Research Project Proposal 7245

DGA Detection with Machine Learning and Deep Learning

Group Members

(Alphabetically ordered)

Haimin Zhang
Lixi Zhou
Shiqi Dai

Topic Description

Domain generation algorithms(DGA) are used in various families of malware, which generate a large plenty of domain names that can be used as rendezvous points with their command and control (C2) servers. Security vendors usually used blacklists to identify malware, but DGA can constantly update domain to evade the blacklist detection. In order to solve this problem, instead of using low-efficient traditional methods, we will use machine learning algorithms to detect DGAs and compare the performance of these algorithms.

Background and Related Work

Internet security vendors have provided several strategies to intercept DGA traffic. In traditional, security providers would first decode the algorithm by applying reverse engineering. Generating a list of domains with a given seed, then preregister, sink-holed or put them into a DNS blacklist to prevent potential C2 traffic. Another common strategy is to find similar domain groups by using their statistical properties to determine if DGA generates a domain. The main disadvantage of traditional strategies is the lack of capability to be used for real-time detection and protection.

Therefore, several strategies based on machine learning are introduced. FANCI is one of them. FANCI stands for Feature-based Automated NXDomain Classification and Intelligence, and it was introduced in 2018 by Schüppen, Teubert, Herrmann, and Meyer. It is a system for detecting infections with domain generation algorithm based malware by monitoring non-existent domain responses. FANCI mainly uses two supervised learning algorithms, random forests and support vector machine. Because of the using of RF and SVM, all the domain data(text) has to be represented by features. Schüppen et al.,(2018) described 21 features and used three different categories to group their features. They are structural features, linguistic features, and statistical features. Structural features have to subcategories as inherent structural features and non-self-explanatory structural features. Non-self-explanatory structural features can be some boolean type features or calculated ratio features. Linguistic features are used to measure the deviations from common linguistic patterns of domain names. Statistical features are n-gram frequency distribution and entropy which are common approaches in the feature engineering of domain data. According to Schüppen et al.,(2018), FANCI is based on supervised learning classifiers. It requires training with labeled data. Thus the first module is a training module. The output of the training module is a trained model, then the next module--classification module will use the model to classify new input data. Before classifying, classification module will also perform some preprocessing like feature extraction. In the end, the intelligence module will supply intelligence based on classification results, in particular, find infected devices and identify new DGAs or unknown seeds. FANCI is a very flexible system. There are two main usage scenarios, all-module using and distributed using.

However, the above traditional machine learning approaches have to use manually picked features to create classifiers like FANCI. They usually have two significant drawbacks: First, hand-crafted features are easy to circumvent by hackers. Second, getting hand-crafted features is relatively time-consuming at the runtime. Thus deep learning/neural network approaches have been taken seriously nowadays. Specific neural networks require less feature engineering and perform better at the run-time. They work directly on raw domain names with a minimal transformation. In other words, if a new family of DGA shows up, the classifier can be retrained right away without the need for manual feature engineering. Also, neural network models act like the black box so it is hard for hackers to reverse and beat. Second, deep learning models have better "true positive"/"false positive" rate and real-time performance. In test cases, neural network classifiers are usually able to achieve satisfying accuracy.

Data Sources

Benign Domains:

- Alexa Top 1 Million Sites: The Alexa Top Sites web service provides access to lists of websites ordered by Alexa Traffic Rank. (https://www.kaggle.com/cheedcheed/top1m)

Malicious DGA Domains:

- Bambenek Consulting provided malicious algorithmically-generated domains (License)(http://osint.bambenekconsulting.com/feeds/dga-feed.txt)
- 360 Lab DGA Domains**:** A collection of domains generated by DGA and it is maintained by 360--a Chinese security vendor. This dataset keeps updated every day. (https://data.netlab.360.com/feeds/dga/dga.txt)

Algorithms and Code Used

AutoML: H2O

H2O has an industry leading AutoML functionality that automatically runs through all the algorithms and their hyperparameters to produce a leaderboard of the best models.

Random Forest(RF)

Random forest is a bunch of decision trees. It can be seen as an ensemble model. A random forest model will take all predicting results from its inner decision trees as a vote.

SVM(Support Vector Machine)

SVM is able to classify input data by using the computed hyperplane which is trained from the training set. In other words, given labeled training data, the algorithm outputs an optimal hyperplane which categorizes new examples.

CNN(Convolutional Neural Network)

Typically, CNN is used on image/audio data. It plays a vital role in cognitive computing like image/voice recognition. But there are some approaches on text classification that use CNN as the classifier.

LSTM(Long Short-Term Memory Neural Network)

Long short-term memory (LSTM) units are units of a recurrent neural network (RNN). An RNN composed of LSTM units is often called an LSTM network.

References

[1] Antonakakis, M., Perdisci, R., Nadji, Y., Vasiloglou, N., Abu-Nimeh, S., Lee, W., & Dagon, D. (2012, August). From Throw-Away Traffic to Bots: Detecting the Rise of DGA-Based Malware. In USENIX security symposium (Vol. 12).

[2] AWS | Alexa Top Sites-Up-to-date lists of the top sites on the web. (n.d.). Retrieved from https://aws.amazon.com/alexa-top-sites/

[3] Domain generation algorithm. (n.d.). Retrieved from https://www.wikiwand.com/en/Domain_generation_algorithm

[4] G. (2018, November 05). Google-research/bert. Retrieved from https://github.com/google-research/bert

[5] H2O – Data Resource Portal. (n.d.). Retrieved from https://www.northeastern.edu/dataresources/h2o

[6] H. A., & J. W. (2018, February 22). Using Deep Learning to Detect DGAs. Retrieved from https://www.endgame.com/blog/technical-blog/using-deep-learning-detect-dgas

[7] Koehrsen, W. (2018, June 02). Automated Feature Engineering in Python – Towards Data Science. Retrieved from https://towardsdatascience.com/automated-feature-engineering-in-python-99baf11cc219

[8] Plohmann, D., Yakdan, K., Klatt, M., Bader, J., & Gerhards-Padilla, E. (2016, August). A Comprehensive Measurement Study of Domain Generating Malware. In USENIX Security Symposium (pp. 263-278).

[9] Schiavoni, S., Maggi, F., Cavallaro, L., & Zanero, S. (2014, July). Phoenix: DGA-based botnet tracking and intelligence. In International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment (pp. 192-211). Springer, Cham.

[10] Schüppen, S., Teubert, D., Herrmann, P., Meyer, U., & Sch, S. (2018, August). FANCI: feature-based automated NXDomain classification and intelligence. In Proceedings of the 27th USENIX Conference on Security Symposium (pp. 1165-1181). USENIX Association.

[11] Tran, D., Mac, H., Tong, V., Tran, H. A., & Nguyen, L. G. (2018). A LSTM based framework for handling multiclass imbalance in DGA botnet detection. Neurocomputing, 275, 2401-2413.

[12] Woodbridge, J., Anderson, H., Ahuja, A., & Grant, D. (2016). Predicting Domain Generation Algorithms with Long Short-Term Memory Networks.

[13] Yu, B., Pan, J., Hu, J., Nascimento, A., & De Cock, M. (2018). Character Level Based Detection of DGA Domain Names.