Machine Learning Engineer Nanodegree Capstone Proposal

Domain Background

Industrial control systems (ICS) are computer systems that control the operation of industrial processes and historically have been designed to be operated in isolated environments. Increasingly over time components of these systems have been integrated into larger corporate networks and connected to the Internet without a corresponding attention to security [1][2]. One domain where this has occurred is water distribution systems with the adoption of smart water technologies [3].

The Battle of the Attack Detection ALgorithms (BATADAL) was a recent competition to compare the performance of attack detection algorithms specifically against water distribution systems [4]. Intrusion detection systems are an important component in assisting in recovery as it is not possible to eliminate all attacks [5].

Problem Statement

Intrusion detection has three primary goals: to identify an attack and raise an alarm in the least amount of time, avoid issuing false alarms, and identify which components of the system have been compromised. Attacks most often take the form of context-anomalous behaviors that can usually be uncovered using techniques such as principal component analysis (PCA) to separate the data into normal anomalous components or convolutional variational auto-encoders to calculate the probability of observing the data. The detection system's performance can be evaluated by determining if an alarm raise by the system is a true alarm or a false alarm and time to detection can be approximated by measuring the false negatives. The optimal system would raise an alert in the shortest possible time and never raise a false alert.

Datasets and Inputs

There are three datasets obtained from the competition website [6]. They provide a simulated attack against a moderately sized real water distribution system. Each was generated by running extended-time hydraulic simulations with EPANET toolkit [7].

The first training dataset is one year of data (06 January 2014 to 06 January 2015) without attacks and can be used to determine the baseline of normal system operations. The second training dataset is six months of data (04 July 2016 to 25 December 2016) containing seven labeled attacks spanning 492 hourly time steps. The test set is three months of data (04 January 2017 to 01 April 2017) containing seven attacks spanning 407 hourly time steps.

Each dataset contains tabular data reporting the time stamp and the observed values from each sensor. Available readings are the water level in meters for each tank, status (binary: 0 for off/closed, 1 for on/open) and flow in liters per second for each pump and valve in the system, and the suction pressure and discharge pressure in pascals for each value and pumping station. The time step for each data set is fixed hourly intervals. The attack labels identify if the system is under attack (binary: 0 for safe, 1 for under attack).

Solution Statement

The proposed solution model is a three stage model. The first stage finds statistical outliers in the data using simple statistical tests, focusing on local anomalies affecting singe sensors. The second stage uses a multi-layer perception (MLP) to detect contextual anomalies that affect multiple sensors. The third stage uses principal component analysis (PCA) to reduce the high-dimensional sensor data into normal and anomalous conditions to detect global anomalies. By combining these three stages, the combined model will ideally be able to detect anomalies that signify an attack and raise an alert.

Benchmark Model

A benchmark model is the competition submission from Chandy et al. [8]. It uses a variational auto-encoder (VAE) based model the learns by maximizing a variational lower bound of the likelihood of the data (a higher lower bound implies a better fit to the observed data). There are two main components: an encoder that learns a low-dimensional representation of the high-dimensional data and a decoder that learns to covert the latent representation back into the original data. The VAE model is trained on the first training set to learn the normal system operation without attacks and then applied to the second training set to compute a logarithmic reconstruction probability of the data and a low value suggests a possible anomalous behavior.

Chandy et al. also needed to specify a probability threshold to classify if the system is safe or under attack. Additionally they added a rule violation model to determine if the data from the sensor readings is operationally, physically, or hydraulically possible as a further sanity check on the VAE model. The performance of the benchmark model can be computed using the confusion matrix generated by comparing the model's predicted attacks against the labeled attacks.

Evaluation Metrics

The imbalance of the classes in the datasets is important to take into account when determining evaluation metrics for the model. Attacks on the system are rare and few components are compromised in each attack. Attacks occur on 11.78% of the time steps in the second training dataset and 19.48% of the time steps in the test dataset, and only 16% of the components are compromised in the attack that compromises the most components. This means that normal metrics such as accuracy and the F1 score are distorted by the class imbalance [9]. Intrusion detection requires a balanced metric, too many false positives make the model worthless just as too many false negatives.

The Matthews correlation coefficient (MCC) will be used to evaluate the model to address this issue as it remains balanced even with classes of very different sizes [10]. It is calculated from the confusion matrix, computed using the true positives (TP), true negative (TN), false positives (FP), and false negatives (FN).

