Cyber Data Analytics - Assignment2

Prerak Mody - 4777042

May 2019

1 Task1 - Familiarization Task

The BATADAL Dataset consists of 43 columns of raw data (X) and one column with the attack flag (Y). The different types of signals (few examples in fig. 1) that exist are :-

- Continuous signals - Eg : tank sensors for water level (L_T1 to L_T7)
- Binary Signals - Eg : Pump Actuator Status (S_PU1 to S_PU11) or ATT_FLAG (1's and 0's)

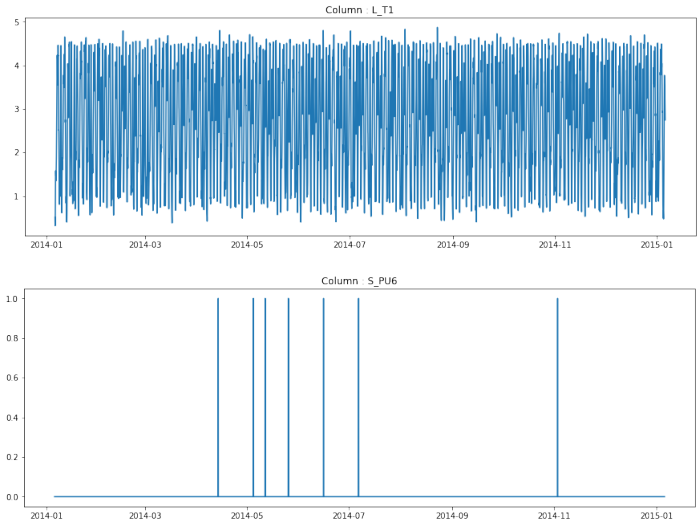

Figure 1: (Q1) Types of signals

We can notice in fig. 2 that there exists the presence of both highly positive and negative correlations between signals such as :-

- Strong Positive Correlation - F_PU2 and S_PU2 with P_J269 (i.e flow and status of pump2 is positively correlated with the pressure)
- Strong Negative Correlation - P_J269 with F_PU1 (i.e pressure is negatively correlated with flow in pump1)

We do a timeseries prediction task using linear regression as can be seen in fig. 3.

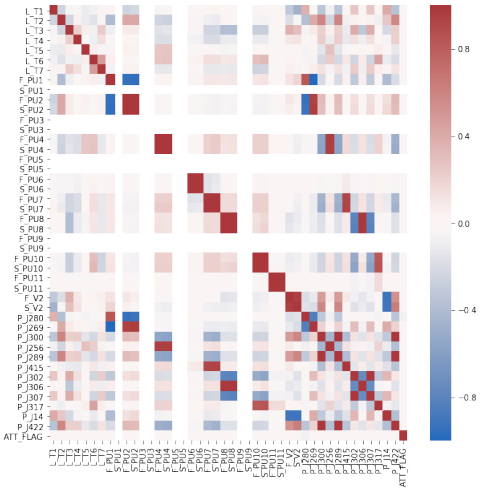

Figure 2: (Q1) Correlation between signals

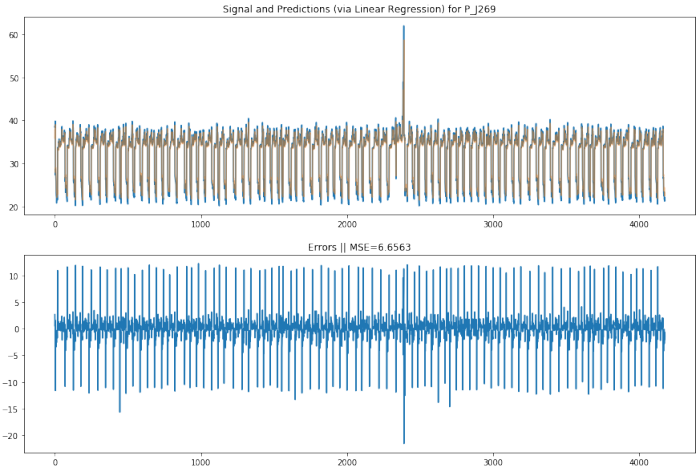

Figure 3: (Q1) Predicting time series values

2 Task2 - ARMA Task

In this task, we use the ARMA model to detect anomalies in the dataset. The datasets loaded are the first and second training sets. We then visualized most of the sensors and the variation in their value over time was looked at. The following five sensors were selected to analyse: ['F_PU1', 'F_PU2', 'F_V2', 'P_J289', 'P_J269', 'P_J307', 'P_J14']. For training we use the first dataset with normal data. For testing we use the second dataset, which contains normal and attack data.

The generation of ARMA models is described next. The initial step if to check if the signal being tested is stationary. For this, the Augmented Dickey-Fuller Test is used. Using the ARMA model requires to test for stationarity. In order to determine the 'd' parameter (differencing) it is imperative to know if the signal is stationary. The autocorrelation and partial-autocorrelation plots and AIC statistics are then used for determining the optimal order (p,q) of the ARMA model. The ARMA model is then created using the first dataset. Based on the parameters of the model created, the ARMA model for the second dataset (test set) is created.

