CS4035 Cyber Data Analytics: Assignment 2
Group 9

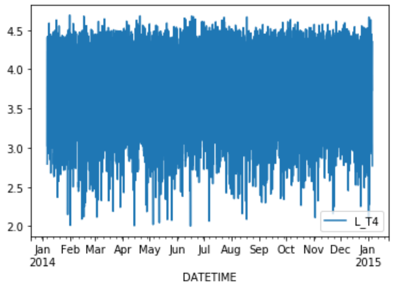
Georgios Dimitropoulos: 4727657
Emmanouil Manousogiannis: 4727517

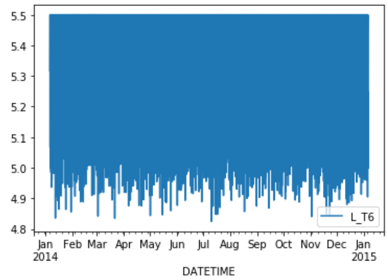
June 4, 2018

Familiarization task

Kind of Signals

A huge number of techniques have been proposed to detect anomalies in traffic data [1], [2], [3], [4] and [5]. The primary goal of this work is to detect the presence of an ongoing attack in the BATADAL dataset. We have one training set and one evaluation set with several measurements of sensor signals which include water tank level(L_T), pressure($P-J$) and flow levels of pumps and valves ($F-PU$,$F-V$).In total we have 31 signals of different sensors and 12 actuators which either indicate that a sensor is active or not.Each signal sample,is accompanied by its corresponding timestamps and their labels.As an example, the signals LT4 and LT6 can be depicted in the below figure.





Correlation of Signals and Cyclic Behavior

It can be depicted by Figure 1 the correlation matrix for all the signals. The Signals that have a red color in this heatmap are highly correlated. It can be observed for instance that some signals indicating pressure are highly correlated with some corresponding sensors indicating water flow levels under normal circumstances.

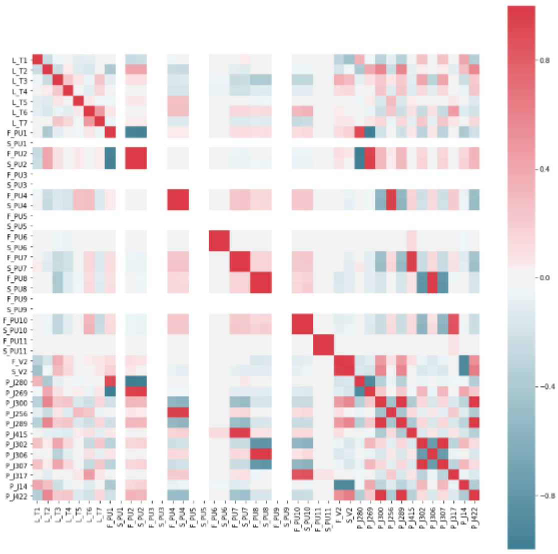


Figure 1: Correlation Matrix as a Heatmap

The frequency domain representation of a signal facilitates the observation of several characteristics of the signal that are not visible time domain. For instance, frequency-domain analysis becomes useful in the examination of the cyclic behavior of a signal. Thus, for this reason we transform the signals to frequency domain through the Fourier Transform. It can be observed in figure 2, that the points with large spectral lines denote a cyclic behavior

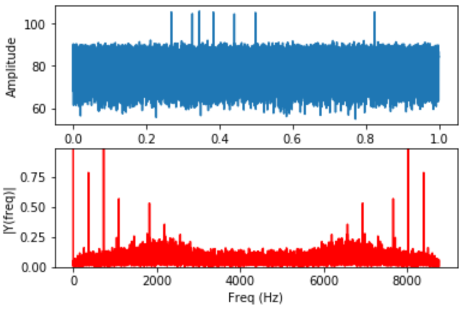


Figure 2: Cyclic Behavior

Difficulty of Prediction

In order to determine if the the prediction of the the next value in a series is easy or hard we implemented the Persistence algorithm. For instance the plot of the predictions and expected results for the FV2 signal can be depicted in figure 3 , where the mean squared error was adopted as metric difference between them and was found to be 0.436 which is a good enough performance.

