The Analytics Edge Competition Report

Group 3
Li Xingxuan 1002189
Mao Liyan 1002190
Sun Yurou 1002202

10th December, 2018

1 Introduction

Modern water distribution systems rely on computers, sensors and actuators for both monitoring and operational purposes. This combination of physical processes and embedded systems—cyber-physical systems (CPSs), improves the level of service of water distribution networks but exposes them to the potential threats of cyber attacks. Thus, the task for the group is to develop an attack detection algorithm for a given water network following a set of rules.

2 Early Development

In order to identify the attacks in the test data set with the list of features, we firstly made attempts to fit the training data directly with models such as logistic regression models, random forest, neural networks, LSTM and gradient boosting. Apart from implementing different models, we also performed feature engineering including normalization, feature aggregation, feature selection and discretization based on our understanding of the information from the related paper. The models varied in performance, with the best accuracy at 55.6% when we added previous and post 2-day tank water levels and pressure as features for the current day and fit the data using a random forest model. The result indicates that the model only successfully detected half of the attacks, while missed the other half because it could not identify the types of attacks that did not appear in the training data. Thus, this inspired us to visualize and summarize the training data set.

3 Approach

After visualizing the training data and reading the related paper, we observed that each attack can be identified by its anomalies in the distribution of certain features including water level, pressure, etc. By comparing the distribution of

each feature between testing data and non-attack training data, we are able to identify anomalies and mark them as attack. Our approach would be to compare the features of the testing data with the distribution of the features of the non-attack training data and mark the testing data as attack if the comparison shows huge difference.

3.1 Random-sized moving window

More specifically, the model has a random-sized moving window ranging from 125 to 133 without overlapping to slice the testing data into batches. The distribution of each slice is compared with the distribution of entire empirical no-attack training data. The reason why the size of the moving window was made randomly generated within a range is that windows without overlapping cannot precisely identify the start and the end of the attacks. The method used to compare the distributions is introduced in the following subsection.

3.2 Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov test is a non-parametric test of the equality of two continuous, one-dimensional probability distributions by computing the distance between them. In this case, one is the distribution of each batch of testing data while the other is the distribution of the entire empirical no-attack training data. With the Kolmogorov-Smirnov statistic, the difference between the two distributions is quantified. The following equation computes the Kolmogorov-Smirnov statistic where $F_{1,n}$ and $F_{2,m}$ are the empirical distribution functions of the testing and training data and *sup* is the supremum function.

$$D_{n,m} = sup|F_{1,n}(x) - F_{2,m}(x)| \tag{1}$$

The null hypothesis is rejected at level α if:

$$D_{n,m} > c(\alpha)\sqrt{\frac{n+m}{nm}} \tag{2}$$

3.3 Random forest

For each iteration, the testing data set is sliced by the random-sized window into batches. Within the iteration, Kolmogorov-Smirnov Test is carried out on each batch and the entire no-attack training set, giving prediction of each data point based on the Kolmogorov-Smirnov statistic. A random forest is then built on top of the model. If more than twenty percents of the iterations show positive result for attack, we mark the data points as attack. With running the predictions for several iterations and passing through the random forest, we are able to optimize the start and end part of each attack.

3.4 Limitation

Since we are using random-sized window without overlapping to predict the testing data batch by batch, there could be a case where a very small portion of the attack data points is cut up into a batch so the Kolmogorov-Smirnov Test would show negative results for this whole batch, thus misclassifying the small portion of attack data as non-attack. In this case, although the data is True, it will be classified as False by this model. Furthermore, the selection of the significance level α is essential in this model because it determines the range within which one batch should be classified as True or False.

4 Results

The best prediction from this approach yielded the F1-score of 0.9093 which means that the model can successfully identify most of the attacks. However, the performance of the prediction might vary, as the size of the windows is randomized. This randomization may also alter some predictions at the heads and the tails of a window, resulting in the variation in the performance of the model prediction.

5 Interpretability

This approach, in practice, could be used to detect the attacks by units. With huge amount of data in normal state, the distribution of the water levels and pressure under normal state could be easily summarized. So if data units fall into abnormal distribution, it's highly possible that the system has been attacked. The basic idea of the model is to detect outliers from the normal data, which is similar to manually checking the behaviour of the system. Human would tell that the system is under attack if the system behaviour is abnormal. Inspired by this, our approach works by quantifying the difference between the two distributions and to determine the safety status of the system by statistics. Thus, the detection would be more statistically accurate. Human could understand the cause of the predictions from the difference and use the prediction as a strong support for practical decisions.