BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2023
http://excel.fit.vutbr.cz

Classification of domain names generated by DGA

Filip Bučko

Abstract
The detection and classification of domain names generated by Domain Generation Algorithms (DGA) is critical in cybersecurity. DGA is a technique malware uses to evade detection and establish command and control channels with the attacker. These domain names are typically generated algorithmically based on various parameters, such as the current date, time, and network information. DGAs make detecting and blocking malicious domains more difficult, as they can generate many seemingly random domain names, making it challenging to identify malicious communication. DGA detection is an essential technique for identifying and mitigating malware threats. This paper presents our feature-based machine learning approach using XGBoost to detect and classify DGA-generated domain names. Our model for detection achieves high accuracy (98.7%) and outperforms many feature-based machines learning existing solutions in the literature

*xbucko05@vut.cz, *Faculty of Information Technology, Brno University of Technology*

1. Introduction

DGAs provide a way for malware to hide communication with the command-and-control (C&C) servers. Moreover, thus the malware evades detection. Therefore, detecting and classifying DGAs is an important step in protecting computer systems and networks from the malware threats.

DGA techniques constantly evolve and change over time, which makes it difficult for traditional detection methods to keep up. In addition, there are many different types of features, and selecting the most effective ones for DGA classification is challenging [1]. In our approach, we implement two classification modules. The basis for implementing these modules is using XGBoost [2] machine learning models.

2. Related works

The machine learning feature-based approaches (e.g., Decision trees) achieve high accuracy, but the performance of these classifiers largely depends on the used feature set [3, 4]. On the other hand, the featureless approach (deep learning) does not require this, as they learn to extract the relevant features automatically during training. The main drawback is the complexity of interpreting the decisions. Different types of deep learning classifiers (e.g., NN3, CNN4) have been proposed for binary DGA domain classification [5, 6, 7] as well as for multi-class DGA classification [8, 9, 7]. Various comparisons in papers [8, 10, 5] show that deep learning classifiers outperform classical approaches. However, the feature-based classifiers compared in these works are often developed with little or no effort devoted to data preprocessing and attribute extraction.

Contributions Our implementation with XGBoost brings many advantages regarding flexibility, speed, and scalability on large datasets and handling imbalanced data. XGBoost also provides a mechanism for ranking feature importance, which helps identify the most critical features for the classification task.

3. Solution

We present the DGA detector, composed of Binary and Multi-class classifiers for detection and classification purposes. The binary classifier is used to filter out benign domains from DGA domains. Domains identified as DGA are further sent to a multi-class classifier. The current implementation of the multi-class classifier consists of Regular Expression Classification (74 regular expressions) and XGBoost classifier (93 families). The model architecture is shown in 1.

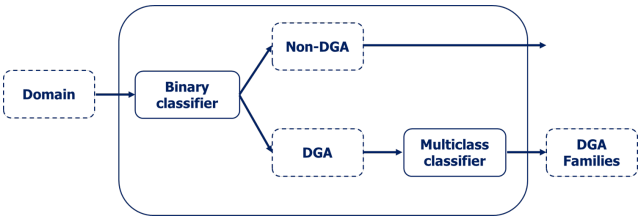

Figure 1. DGA detector architecture

3.1 Data preprocessing

Data preprocessing plays a vital role in the overall performance of the model. In the context of DGA classification using XGBoost, we preprocessed the data as follows: 1) Data Collection; 2) Feature Extraction; 3) Data Cleaning; 4) Splitting the Data into training and testing sets, and 5) Balancing the Data combining proportion pick from DGA families and SMOTE (distribution of the classes in the dataset is significantly uneven 5:22 000 000). As part of the data collection, we managed to collect 230 000 verified benign domains. We got a dataset of 123 million DGA domains of 93 DGA families provided by Fraunhofer-Institut für Kommunikation, Informationsverarbeitung und Ergonomie FKIE.

3.2 Feature engineering

Combining findings from our experimental results and observations from existing research in the area, we eventually decided to use 18 machine-learning features. Those cover the domain structure, character frequencies, n-gram occurrence, and matches with selected dictionary words. The top 5 features that gave the best score for binary and multi-class classification are mentioned in Section 4.

