63rd FTTCE International Congress
26 - 28 September 2024, Kraków, Poland

Enhancing DGA Detection with Machine Learning Algorithms

Hubert Biros and Mirosław Kantor

Institute of Telecommunications, AGH University of Krakow, Kraków, Poland

Abstract

The domain generation algorithm (DGA) is a popular technique malware uses to reliably establish a connection to a command and control (C&C) server. Pseudo-random domain names generated by DGAs are used to bypass security measures and allow attackers to maintain control over malware-infected devices.

State-of-the-art solutions for detecting DGA domain names are based on machine learning methods to create DGA classifiers. Machine learning-based classifiers can be developed using two main approaches. The first is the featureful approach, where a set of features is extracted from the domain name or other auxiliary information and used to construct a detection model [1]. The second is the featureless approach, which leverages deep learning techniques, such as neural networks, convolutional networks, or LSTM networks, to create classifiers directly from the domain name without feature extraction [2]. Some works propose a hybrid approach, combining both feature-based and featureless methods [3].

Many studies focus on developing universal classifiers capable of detecting both character-based and word-based DGA domains. However, as highlighted in the findings of [4], some such models only work well in detecting specific types of DGA domains. In our study, we propose a two-pronged approach to detect character-based and word-based DGA domains. Seven machine learning algorithms, namely support vector machine (SVM), extremely randomized trees (ET), logistic regression (LR), gaussian naive Bayes (GNB), nearest centroid (NC), random forests (RF), and k-nearest neighbors (KNN), were employed to detect character-based DGA domains using a featureful approach. We utilized a set of 25 features, 4 of which were newly proposed in our work, enhancing detection model performance. To classify wordbased DGAs, we applied the featureless approach using convolutional neural networks (CNN) and long short-term memory (LSTM) models. These models were designed to embed words that constitute domain names, providing a unique perspective for this type of DGA detection.

We used a dataset containing both training and testing subsets to develop DGA classifiers. Each subset included an equal ratio (1:1) of benign and DGA domains. For non-DGA domains, 450,000 samples were collected, with 400,000 used for training and 50,000 for evaluation. These benign domains were sourced from the top one million domains ranked by Majestic. The same benign dataset was employed for both word-based and character-based DGA classifiers. As for the DGA domains, samples from 56 malware families were used to create a collection of 450,000 domains for the training and evaluation of character based DGA detectors. For word-based DGA classifiers, the DGA dataset contained the same number of domains, but they were generated by 8 malware families. These collections were split similarly to the

benign domain set for training and evaluation purposes. The DGA domains were obtained by executing reverse-engineered DGA code snippets available online or using predefined domain lists.

During the development of character-based DGA classifiers, a set of 35 domain-based features was proposed, including 8 novel features that had not been used in previous studies. Feature selection techniques were employed to reduce the number of features, eliminating redundant features that conveyed the same type of information. Ultimately, 10 features were removed. For word-based DGA classifiers, the process began by preparing domain names—removing TLDs and breaking the names into word components.

After training the classifiers, we evaluated their performance using a dedicated test dataset of 100,000 domain names. To measure the performance of the models, we used seven key metrics: ACC (overall accuracy), PPV (positive predictive value) or precision, TPR (true positive rate) or recall, FPR (false positive rate), FNR (false negative rate), F1 score and AUC (area under the ROC curve). The detailed performance results for each model in detecting character-based and word-based DGAs are summarized in the Table 1:

Table 1: Performance results for models in detecting character-based and word-based DGAs

Classifier	PPV	TPR	FPR	FNR	F1	ACC	AUC
	Character-based DGA Classifiers						
ET	96.84%	95.48%	3.12%	4.52%	96.16%	96.18%	0.9937
SVM	94.39%	93.08%	5.53%	6.92%	93.73%	93.77%	0.9846
LR	94.28%	93.18%	5.66%	6.82%	93.72%	93.76%	0.9847
GNB	83.28%	91.27%	18.33%	8.73%	87.09%	86.47%	0.9278
NC	85.36%	86.50%	14.83%	13.50%	85.93%	85.83%	0.9380
RF	97.60%	96.43%	2.37%	3.57%	97.01%	97.03%	0.9954
KNN	96.33%	94.98%	3.62%	5.02%	95.65%	96.00%	0.9901
	Word-based DGA Classifiers						
LSTM	94.34%	96.50%	5.78%	3.50%	95.41%	95.36%	0.9905
CNN	97.84%	98.78%	2.19%	1.22%	98.31%	98.30%	0.9975

Bibliography

1. Hoang X.D. & Vu, X.H. A Novel Machine Learning-based Approach for Detecting Word-based DGA Botnets. 2021, Journal of Theoretical and Applied Information Technology, Vol. 99, No. 24.
2. Highnam K., Puzio D., Luo S., & Jennings N. R. Real-Time Detection of Dictionary DGA Network Traffic Using Deep Learning. 2021, SN Computer Science, 2(2). (DOI:10.1007/s42979-021-00507-w).
3. Sivaguru R., Peck J., Olumofin F., Nascimento A. & Cock M. D. Inline Detection of DGA Domains Using Side Information. 2020, IEEE Access (Volume: 8); pp 141910-141922. (DOI:10.1109/access.2020.3013494).

4. Xuan Hanh Vu, Xuan Dau Hoang, Thi Hong Hai Chu, A Novel Model Based on Ensemble Learning for Detecting DGA Botnets, 2022.