Cyber Data Analytics Assignment 2

INTRODUCTION

Most data in cyber data analytics is sequential in nature. Applying machine learning to sequential data is difficult because past data (rows) provide information for future data (rows). The data points are thus not i.i.d (independently and identically distributed). Learning from sequential data and time series is a large domain covering many problems, solutions, and algorithms.

In this exercise, you will apply the techniques taught in class to the problem of anomaly detection in SCADA systems. Anomaly detecting is typically harder than classification because the data are unlabeled. We have to rely on statistics such as occurrence counts or value ranges to find anomalies, rendering many machine learning methods inapplicable. Securing SCADA system is considered one of the most important problems in cyber security.

LEARNING OUTCOMES

After completing this assignment, you will be able to:

1. Correctly apply machine learning methods to sequential data
2. Detect anomalies in multivariate data
3. Detect anomalies in continuous and discrete sequential data
4. Evaluate the performance of anomaly detection methods

INSTRUCTIONS

For this lab assignment you will use the SWaT dataset to detect anomalous behavior. You will divide the LOF / PCA and ARMA / N-gram tasks between the two of you, you may not work together on these parts! Together you will explore the data and compare the performance of the 4 methods.

Try to train on the training data (should not contain anomalies) and test on the test data (contains anomalies)!

Data exploration (5 points)

- Load the SWaT (train) sensor data into a Jupyter Notebook.
- What different kinds of signals do you see? Show 3 examples
- Are the signals clearly correlated? Do they show cyclic behavior? If so, show an example of when this happens.

- Use any method (preferably something simple) for predicting the next value for each of the time series and measure its performance. Is it easy to predict this? Which series are easy, which are hard?

The following 4 tasks are individual, this means every group member makes 2, one of LOF/PCA, one of Ngram/ARMA.

LOF (5 points)

- Perform LOF-based anomaly detection on the signal multivariate data points (do not take sequential context into account), use a distance of your choice.
- Plot the LOF scores on the training data 1 as a signal for several numbers of neighbors. Select a number to use and justify this choice using the obtained LOF scores and detected anomalies.
- Do you see large abnormalities in the training data? Can you explain why these occur? It is best to remove such abnormalities from the training data since you only want to model normal behavior.
- What kind of anomalies can / can you not detect using LOF?

PCA (5 points)

- Perform PCA-based anomaly detection on the signal multivariate data points (do not take sequential context into account).
- Plot the PCA residuals for different number of components on training data 1 in one signal. Choose the number of components based on the residuals and detected anomalies.
- Do you see large abnormalities in the training data? Can you explain why these occur? It is best to remove such abnormalities from the training data since you only want to model normal behavior.
- What kind of anomalies can / can you not detect using PCA?

ARMA (5 points)

- Learn an autoregressive moving average model (see Wikipedia for an introduction if unfamiliar) for 5 (different looking) signals. Most statistical packages (statsmodels in Python) contain standard algorithms for fitting these models to training data. Note that there exists a wide range of ARMA variants; you only have to use the basic model.
- Use autocorrelation plots in order to identify the order of the ARMA models. The parameters can be determined using Akaike's Information Criterion (AIC) or another model selection method.
- Plot the residual errors and study some of the detected anomalies. What kind of anomalies can / can you not detect using ARMA models?
- Which sensors can be modeled effectively using ARMA?

N-grams (5 points)

- Discretize the sensor data for 5 (different looking) signals using percentiles. Visualize the discretization.

- Apply N-grams (on 5 signals again) to sliding windows with a length of your choosing in order to find anomalies. Choose a value for N, and a value for a larger sliding window containing the N-grams. Count the occurrence frequencies of the N-grams in each window. Make a table with the different windows as rows and n-grams as columns, in each cell you put the counts for that n-gram in that window. Use a distance measure of your choice (tip: cosine) and detect anomalies using a simple nearest neighbor approach.
- Plot the anomalies you find. What kind of anomalies can / can you not detect?
- Which sensors can be modeled effectively using N-grams?

Comparison task (10 points)

- Compare the performance of the four implemented methods. It is ok if some method's implementations are less thorough. The goal of this task is to setup a sound comparison and evaluation, not implement new methods. Evaluating anomaly detection methods is not straightforward, and different research studies frequently use different measures. You can either:
  - Test point-wise precision and recall
  - Overlap-based false and true positives
  - Count a true positive if it detects at least one anomaly in an anomalous region
  - Compare the top-k detected anomalies,
  - Describe in a few lines which comparison method you chose for this data and why
- Keep in mind that in practice an analyst has to take action on every positive detected but will not study every detected data point. Which methods do you advice to use for the data?