The residuals for the prediction model are then obtained. In order to categorize sample as anomalies or not requires the setting up of the threshold value. The threshold in this case is set as twice the standard deviation. The samples with residuals higher than the threshold are marked as anomalies.

3 Task3 - Discrete Models Task

We chose the Symbolic Approximate Aggregation (SAX) method [1] which transforms a time series input to a high level representation (i.e strings). This method is quite useful since it reduces the dimensionality of the data and makes it easier to process. The original time series is first transformed into its Piecewise-Aggregate-Approximate (PAA) format (for dimensionality reduction using windows of size=10) and then utilizes a lookup table to convert the PAA data into a string using a predefined alphabet as can be seen in fig. 4

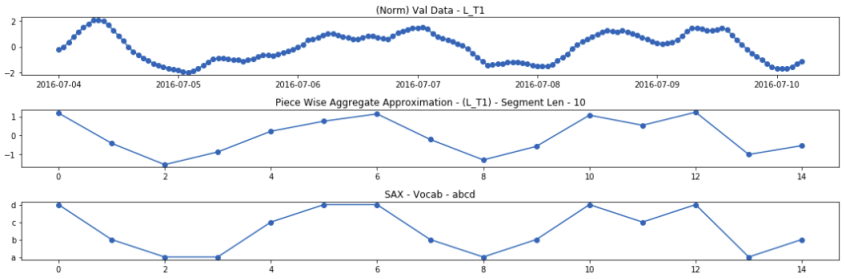

Figure 4: (Q3) Discretization using SAX

We apply the N-grams technique (N=3) to identify any anomalies. Such an approach is able to detect volumne anomalies very well in signals such as ['F_PU6', S_PU6, F_PU7, S_PU7, 'F_PU10', 'S_PU10', 'F_PU11', 'S_PU11', 'P_J14', 'P_J302', 'P_J307']. Vizualizations for these signals can be found in *03_Discretization.ipynb*

4 Task4 - PCA Task

To carry out PCA, certain preprocessing steps are carried out. This involves removing labels from the train dataset and normalizng the signals. Certain abnormalities are removed from the datasets since these might affect the principal components. The figure showing the abnormlties can been in the Jupyter Notebook - 04_PCA_Task.ipynb. The netwrok has 44 signals. In [**?**], the vast majority of variance is captured atmost by 14 or 15 signals as shown in Figure 4.

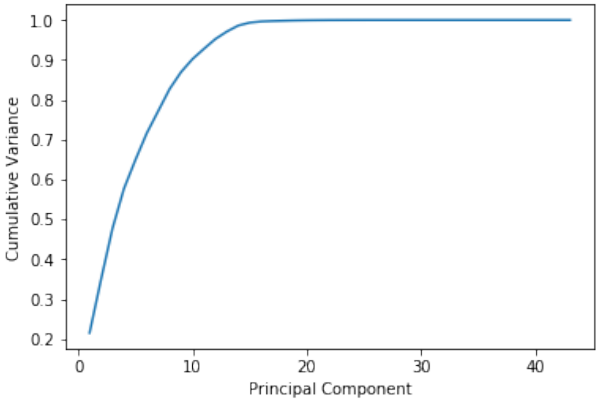

Figure 5: Cumulative variance captured by Principal Components

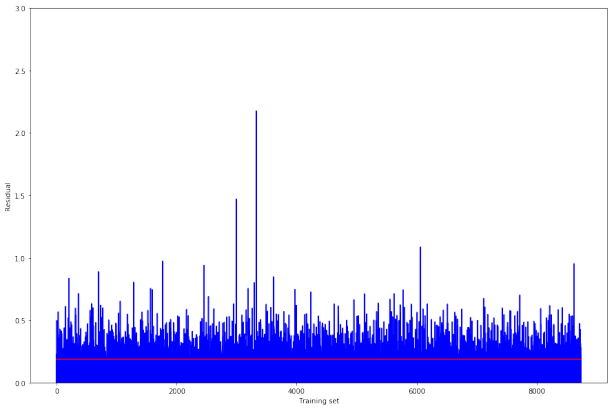

Figure 6: Residuals for training data with threshold

A function can also be found in the notebook which returns the number of principal components required. While the threshold is set using the method described in [**?**], it seems to raise higher false positives. Therefore, the threshold was manually calibrated a bit. The PCA model was then tested in the evaluation dataset (dataset 2). The metrics are as follows.

PCA model is able to detect point-wise anomalies but it has the drawback that it is not able to find the corresponding sensor.

Metric	Value
TP	87
FP	33
Accuracy	96.05
Precision	72.50
Recall	39.73
F_score	51.33

Table 1: Table 1: Metrics for PCA

5 Task5 - Comparison Task

Appendix

5.1 Task2

Here we show a table of the AIC values of all sensors to pick which sensors we are able to model effectively.

References

[1] J. Lin, E. Keogh, L. Wei, and S. Lonardi, "Experiencing sax: a novel symbolic representation of time series," *Data Mining and knowledge discovery*, vol. 15, no. 2, pp. 107–144, 2007.