System Call Based Malware Detection Using Hybrid ML Methods Trained on the AWSCTD Dataset

Nana Kwame Gyamfi, Nikolaj Goranin, Dainius Čeponis and Antanas Čenys
Department of Information Systems, Faculty of Fundamental Sciences, Vilnius Tech

Introduction

- Automatic malware detection methods can be classified into two main classes: signature-based and anomaly-based. The signature-based approach is considered to be reliable and having low false-positive rate for known malware types, but is not able to detect new and zero-day attack. It also has other drawbacks, like signature database increase, as well. Anomaly-based approach is considered as a perspective for detecting new and zero-day malware. Typically, they are statistical analysis or machine learning based and requires training on a classified data.
- The limitations of signature-based attack detection on the operating system level (inability to detect minimal attacks and composable malware, constantly growing signature base, and detection speed issues) have prompted research into anomaly-based solutions, which could reduce the number of false negatives.
- Earlier research was utilizing non-hybrid ML methods on the AWSCTD or other datasets.

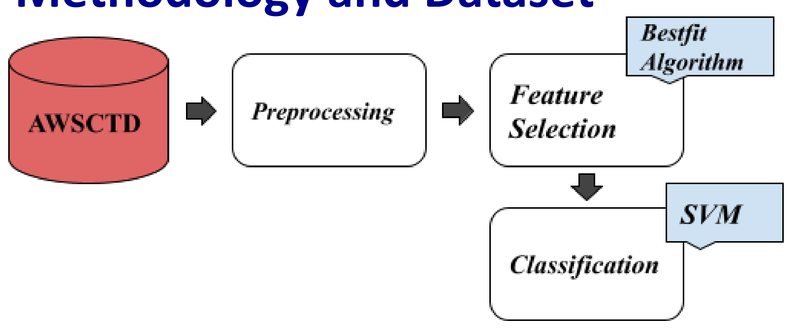
Experimental Setup and Metrics

Experiments were performes on: Google Collaboratory.

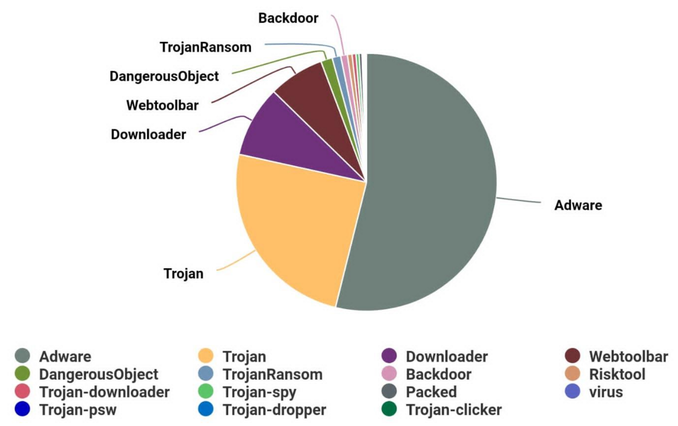
Motivation

- A system call is an API used by the computer program to request a service from the kernel of the operating system. Because of that origin (a primary artifact of the OS kernel), system call data is a very popular choice for malware research and detection.
- System call sequences are another popular method to represent features, which is costly but generates strong detection metrics. Since tasks of different application differ, that information can be used to classify them correctly. Hidden Markov Models (HMM), Recurrent Neural Networks (RNN) and Convolutional Neural Networks (RNN) are the top ML methods used for this task type.
- Majority of research on anomaly-based malware detection is currently concentrated on the use of system call sequences generated by the analyzed suspicious programs as an input data. Although formally the achieved results seem promising and the obtained accuracy rate is higher than 99%, still it is necessary to mention that it is possible to get such accuracy only in case when a long sequence of system calls (>600) is provided for analysis. Such data input prompts, that in a real environment malware would be able to perform the all-planned actions before being detected.

