PROJECT PROPOSAL FORM

This document serves as an example. You can either fill out and submit, or use it as a guideline for writing your own document.

The purpose of the proposal is to demonstrate that you have put time into planning. Your final project can be different if you change your mind. The guidelines here are flexible, not hard constraints.

Student 1: Maryam Al Shami

Student 2: Chae Dickie-Clark

Student 3: François LaBerge

Student 4:

Date submitted: November 5 2022

Propose a title for your project. If your project were written up as a research paper, what title would you give it? A good paper title will help each individual reader to know whether they should or should be interested in reading the paper. For example, the title *Intriguing properties of neural networks* (Szegedy *et al*. 2014) is a title that, although a little too vague, at least suggests that the nature of the work is an investigation, and that the focus was neural networks, and that the results are surprising. As another example, *The fastest pedestrian detector in the West* (Dollar *et al*. 2010) is a fun title indicating that the goal is "pedestrian detection" and that the nature of the contribution is "speed."

Cyberattack detection for SCADA control systems: Comparative assessment of machine learning models

Describe the goal of your project. What are you trying to achieve? What "main question" are you trying to answer, or at least to provide evidence for? Secondary goals are OK, but you should still have a clear "main goal" or "main question." From your description, it should also be clear whether your project is about: making better predictions for some application? speeding up training and/or predictions? simply comparing predictive performance and/or speed of several methods? assessing or comparing interpretability? understanding failure modes or sensitivities of some methods? Etc.

SCADA is network architecture widely used for industrial control systems. These systems are increasingly being connected to the internet to allow remote monitoring and reduce operational costs. As one might expect, this exposes them to cyber attacks. For example, an attack against a power grid system built on SCADA can be devastating. As such, the need for attack detection of such systems is increasing. Our objective is to determine if classical machine learning methods can be used to detect such attacks. We narrow down the scope of our project by focusing specifically on gas pipeline systems. Our project will compare the performance of several methods. First, we will apply anomaly detection algorithms with k-means, GMM and PCA. Second, if time permits, we will further try classification algorithms such as SVM, random forest and ensemble methods. Third, we will employ a novel feature engineering method using PCA, independent component analysis (IDA), linear component analysis (LDA), and bloom filters.

Describe the data you plan to use. One of the hardest steps for a good machine learning project is to find data that is truly suitable for your goals. Finding good data not the most fun part, but it's one of the most important—after all, for machine learning it is "garbage in, garbage out". Here are some things you should ideally know:

- What are the 'modalities' that apply to the data? (images, video, speech, text, tabular, categorical, numerical, time series, experimental measurements, etc.)
- What does an input look like? (show an example if possible, like an image, or a sound wave, or some features, or at least try to describe)
- For an example input, what does the desired output look like?
- How many training and testing samples will there be? Can some be realistically trained on a laptop, or is something more powerful needed?
- Anything special about how the training and testing data should be split?
- Might the data need preprocessing before you can use it?

We will use an openly available dataset from the Mississippi State University SCADA Laboratory available at https://sites.google.com/a/uah.edu/tommy-morris-uah/ics-data-sets. The dataset is listed as "Raw Data Gas Pipeline". The dataset contains 26 features extracted from a MODBUS frame plus the frame's label indicating if the frame was bening traffic or originated from an attack. MODBUS is a ubiquitous communication protocol used in SCADA systems. The protocol can be employed in various mediums, including the internet. The features are defined as follows,

Feature	description	example
command_address	undefined	4
response_address	undefined	4
command_memory	undefined	183

response_memory	undefined	233
command_memory_count	undefined	9
response_memory_count	undefined	18
comm_read_function	undefined	3
comm_write_fun	undefined	10
resp_read_fun	undefined	3
resp_write_fun	undefined	10
sub_function	undefined	0
command_length	undefined	41
resp_length	undefined	19
gain	PID gain	115
reset	PID reset rate	0.2
deadband	PID dead band	0.5
cycle time	PID cycle time	1
rate	PID rate	0
setpoint	the target pressure	20
control_mode	control mode of the system, automatic, manual or off	2
control_scheme	scheme to control the setpoint	1
pump	pump control state	0
solenoid	valve control state	0
crc_rate	undefined	0
measurement	pressure measurement	0.528735637664795
time	time stamp	1.10686765499064

