ISSN 2347 - 3983
Volume 8. No. 5, May 2020
International Journal of Emerging Trends in Engineering Research
Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter53852020.pdf
https://doi.org/10.30534/ijeter/2020/53852020

Malicious domain detection based on DNS query using Machine Learning

Cho Do Xuan[1], Tisenko Victor Nikolaevich[2], Nguyen Quang Dam[3], Nguyen Quoc Hoang[4], Do Hoang Long[5]
[1,3,4,5]Information Assurance dept. FPT University, Hanoi, Vietnam, chodx@fe.edu.vn, damnqse05820@fpt.edu.vn, hoangnqse06012@fpt.edu.vn, longdhse05220@fpt.edu.vn
[2]Department Quality Systems, Peter the Great St. Petersburg Polytechnic University, Russia, St.Petersburg, Polytechnicheskaya, 29, v_tisenko@mail.ru

ABSTRACT

Currently, cyber-attacks have increased rapidly in both the number of attacks and the extent of their damage to organizations and businesses. In particular, cyber-attack techniques based on user-side vulnerabilities are developing very strongly. One of the methods that are commonly used by attackers is distributing malicious domains into users' machines. Because of the serious consequences of the distribution of malicious domains, the problem of early detection of malicious domains is very necessary today. In this paper, we propose a method of detecting malicious domains based on the connection behavior analysis technique using machine learning algorithms. The difference between our research and other studies is shown in looking for and extracting features that accurately represent the behavior of malicious domains and normal domains. Besides, in order to classify the normal domain and malicious domain, we select Random Forest (RF) supervised learning algorithms. In the experimental results, we change the parameters of the RF algorithm to seek the most optimal parameter for the algorithm when applying them to the problem of detecting malicious domains.

Key words: domain, DNS query, machine learning, malicious domain detection.

1. INTRODUCTION

Domain Name System (DNS) is a system that helps convert domain names to corresponding physical IP addresses [1]. The documents [1], [2], [3], [4] presented characteristics and main components of DNS. Because of the characteristics of the DNS operation process, attackers often exploit DNS to attack the system. The documents [1], [2], [3], [4] listed a number of attacks on the system and users through DNS vulnerabilities including DNS cache poisoning, Fast flux DNS, phishing, etc. In addition, the publication [5] has listed the extent of the damage of cyber-attacks in general and attacks on users through malicious domain distribution. According to the research of the trend of network attacks in 2020 [6], the techniques of attacks on users by spreading malicious domains are predicted to have sophisticated transformations and serious consequences. Therefore, the problem of researching and proposing detection in order to early warning about how the malicious domain works is necessary today.

The document [1], [2] classified the different types of malicious domains. Accordingly, based on the classification of these malicious domain types, the studies [3], [4] presented three methods of detecting malicious domains. In particular, the method of detecting malicious domains based on behavior analysis techniques using machine learning and deep learning techniques is highly effective because it has the ability to detect new malicious domains. In this paper, we propose a method of detecting malicious domains using machine learning based on features that represent abnormal behavior of the domain on both object-based approaches and association-based approaches..

2. RELATED WORKS

In the studies [1], [7], the authors listed the feature groups that could be selected and used to detect malicious DNS. The feature sets used including: lexical, link popularity, webpage content, DNS answers, DNS fluxiness, network features, etc. In the study [8], Bilge et al. proposed a method to detect malicious domain based on 4 main feature groups consisting of Time Based Features, DNS Answer Based Features, Time To Live Based Features, and Domain Name Based Features. To classify malicious domain and normal domain, the authors use two main algorithms: J48 and C4.5. The study [9] proposed a method of detecting malicious domains based on the decentralization of DNS using statistical features. The publication [10] proposed the idea of detecting malicious domains using Graph Inference based on the HTTP proxy log. Segugio [11] focuses on the who is querying what information and constructs a machine-domain bipartite graph based on DNS traffic between clients and the resolver. Khalil et al. [12] build a domain-IP graph based on a passive DNS dataset and then simplify it to a domain graph for detection. Futai Zou et al. [13] try to utilize both the client-query-domain relation and the domain-resolve-IP relation by constructing a DNS query response graph and a passive DNS graph. In the study [14], the authors focused on

detecting malicious domains based on feature groups (consisting of construction-based, IP-based, TTL-based, and WHOIS-based) using Extreme Learning Machine. In the publication [15], in order to detect malicious domain, the authors use the Random Forest algorithm and three main feature groups: domain name lexical features, ranking features, and DNS query features. In addition, studies [1], [2], [3], [4] also list studies of malicious domain detection based on DGA and Fast-Flux techniques.

