Cyber Data Analytics
Lab 2 - Anomaly Detection

Richard Vink (4233867)

June 5, 2019

1 Familiarization task

In order to better protect water distribution systems, machine learning can be used to create detection algorithms. BATtle of the Attack Detection ALgorithms (BATADAL) is a website which challenges developers to come up with the best algorithm to detect attacks on those infrastructes. We will be using their data[1] in order to create our own algorithm. The data consists of several attributes. All the L signals represents the water levels of the corresponding tank numbers. S indicates the status of pump 1. This can be either 0 (OFF/CLOSED) or 1(OPEN). The P values indicate the suction pressure as well as the discharge pressure for the specified junction. F indicates the flow through each pump and valve.

To show correlations between attributes. We will create a correlation matrix. The correlation matrix can be seen below in Figure 1. Also one of the correlations is taken from the matrix and plotted in Figure 6 located in the Appendix.

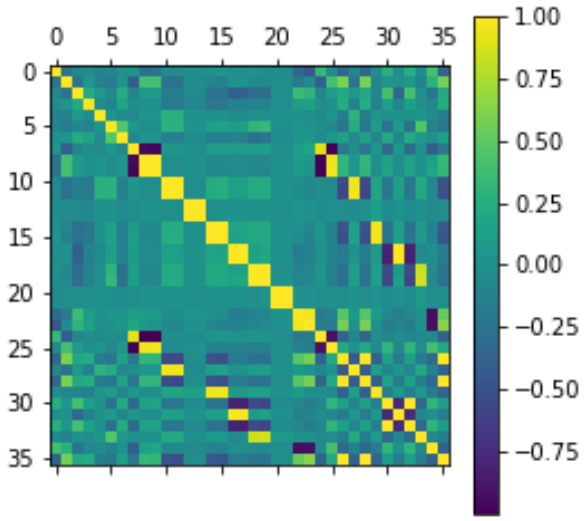

Figure 1: Correlation matrix for Training Dataset 1

As you can see in both figures. There is a correlation. There is also some sort of cyclic behaviour as seen in Figure 6.

One can now create a prediction model in order to predict the cycle of F_PU1. We will use an Autoregressive (AR) model. The outcome of the first 500 values are visualized in figure 7 which can be found in the Appendix.

What kind of signals are there The data was loaded

2 ARMA task

For this task we will learn ARMA for flow variables for flow values of 5 different pumps. These ARMA values all have an order of (4,4). It turned out Pump 7 showed the last attack and Pump 10 the first. The plots for these ARMA values can be seen below. The other 5 flow pump values can be found in the appendix.

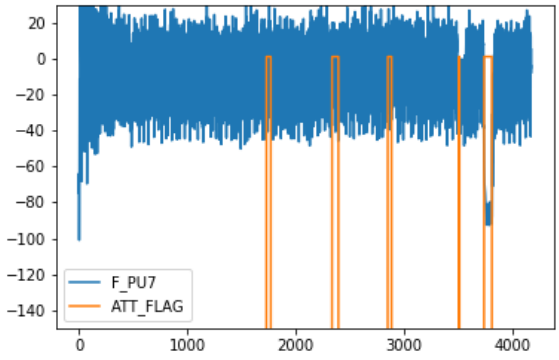

Figure 2: ARMA for F_PU7

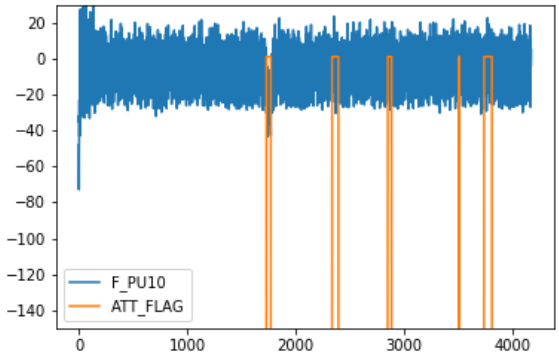

Figure 3: ARMA for F_PU10

3 Discrete Models task

Now the data will be discretized. This will be done using SAX as a discretion method combined with n-grams. The code saves all the timestamps for the

anomalies. If an anomaly occurs in the data, it will return in a True Positive. For L_T1. The Truth Positives are 41, whilst the false positives are 98. This means that 41/139 anomalies could be found. For L_T2 this is 17/125. For L_T3 this is 1/20. For L_T4 this is 45/556. For L_T5 this is 14/85. For L_T6 this is 15/291. For L_T7 this is 23/429. L_T1 has the highest TP rate, meaning that most of the anomalies were found.

