Classifying computer processes in the DARPA OpTC dataset

Literature Review Andrew Veal

General problem/task definition

The detection of malware and malicious activity in enterprise networks is an ongoing challenge in cybersecurity. Our objective is to build a system that will classify activity associated with a computer process as benign or malicious, using host-based logging data from the computer. We will use supervised learning and 'ground truth' labelled data to build a classification model. We shall use data from the DARPA Operationally Transparent Cyber (OpTC) dataset **[1]**, **[3]** which has event logs associated with processes, files, registries and network connections on computers in an enterprise-scale network. The dataset represents real-world networks and also contains malicious events associated with a "red team" who introduced malware and orchestrated malicious activity over three days on a small number of computers.

We have selected four papers to review that apply supervised learning to problems of classifying computer processes **[2]**, machine activity **[4]**, software system calls **[5]** and computer network activity **[6]** as benign or malicious. The papers use different datasets and methods but all use machine learning to distinguish benign and malicious activity in computer event logs.

Reference **[6]** addresses the sub-task of pre-processing techniques that may be applied to datasets with class imbalance between normal and malicious examples.

Concise Summary of papers

Anjum et al (2021) **[1]** describe the DARPA OpTC dataset. They detail the objects in the dataset, give some statistics for malicious and benign events and propose several research directions where this dataset could be used. We will refer to this paper but will not review it.

Cochrane et al (2021) **[2]** develop a new classification algorithm 'SK-Tree' for detecting malware in computer process event logs. They model event data using a streaming tree data structure that captures the hierarchy of parent and child processes and their sequences of events. 'SK-Tree' uses kernel methods to create a binary classifier for streaming tree data that predicts whether computer processes are malicious or benign. Results of applying 'SK-Tree' to a small sample of the DARPA OpTC dataset are given, with an area under ROC score of 98%.

Rhode et al (2018) **[4]** use machine activity metrics recorded during the execution of malicious and benign software to build binary classification models. They created a dataset by executing benign and malicious software in a sandbox environment and logging 10 machine activity metrics. Although the focus of the paper is on developing a recurrent neural network (RNN) model, they report the accuracy of classifying software as malicious or benign for 9 machine learning algorithms on their dataset.

Walker et al (2019) **[5]** use the frequency counts of Windows API system calls made by malicious and benign software to build binary classification models. They created datasets by executing benign and malicious software in a sandbox environment and logging 264 different Windows API system function calls. Results for the accuracy of classifying software as malicious or benign using Windows API call frequencies are given for 6 machine learning algorithms on datasets of varying size.

Wheelus et al (2018) **[6]** evaluate pre-processing techniques to mitigate the class imbalance between normal and malicious traffic present in cyber security datasets. They present results for the effect of over sampling, under sampling, balanced class weights and bagging on the performance of 5 classification algorithms on the computer network traffic dataset UNSW-NB15.

Compare and contrast

The four papers we have reviewed all apply supervised learning to build binary classifiers to classify activity in computer event logs as benign or malicious.

The authors use different types of data and look at different aspects of the malware detection problem:

- Cochrane et al (2021) **[2]** classify computer processes, using 20 types of events associated with process activity from the DARPA OpTC dataset.
- Rhode et al (2018) **[4]** classify machine activity associated with software execution, using 10 input features measured in a sandbox environment.
- Walker et al (2019) **[5]** classify frequency counts of software system calls, using 264 types of Windows API calls measured in a sandbox environment.
- Wheelus et al (2018) **[6]** classify computer network traffic, using 11 attributes associated with sessions recorded in the UNSW-NB15 dataset.

