GR-23 Machine Learning Techniques for Malware Network Traffic Detection

Abstract

The objective of this research is to investigate and compare multiple machine learning classification techniques with an optimized data set for application in malware network traffic detection. The MTA-KDD'19 data set for malware network traffic analysis contains 65,500 instances and is used for constructing the classification models herein. Cross-Validation and single hold out data portioning techniques are applied during model evaluation to validate accuracy results. The following machine learning classification models are applied for this investigation: Multilayer Perceptron, Decision Tree, Support Vector Machine, and K-Nearest Neighbors.

Introduction

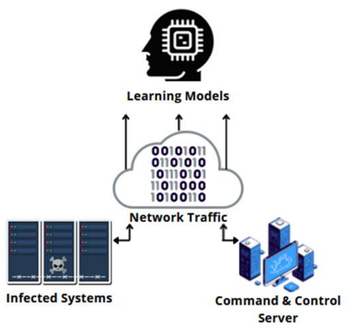
Traditional malware network traffic detection techniques focus on analyzing raw network traffic packets and have not deterred the surge of persistent obfuscated malware. A very effective attribute of some of the most pervasive malware is injection within legitimate network-related system processes. Such tactics continue to evade traditional malware detection systems. Therefore, it is important to develop new research techniques that are focused on analyzing optimized metadata from network traffic to effectively identify an ever-increasing amount of malware. The principal tool utilized for this research project is the Waikato Environment for Knowledge Analysis (WEKA). The characteristics of the MTA-KDD'19 dataset and the functionality within WEKA make the scope and scale of this research feasible.

Research Question

New information in the area of malware network traffic analysis is pursued through this research. The objective research question is: Can machine learning techniques produce highly accurate classification models for malware network traffic detection based on a statistically optimized data set?

Methodology

The MTA-KDD'19 data set is composed 33 numerical features. The features are grouped into categories that include: TCP Flag Type, IP Protocol Type, Packet Stats, Packet IO Ratios, First Packet Length, Repeated Packet Ratios, Connection Volumes, and HTTP Requests. Two data portioning techniques are used to split the data into training and testing sets. Cross Validation is used with a parameter set to 10-Fold. Hold out is used to split the data into 80% training and 20% testing sets. The following graphic displays an overview of the framework.



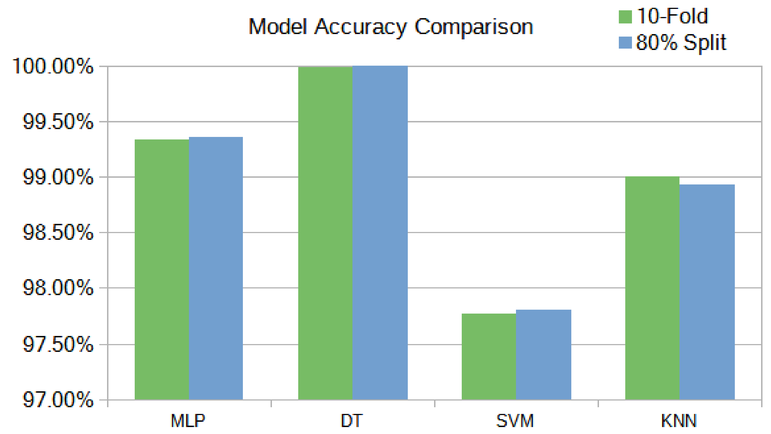
Results

Decision Tree: The WEKA J48 implementation of C4.5 (Quinlan 1993) decision tree was utilized. This classifier produced the best result with 100% accuracy when set to run an 80% train 20% test split and 99.99% accuracy when set to run 10-fold cross-validation. This result has incrementally improved upon the results observed in the experiment presented by the authors of the MTA-KDD'19 dataset.

Multilayer Perceptron: The WEKA hidden layer definition of a = (attributes + classes)/2 was utilized. This classifier produced the second-best result with 99.36% percent accuracy when set to run an 80% train 20% test split and 99.34% accuracy when set to run 10-fold cross-validation.