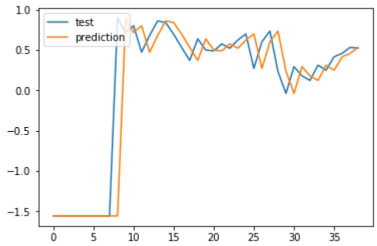
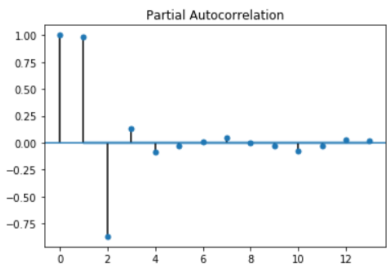
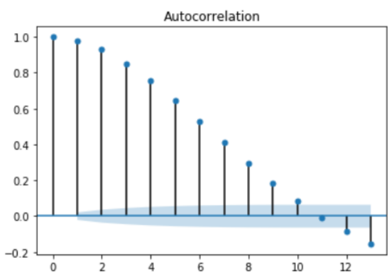


Figure 3: Implementation of Persistence Algorithm

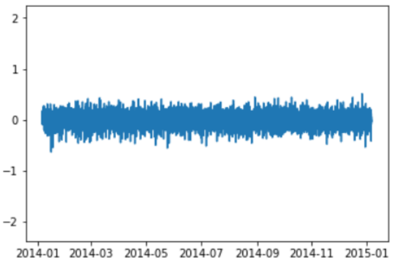
ARMA task

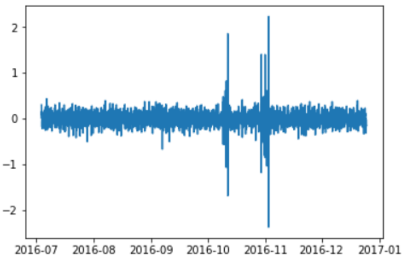
In this section we are implementing the autoregressive moving average model (ARMA).ARMA model uses the values of the previous p values of a series as well as the q previous residuals in order to predict the next value of the series.The first thing in order to plan our actions was to check the stationarity of our signal through a Dickey-Fuller test.After confirming that, we tried to determine the best order (p and q of our model), we initially used the autocorrelation (ACF) and partially autocorrelation (PACF) plots of our signals. An example is shown below. As can be seen





there is a strong positive autocorellation over the first 10 lags, which would indicate a selection of MA order between 1 or 2 where this is strongest.Similarly we can make an estimation for the AR order,based on the PACF plot. However, in order to confirm our initial estimations and also in order to tune the AR and MA coefficients, we also performed an AIC test. We compared a variety of different parameters and selected the model which scored the lowest AIC score.
After training our model on the training dataset (excluding all the actuators which contain binary data), where no attacks are considered, we used the same model on our evaluation dataset.In order to detect any anomalies, we plotted the residuals (actual signal values minus the model prediction) for both models of the training and evaluation data.A typical example of sensor L_T1 is added below. As can be easily noticed,there is a huge difference between the predicted value

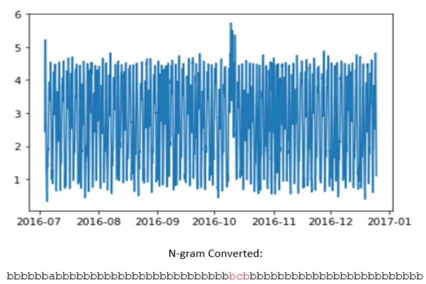


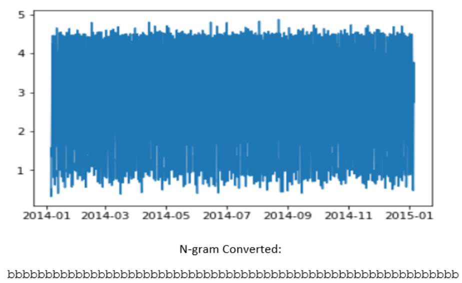


of the ARMA model and the actual value between October and November 2016. This identified anomaly, is clearly suspicious and should raise an alarm for a possible attack.In order to determine the threshold above which we raise this alarm, we start with a typical value of $2 * \sigma$, where σ is the standard deviation of the residual.Then we steadily increase the threshold until no sample from the training data is considered an anomaly. (as there were no attacks there). Of course we set a maximum limit of $3.5 * \sigma$ to avoid a really high threshold in some cases.
ARMA is a very efficient model, in case we want to detect sudden anomalies of points that do not last for long. In case an anomaly lasts for a long time period, it is hard for ARMA to detect it after a while,as it is using the previous actual values to predict future ones.