The scale of the datasets and the class imbalance between benign and malicious examples in the datasets varied significantly across the papers:

- Cochrane et al (2021) **[2]** selected event logs for the host computer with the greatest proportion of malicious activity on a single day; after modelling and filtering, the final dataset contained 4199 streaming trees. The class imbalance is not recorded, but the vast majority of the activity is benign **[1]**.
- Rhode et al (2018) **[4]** created a dataset with 2345 benign and 2286 malicious examples of software.
- Walker et al (2019) **[5]** created a dataset with 7400 malicious and 30 benign examples of software. They randomly down-sampled the malicious examples to create a ‘medium’ sample with 853 malicious and 30 benign examples, and a ‘small’ sample with 30 malicious and 30 benign examples.
- Wheelus et al (2018) **[6]** created a dataset with 94.44% benign and 5.56% malicious examples from the UNSW-NB15 dataset which has over 2.5 million data points.

Most of the papers used several different machine learning algorithms:

- Cochrane et al (2021) **[2]** report results for their ‘SK-Tree’ algorithm.
- Rhode et al (2018) **[4]** report results for 9 algorithms: Random Forest (RF), Multi-Layer Perceptron (MLP), K Nearest Neighbours (KNN), Support Vector Machine (SVM), Decision Tree (DT), AdaBoost, Naïve Bayes (NB), Gradient Boosted Trees (GBT) and Recurrent Neural Networks (RNN).
- Walker et al (2019) **[5]** report results for 8 algorithms: Logistic Regression (LR), Linear Discriminant Analysis (LDA), K Nearest Neighbour (KNN), Classification and Regression Trees (CART), Naïve Bayes (NB), Support Vector Machine (SVM), Decision Tree (DT) and Random Forest (RF).
- Wheelus et al (2018) **[6]** report results for 5 algorithms: Logistic Regression (LR), Decision Tree (DT), Multi-Layer Perceptron (MLP), Random Forest (RF) and a feed-forward Neural Network (NN).

We shall now compare the results presented in each paper, noting the training-validation-test splits and the metrics used to evaluate success:

- Cochrane et al (2021) **[2]** use 5-fold cross validation to determine performance and hyper-parameters for the ‘SK-Tree’ binary classifier. They show classifier results for each fold in a Receiver Operator Characteristic (ROC) curve. The Area Under ROC (AUROC) metric for the classifier is 0.98.
- Rhode et al (2018) **[4]** split out a test set of 500 (206 benign, 316 malicious) examples and use 10-fold cross validation on the training set of 4131 examples. They report the accuracy, false positive rate and false negative rate for training-validation and test sets for all 9 classifiers. The RNN classifier had the highest accuracy of 96% on the test set and an accuracy of 88% on the training-validation set. The Gradient Boosted Tree classifier had the highest accuracy of 96% on the training-validation set and an accuracy of 93% on the test set.
- Walker et al (2019) **[5]** used a 70/30 split for training and validation sets and used 10-fold cross validation to determine the accuracy of classifiers. Results were given for ‘small’, ‘medium’ and ‘large’ samples from the full dataset. As noted above, smaller samples under-sampled the majority class of malicious examples to give a more balanced dataset. The ‘medium’ sample appeared to have the best trade-off between number of training examples and class imbalance. Most of the classifiers in the ‘medium’ sample had an accuracy of 96% but had more false positives than true negatives (hence poor performance on the minority benign examples).
- Wheelus et al (2018) **[6]** used cross validation and the Area Under ROC (AUROC) metric to measure classifier performance. They present results for the effect of over sampling, under sampling, using balanced class weights and bagging on classifier performance for a dataset with class imbalance. Over sampling was achieved by synthesizing instances of the minority class and under sampling by randomly removing examples of the majority class until the majority class had a ratio (in this case) of 2:1 of the minority class. With no pre-processing, the baseline performance of the 5 classification algorithms was in the range 0.691 – 0.806 AUROC. Each of the pre-processing techniques provided a lift in performance above the baseline level for most

classification algorithms but the improvement was dependent on the algorithm and in some cases the lift was negative.

