Data Protection, Protection by Data – Project

Year : 2022-2023
Lecturer : Prof. Pierre Parrend

Organisation

The planning is as follows:

Quickoff:	5/10
Intermediate presentation:	25/10
Defence:	30/11

Organize your work with best practices:

- Each scenario is taken by half of the student groups
- Work in groups of 3 or 4
- Share the work (and don't forget to mention task allocation in your presentations)

Project objectives

The objective of this project is to develop a python module for intrusion detection, for 1 of the two following scenarios:

1) Detect malware at runtime
2) Detect Advanced Persistent Threats

Your software must support application in Edge Computing environment, i.e. in constraint resource environments.

To demonstrate this point, you will:

- Evaluate required resources consumption (time, CPU, RAM) for learning and for detection.
  If you have suitable material available home, you can evaluate several execution environments: Cloud environment (Colab, Kaggle or other), PC, Raspberry, Arduino (yes plan some fun for the installation!).

- Compare following algorithms (given as `model_name`): KNN, CART, Random Forrest, XGBoost, SVM, MLP
  - Which of these algorithms support multi-class classification?

  - For algorithms supporting binary classification, extract true positive, true negative, false negative, false negative, for the capability of detecting an attack.

  - For algorithms supporting multi-class classification, extract true positive, true negative, false negative, false negative, for each attack class.

- Evaluate the detection performance for each attack type, according to the size of learning dataset (use number of rows in power of ten (100, 10^3, 10^4, 10^5, 10^6, 10^7 is the dataset supports it)).
  Compare the metrics for balanced data and metrics (precision, recall=TPR, TNR, accuracy) for unbalanced data (F1-score, unbalanced accuracy, Matthews Correlation Coefficient) for each algorithm, and each attack class.

- Compare the detection on the reference dataset to two similar datasets.

To setup your benchmark environment, install a Jupiter Notebook on your own machine. Perform all performance evaluation on a single environment.

Scenario 1: Detect malware at runtime

The reference dataset is:

- https://www.unb.ca/cic/datasets/malmem-2022.html

The datasets for performing complementary evaluation of your approach are:

- https://research.unsw.edu.au/projects/adfa-ids-datasets
  use ADFA-LD
- https://www.cs.unm.edu/~immsec/systemcalls.htm

The detection library should entail at least following methods:

```
def init_model(model_name, train_and_test_dataset):
       return model

def is_malware(model, malware_features):
       return (Boolean, probability)
```

Scenario 2: Detect Advanced Persistent Threats

The reference dataset is:

- pwnjutsu-dataset: https://pwnjutsu.irisa.fr/dataset/

The datasets for performing complementary evaluation of your approach are:

- https://ipsr.ynu.ac.jp/aptgen/index.html
- mscad-intrusion-detection:

https://drive.google.com/file/d/1-Z1_o7XWq7GVVE50UwCAPjFgsGAtaQ6e/view?usp=sharing

The detection library should entail at least following methods:

```
def init_model(model_name, train_and_test_dataset):
       return model

def is_apt(model, malware_features):
       return (Boolean, probability)
```

Intermediate presentation

For the intermediate presentation:

- Perform the analysis of the reference dataset
- With at least two classification algorithms
- And all evaluation metrics
- For 10^5 entries in the dataset

You have a 10 minute slot for presentation + demonstration of your code

Final presentation

For the intermediate presentation:

- Evaluate the resource consumption
- For all classification algorithms
- Evaluate the scalability of the models according to the size of the learning dataset
- Compare the evaluation with the 2 complementary datasets

Project deliverables

The deliverables of the project are:

- The detection library as python module
- A Jupiter notebook, including manual for loading suitable datasets and launching the execution of the complete evaluation
- Project report (20 pages)
- Presentation (10 minutes)
- Demonstration (5 minutes)