Methodology and Dataset

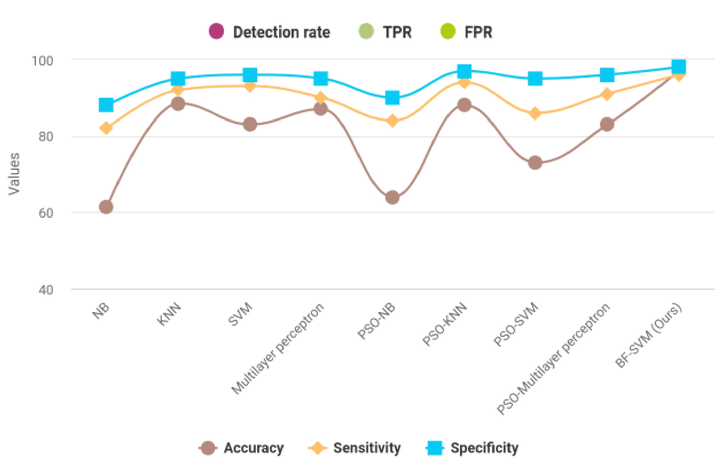


1. Data Collection (AWSCTD dataset was used)
2. Preprocessing for data cleansing, filtering, and transformation
3. Feature selection for dimensionality reduction
4. Classification stage to produce the classified results.



Results

Usage status (on the dataset chosen)	Model	Accuracy	Sensitivity	Specificity	Detection rate	TPR	FPR	Precision	Recall	F1-score
Previously used	Naïve Bayes	61.55	82.6	88.30	86	87	13	72	65	78
Previously used	SVM	83.47	93	96.20	85	87	13	77	72	83
Previously used	Multilayer Perceptron	87.63	90.2	95	88	90	10	83	75	81
First time used	KNN	88.55	92.45	95.80	89	91	9	75	68	80
First time used	Pso-naïve Bayes	64.96	84.1	90	84	86	14	84	78	85
First time used	Pso-knn	88.18	94.10	97	88	90	10	82	77	84
First time used	Pso-SVM	78.67	86.2	95	90	92	8	88	80	86
First time used	Pso-multilayer perceptron	83.26	91.3	96	93	93	7	90	80	89
First time used	BF-SVM	97.35	96.52	98.44	95	96	4	95	82	94



Conclusions

- The experimental evaluation of several classical (KNN, Naïve Bayes, SVM) and hybrid (Pso-naïve Bayes; Pso-KNN; Pso-SVM; Pso-multilayer perceptron; BestFirst-SVM) ML methods trained on our previously generated AWSCTD dataset was performed.
- The best accuracy (97.35%) results were achieved with the BestFirst-SVM method, which outperformed earlier used ML methods on the AWSCTD dataset.
- Still, even hybrid ML methods lack behind the earlier tested deep-learning methods, as AWSCTD-CNN-S, by accuracy, although winning the speed competition.
- Later optimization can be concentrated on: utilization of adjacent (and metadata) data to minimize the length of system call sequence needed for reliable attack detection, thus minimizing the reaction time; optimization of data structures used; optimization of parameters of currently utilized artificial intelligence methods; search and/or development of the new artificial intelligence methods/architectures, mostly suitable for the anomaly detection task.

References

[1] Ceponis, D., & Goranin, N. (2018). Towards a Robust Method of Dataset Generation of Malicious Activity on a Windows-Based Operating System for Anomaly-based HIDS Training.
[2] Dogo, E. M., Nwulu, N. I., Twala, B., & Aigbavboa, C. (2019). A survey of machine learning methods applied to anomaly detection on drinking-water quality data. In Urban Water Journal (Vol. 16, Issue 3, pp. 235–248).
[3] W. J. D. (2015). Classifying System Call Traces using Anomalous Detection By.
[4] Kemmerer, R. A., & Vigna, G. (2002). Intrusion detection: A brief history and overview. Computer, 35(SUPPL.), 27–30.
[5] Liao, H.-J., Lin, C.-H. R., Lin, Y.-C., & Tung, K.-Y. (2013). Intrusion detection system: A comprehensive review. Journal of Network and Computer Applications, 36(1), 16–24.

DAMSS DATA ANALYSIS METHODS FOR SOFTWARE SYSTEMS
13th Conference
Druskininkai, Lithuania, Hotel „Europa Royale“
December 1 – 3, 2022
http://www.mii.vu.lt/DAMSS
VILNIUS TECH
Fundamentinių mokslų fakultetas