The example values are taken from the first instance in the dataset. Even though the dataset has multiple labels for the different attacks, we will only use those labels as binary classes. 0 will represent a bening packet and anything else will represent an attack. Therefore, the output of our models will be binary. An important thing to note about the dataset is that class labels are unbalanced, there are far more benign packets than attack. For the anomaly detection algorithms, the labels won't be used for training, but will be used for performance evaluation. The dataset contains roughly 97000 instances and will be split with a 80/20 scheme for training and testing. While splitting the dataset, we will have to make sure to keep the same distribution of benign/attack instances on both sets to maintain class proportionality. The dataset should be small enough for training on our machines, however if it reveals to be too large, we will select a class-proportional subset of it. Lastly, some preprocessing on the dataset will be necessary, for example, some data points might be missing some features and will be removed. Furthermore, as previously stated, we will use PCA, ICA, CCA, and Bloom filtering to preprocess our features before training on the models.

Describe how you will measure "success." You should explain how you will know whether you have achieved the goal(s) that you described earlier. What does "success" look like? What does "failure" look like? Keep in mind that your project can still succeed (in the sense of a good grade!) even if the experimental results are bad—what is important is that your experimental results are *conclusive*! A bad project is one in which you cannot even tell whether the goal was achieved or not.

When dealing with an unbalanced amount of class labels, as is our case, classification accuracy alone might be deceptive. Hence, we are going to use a confusion matrix to summarize the performance of our models. We will also measure performance using an f-score. Although the specific β value is undetermined as of now, we expected to pick a value larger than 1 to minimize false negatives. As a preliminary metric goal, we will aim for an f-score of at least 0.8.

Describe how work will be divided. It is very important for everyone to have a meaningful role in the project. If one person (the most experienced person) does all the programming or writing, then everyone else in the group loses this important chance to gain experience. For example, if there is no way to "happily divide" the work because two group members want to work on the same part, that is totally OK and no one should feel guilty for wanting that; both group members can do their own version of that part of the project, and then the final report can say "two group members each implemented did this part, and their results {matched, didn't match}" When two people attempt and come to different conclusions, that is interesting and a chance for everyone to learn!

First, we expect the project to be divided in the following subtasks

- dataset
    - dataset loading
    - data cleaning
    - feature scaling
    - feature engineering
        - PCA
        - ICA
        - LDA
        - Bloom filter
- Model selection
    - anomaly detection
        - K-means
        - Gaussian Mixture Models
        - PCA
            - kernel PCA
    - SVM
        - Gaussian RBF kernel
    - random forest
    - ensemble learning
- Hyperparameter Tuning
- Report writing

The tasks will be divided between the team members in the following way,

Chae:

- Anomaly detection: K-means + Performance measurement
- Anomaly detection: Gaussian Mixture Models + Performance measurement
- Hyperparameter Tuning
- Report writing

Maryam:

- Dataset loading
- Dataset cleaning
- Normalization & Class balancing
- Anomaly detection: PCA + Performance measurement
- SVM + Performance measurement
- Report Writing

François:

- Feature scaling
- Feature engineering
- Random forest + Performance measurement
- Ensemble learning + Performance measurement

- Report writing

All members will work on writing the final report, although the specific work division for this task has not been chosen yet.

List the main Python packages you expect to use. PyTorch? TensorFlow? Scikit-learn? Special packages for working with your data? (It is OK if this list is incomplete or changes for the final project.)

- sci-kit learn
- numpy
- matplotlib
- pandas
- imblearn