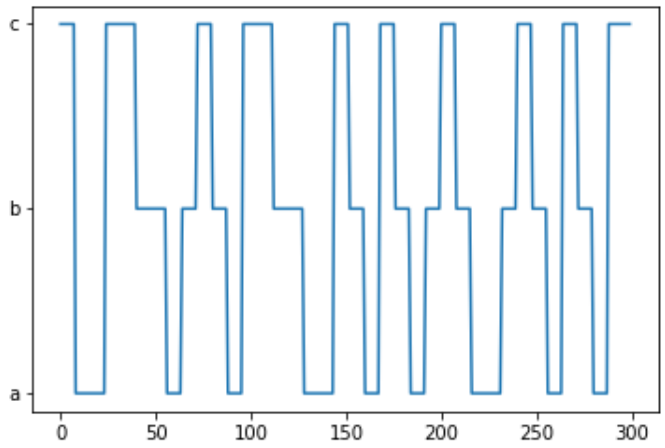

Figure 4: Visualization of Discretion L_T5

4 PCA task

This task is all about PCA-based anomaly detection on the signal data. The plot of the PCA residuals can be seen in figure 5. Training 2 consists of some high spikes. These are large abnormalities. PCA is used to explain the variance. Thus, it shows anomalies at places in the series where the variance is different.

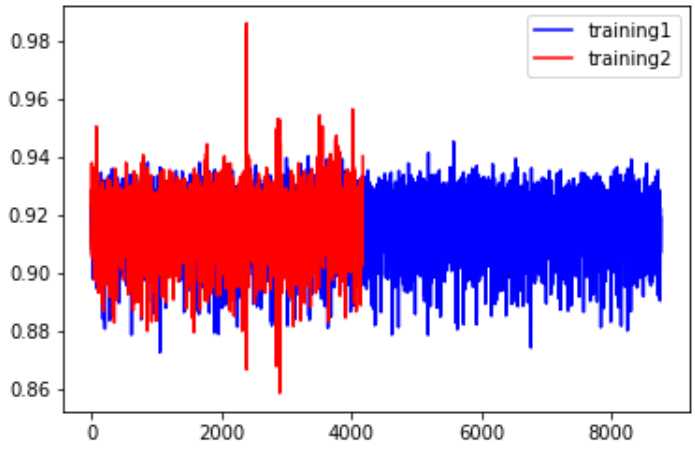

Figure 5: PCA of Training 1 + Training 2

5 Comparison task

This tasks compares the performance of each of the model. This will be done by computing the precision for all sets.

1. **ARMA**: F7 was the best performing signal here. Which had 3813−3741 True Positives which is 72. Out of 4177 means a precision of 72/4177 which is around 0.017.

2. **Discrete Models**: Most True Positives were found at L_T4 namely: 45. Thus 45/4177. Which is 0.01077.

3. **PCA**: 12 True Positives were found. Thus 12/4177 is 0.00287.

Concluding, it seems that ARMA has the best precision, however as shown in the figures it only detected the last part of the anomalies. To detect all sort of anomalies, a different model should be used. The second highest precision is the one from the discrete models. These models also detect anomalies at all parts of the sequence. Though, for a better understanding all different prediction models should be used.

