Data Protection, Protection by Data – Project

Year : 2023-2024
Lecturer : Côme Frappé - - Vialatoux

Organisation

The planning is as follows:

- Quickoff: 12/10
- Advancement point: 25/10
- Defence: 22/11

Organize your work with best practices:

- Work in groups of 3 or 4 (total of 7-8 groups for the whole promo)
- Share the work (and don't forget to mention task allocation in your report)

Project objectives

The objective of this project is to apply the data analysis chain on a cyber-physical dataset:

1) Using only network data
2) Using only physical data
3) Using both at the same time

Your code must run on an average laptop (16Go ram – no gpu) but optimisation (Ex: GPU support, memory optimisations, etc.) are welcome.

Evaluation required:

- Compare following algorithms: KNN, CART, Random Forrest, XGBoost, SVM, MLP (and additionnal ones if you want)

- 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, etc.)).

- Compare the metrics for balanced data (precision, recall, TPR, TNR, accuracy), metrics for unbalanced data (F1-score, balanced accuracy, Matthews Correlation Coefficient) and confusion matrices for each algorithm, and each attack class.

- Evaluate resources consumption for learning and for detection. (Fit time, prediction time, RAM)

- Compare the performance of your models to the ones published in the paper associated.

- You are free to use oversampling and/or undersampling in order to increase your models performances

- You are free and encouraged to test any idea you have of novel detection methods. Even if the results are bad, do not hesitate to include them ! (make sure to have the mandatory models working in priority)

The dataset :

The dataset and associated paper can be accessed there:

- Link

/!\ Network datas are heavy !

Project deliverables

The deliverables of the project are:

- A streamlit webapp providing an interactive interface to explore your results (models performances, data visualisations, exploratory data analysis results worth showing, etc.)

- For treatments outside the webapp, the associated notebook containing said treatments (Ex: model training, data prep etc.)

- Project report (10-20 pages) explaining :
    - The results of your exploratory data analysis and their consequences on how you handled the data
    - How you combined network and physical data (and the benefits associated, if any)
    - for each model
        - Your data preparation steps (can be identical for multiple models)
        - How you trained your models (parameter tuning, improvements made, model architectures, computational resources, etc.)
        - Their performances
    - Your novel detection methods (if any)

- Conclusion + personal sections (1 per member) on what were your contribution and your takeaways on this project.

Advancement point:

At this date you must have :

- Performed the exploratory data analysis of the network and physical datasets (seperately at least)
- Data preparation + at least two classification algorithms for at least one dataset
- All evaluation metrics asssociated
- First version of streamlit webapp to present/explore the results

  Each one of this points will give 0,5 points on your final grade for the project for a total of 2 points.

Final presentation

For the final presentation:

- For all datasets
  - Data exploration highlight + dataprep
  - Compare algorithms performances
  - Evaluate the algorithms resource consumption
- Evaluate the scalability of the models according to the size of the learning dataset
- Conclude on the benefits of coupling the physical and network datasets

  And as always, nice and clear visualisations to support what you are saying !

It is recommended to use your streamlit webapp directly as a support for your presentation, but you are not bound to it and can still use slides. If your presentation is on slides, then you will need to reserve part of your presentation time to do a demonstration of the streamlit webapp.

Presentation time : 10 minutes + ~3 minutes questions

Evaluation criteria

The final grade is evaluated as follow

Presentation	Dynamics of oral presentation	2
Results	EDA of datasets	2
	Algorithms application	3
	Performance Analysis	3
	Handling of datasets	2
Webapp	Completeness	2
	User Experience	2
	Pertinence	2
Advancement points		2
Total	Evaluation	20

The global mark is conditioned to having sent the deliverables.