Malicious Domain Detection based on Decision Tree

Thin Tharaphe Thein[1] Yoshiaki Shiraishi[1] Masakatu Morii[1]

Abstract: Different types of malicious attacks have been increasing simultaneously and become a serious issue for cybersecurity. Most attacks leverage domain URLs as an attack communications medium and compromised users into a victim of phishing or spam. In this paper, we take advantage of the machine learning methods to detect the maliciousness of a domain by using three features: DNS-based, lexical-based, and semantic-based features. The proposed approach exhibits high performance based on a small experiment dataset. The detection results achieved an accuracy of 93%, 90%, and 92% by using Random Forest, AdaBoost, and XGBoost respectively.

Keywords: Malicious domain detection, machine learning, Domain Name System (DNS)

1. Introduction

With the increasing advancement of information technology, the risk and complexity of cybersecurity threats are increasing at an alarming rate, and various malicious cyber-attacks emerge endlessly on a daily basis. Generally, malicious domains are key components for attackers to run malicious activities over the Internet. They can made users into victims of spam, phishing, drive-by-download which may compromise the privacy of users, or install malware or incur a financial loss. It is therefore critical to be able to discover and block these kinds of malicious activities. Though the existing domain and IP blacklists can be used to block malicious domains, these blacklists cannot keep up with the continuous increase in newly registered domains and therefore another effective approach for detecting malicious domains is desirable.

Domain Name System (DNS) data is one of the most remarkable resources to utilize for detecting malicious domains and extensive research on detecting malicious domains using DNS data has been performed. In this paper, using the DNS-based features, lexical features, and semantics features, we build the classification models with Random Forest, XGBoost, and AdaBoost to estimate the maliciousness of a domain.

2. Domain Name System (DNS)

The Domain Name System (DNS) is one of the core protocols of the Internet. Domain Name System (DNS) is a decentralized and hierarchical naming system for resources connected to the Internet, which maps human-readable domain names into their respective IP (Internet Protocol) addresses, and the reverse mapping (IP to domain) is also possible. In short, DNS is a simple service for lookup and name to address resolution of URL into IP address and vice versa [8].

DNS uses a hierarchy, which is an inverted tree structure, to manage its distributed database system. The DNS tree starts with a root domain at the top which is represented by a dot(.). Under the root domain are the Top Level Domains (TLDs) which contains generic Top Level Domains (gTLDs) or country code top level domain (ccTLDs). Figure 1 shows an overview of how DNS resolver works to route a request.

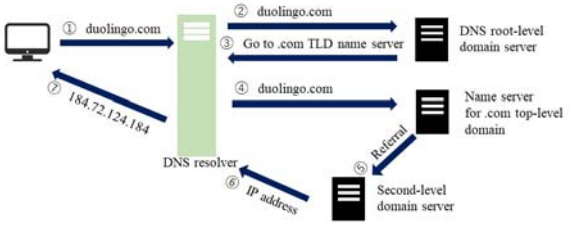

Figure 1 Domain Name System (DNS)

3. Related Work

3.1 Approaches for Malicious Domain Detection

The detection methods for malicious domains in previous studies can mainly be divided into two categories: classification-based approach and graph-based approach.

Classification-based Approach

This approach mainly uses machine learning algorithms with malicious features e.g., domain length, number of characters, etc. and DNS features. Leyla et al proposed a system called EXPOSURE which can detect malicious domain names by using a decision tree algorithm with features extracted from passive DNS analysis [1]. Similarly, in [2], it proposed malware detection based on DNS records and domain name features to identify

1 Kobe University

malicious domains by using decision tree classifiers. Chiba et al proposed a DomainProfiler system that use a random forest classifier to detect newly registered malicious domain names with time-series domain features [3].

Graph-based Approach

Previous studies related to this approach use the association between domains and IP addresses or clients to form domain graphs, and then apply graph-based learning algorithms such as belief propagation, label propagation, and graph convolutional networks. Khalil et al construct a domain-IP bipartite graph from the association between domains and IPs and then used a path-based algorithm to discover potential malicious domains [4]. Kazato et al proposed a graph convolution networks-based domain maliciousness estimation by building a domain relation graph [5]. This approach makes use of domain-IP relationship, domain owner information, and autonomous system number (ASN) to construct the domain graph. In graph-based domain classification approach, the association between domains plays an important role in determining the accuracy of the detection rate.

3.2 DNS Traffic Analysis

Because domain names are usually affiliated with the domain name system (DNS), the DNS data can be used to make an analysis of a domain. Generally, there are two ways to collect DNS traffic data: Active DNS data and Passive DNS data. Active DNS is a real-time system where we query DNS servers and resolvers to map domain names into IP addresses. It mostly captures the DNS records of a given domain, such as address(A) records, Name Server (NS) records, and Mail Exchange (MX) records, etc. Active DNS data does not have privacy problems since it does not include the information of user's query domains. Passive DNS data provides a summarized view of domain queries have richer information than active DNS but the collection of passive DNS is a lot more complex and active Passive. Though there are some services that offered passive DNS data, apparently it is not free. According to this [6], using active DNS data can discover newly created potentially malicious domains. In this research, only active DNS is used due to certain limitations.

