CS 4035 LAB2 - ANOMALY DETECTION GROUP 29

Yuhang Tian
5219728

Mingyu Gao
5216281

May 26, 2021

1 Part1 - Visualization

SCADA is a system to collect data from sensors and actuators with the network of software and hardware components, which is vulnerable to cyber-physical attacks.

1.1 What different kinds of signals do you see?

The data collected by the 51 sensors and actuators can be divided into three main kinds - periodic change within a certain range, continuous drift, and level-off value, which can be seen in the table1.

Category	Signals	Description
periodic change within a certain range	FIT101 LIT101 MV101 P101 FIT201 MV201 P203 P205 FPIT301 FIT301 LIT301 MV301 MV302 MV303 MV304 P302 LIT401 FIT601 P602	1) Point Anomaly Detection: Hard 2)Contextual Anomaly Detection: Easy 3) Collective Anomaly Detection: Easy
continuous drift	AIT201 AIT202 AIT303 AIT402 FIT401 AIT501 AIT502 AIT503 AIT504 FIT501 FIT502 FIT503 FIT504 PIT501 PIT502 PIT503	1) Point Anomaly Detection: Hard 2)Contextual Anomaly Detection: Hard 3) Collective Anomaly Detection: Hard
Maintain at a certain value	P102 P201 P202 P204 P206 P301AIT401 P401 P402 P403 P404 UV401 P501 P502 P601 Time Domain Signal	1) Point Anomaly Detection: Easy 2)Contextual Anomaly Detection: Easy 3) Collective Anomaly Detection: Easy

Table 1: Categorization of all signals

1.2 a) Are the signals clearly correlated? b)Do they show cyclic behavior?

a) We draw a heat-map to visualize the correlation between each signal as shown in Figure 2 in the visualization part. Checking the correlation between two signals can help us determine the correlation between any columns in the data set.
Most signals have no obvious correlation (value +-0) with each other, while there exist some perfect correlated signals (excluding the signals with themselves) and perfect negatively correlated signals (signals move in the opposite direction).
As for the perfectly correlated signals, for example, we can see in Figure 3 that signals 'P403' and 'P404' are perfectly correlated. It is not strange as the signal P404 is the backup of P403. We can also conclude from the heat-map that perfect correlated signals are all the third kind of signals in table1, which have stable values at all times.
We can see that the signal AIT504 is negatively correlated to many signals like P501, with a correlation value of 0.92.
b) As for the cyclic behavior, the first kind of signals in table1 repeat their values in certain periods. And some of the second shows cyclic behavior along with fluctuations, like AIT502, repeat its periods five times from point 180000 to 210000 in the data set. For those signals which maintain a certain value, their stable status does not show cyclic behavior.

1.3 a) Is it easy to predict this? b)Which series are easy, which are hard?

a) We implemented a sliding window to predict some signals. By calculating values in a window(size of 30), we can use the value at the current point to predict the value of the next point, which transforms the sequence into a supervised learning problem.

b) Our sliding window method predicts four signals in figure 5, and we can see that the predicted curve almost coincides with the true collected values. In the four signals we chose, the last signal - 'FIT502' is much harder as its value fluctuates more frequently.

2 Part2 -Individual Task - LOF

2.1 Do you see large abnormalities in the training data?

The number of abnormalities relies on the number of neighbors that are chosen to estimate a sample point. If the neighbors are few, a great number of sample points will be sentenced to be outliers, but with the increase of neighbors, the number of abnormalities in the training set becomes less. The number of neighbors cannot be set too large, as it will be less sensitive to the truth outliers in the testing set. However, the number cannot be set too small either, as it will become too sensitive to the local changes.

2.2 Can you explain why these occur?

Even in the data set where no abnormalities are included, some segments in the signals may still not rigorously follow the normal behaviors. In consequence, some points may form outlier clusters in which points are normal but are abnormal globally when taking all segments into consideration.

2.3 It is best to remove such abnormalities from the training data since you only want to model normal behavior?

If a pattern that is normal shows abnormal in the training set, when a similar pattern appears in the testing set, the similar one would be predicted as an abnormality because the pattern originally in the training set is sentenced to be abnormal. If the same pattern is removed from the training set, when a similar pattern appears in the testing set, this coming one would be predicted as an abnormality because the model has no idea about this "new" pattern. Therefore, deleting the abnormalities in the training set seems to impact less on the predictions.

2.4 What kind of anomalies can or can you not detect using LOF?

LOF can detect point anomalies by measuring the local deviation of density for a given sample with respect to its neighbors. Point anomaly is defined as *A single instance of data that is anomalous since it deviates largely from the rest of the data points* which is consistent with the measurement of LOF *how isolated the object is with respect to the surrounding neighborhood*.

3 Part3 -Individual Task - PCA

3.1 Do you see large abnormalities in the training data? Can you explain why these occur?

Using PCA to detect the anomaly is quite similar to auto-encoder. PCA, trained by the training set, is used for reducing the feature dimensions of samples and recovering them back. If the samples can be recovered from the loss compression, they obey normal structure and are normal data; otherwise, they are abnormalities. In this situation, the recovering threshold is quite important to decide whether the recovered points belong to normal ones. If the threshold is set too small, it will have a considerable amount of abnormalities in the training data. Thus, in order to make the judgment of anomaly fair, the threshold should be set suitable large.