In this paper, we propose several feature groups including domain name lexical features, ranking features, DNS query features, etc.

3. MALICIOUS DOMAIN DETECTION USING MACHINE LEARNING

3.1 Feature selection and extraction

Some features of abnormal behavior of domain are shown in table 1. All features that marked with an asterisk "*" in table 1 are newly extracted and selected in this research. And most studies are based on feature groups including:

- **Domain name lexical features**: The features are extracted from the domain name. It is the basic information of the domain. We use this information to predict malicious domains.
- **Ranking features:** The features are extracted from the famous websites such as alexa.com's database.
- **DNS query features:** is the server's response information (namely IP address, mail exchange, etc.) when we send DNS query packets.
- **Other features:** The features which register in the whoi

Table 1: The list of Domain feature

No.	Group	Feature	Data type	Description
1	Domain name lexical features (L)	Domain name length	Integer	Length of domain name
2		Domain name token count	Integer	The number of tokens that are separated from the domain name by the character '.'
3		Average domain token length	Real	The average length of tokens
4		Longest domain token length	Integer	The longest length of tokens
5		Number of IP address in domain name	Integer	The number of IP addresses in the domain name
6		Number of special characters	Integer	The number of special characters in the domain name
7		Number of digits	Integer	The number of digits in domain
8		Number of continuous digits	Integer	The number of continuous digits in the domain name
9		Longest continuous digits length	Integer	The longest length of continuous digits
10		Number of continuous letters	Integer	The number of continuous letters in the domain name
11		Longest continuous letters length	Integer	The longest length of continuous letters
12		Maximum Levenshtein ratio	Real	Maximum Levenshtein ratio with popular domain
13		Brand name presence	Binary	Whether or not there exists a brand name in the domain name

14	Ranking features (R)	Rank in Alexa host	Integer	The rank of the domain name in the list of 1 million popular domain names from Alexa host
15		Rank in Alexa country	Integer	The rank of the domain name in the list of 1 million popular domain names from Alexa country
16		Rank in Domcop	Integer	The rank of the domain name in the list of 10 million popular domain names from Domcop
17	DNS query features (D)	Resolved IP count	Integer	The number of IP addresses that are returned in DNS queries
18		Distinct country count	Integer	The number of countries from IP addresses
19		Silent IP ratio	Real	The ratio of the silent IP address
20		HTTP response status	Integer	HTTP response status
21		Name server count	Integer	The number of name servers that are returned in DNS queries
22		Name server IP count	Integer	The number of IP addresses of name servers in DNS queries
23		Name server Country count	Integer	The number of countries where the name server is located
24		Mail exchange server count	Integer	The number of mail exchange servers that are returned in DNS queries
25		NS Count*	Integer	The number of NS record in DNS queries
26		CName count *	Integer	The number of CName record in DNS queries
27		Time to live (TTL)	Integer	Time to live (TTL) of cache record for the domain name at the name server
28	Other Feature	SSL Certificate*	Boolean	Does the domain register an SSL certificate?
29		Age Domain*	Integer	The registration time of the domain until the current time
30		Domain registration*	Boolean	Has the domain been registered?

3.2 Domain classification method

Based on the features presented in Section 3.1, further processing steps are needed to discriminate normal domains and abnormal domains. In this paper, Random Forest classifiers [16] are applied to distinguish between abnormal and normal requests. Random Forest is an ensemble classification method [17]. This algorithm is based on an ensemble of classifiers, which normally are Decision Trees to make the final prediction. The theoretical foundation of this algorithm is based on Jensen's inequality [17]. According to Jensen's inequality applied to the classification problems, it is

shown that the combination of many models may produce less error rate than each individual model

4. EXPERIMENTS AND EVALUATION

4.1 Experimental data

Experimental dataset in this paper consists of 164077 domains that are collected at [18], [19], [20], [21], [22]:

- Dataset of Benign domains: This dataset consists of 79910 benign domains that have been collected from the most well-known domain names on the Internet.
- Dataset of unknown and malicious domains: This dataset covers 84167 phishing domains derived from PhishTank, C&C domains, Malicious domains list.

The data is divided into 2 datasets: 80% of the data is used for training the classification model, 20% of the data is used for testing. The unknown and malicious domains are labeled positive and benign domains are labeled negative.