6 Appendix

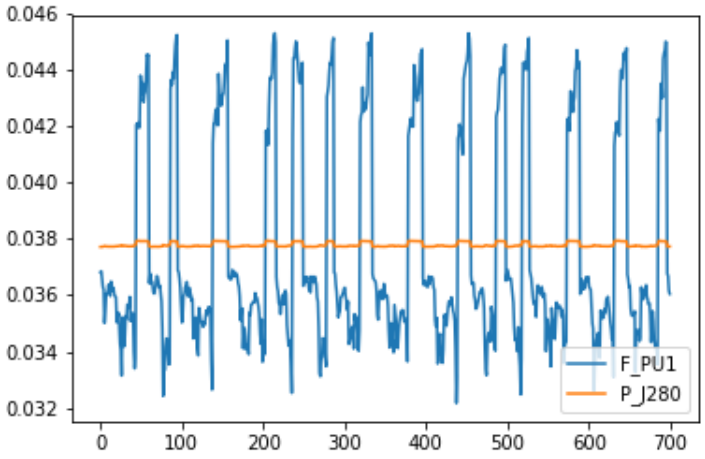

Figure 6: Cycle + Correlation of FPU1 and PJ280

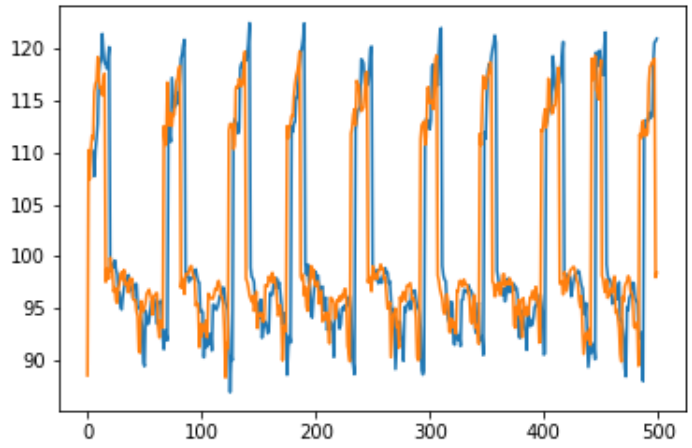

Figure 7: First 500 values of predicting F_PU1 with AR(5) model

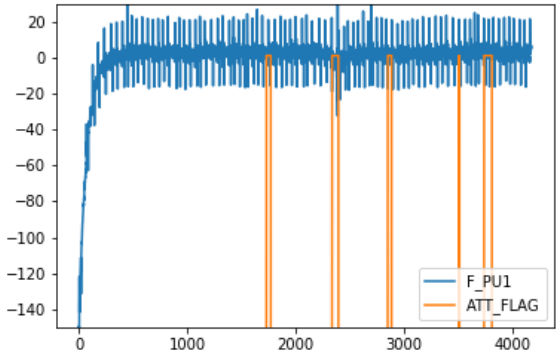

Figure 8: ARMA for F_PU1

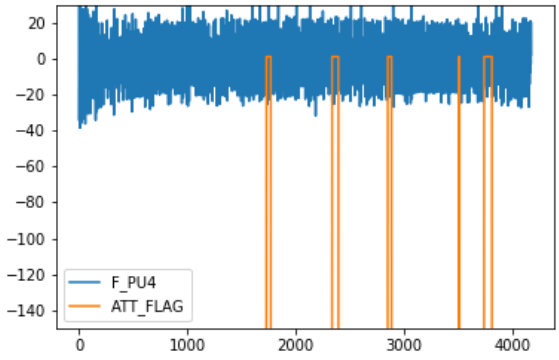

Figure 9: ARMA for F_PU4

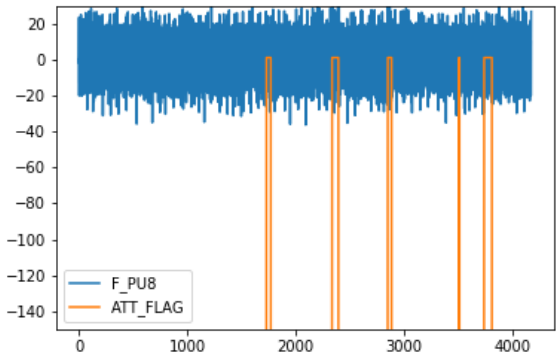

Figure 10: ARMA for F_PU8

References

[1] Riccardo Taormina, Stefano Galelli, Nils Ole Tippenhauer, Elad Salomons, Avi Ostfeld, Demetrios G Eliades, Mohsen Aghashahi, Raanju Sundararajan, Mohsen Pourahmadi, M Katherine Banks, et al. Battle of the attack detection algorithms: Disclosing cyber attacks on water distribution networks. *Journal of Water Resources Planning and Management*, 144(8):04018048, 2018.