Future Work

It is an open question how well the methods of Rhode et al (2018) **[4]** and Walker et al (2019) **[5]** developed on datasets from sandbox environments will transfer to enterprise networks where the vast majority of events are benign. We would like to take the approaches of creating input features from machine activity metrics and using aggregated frequency counts of events and apply them to computer process events in the DARPA OpTC dataset.

Cochrane et al (2021) **[2]** applied a complex algorithm to a collection of events associated with computer process activity in the DARPA OpTC dataset. Their early results using event logs from a single host computer indicate strong classifier performance – it would be interesting to see how the 'SK-Tree' classifier performed on data from a larger number of computers from the dataset.

We would like to understand whether simpler methods can be used to profile malicious computer process activity. Anjum et al (2021) **[1]** show that the aggregate distributions of benign and malicious event counts are different in the DARPA OpTC dataset. Can we distinguish between malicious and benign processes using a frequency count of the (object, action) events associated with each process as a feature vector? If so, can we identify which event types are more important for classification of malicious processes?

Coming up with an effective data selection and partition strategy for experiments on the DARPA OpTC dataset will be important, as there are only 29 computers with malicious activity in the dataset and the ratio of malicious to benign events is very small. None of the references discussed the experimental protocol around forming training-validation sets in detail and only Rhode et al (2018) **[4]** used a test set.

We shall need to consider pre-processing techniques for data to mitigate the extreme class imbalance between benign and malicious samples. Wheelus et al (2018) **[6]** gave no clear recommendations – we anticipate using under-sampling and balanced class weighting methods but we shall look for further references.

None of the references used recall and precision as success metrics for the classification of malicious examples. In a setting where malicious examples are rare, it is important to account for the impact of false positives: *recall* measures the coverage of malicious examples in the predicted class and *precision* measures how accurate the malicious predictions were.

References

[1] Md. Monowar Anjum, Shahrear Iqbal and Benoit Hamelin (2021) *Analysing the Usefulness of the DARPA OpTC Dataset in Cyber Threat Detection Research* **arXiv:2103.03080v2**

[2] Thomas Cochrane, Peter Foster, Varun Chhabra, Maud Lemercier, Terry Lyons and Cristopher Salvi (2021) *SK-Tree: a systematic malware detection algorithm on streaming trees via the signature kernel* **arXiv:2102.07904v3**

[3] DARPA (2020) Operationally Transparent Cyber data release – http://github.com/FiveDirections/OpTC-data

[4] Matilda Rhode, Pete Burnap and Kevin Jones (2018) *Early-stage malware prediction using recurrent neural networks* **Computers & Security 77**

[5] Aaron Walker and Shamik Sengupta (2019) *Insights into Malware Detection via Behavioral Frequency Analysis using Machine Learning* **MILCOM 2019 IEEE Military Communications Conference**

[6] Charles Wheelus, Elias Bou-Harb and Xingquan Zhu (2018) *Tackling Class Imbalance in Cyber Security Datasets* **2018 IEEE International Conference on Information Reuse and Integration for Data Science**

Extra Credit Slack Post – screen shot

nexus.veal 8:31 PM

Here is the literature review for my project on "Classifying computer processes in the DARPA OpTC dataset"

PDF

LiteratureReviewClassifyingComputerProcesses Upload.pdf

98 kB PDF

Classifying computer processes in the DARPA OpTC dataset

Literature Review Andrew Veal

General problem/task definition

The detection of malware and malicious activity in enterprise networks is an ongoing challenge in cybersecurity. Our objective is to build a system that will classify activity associated with a computer process as benign or malicious, using host-based logging data from the computer. We will use supervised learning and 'ground truth' labelled data to build a classification model. We shall use data from the DARPA Operationally Transparent Cyber (OpTC) dataset **[1]**, **[3]** which has event logs associated with processes, files, registries and network connections on computers in an enterprise-scale network. The dataset represents real-world networks and also contains malicious events associated with a "red team" who introduced malware and orchestrated malicious activity over three days on a small number of computers.