Université Rennes 1

M2 CSE cybersecurity

Project 2

Cluster analysis

December 2, 2020

1 Introduction

Historically, malware protection has heavily relied on manual reverse engineer for analyzing and creating signature creation against suspicious programs. However, with the advent of techniques such as polymorphism/metamorphism, malware started to grow exponentially in number and quickly evolve in complexity, making the manual approach infeasible. In this scenario, it is imperative to rely on techniques that can autonomously analyze and help to better respond to cyberthreats. Towards this goal, in this project we are going to practice the cluster analysis for *malware classification*, i.e., assigning the correct malware type and family/strain to malicious samples.

2 Instructions

The project includes features (e.g. *strings*, *file type*, etc.) extracted from different binaries. The goal is to to compute clustering based on these features by resorting to *unsupervised learning algorithms*.

The dataset is serialized in *pickle* format (see here how to read this file) and saved in the file *dataset.pkl* at the root of the project's repository (access here for the invitation).

In this project it is important to follow the workflow for cluster analysis, i.e.:

1. Data gathering
2. Data processing
3. Feature engineering
4. Algorithm selection
5. Cluster analysis

Recall that defining the relevant *features* to be used in a cluster analysis is key factor for its success. Therefore, you can complete the original dataset with external data (i.e. data gathering), modify/clean up data (i.e. data processing) and encode data (i.e. feature engineering) as you deem necessary.

Comparing different algorithms, specially when their working principles differ, is strongly suggested. In the report, explain the overall concepts of the chosen algorithms and provide details on how to compare them. As long as necessary, repeat the workflow so as to refine your results. Finally, select the best clustering results that are able to unveil the malware families present in the dataset.

2.1 Evaluation

The project will be documented in a *final report* and presented in two occasions: a *mid presentation* and a *final presentation*.

The criteria that will be observed in the project evaluation and the due dates are presented below.

Activity	due date	eval. weight
Report	14/12	40%
Mid presentation	09/12	20%
Final presentation	14/12	40%

Report

The final report should contain a clear description of the work produced during the project, including details about the whole analysis process. This includes the analysis methodology, intermediary results (even those that do not participate in the final results), any generated code, etc. **You may want to use Jupyter notebook for you report**.

The final report must be in *PDF* format or as **Jupyter notebook** (i.e. *.ipynb*) and should be pushed in the Github repository assigned for this project. This repository can include additional files, furthermore you are free to organize and use it as you deem most appropriate for your project repository, however in this case you should create a readme file for describing the project organization (and where to find the final report).

Use this link for pulling the project repository.

Mid presentation

The *mid presentation* is intended for exposing the state of the ongoing project and engendering valuable discussions for its continuation. Therefore, your participation in the discussions and exchanges about your colleagues' projects will be considered as important as your project exposition in the evaluation.

The presentation will be take place remotely in a video-conference and slides as well as live demo or recorded video can be used. It should be planed to take 30 minutes, where 15 minutes will be dedicated for the project presentation and 15 for questions & answers.

Final presentation

The *final presentation* is intended for exposing the the work produced in the project, detailing its most important aspects and findings. As in the *mid presentation* your participation in the discussions and exchanges is desired and will be taken into account in the final evaluation.

The presentation will be take place remotely and it is expected to contain descriptive slides - supplementary live demo/recorded video is a plus. It should be planed to take 30 minutes, where 25 minutes will be dedicated for the project presentation and 5 for questions & answers.