4.2 Calculation parameter

To evaluate the effectiveness of features and machine learning algorithms that are selected, we use some parameters in Table 2.

Table 2: Calculation parameter

Parameter	Notes	Calculation process
TP	True Positive – result of predicting domain correctly	Count number of domains predicted is domain and it is correct
TN	True Negative – result of predicting domain correctly.	Count number of domains predicted is normal and it is correct
FP	False Positive - result of predicting domain incorrectly.	Count number of domains predicted is phishing and it is incorrect
FN	False Negative – result of predicting domain is normal incorrectly .	Count number of domains predicted is normal and it is correct
FPR	False Positive Rate.	False alert rate.
FNR	False Negative Rate.	Miss rate.
TPR	True Positive Rate.	Accuracy rate of predicting domains which have true label is 'phishing'.
TNR	True Negative Rate.	Accuracy rate of predicting domain which have true label is 'normal'.

Accuracy: the ratio between the number of points correctly predicted and the total number of points in the test dataset.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\ \%$$

Precision: the ratio of the number of true positive points among those classified as *positive* (TP + FP). High Precision value means that the accuracy of the found points is high.

$$precision = \frac{TP}{TP + FP} \times 100\,\%$$

Recall is defined as the ratio of the number of true positive points among those that are actually *positive* (TP + FN). High Recall value means that the TPR is high meaning that the rate of missing the actual positive points is low.

$$\text{Re } call = \frac{TP}{TP + FN} \times 100\,\%$$

$$F1 = \frac{2 \times precision \times \text{Re } call}{precision + \text{Re } call}$$

$$FRP = \frac{FP}{FP + TN} \times 100\ \%$$

$$PPV = \frac{TP}{TP + FP} \times 100\ \%$$

$$TRP = \frac{TP}{TP + FN} \times 100\ \%$$

4.3 Experimental detection of malicious domain

Table 3 below describes the experimental results of detecting malicious domains using the RF machine learning algorithm.

Table 3: Experimental results of detecting malicious domains when changing the number of trees of RF

Num of trees	TPR	TNR	FPR	FNR	Recall	Precision	F1	Accuracy
10	96.87%	93.47%	6.53%	3.13%	96.87%	94.88%	95.87%	95.36%
20	97.09%	96.23%	3.77%	2.91%	97.09%	96.98%	97.04%	96.71%
40	97.23%	90.93%	9.07%	2.77%	97.23%	93.05%	95.09%	94.43%
60	97.24%	91.17%	8.83%	2.76%	97.24%	93.22%	95.19%	94.54%
80	**96.80%**	**97.07%**	**2.93%**	**3.20%**	**96.80%**	**97.63%**	**97.21%**	**96.92%**
100	96.83%	91.88%	8.12%	3.17%	96.83%	93.70%	95.24%	94.63%

From Table 3, we can see that the algorithm has the highest Accuracy and Precision respectively 96.92% and 97.63% when the number of decision trees is 80. Meanwhile, the false detection rate of False Alarm was only 2.93%. Besides, when changing the number of decision trees from 10 to 100, the accuracy of the algorithm doesn't change much. This shows that with a dataset that balanced the ratio of normal and abnormal records, the RF algorithm detects well and steadily. However, when the number of decision trees increases, training and testing time also increases

5. CONCLUSION

The problem of detecting and warning malicious domains is one of the most current urgent issues for the task of preventing phishing attacks. In this paper, with the support of the RF algorithm and the proposed features of abnormal behavior of the domain, we processed, analyzed, and detected successfully malicious domains. The innovation of our approach is looking for and extracting characteristic abnormal behavior of domain on both object and association-based. Experimental results presented in Table 3 show that the RF algorithm brings good and stable results. However, in our research, we still encounter problems related to extracting the features of a malicious domain based on DNS queries which lead to time-consuming processing. In the future, we will improve this problem by pushing domain flows in batches based on parallel processing technologies and big data. However, in fact, the application of machine learning algorithms and analysis of abnormal domain behaviors could help identify suspicious domains, but to prevent attacks through malicious domains, human factors related to morality and information security awareness is still the most important.

