Mid-term Project

Spoofing attacks detection for industrial control systems using machine learning-based anomaly detection

Background

The protection of industrial control systems (ICSs) is of utmost importance due to devastating physical damage a potential cyber attack may cause. As shown in Fig. 1, the values of control (readings from actuators) and the values of state (readings from sensors) could be manipulated by malicious attackers. In this project, we have a case of a chemical process. Please read Sec. 4.1.1 of the reference paper and the README.txt in the folder to have some background information. Note that no domain knowledge in chemistry and process control is required to complete the project.

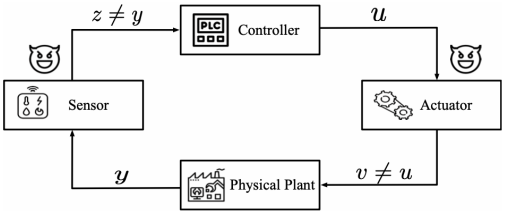

Figure 1 A general control system under spoofing attacks. *u*: true control action, *y*: true output observation, *v* and *z*: manipulated values of control and output.

Dataset

We have three datasets for training, validation, and testing. We collect sensor and actuator readings from the whole process. Each row is from a measurement at one time step. The explanation for each column (feature) can be found in README.txt. You are free to choose which features in your machine learning model.

Training: all data are benign, i.e., collected when the system is in normal status.

Validation: Each CSV file is a scenario. The data (rows) are sorted based on timestamp. We need to identify a csv file (a scenario) is normal or abnormal. In the validation set, validation-key.txt file has been provided. In this file, the first column is for scenario number, the right column is for label (0: normal, 1: abnormal).

Testing: You need to label 0 or 1 for each csv file in the testing. You **MUST submit your solution in .txt file with the same format as the key file in the validation set: first column – scenario number, right column – label, use ONE space to separate them. ONE line for one scenario.** We will run the script to evaluate your submission.

Report:

Task 0 (5 points): Give the contribution of each team member. For example, if you two contribute equally, just mention 50% for student A, and 50% for student B.

Task 1 (20 points): Get familiar with the dataset. You can do some statistic analysis (table or plotting) for each feature. You can also compare the feature difference between the normal data and abnormal data. **Put your analysis within 2 pages in the report.**

Task 2 (30 points): Using independent Gaussian analysis for anomaly detection. The threshold can be tuned using the validation dataset. Report the True Positive, False Positive, True Negative, and False Negative, Precision, Recall, and F1 score.

Task 3 (45 points): Using Multi-variate Gaussian analysis for anomaly detection. The threshold can be tuned using the validation dataset. Report the True Positive, False Positive, True Negative, and False Negative, Precision, Recall, and F1 score.

Hint: You don't need to submit the label for each row of datum. We only need your conclusion for a whole CSV. You can do majority vote or aggregation of the data, e.g., you can say if more than 50% (the threshold I give here is just random) rows are abnormal, the whole CSV is more likely to be abnormal.

VERY IMPORTANT THINGS:

1. **ONLY SUBMIT 1) code, 2) report, 3) solution of the testing data in .txt format (not following the format requirement will lose points). DON'T SUBMIT DATASETS!!!**
2. **We just need one submission from a team.**