3.3 Classification pipeline

1. The input to the DGA classifier is an unclassified domain
2. Features are extracted from the domain
3. A vector of extracted features are sent to the Binary Classifier
4. In case of positive DGA prediction, we send the domain and vector to the Multi-class Classifier.
5. Value vector is sent to the input of XGBoost model for multi-class classification
6. The output of the XGBoost model is a list of DGA families ordered from the highest classification probability of each DGA family. If a given family is in the regex database, then the domain enters the module for regex classification. If the regex classification does not match the most probable family from the XGBoost model, then the next family in the sequence is selected from the list of DGA families. The multi-class classification process repeats.
7. The output of the classification is the DGA family

4. Results

Our XGBoost Binary classifier outperforms many existing solutions in the literature, achieving an accuracy of 98% in detecting DGA-generated domains. With Multi-class classification, we managed to achieve an accuracy of 88%. Other tested metrics are displayed in tables 1 and 2.

Evaluation Metrics	Percentage
Accuracy	98.1%
Precision	98.4%
F1 Score	98.1%
Recall	97.8%

Table 1. Metrics of Binary classifier

Evaluation Metrics	Percentage
Accuracy	87.8%
Precision	87.3%
F1 Score	87.8%
Recall	86.2%

Table 2. Metrics of Multi-class classifier

In the Binary classifier, the Non-DGA N-gram ratio, Number of subdomains, and the Total domain length appeared to be the most profitable features for classification. Within the Multi-class classification, the most important features are the SLD length, Total domain length, and Hexadecimal ratio

5. Conclusions

This paper presented our feature-based machine learning approach using XGBoost to classify DGA-generated domain names. Our model achieved high accuracy in detecting DGA-generated domains and showed robustness against new DGA families not seen during training. Our approach can serve as a valuable tool in detecting and mitigating DGA-based threats. The subject of future work will focus on the featureless approach using a Transformer deep learning model.

Acknowledgements

I would like to thank my supervisor Radek Hranický and the whole research group for their willingness, help, and friendly approach during the solution of this work. The research is supported by the "Flow-based Encrypted Traffic Analysis" project, no. VJ02010024 granted by the Ministry of the Interior of the Czech Republic and "Smart information technology for a resilient society" project, no. FIT-S-23-8209 granted by the Brno University of Technology.

References

[1] STTELARCYBER. What are dgas and how to detect them? [online], 2018. vid. [2022-11-15].

[2] NVIDIA. XGBoost, 2023. [Online; Accessed: 2023-3-18].

[3] Yoav Freund and Robert E. Schapire. Experiments with a new boosting algorithm. In *Proceedings of the Thirteenth International Conference on International Conference on Machine Learning*, ICML'96, page 148–156, San Francisco, CA, USA, 1996. Morgan Kaufmann Publishers Inc.

[4] Leyla Bilge, Sevil Sen, Davide Balzarotti, Engin Kirda, and Christopher Kruegel. Exposure: A Passive DNS Analysis Service to Detect and Report Malicious Domains. *ACM Transactions on Information and System Security*, 16(4), apr 2014.

[5] Bin Yu, Jie Pan, Jiaming Hu, Anderson Nascimento, and Martine De Cock. Character Level based Detection of DGA Domain Names. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2018.

[6] Joshua Saxe and Konstantin Berlin. expose: A character-level convolutional neural network with embeddings for detecting malicious urls, file paths and registry keys. *arXiv preprint arXiv:1702.08568*, 2017.

[7] Jonathan Woodbridge, Hyrum S Anderson, Anjum Ahuja, and Daniel Grant. Predicting domain generation algorithms with long short-term memory networks. *arXiv preprint arXiv:1611.00791*, 2016.

[8] Raaghavi Sivaguru, Chhaya Choudhary, Bin Yu, Vadym Tymchenko, Anderson Nascimento, and Martine De Cock. An Evaluation of DGA Classifiers. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 5058–5067, 2018.

[9] Duc Tran, Hieu Mac, Van Tong, Hai Anh Tran, and Linh Giang Nguyen. A lstm based framework for handling multiclass imbalance in dga botnet detection. *Neurocomputing*, 275:2401–2413, 2018.

[10] Jan Spooren, Davy Preuveneers, Lieven Desmet, Peter Janssen, and Wouter Joosen. Detection of Algorithmically Generated Domain Names Used by Botnets: A Dual Arms Race. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, SAC '19, page 1916–1923, New York, NY, USA, 2019. Association for Computing Machinery.