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

Familiarization task (5 points)

Load the SWaT sensor data (train with training data, test with test data) into a Jupyter Notebook and understand the data. Answer the following questions:

1. What types of signals are there?
2. Are the signals correlated? Do they show cyclic behavior?

Visualize these types and the presence or absence of correlation.

The following tasks are individual, this means every group member makes does either LOF/Ngram or PCA/Regression.

LOF/Ngram task – individual (10 points)

Perform LOF-based anomaly detection on the signal multivariate data points (do not take sequential context into account), use a distance metric of your choice. Plot the LOF scores on a sample of the training data (it can take a long time to run) 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. Describe the kind of anomalies you can detect using LOF.

Choose one signal that displays interesting temporal behavior. Discretize the sensor data using percentiles. Apply N-grams to sliding windows with a length of your choosing to find anomalies. Choose a value for N, and a value for a larger sliding window M containing the N-grams. Count the occurrence frequencies of the N-grams in each length M 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 (a subsequence) in that window. Use a distance measure of your choice (tip: cosine) and detect anomalies using a simple nearest neighbor approach. Try to add differencing before the discretization pipeline. Does it improve performance? Explain. Plot the residuals. What kind of anomalies can you detect? Try your approach on different sensors. Which kind of sensors can be modeled effectively using N-grams?

PCA/Regression task – individual (10 points)

Perform PCA-based anomaly detection on the signal multivariate data points (do not take sequential context into account). For PCA, a residual is the distance (your choice) between the original and reconstructed data points. Plot the PCA residuals for different number of components on the training data 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. Describe the kind of anomalies you can detect using PCA.

Choose one signal that displays interesting temporal behavior. Make partial autocorrelation plots and use these to choose sliding window size for a linear regressor. Study the performance of a linear regressor for predicting the next data point using a range of sliding window lengths. Does the autocorrelation analysis agree with the obtained results? Explain why (not). Apply differencing. Does the predictive performance increase? Explain why (not). Plot the residual errors with a setting of your choice and study some of the detected anomalies. What kind of anomalies can you detect using linear regression models? Try your approach on different sensors. Which kind of sensors can be modeled effectively using linear regression?

Comparison task – 1 A4

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 in not straightforward, and different research studies frequently use different measures. You can either:

- test point-wise precision and recall, or
- overlap-based false and true positives, or /and
- count a true positive if it detects at least one anomaly in an anomalous region, or
- compare the top-k detected anomalies,
- or...

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 SWaT data? Use the validation set to evaluate your performance. The test data without labels is used to upload a solution to the Kaggle competition (which naively uses point-wise F1 score as evaluation metric). Apply a method of your choice to the Kaggle competition and try to outperform our baseline.

Bonus task – 1 A4

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. Feel free to use this solution in the Kaggle competition.

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

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 all parts of the assignment (including individual tasks). The word count should not exceed 1600 words (see first cell). Include libraries used to run the code other than numpy, scipy, pandas, and scikit-learn.

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 word count requirements for the different parts.

Submissions submitted after the deadline will not be graded, **deadlines are strict!**

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*
Linear regression / Ngram	*The sliding window lengths and parameters are set reasonably 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. Performance is ok.*	*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.

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 120 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 Lab1, 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.

Lab sessions are Friday's 10:45-12:45 physically in different locations at campus (check mytimetable for locations), or virtually on gather.town. Please see Brightspace for details.

SUBMISSION AND FEEDBACK

Submit your work in Brightspace, under assignments. Also submit it on peer.tudelft.nl. 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 review of your work, up to 3 days after receiving the completed forms. You should do so via the response function on peer.tudelft.nl.