Discrete Models task

In this section, we tried to discretize the sensor data using the N-gram sequential data mining method.Using Symbolic Aggregate Aprroximation algorithm (SAX), we converted our sensor signals so that they can be represented from a letter series.More specifically, the procedure is as follows. We divided our signals in n 'batches' and computed the mean value of each batch. We also divided the values of they y-axis in 3 areas. 'A' indicatin low y-values,'b' representing medium values and 'c' representing high values.Each of the afformentioned batches then, is assigned with one of the mentioned letters based on its mean value.An indicative example will be ploted below. We have to mention here, that we decided to split our signals in 60 subsets(batches), as every one of them should represent a region of our signal,without being too large or too small either. As can be noticed,the mean value of the sensor on the right side stays constant throughout the

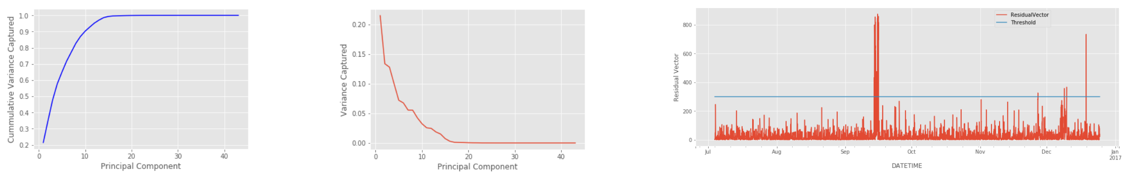




whole plot. On the other hand, there is a significant spike in the middle of the plot, which lasts for quite long.As a consequence, the mean value of the signal is increased and this is captured by the discretization technique we applied.We can conclude then, that in an anomaly detection problem this can be quite helpful as it can capture anomalies that happen for a significant amount of time. On the other hand, it is not very efficient in detecting 'instant' anomalies, as the mean value is not affected a lot.
In order to classify an anomaly we examined all thre created *trigrams* on the training dataset for each sensor signal. We computed the probability of the signal to be present at the training set, where no attacks were mentioned.After that, we counted the same probabilities on the created trigrams of the evaluation data signals.Since many trigrams had probabilities quite close to 0, we decided to set the threshold to 0,meaning that we classified a detected anomaly as an attack, if this trigram never appeared in the training set.

PCA task

PCA [6] is a coordinate transformation method that maps a given set of data points onto new axes which are called principal components and are linearly uncorrelated in a way that the first of them explains the maximum variance in the data and so on. As a preprocessing step we normalize the data in order to have zero mean and unit variance. It can be observed from the below figure (left and center) that the first 15 principal components are able to explain 99% of the variance in the data and more specifically the first 10 principal components are able to explain 90% of the variance in the data. Hence, for this reason the first 10 principal components are chosen to model the Normal Subspace through following the methodology proposed in [7] and [8]. By plotting the projection of the signals onto the next 5 (principal components 11-15), we depict that the signals show a large number of spikes that can be used to model abnormalities in the data. Thus, principal components 11-15 are chosen to model the Anomalous subspace. After performing the modeling of the normal and anomalous subspaces, we project the test data onto the these subspaces. We use the the abnormal changes in the residual traffic as an indication of the presence of anomalies. In order to be able to do this, we compute the Squared Prediction Error (SPE) for the residual vector as described in [9] . Data are classifying as anomaly if the the SPE is larger than a specific threshold. The residual vector obtained though projecting data into the anomalous subspace can be depicted in the below figure (right). We could say also that some abnormalities in the beginning may occur as the system has not been stabilized yet. The threshold is set on a way that results in few false positives. Q-statistic test [10] was examined in order to determine the right value for that threshold. However, we found that the threshold was not optimal for our case and for this reason we set it through experimentation and cross-validation. The confusion matrix and the evaluation metrics that was found can be depicted in table 3. It is worth mentioning that a possible variation in our method could be to extend the subspace method to diagnose anomalies in a broader variety of traffic data as described in [11]. Finally, we could say that the PCA model is able to detect point-wise anomalies but it is not able to find for instance the corresponding sensor.



Evaluation Metric	Value
Accuracy	95.04
Recall	10.05
Precision	68.75
F-Measure	17.53
TP	22
FP	10
TN	3948
FN	197

Table 1: Evaluation Metrics

Comparison task

In this question we tried to compare the performance of the PCA method with the ARMA and the discrete model.The comparison method was not straightforward at all due to the characteristics of our problem.There are very few data labeled as positive and our three methods are also quite different in their implementation and in the kind of anomalies they can detect.

Below we are presenting a summary of all the performance metrics that we could get, namely accuracy , precision, recall and f score.

