Cyber Data Analytics Assignment 2

Study:

1. https://drive.google.com/drive/folders/0Bz_wR-rAQ5aRVlR2RnducGk5TDA
2. A Dataset to Support Research in the Design of Secure Water Treatment Systems, Jonathan Goh, et al
3. Anomaly detection in Cyber Pyhsical Systems Using Recurrent Neural Network, Jonathan Goh, et al
4. Anomaly Detection: A Survey, V. Chandola et al.
5. Diagnosing Network-Wide Traffic Anomalies, A. Lakhina et al.
6. Modeling Heterogeneous Time Series Dynamics to Profile Big Sensor Data in Complex Physical Systems, Liu et al.

Familiarization task (5pt) – ½ A4

Load the SWaT **sensor data** into your favorite analysis platform (R, Matlab, Python, Weka, KNIME, …) and understand the data. Answer the following questions:

- How many different types of signals are there? What types of signals are these?
- Are the signals correlated?

Visualize these types and the presence or absence of correlation. Separate the data into training and testing. Use the data until the day of the first anomaly as training.

PCA task (5 pt) – 1 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. 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.

ARMA task (5 pt) – 1 A4

Learn an autoregressive moving average model (see Wikipedia for an introduction if unfamiliar) for each individual sensor. Most statistical packages (R, statsmodels in Python) contain standard algorithm for fitting these models from training data. Use AIC or autocorrelation plots for identifying the order and parameters of the ARMA models. For the first three sensors, study the anomalies detected. What kind of patterns do the ARMA models detect? Think of a way (study the papers) to combine the predictions of all the individual models into a single anomaly detection method. Implement it.

Comparision task (5 pt) – 1 A4

Compare the PCA and ARMA models on the SWaT dataset. 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 or ARMA models? Or do you prefer a combination of techniques?