of multiple classifiers, in this case, individual decision trees, allows us to obtain better results compared to a single decision tree. For better results we picked the number of trees equals to 100. After training each individual decision tree, overall random forest predictions are made by taking the statistical mode of individual tree predictions for classification trees [10].

Another good machine learning approach that has been examined in this study is multilayer perceptron. Neural networks have been good recommended themselves in solving a variety of applied problems which also includes cybersecurity issues [10]. Taking into account that dataset is not very big, it was decided that one hidden layer in a feed-forward neural network will be enough for desirable performance. Forty hidden neurons with Rectified Linear Unit (ReLU) activation function has been selected [11]. The learning late parameter has been picked up as 0.001 and the coefficient of Tikhonov regularization has been chosen equals to 1.

In this paper, we proposed to utilized gradient-boosted decision trees that are one of the state-of-the-art methods for regression and numerical data classification [12]. This approach uses a smarter way of decision trees combination for better predictions compared to a simple ensemble approach in the random forest method [10]. In this experiment, we restricted the maximum depth of each tree to 5 and set up a learning rate parameter equals to 0.1.

Results

For the random forest approach, we received overall accuracy - 0.957, and f1-score [13] equals to 0.98 and 0.81 for benign and malicious websites, respectively. In addition, we could observe the importance of the features provided by the random forest method (fig.1). The most important features for discrimination malicious and benign websites are WHOIS_REGDATE and WHOIS_UPDATED_DATE.

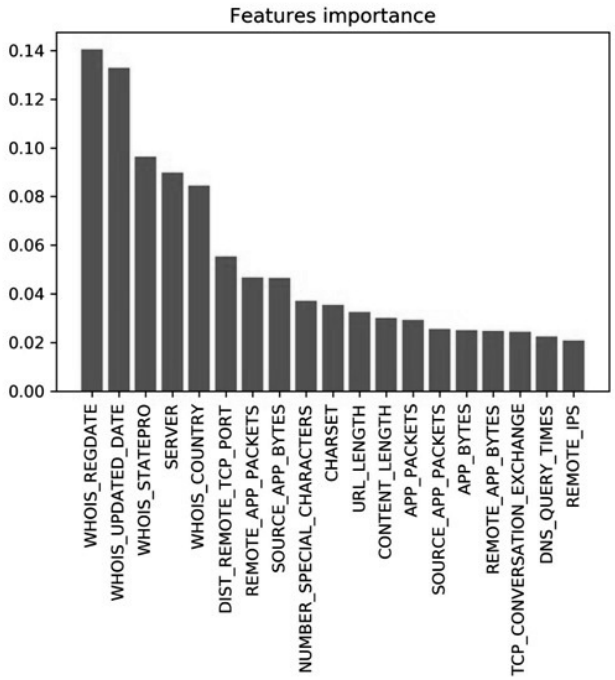

Fig. 1. The random forest features importance

Multilayer perceptron achieved overall accuracy equals to 0.9626, and f1-score equals to 0.98 and 0.85 for benign and malicious websites, respectively. Gradient-boosted decision trees provide 0.9701 overall accuracies on an independent test set, and f1-score equals to 0.98 and 0.88 for benign and malicious websites, respectively.

Conclusion

The results have shown that machine learning methods are a powerful instrument and are suitable for solving a very important cybersecurity task, such as detection malicious websites. All examined methods provided desirable quality with overall accuracy higher than 95%. In addition, all machine learning methods provided very high f1-score for benign websites and a little bit lower f1-score for the malicious website due to unbalanced training and test datasets. At the same

time, proposed gradient-boosted decision trees, which are state-of-the-art machine learning method for such tasks, outperformed random forest method and multi-layer perceptron in this task in both overall accuracies as well as in f1-score.

References

[1] C. Jian, "Analyzing and defending against web-based malware," *ACM Computing Surveys (CSUR)*, p. 49, 4 2013.

[2] I. S. 2019, "Internet world stats." `http://www.internetworldstats.com/stats.htm`.

[3] M. Uma and G. Padmavathi, "A survey on various cyber attacks and their classification," *IJ Network Security*, no. 15, 2013.

[4] I. S. T. Report, "Vol. 23." Symantec. `https://www.symantec.com/content/dam/symantec/docs/reports/istr-23-2018-en.pdf`.

[5] D. Sahoo, L. Chenghao, and H. S. CH, "Malicious url detection using machine learning: a survey," pp. 1–37, 2017. arXiv preprint arXiv:1701.07179.

[6] R. K. Nepali and W. Yong, "You look suspicious!!: Leveraging visible attributes to classify malicious short urls on twitter," *49th Hawaii International Conference on System Sciences (HICSS)*, 2016.

[7] Y. Alshboul, N. Raj, and W. Yong, "Detecting malicious short urls on twitter," pp. 1–7, 2015.

[8] `https://github.com/urcuqui/WhiteHat/tree/master/Research/Web%20security`.

[9] C. Urcuqui, A. Navarro, J. Osorio, and M. García, "Machine learning classifiers to detect malicious websites," *In SSN*, pp. 14–17, 2017.

[10] C. Chio and D. Freeman, "Machine learning and security: Protecting systems with data and algorithms," *O'Reilly Media, Inc.*, p. 385, 2018.

[11] I. Goodfellow, Y. Bengio, and A. Courville, ""deep learning"," *MIT press*, p. 800, 2016.

[12] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016.

[13] Y. Sasaki, "The truth of the f-measure," *Teach Tutor mater*, pp. 1–5, 5 2007.