$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Possible values for MCC range from -1.0 (perfectly wrong classification) to +1.0 (perfect classification), but in practice the possible range is 0.0 (random guessing) to 1.0, as a negative MCC values means that inverting the binary classification results in a positive score. The benchmark model achieved a MCC of 0.429 on the test dataset (computed using the confusion matrix presented in [4]). It's main failing was a large number of false positives which was partially obscured by the competition's chosen metric (the mean of the true positive ratio and true negative ratio).

Project Design

The work flow begins with applying attack labels to the datasets as the competition datasets were unlabeled. Labeling allows for the computation of the evaluation metric. There are three proposed stages to the combined model and each requires certain steps. It may turn out to unnecessary to combine multiple stages, if one is all that is needed. The choice of stages and sequence of work is chosen to maximize the likelihood of the stages being complimentary and are arranged by the least amount of work to implement to the most.

The first stage classifier is simplified if the data is normalized with a mean of zero. There are three possible strategies for achieving this. First, compute the mean and standard deviation using the first training set (the baseline without any attacks) and apply that to the second training set and test sets (with attacks). Second, compute the running mean and standard deviation online using Welford's algorithm as the sequence of changes is important. Third, using some combination of the first two strategies, compute the mean and standard deviation for each hour of the day (0 is midnight, 23 is 11:00 PM) and normalize the data based on the time stamp. The normalized data should then quite clearly show local anomalies of single sensors.

The second stage classifier requires an artificial neural network, most likely a simple multi-layer perceptron. It can be trained using the second dataset and k-fold cross validation. Another, more complicated method would be to take advantage of the time series nature of the data by using a Long Short-Term Memory (LSTM) recurrent neural network architecture or a convolutional architecture.

The third stage classifier uses unsupervised learning methods to separate the observed sensor data into normal and anomalous clusters to detect global anomalies. A data point with high variation of the principal directions will be an abnormal instance [11]. Incremental principal component analysis is a good choice for online detection and an alternative approach might be a convex optimization routine as described in Mardani et al. [12].

Refrences

[1] Stouffer et al. (2013) "NIST Special Publication 800-82 Guide to Industrial Control Systems (ICS) Security" PDF (https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-82r2.pdf)

[2] Rose Tsang (2010) "Cyberthreats, Vulnerabilities and Attacks on SCADA Networks" PDF (https://web.archive.org/web/20120813015252/http://gspp.berkeley.edu/iths/Tsang_SCADA%20Attacks.pdf)

[3] Rasekh et al. (2016) "Smart Water Networks and Cyber Security." DOI: 10.1061/(ASCE)WR.1943-5452.0000646 (https://ascelibrary.org/doi/10.1061/%28ASCE%29WR.1943-5452.0000646)

[4] Taormina et al. (2018) "The Battle of the Attack Detection Algorithms: Disclosing Cyber Attacks on Water Distribution Networks" PDF (https://www.researchgate.net/profile/Stefano_Galelli/publication/323512338_THE_BATTLE_OF_THE_ATTACK_DETECTION_ALGORITHMS_DISCLOSING_CYBER_ATTACKS_ON_/links/5a991053a6fdccecff0dde81/THE-BATTLE-OF-THE-ATTACK-DETECTION-ALGORITHMS-DISCLOSING-CYBER-ATTACKS-ON-WATER-DISTRIBUTION-NETWORKS.pdf?origin=publication_detail)

[5] Anderson, R. J. (2010) Security Engineering: A Guide to Building Dependable Distributed Systems, Second Edition. ISBN-13: 978-0470068526 (https://isbnsearch.org/isbn/9780470068526)

[6] BATANAL Competition Datasets. URL: https://www.batadal.net/data.html (https://www.batadal.net/data.html)

[7] EPANET MATLAB Toolkit Github. URL: https://github.com/OpenWaterAnalytics/EPANET-Matlab-Toolkit (https://github.com/OpenWaterAnalytics/EPANET-Matlab-Toolkit)

[8] Chandy et al. (2017) "Cyberattack Detection using Deep Generative Models with Variational Inference" arXiv: 1805.12511v1 (https://arxiv.org/abs/1805.12511)

[9] David M W Powers (2007) "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation" PDF (http://www.flinders.edu.au/science_engineering/fms/School-CSEM/publications/tech_reps-research_artfcts/TRRA_2007.pdf)

[10] Boughorbel, S.B (2017). "Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric".

[11] Lee et al. (2013) "Anomaly detection via online oversampling principal component analysis." DOI: 10.1109/TKDE.2012.99 (https://doi.org/10.1109/TKDE.2012.99)

[12] Mardani et al. (2013). "Recovery of low-rank plus compressed sparse matrices with application to unveiling traffic anomalies." DOI: 10.1109/TIT.2013.2257913 (https://doi.org/10.1109/TIT.2013.2257913)