4. Proposed Scheme

The overview of the proposed approach is shown in Figure 2. The DNS data and additional information related to domains are firstly collected by data collector. Three groups of features (DNS-based, Lexical-based and Semantics-based) are then extracted from the result of the data collector for a given domain name. The maliciousness estimation is then performed by ensemble classifiers.

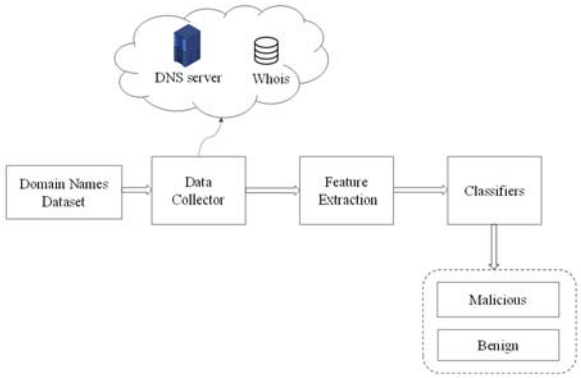

Figure 2 Overview of the proposed scheme

Table 1 Domain Feature Set

DNS-based Features	Number of A records
	Number of NS records
	Number of MX records
	Time-to-live (TTL)
	Active time of domain
	Lifetime of domain
Lexical Features	Number of consecutive characters
	Number of digits
	Length of domain
	Number of words
Semantic Features	Domain reputation score

4.1 Data Collector

The DNS traffic data related to domains are actively query and after the query requests are processed by DNS server, the corresponding response results are returned. Examples of response data include domain's A records, NS records, Time-to-live (TTL), etc. The result dataset is further enriched by whois information which includes domain registration date, expiration date, and updated date.

4.2 Features Extraction

In this step, the collected DNS traffic data are processed in order to extract features that can effectively distinguish malicious and benign domains. From the observation and analysis of the large amount of DNS data obtained from data collector, 11 features as shown in Table 1 are extracted to build the classification model for malicious domain detection. The following section discusses how these features can be used to differentiate between benign and malicious domains.

DNS-based Features

The DNS response records of malicious domains are very different from that of benign domains. There are more A (address) records in malicious domains since malicious domains are usually hosted on many different IP addresses. Also, there are more distinct NS records and fewer MX records in malicious domains compared to benign domain. The TTL values for malicious domains tend to be short-lived compared to benign domain. Also, the lifetime and active time of the benign domain are typically much longer than malicious domain. Lifetime and active time of the domain are calculated as Equation (1) and (2). The domain's lifetime is the interval between the domain's expiration date and registration date. Similarly, the active time of the domain is the interval between the updated date and registration date of a domain. Based on this information, the following characteristics are selected for DNS-based features: number of A records, number of NS records, number of MX records, TTL, active time, and lifetime of a domain.

$$Lifetime = Date_{Expire} - Date_{Create} \quad (1)$$

$$Activetime = Date_{Update} - Date_{Create} \quad (2)$$

Lexical Features

Usually, the names of the benign domain are easily pronounceable and able to be recognized with no trouble while malicious domain names are mostly non-pronounceable by humans. From the observation on a large number of malicious domains, it is found that malicious domains contain more numbers, and the confusing mixture of numbers and words make it hard to pronounce malicious domain. Therefore, the following characteristics are selected for domain's lexical features: length of domain, number of digits, number of words, and number of consecutive characters.

Semantic Features

The previous work by Lui et al [7] introduced detecting malicious domain using semantics features in which domain with the highest accessed rate are chosen as a benign domain. Each domain name is then segmented by N-Gram method to build whitelist domain name substrings and the whitelist substrings is used to calculate the reputation(maliciousness) of a domain. This research follows a similar approach as the previous approach in calculating the reputation value of a domain.

Firstly, as ground truth to build whitelist domain substring, the top 100,000 domain names in Alexa[11] are collected and segmented by N-Gram method by setting the lengths of N to 3,4,5,6,7. A total of 344,503 domain name substrings are extracted from top 100,000 domain names and used as a whitelist domain names substring. Then the reputation score of the testing domain is calculated according to equation (3):

Table 2 Domain Reputation Score

Domain Name	Reputation Score	Label
duolingo.com	44.064	benign
discord.com	62.8	benign
dkdrlah12.0pe.kr	7.347	malicious
dqy.qyuyu.com	0.567	malicious
facebook.com	63.412	benign
douate.com	20.185	malicious

$$Reputation\ Score_{domain} = \sum_{i=1}^{k} \log_2\left(\frac{S_N(k)}{N}\right) \quad (3)$$

where $S_N(k)$ is the total number of occurrences of k-th domain name substrings in whitelist domain name substrings. N is the length of N-Gram (N = 3, 4, 5, 6, 7). Some results of the domains' reputation score are shown in Table 2. It can be observed that the reputation score of the benign domain tends to be larger than the malicious domain because the segmented benign domain substrings occur frequently in whitelist domain name substring.

