Classifying computer processes in the DARPA OpTC dataset

Andrew Veal

What is the real-world problem?

The detection of malware and malicious activity in enterprise networks is an ongoing challenge in cybersecurity.

What should your system do?

The system should classify activity associated with a computer process as benign or malicious, using host-based logging data from the computer.

The system will use supervised learning and “ground truth” labelled data to build a classification model.

What are its inputs and outputs?

The inputs are event logs that record events associated with processes, files, registries and network connections on computers in a network. Feature vectors for the activity associated with a computer process will be created by joining and aggregating event data from multiple tables.

The outputs are predicted class labels for each computer process (benign or malicious).

What datasets are you going to use?

We shall use the DARPA Operationally Transparent Cyber (OpTC) dataset **[1]**.

The dataset comprises over 17 billion events from an isolated enterprise network of 1000 host computers. Sensors logged 11 object types, including file, netflow, process and registry events, over 14 days. On 3 days a “red team” introduced malware and orchestrated malicious activity on 29 computers. A “red team” ground truth document identifies o(100) unique malicious parent processes.

How will you measure the success of your system?

We shall use the Precision, Recall and Accuracy of the classification of computer processes as our measure of success.

We are seeking to build up tools, techniques and a baseline from simple techniques.

What are the challenges of building the system?

There are a very small number of malicious examples in the dataset, so it is marginal whether there are sufficient examples for supervised learning.

There is extreme class imbalance – almost all the activity in the dataset is benign.

The volume and variety of information in the dataset - spanning file access, network connections, dynamic library loads – presents a challenge for feature engineering.

What is the phenomena in the data that you are trying to capture?

We are trying to capture an aggregated summary of activity associated with a computer process that differentiates malicious behaviour from benign behaviour.

Which topics might address these challenges?

To address the class imbalance, oversampling (including multiple copies) of the positive (malicious) examples may be used to balance the training and test datasets.

We shall need a strategy to maximize the use of the training and test data through cross validation, taking care to ensure there are positive examples in all folds.

References

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