3.2 It is best to remove such abnormalities from the training data since you only want to model normal behavior.

If the abnormalities are removed from the training data, we will have no idea where to put the threshold. The abnormalities in the training data tell people which points cannot be recovered back.

3.3 What kind of anomalies can or can you not detect using PCA?

PCA detects point outliers that have an obviously different variance than the regular distribution in the direction of the principal components.

4 Part4 -Individual Task - ARIMA

4.1 What kind of anomalies can / can you not detect using ARMA models?

ARMA has the ability to identify contextual anomalies because ARMA regress on previous values then uses them to predict the next value. According to its results, we observed that when the tendency of the predicting signal gradually changes, ARMA can abide by the signal well, but when there is a sudden change, it cannot perfectly catch and follow this change. Since ARMA is quite sensitive to local change, it is suitable for detecting contextual anomalies.

4.2 Which sensors can be modeled effectively using ARMA?

The prediction using ARMA is quite time-consuming. In this task, we adopted ARMA as an off-line model - given training set and to predict testing set. However, we think ARMA is much better to be designed as an online model. ARMA is quite suitable for the sensors which monitor the moderately changing signals such as AIT402, AIT501, and AIT502. Other signals like MV301 are not appropriate for it, since it often alarms abnormalities for them when observes dramatic changes.

5 Part5 -Individual Task - NGram

5.1 Plot the anomalies you find. What kind of anomalies can / can you not detect?

For signal 'LIT101', the N-gram model finds the contextual anomaly from point around 230000 to 260000. And it also finds three-point anomalies around the points 360000 and 380000 in the test data set. However, it mistakenly marks many normal points as anomalies.
For signal 'LIT301', the N-gram model finds the contextual anomaly around the point from point around 230000 to 260000 and two-point anomaly around the point 10000 and 460000 in the test data set. It fails to find another two obvious point anomaly - around the point 95000 and 350000 in the test data set.
For signal 'AIT202', the N-gram model marks several points in the contextual anomaly from point around 230000 to 260000 which reflects that it successfully judges them as collective anomalies.
For signal 'AIT402', the N-gram model distinguishes several points in the contextual anomalies from point around 230000 to 260000, however, the number of its marking is quite smaller than the real number of outliers.
For signal 'AIT503', the N-gram model marks some point anomalies while it fails to detect the obvious contextual anomaly from point 230000 to point 260000. This signal performs worst using N-gram model.

5.2 Which sensors can be modeled effectively using N-grams?

As the classification we make in table1, we implement the N-grams model on three kinds of signals. For the first type of signal- periodic change within a certain range, signal 'LIT101' and 'LIT301', this model marks more points as an anomaly which increases the false positive value. For the second type - continuous drift, the signal 'AIT202' and 'AIT402', the N-gram model marks fewer points as anomalies, which results in a lower True Positive value. What's more, the signal 'AIT503' performs worst.
The N-grams model is more suitable for detecting anomalies in the signal which changes periodically within a certain range.

6 Part6 -Comparison Task

The comparison task involves comparing the four detection methods. However, the assignment was interpreted in different ways by the team partners. Therefore, we separately compare the LOF and PCA methods, and the ARIMA and NGRAM methods against each other.

6.1 Which methods do you advise using for the data?

LOF v.s. PCA
To compare the detection precision of LOF and PCA, we plot the confusion matrix to show their performance as we run all test data set in these two methods. Confusion matrix can evaluate the different models by breaking positives into true positives and false positives.
To perfect the comparison, we execute some selection and process these two models. For the LOF model, we choose the signal 'LIT301' to draw the confusion matrix as it performs better than any other signal or signal combination. In the PCA model, we reduce the 51-dimension features to 12 components according to the residual plot.
From the perspective of positives, LOF seems better achieving normalized TP of 0.99 while keeping the FP to 0.01. The confusion matrix also allows us to see the False Negatives at the same time. FNs are important to keep low since it means that there is a breach in the system that we could not recognize which is more dangerous than having FPs. PCA performs better in the FN value.
According to our results, PCA seems the better strategy as its TP value decreases by 18 percent while the FN value increases by 41 percent compared with the result of LOF. Of course, we could definitely achieve better results if we could combine these methods to detect many different kinds of anomalies.

Method	TP	difference of TP	FN	difference of FN
LOF	0.99		0.41	
PCA	0.81	-18%	0.24	+41%

Table 2: Comparison of LOF and PCA

ARIMA v.s. N-gram
Firstly, the N-gram is better than the ARIMA model from the perspective of speed as the former one could run faster and predict for the

whole test data set. However, the ARIMA model runs for half an hour for predicting 300 points.
Secondly, the ARIMA model seems to predict the signal too well and it could only detect point anomalies or a shorter sequence of anomalies as these points deviate normal data too far. Or it needs more tuning to get better results.
However, the ARIMA model is not useless. We select one region which contains 50 positive and 250 negative points from point 11160 to point 11460. As shown in part6 in the .ipynb file, the ARIMA model finds 23 anomalies in total 50 anomalies while all five signals of the N-gram model find nothing in this region even some signal's results mark a lot of anomalies in the total test data set.
To sum up, the N-gram model performs better than the ARIMA model from the perspective of effectiveness and speed while the ARIMA model also performs better in some anomalous regions.