REFERENCES

1. Yury Zhauniarovich, Issa Khalil, Ting Yu, Marc Dacier. **A Survey on Malicious Domains Detection through DNS Data Analysis**. *ACM Computing Surveys (CSUR*). Volume 51 Issue 4, Article 67, pp.36 -73, 2018. https://doi.org/10.1145/3191329
2. S. Torabi, A. Boukhtouta, C. Assi and M. Debbabi. **Detecting Internet Abuse by Analyzing Passive DNS Traffic: A Survey of Implemented Systems**, *IEEE Communications Surveys & Tutorials*, vol. 20, no. 4, pp. 3389-3415, 2018. https://doi.org/10.1109/COMST.2018.2849614
3. Manmeet Singh, Maninder Singh, Sanmeet Kaur. **Issues and challenges in DNS based botnet detection**: **A survey**. *Computers & Security*, Volume 86, pp. 28-52, September 2019. https://doi.org/10.1016/j.cose.2019.05.019
4. Hin Dom: **A Robust Malicious Domain Detection System based on Heterogeneous Information Network with Transductive Classification**. *arXiv.org*, arXiv:1909.01590.
5. Imperva, 2019 Cyberthreat Defense Report, 2019. [Online]. Available: https://www.imperva.com/resources/reports/CyberEdge-2019-CDR-Report-v1.1.pdf [Accessed 1 April 2020].
6. Sophos, Sophos 2020 Threat Report, [Online]. Available: https://www.sophos.com/en-us/medialibrary/PDFs/technical-papers/sophoslabs-uncut-2020-threat-report.pdf [Accessed 11 April 2020].
7. Doyen Sahoo, Chenghao Liu, Steven C.H. Hoi. **Malicious URL Detection using Machine Learning: A Survey**. *arXiv*:1701.07179v2. 2017
8. Bilge L, Kirda E, Kruegel C, et al. **Exposure: Finding Malicious Domains Using Passive DNS Analysis**[C]. in *Proc. Network and Distributed System Security Symposium*, San Diego, California, USA, February 2011.
9. Manos Antonakakis, Roberto Perdisci, Wenke Lee, Nikolaos Vasiloglou, and David Dagon. **Detecting malware domains at the upper dns hierarchy**, **USENIX security symposium**, volume 11, pp. 1–16, 2011.
10. Pratyusa K Manadhata, Sandeep Yadav, Prasad Rao, and William Horne. **Detecting malicious domains via graph inference**. in *Proc 19th European*

Symposium on Research Wroclaw, Poland, September 7-11, 2014., pp. 1–18. https://doi.org/10.1007/978-3-319-11203-9_1

11. Babak Rahbarinia, Roberto Perdisci, and Manos Antonakakis. Segugio: **Efficient behavior-based tracking of malware-control domains in large isp networks**. in *Proc. 45th Annual IEEE/IFIP International Conference on Dependable Systems and Networks*, Washington, DC United States, 2015 pp. 403–414. https://doi.org/10.1109/DSN.2015.35
12. Issa Khalil, Ting Yu, and Bei Guan. Discovering malicious domains through passive dns data graph analysis. in *Proc. the 11th ACM on Asia Conference on Computer and Communications Security*, Xi'an China May, 2016, pp. 663– 674.
13. Futai Zou, Siyu Zhang, Weixiong Rao, and Ping Yi. **Detecting malware based on dns graph mining**. *International Journal of Distributed Sensor Networks*, vol. 11, pp.102-1015, 2015.
14. Yong Shi, Gong Chen, Juntao Li. **Malicious Domain Name Detection Based on Extreme Machine Learning**. *Neural Processing Letters*, vol 48, pp. 1347–1357, 2018. https://doi.org/10.1007/s11063-017-9666-7
15. Do Xuan Cho; Ha Hai Nam. **A Method of Monitoring and Detecting APT Attacks Based on Unknown Domains.** *Procedia Computer Science*, vol 150, pp. 316-323. 2019**.**
16. Shai, S.S., Shai B.D.: **Understanding Machine Learning: From Theory to Algorithms**. Cambridge University Press. 2014. 332 pp.
17. LEO BREIMAN. **Random Forests.** *Machine Learning*. vol. 45, Issue 1, pp 5–32. 2001 https://doi.org/10.1023/A:1010933404324
18. OpenDNS public domain lists of domain names for training/testing classifier. https://github.com/opendns/public-domain-lists. [access date 1/4/3018]
19. Malware Domain List. http://www.malwaredomainlist.com/ [access date 1/4/2020].
20. Join the fight against phishing. https://www.phishtank.com/. [access date 1/4/2020]
21. Alexa - Top Sites for Countries. https://www.alexa.com/topsites/countries. [access date 1/4/2020]
22. Public-domain-lists. https://github.com/opendns/public-domain-lists. [access date 3/4/2020].