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:

- *Correctly apply machine learning methods to sequential data*
- *Detect anomalies in continuous and discrete sequential data*
- *Detect anomalies in multivariate signal data*
- *Evaluate the performance of anomaly detection methods*

INSTRUCTIONS

Familiarization task – 1 A4

Load the BATADAL sensor data (start with training data 1, optimize using training data 2, test with test data) into your favorite analysis platform (R, Matlab, Python, Weka, KNIME, ...) and understand the data. Answer the following questions:

- What kinds of signals are there?
- Are the signals correlated? Do they show cyclic behavior?
- Is predicting the next value in a series easy or hard? Use any method from class.

Visualize these types, the presence or absence of correlation, and the performance of prediction.

ARMA task – 1/2 A4

Learn an autoregressive moving average model (see Wikipedia for an introduction if unfamiliar) for at least 5 individual sensors (pick them in a sensible way!). Most statistical packages (R, statsmodels in Python) contain standard algorithm for fitting these models from training data. 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. Note that there exists a wide range of ARMA variants; you only have to use the basic model.

Decide how to set the detection threshold sensibly. Study some of the anomalies detected anomalies. What kind of anomalies can you detect using ARMA models? Which sensors can be modeled effectively using ARMA?

Discrete models task – 1/2 A4

Discretize the sensor data using any of the methods discussed in class. Explain why you choose this method and why the obtained discretization makes sense. Visualize the discretization.

Apply any of the sequential data mining methods (N-grams, Sequence alignment with kNN, …) to sliding windows with a length of your choosing in order to find anomalies. Whenever an observed N-gram's probability is too small, or the sequential data is too distant from any training sequence, raise an alarm. Set your thresholds sensibly. What kind of anomalies can you detect using the sequential model? Which sensors can be modeled effectively?

PCA task – 1/2 A4

Perform PCA-based anomaly detection on the signal data. Set the threshold on training data to a value that results in few false positives on the training data. Plot the PCA residuals in one signal. 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 you detect using PCA?

Comparison task – 1 A4

Compare the performance of the PCA method with the ARMA and discrete models. Comparing 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. Do you recommend using PCA, ARMA, or discrete models?

Bonus task – 1 A4

Use PyTorch (or another framework) to learn Deep Neural Networks for anomaly detection as discussed in class and in the literature. Try out different learning rates and detection thresholds. Compare the performance with the three approaches tested above. Do you recommend using Deep Learning? Why (not)?

RESOURCES

Slides from Lectures 4, 5

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

All are made available through Brightspace

Your favorite analysis platform (R, Matlab, Python, Weka, KNIME, ...)

Data from https://batadal.net/

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

Links on Brightspace to online tutorials.

Code samples available on Brightspace.

PRODUCTS

A small report (max 4 pages, 5 including bonus), and the code used to obtain the results. Both will be assessed using the below criteria.

ASSESSMENT CRITERIA

The assignment will be assessed by peer review. The form will be made available directly after the assignment 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 and a Linux operating system, possibly a virtual machine. 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. When working in a data analysis notebook, you have to copy and paste the text and results into a printable document satisfying the requirements.
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*
ARMA	*The ARMA order 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*
Discrete	*The discretization is sound. The discrete method and way to set thresholds is explained clearly. The anomaly types and sensors are identified.*	*0-5 points*
PCA	*PCA is used correctly, with explanations for the number of used principal components. The detection threshold is determined soundly. The anomaly types are identified.*	*0-5 points*
Comparison	*The comparison is correct and the conclusions are reasonable. Sound reasons are provided for the used evaluation metric. The conclusions are relevant for anomaly detection in practice.*	*0-5 points*
Report and code	*The data-detection flow is clearly described, including preprocessing and post-processing steps.*	*0-5 points*
Bonus	*Correct application of deep learning to anomaly detection, sound experimental setup and analysis of the results.*	*0-5 points*

Your total score will be determined by summing up the points assigned to the individual criteria, and averaging to account for the number of peer reviews. In total 35 points can be obtained in each course assignment, the total number of obtained points will be divided by 90 to determine the final grade. In case one of the reviews is significantly worse than (at least 10 points difference) the others, this review score is not taken into account.
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

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.

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, and in the peer reviewing system. 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.