Evaluation Metric	PCA	ARMA	Discretization
Accuracy	95.04	91.08	83.17
Recall	10.05	8.17	15.42
Precision	68.75	24.26	16.21
F-Measure	17.53	4.37	14.79

Table 2: Evaluation Metrics of all methods

As far as the the PCA model is being considered, *point-wise detection* of true and false positives was adopted. As we mentioned before in this case a point is labeled as a TP if the detected residual value lies in the anomalous region.This was quite straightforward for this method as there is one prediction for each data sample. Regarding ARMA and the discretization method, this was not easy. Those models run for each sensor separately, so it is difficult to extract a point-wise evaluation system. However, we tested our models on all sensors and counted their evaluation metrics additively. Of course, this is not in favor of those two methods as, even if an attack is present,not all specific sensors are aware of it. For this reason we kept tract of all points that were classified as an anomaly by any sensor and calculated our metrics based on this.
However, we have to mention that regarding Discretization method, even this point wise evaluation , is not really fair.This is probably one reason that the above results are pretty low for this model.Discrete models, will classify a whole region as anomalous if its mean value is above a certain threshold. When classifying all of this region points as positives, this is not actually true, as only one part of this region will probably be part of an attack. A more fair way to extract discretization method evaluation metrics would be *to assign one true positive,if at least one point of the detected anomaly region is part of an attack*. Similarly, one false positive should be assigned every time there is no attack mentioned in one anomalous region.

In general,from the table above we can see that *PCA* model is by far the best in terms of performance. Especially in terms of precision which is the most crucial evaluation metric in our case, it achieves the highest score between the tree so it is probably the best choice.

Bonus task

In this section we tried to combine our three model in one and find out if our implementation improved our results or not.There are many different approaches we could follow. Our presented method combines those 3 methods in one ,based on a *majority voting* system.
More specifically , our algorithm is as follows. For every data point in our evaluation dataset, we calculated the labels assigned by each model. For ARMA and dicrete methods, this means that if one point is classified as an anomaly by one sensor, we consider it an attack.After creating an array of all the detected attacks indices (dates) of each system, we compared their predictions and if one index (date) was present in more than one model, then it was classified as attack from our majority voting system. The results of the above method are presented below.

Evaluation Metric	Value
Accuracy	85.04
Recall	10.05
Precision	23.29
F-Measure	14.09

Table 3: Evaluation Metrics

As can be seen from the table however, our results are a bit disappointing.The combination method performs worse than the PCA itself. So using PCA is a more accurate model than the combination we implemented.Probably, this means that the effect of the other two methods is negative and we need to further optimize them.

References

[1] P. Barford, J. Kline, D. Plonka, and A. Ron. A Signal. *Analysis of Network Traffic Anomalies. In Internet Measurement Workshop, 2002.*

[2] F. Feather, D. Siewiorek, and R. Maxion. *Fault Detection in an Ethernet Network Using Anomaly Signature Matching. In ACM SIGCOMM, 1993*

[3] J. Brutlag.*Aberrant Behavior Detection in Timeseries for Network Monitoring. In USENIX Fourteenth Systems Administration Conference (LISA), 2000*

[4] B. Krishnamurthy, S. Sen, Y. Zhang, and Y. Chen. *Sketch-based Change Detection: Methods, Evaluation, and Applications. In Internet Measurement Conference, 2003*

[5] M. Roughan, T. Griffin, M. Mao, A. Greenberg, and B. Freeman. *Combining Routing and Traffic Data for Detection of IP Forwarding Anomalies (Poster). In ACM SIGMETRICS, 2004.*

[6] Jolliffe IT, Cadima J. *Principal component analysis: a review and recent developments. Philosophical transactions Series A, Mathematical, physical, and engineering sciences. 2016*

[7] R. Dunia and S. J. Qin.*Multi-dimensional Fault Diagnosis Using a Subspace Approach. In American Control Conference, 1997*

[8] R. Dunia and S. J. Qin. *A Subspace Approach to Multidimensional Fault Identification and Reconstruction. American Institute of Chemical Engineers(AIChE) Journal, pages 1813–1831, 1998*

[9] Anukool Lakhina, Mark Crovella, and Christophe Diot. *Diagnosing network-wide traffic anomalies". In: ACM SIGCOMM Computer Communication Review. Vol. 34. 4. ACM. 2004, pp. 219-230*

[10] J. E. Jackson and G. S. Mudholkar.*Control Procedures for Residuals Associated with Principal Component Analysis. Technometrics, pages 341–349, 1979*

[11] A. Lakhina, M. Crovella, and C. Diot. *Characterization of Network-Wide Anomalies in Traffic Flows. Technical Report BUCS-2004-020, Boston University, 2004*

[12] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen.*Metrics for polyphonic sound event detection". In: Applied Sciences 6.6 (2016), p. 162.*

[13] Augustin Soule, Kave Salamatian, and Nina Taft. *Combining ltering and statistical methods for anomaly detection". In: Proceedings of the 5th ACM SIGCOMM conference on Internet Measure- ment. USENIX Association. 2005, pp. 31-31.*

[14] Mei-Ling Shyu et al. *A novel anomaly detection scheme based on principal component classier. Tech. rep. DTIC Document, 2003.*