Explanations and code quality (5 points)

- Remember to write clear code and explain your results understandably. During peer-review, your fellow students will be asked to run and understand your code!
- More text is not always better!

Bonus! (5 points)

Think of a way (study the papers) to combine the predictions of all the individual models into a single anomaly detection method. Implement it and evaluate its effectiveness compared to each of the methods individually.

RESOURCES

Slides from Lectures 3, 4

The paper "Characterizing Cyber-Physical Attacks on Water Distribution Systems" by Toarmia et al.

All are made available through Brightspace

Data from https://itrust.sutd.edu.sg/itrust-labs_datasets/dataset_info/

Wikipedia for excellent explanations of the used methods (ARMA, N-gram, …)

Links on Brightspace to online tutorials.

Code samples available on Brightspace.

PRODUCTS

A zip containing:

- A Jupyter Python notebook for the collaborative parts of the assignment. The word count should not exceed 1000 words more than the original count (see first cell). Include libraries used to run the code other than numpy, scipy, pandas, matplotlib and scikit-learn.
- Separate Jupyter notebooks for the individual assignments, each not exceeding 300 words more than the original count.
- The notebooks will be assessed using the below criteria.

ASSESSMENT CRITERIA

The assignment will be reviewed by your peers, and you are expected to individually review 2 reports. The estimated time you should spend on a review (including code review) is 1 hour. The login details will be provided in the week of the deadline.

Knockout criteria (will not be evaluated if unsatisfied):

Your code needs to execute successfully on computers/laptops of your fellow students (who will assess your work). You may assume the availability of 4GB RAM. Please test your code before submitting. In addition, the flow from data to prediction has to be highlighted, e.g., using inline comments.

Your report needs to satisfy the page limit requirements for the different parts. Submissions submitted after the deadline will not be graded.

The report/code will be assessed using these criteria:

Criteria	*Description*	*Evaluation*
Visualization	*Shows the behavior of one-two signals from the SCADA system. Provides useful input for further tasks.*	*0-5 points*
PCA/LOF	*PCA/LOF is used correctly, with explanations for the number of used principal components/neighbors. The kinds of anomalies detected are identified correctly.*	*0-5 points*

ARMA/Ngram	*The ARMA order/sliding window lengths and parameters are set correctly using only the training data. The residual errors are explained and visualized. The anomaly types and sensors are identified.*	*0-5 points*
Comparison	*Different properties of the algorithms are compared. The comparison is sound and the conclusions are reasonable.*	*0-5 points*
Evaluation	*Sound reasons are provided for the used evaluation metric. The conclusions are relevant for anomaly detection in practice.*	*0-5 points*
Bonus	*Creative solution, correctly implemented.*	*0-5 points*
Report and code	*The data-detection flow is clearly described, including preprocessing and post-processing steps.*	*0-5 points*

Your total score will be determined by summing up the points assigned to the individual criteria. Your report and code will be graded by the teacher and assistants, and the peer reviews are used as guidance. averaging to account for the number of peer reviews. In total 35 points (including bonus) can be obtained in each lab assignment, of which 10 are for the individual parts. In total, 140 points (including bonus) can be obtained in the 4 lab assignments, of which 40 are individual. The total number of obtained points will be divided by 12 to determine the final course grade.

You will receive a penalty of 5 points for each peer review not performed. Significantly different reviews will be subject to investigation. If deemed badly done by the teacher or TA, you will also receive 5 penalty points.

SUPERVISION AND HELP

We use Mattermost for this assignment. Under channel Lab2, you may ask questions to the teacher, TAs, and fellow students. It is wise to ask for help when encountering start-up problems related to loading the data or getting a machine learning platform to execute. Experience teaches that students typically answer within an hour, TAs within a day, and the teacher the next working day. When asking a question to a TA or teacher, your questions may be forwarded to the channel to get answers from fellow students. Important questions and issues may lead to discussions in class.

There is no separate lab session hosted at the university, it is your own responsibility to start and finish on time.

SUBMISSION AND FEEDBACK

Submit your work in Brightspace, under assignments. Within a day after the deadline, you will receive several (typically two) reports to grade for peer review as well as access to the online peer review form. You have 5 days to complete these reviews. You will then receive the anonymous review forms for your groups report and code.

There is the possibility to question the amount of points given to your work, up to one week after receiving the completed forms. You should do so via a private message to the teacher and TA in Mattermost.