Support Vector Machine: The WEKA sequential minimal optimization implementation of SVM was utilized. This classifier produced 97.81% accuracy when set to run an 80% train 20% test split and 97.78% accuracy when set to run 10-fold cross-validation.

K-Nearest Neighbor: The K value was set to 1 as an evaluation of other K values (i.e. 3,5,7,9) did not provide an accuracy improvement. This classifier produced 98.93% accuracy when set to run an 80% train 20% test split and 99.00% accuracy when set to run 10-fold cross-validation.



MODEL PERFORMANCE METRICS

Model	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
DT 10-Fold	100%	0%	100%	100%	100%	100%
DT 80/20	100%	0%	100%	100%	100%	100%
MLP 10-Fold	99.3%	0.7%	99.3%	99.3%	99.3%	99.9%
MLP 80/20	99.4%	0.6%	99.4%	99.4%	99.4%	99.9%
SVM 10-Fold	97.8%	2.4%	97.8%	97.8%	97.8%	97.7%
SVM 80/20	97.8%	2.4%	97.8%	97.8%	97.8%	97.7%
KNN 10-Fold	99.0%	1.0%	99.0%	99.0%	99.0%	99.1%
KNN 80/20	98.9%	1.1%	98.9%	98.9%	98.9%	99.0%

Conclusion

In this research study, multiple machine learning techniques were investigated and applied to an optimized data set for malware network traffic detection. The results of this research provide affirmative evidence that classification models based on a statistically optimized data set can assist in detecting the presence of malware activity in network traffic. This area of research has direct application to cybersecurity industry solutions such as network traffic analysis software and cyber threat intelligence services. This research can help deliver important capabilities including anomaly detection, threat investigation, and network traffic surveillance.

Acknowledgments

Learning opportunities with the following organizations assisted with and motivated the completion of this research:

- Kennesaw State University Cybersecurity Institute
- Offensive Security Research Club at Kennesaw State University
- Evidence Based Cybersecurity Research Group at Georgia State University

The following open-source tools were utilized:

WEKA GitHub

Contact Information

Author:
Jermaine Cameron jcamero6@students.kennesaw.edu

References

[1] Letteri, I., Della Penna, G., Vita, L., Grifa, M. (2020): MTA-KDD'19: A Dataset for Malware Traffic Detection. In: Loreti, M., Spalazzi, L. (eds.) Proceedings of the Fourth Italian Conference on Cyber Security, Ancona, Italy, February 4th to 7th, 2020. CEUR Workshop Proceedings, vol. 2597, pp. 153–165. CEUR-WS.org (2020).

[2] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I., (2009). The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.

[3] Letteri, I., Di Cecco, A., & Della Penna, G. (2020). Dataset Optimization Strategies for Malware Traffic Detection. arXiv preprint arXiv:2009.11347.

[4] Witten, I., Frank, E., Hall, M. (2011). Data Mining: Practical Machine Learning Tools and Techniques: Vol. 3rd ed. Ian H. Witten, Frank Eibe, Mark A. Hall. Morgan Kaufmann.

[5] Mirsky, Y., Doitshman, T., Elovici, Y., & Shabtai, A. (2018). Kitsune: an ensemble of autoencoders for online network intrusion detection. arXiv preprint arXiv:1802.09089.

[6] Meidan, Y., Bohadana, M., Mathov, Y., Mirsky, Y., Shabtai, A., Breitenbacher, D., & Elovici, Y. (2018). N-baiot—network-based detection of iot botnet attacks using deep autoencoders. IEEE Pervasive Computing, 17(3), 12-22.

KENNESAW STATE UNIVERSITY
COLLEGE OF COMPUTING AND SOFTWARE ENGINEERING

Author: Jermaine Cameron
Advisor: Dr. Seyedamin Pouriyeh