5. Evaluation

The performance of the proposed scheme is evaluated by three classification methods: Random Forest, XGboost, and Adaboost. The benign and malicious domain names published on the Internet are collected and labeled for ground truth. The dataset is then used as a training and testing data for evaluating the effectiveness in classifying domain as benign or malicious.

5.1 Dataset

The dataset contains a total of 1457 domain names of which 680 domains are malicious and 777 domains are benign. The benign domain is collected from Alexa Top Sites [11], which is a ranking system based on website's popularity. Therefore, the top-ranked websites can be regarded as benign domains. Malicious domain names are gathered from publicly published domain backlist services. Resolvable malicious domains are randomly selected from malwaredomainlist.com [9] and compromised domain list [10]. These domains are likely to be compromised for malware, command and control communication and phishing.

5.2 Evaluation Results

The domain classification is made by three methods: Random Forest, AdaBoost, and XGboost. The comparison of the

Table 3 Experiment Results

Classifier	Accuracy	Precision	Recall
Random Forest	0.927	0.906	0.899
AdaBoost	0.902	0.893	0.891
XGBoost	0.919	0.901	0.900

experimental results is shown in Table 3. All three classifiers achieved about 90% accuracy in detecting malicious domains even though the domain name dataset is quite small. All classifiers are trained and evaluated on 10-fold cross-validation. Random Forest performed the best out of three methods in accuracy and precision while XGBoost got the highest recall.

6. Conclusions

In this paper, we proposed an approach to classify the domain as malicious or benign by using active DNS traffic data and whois information. Moreover, we incorporate a semantic feature in addition to normally used lexical and DNS-based features in an attempt to improve the detection of the malicious domain. The experimental results show that the proposed approach achieved accuracy as high as 93% on Random Forest Classifier using a small domain dataset.

The current classification only recognized domains as either malicious or benign. We can further categorize the maliciousness of a domain as spam, phishing, command and control, or malware, making it a multi-class classification problem. Moreover, using a combination of Passive DNS and active DNS data might play a role in the enhancement ability to detect bad domains.

Acknowledgements This research was conducted under a contract of "Research and development on IoT malware removal / make it non-functional technologies for effective use of the radio spectrum" among "Research and Development for Expansion of Radio Wave Resources(JPJ000254)", which was supported by the Ministry of Internal Affairs and Communications, Japan.

References

[1] Bilge L, Kirda E, Kruegel C, Balduzzi M. “EXPOSURE: Finding Malicious Domains Using Passive DNS Analysis”. In NDSS 2011 Feb 6 (pp. 1-17).

[2] Messabi KA, Aldwairi M, Yousif AA, Thoban A, Belqasmi F. “Malware detection using dns records and domain name features.” In Proceedings of the 2nd International Conference on Future Networks and Distributed Systems 2018 Jun 26 (pp. 1-7).

[3] Chiba D, Yagi T, Akiyama M, Shibahara T, Yada T, Mori T, Goto S. “DomainProfiler: Discovering domain names abused in future.” In 2016 46th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN) 2016 Jun 28 (pp. 491-502). IEEE.

[4] Khalil I, Yu T, Guan B. “Discovering malicious domains through passive DNS data graph analysis.” In Proceedings of the 11th ACM on Asia Conference on Computer and Communications Security 2016 May 30 (pp. 663-674).

[5] Kazato Y, Nakagawa Y, Nakatani Y. “Improving Maliciousness Estimation of Indicator of Compromise Using Graph Convolutional Networks.” In 2020 IEEE 17th Annual Consumer Communications & Networking Conference (CCNC) 2020 Jan 10 (pp. 1-7). IEEE.

[6] Zhauniarovich Y, Khalil I, Yu T, Dacier M. “A survey on malicious domains detection through DNS data analysis.” ACM Computing Surveys (CSUR). 2018 Jul 6;51(4):1-36.

[7] Liu Z, Zeng Y, Zhang P, Xue J, Zhang J, Liu J. “An imbalanced malicious domains detection method based on passive dns traffic analysis.” Security and Communication Networks. 2018 Jan 1;2018.

[8] “Wikipedia: Domain Name System”. https://en.wikipedia.org/wiki/Domain_Name_System

[9] “Malware Domain List”. https://www.malwaredomainlist.com/

[10] “Compromised Domain List”. https://zonefiles.io/compromised-domain-list/

[11] “Alexa Top Sites”. https://www.alexa.com/topsites