Brett Luskin

DATA 606 - Deliverable 2

Malware Detection with Machine Learning

Literature Review

There are two papers available that I researched on the subject of feature selection in the UNSW-NB15 data set. The first, by the original creators of the data set, Moustafa and Slay, identified eight features of high importance to for detection of network intrusion. The second identified five of high importance. One feature overlapped in both papers: source time to live (STTL). “Time-to-live (TTL) is a value in an Internet Protocol (IP) packet that tells a network router whether or not the packet has been in the network too long and should be discarded” [6]. After doing an initial EDA on the entire dataset, I used this research as a guide to explore some of the features more deeply, particularly STTL.

There are five documented methods that I researched in papers that give accuracy and False Alarm Rates with their Network Intrusion Detection System given in table 1 below:

Algorithm	**Accuracy**	**False Alarm Rate**
Logistic Regression	83.15%	18.48%
Naive Bayes	81.20%	18.30%
Artificial Neural Network	81.50%	22.10%
EM Clustering	78.40%	23.70%
Feed Forward Neural Net	99.50%	0.47%

Table 1

Exploratory Data Analysis

My analysis aligned with the research papers on STTL as the most important feature. When STTL is greater than 50 seconds, the ratio of network traffic that are attacks is 89.73%. When less than 50 seconds, the percentage drops to 0.71%. This means that some of the Machine Learning methods used in earlier research papers should be revisited because they seem like they are achieving very poor results in comparison with random guessing based on this one feature. I will be reimplementing some of these methods for the sake of reproducibility and optimization.

References

1. T. Janarthanan and S. Zargari, "Feature selection in UNSW-NB15 and KDDCUP'99 datasets," 2017 IEEE 26th International Symposium on Industrial Electronics (ISIE), Edinburgh, 2017, pp. 1881-1886.
2. D. Jing and H. Chen, "SVM Based Network Intrusion Detection for the UNSW-NB15 Dataset," 2019 IEEE 13th International Conference on ASIC (ASICON), Chongqing, China, 2019, pp. 1-4.
3. N. Moustafa and J. Slay, "The Significant Features of the UNSW-NB15 and the KDD99 Data Sets for Network Intrusion Detection Systems," 2015 4th International Workshop on Building Analysis Datasets and Gathering Experience Returns for Security (BADGERS), Kyoto, 2015, pp. 25-31.
4. N. Moustafa, J. Slay and G. Creech, "Novel Geometric Area Analysis Technique for Anomaly Detection Using Trapezoidal Area Estimation on Large-Scale Networks," in IEEE Transactions on Big Data, vol. 5, no. 4, pp. 481-494, 1 Dec. 2019.
5. L. Zhiqiang, G. Mohi-Ud-Din, L. Bing, L. Jianchao, Z. Ye and L. Zhijun, "Modeling Network Intrusion Detection System Using Feed-Forward Neural Network Using UNSW-NB15 Dataset," 2019 IEEE 7th International Conference on Smart Energy Grid Engineering (SEGE), Oshawa, ON, Canada, 2019, pp. 299-303.
6. https://searchnetworking.techtarget.com/definition/time-to-live