Comprehensive Exam - Question #1

Rob Gillen

Question

Select one of these 2 papers (*Anomaly detection in cyber physical systems using recurrent neural networks*[1] or *Checking is believing: Event-aware program anomaly detection in cyber-physical systems*[2]), and critique it and the work described in it. Then describe how the potential methods and measurements you want to investigate would aide in evaluating the sceptibility of their proposed approach. Then, in one of the real-world domains you want to explore, discuss how the author's proposed approach compares and contrasts to what you are proposing to do.

Anomaly detection in cyber physical systems using recurrent neural networks

Summary

In this paper, the authors present an approach to detecting anomalous traffic in a cyber-physical system. They begin by describing the problem and providing a broad overview of their approach - Long Short Term Memory based Recurrent Neural Networks (LSTM-RNN) with CUSUM for anomaly detection. They describe what they believe are their contributions and novelty: unsupervised approach based on LSTM-RNN + CUSUM, low false positive rate, the demonstration of their techniques in a crticial infrastructure context (water treament plant), and the ability to not only detect the presence of an anomaly, but also the sensor that is being tampered with. They follow this with a description of LTSM-RNN and CUSUM and the equations they used. They describe their test dataset - the Secure Water Treament Plant (SWaT) from iTrust[3], their attack scenarios, and finally their experiments and results.

The author's detection approach can be decomposed into two distinct phases or steps. First, they utilize LSTM-RNN to provide a time-series informed prediction of what a sensor's subsequent value should be. This prediction value is then compared against the actual/provided value and evaluated for suitability via CUSUM[4]. They build these models and evaluate the output of each sensor in an attempt to validate each sensor's readings. For the prediction, they appear to follow the model in [5] with the primary difference being the approach used in evaluation (CUSUM vs. probablity error above a fixed threshold).

Their results demonstrate the ability to capture the majority of the leavied attacks. The one not captured was one in which the system behaved in a normal fashion and the "attack" was assumed to be part of regular operations. They discuss a few false positives that were triggered and present as future work the expanding of their training and detection to the entire testbed (rather than just the first process stage).

Comments

This paper was an interesting read and presented a unique approach to detecting anomalous traffic on a control system network for a water treatment plant. I find it interesting that they combined a historic statisical approach (CUSUM) with a more recent learning-based approach (LSTM-RNN). I would have found

the paper more interesting had they provided a comparison of how their system performed against the CUSUM-only approach (variance is calculated relative to the mean of the data influenced by the standard deviation rather than the predictions from the LSTM-RNN). Such a comparison would have established their work as moving the state of the art *forward* rather than just an application of a new technique to a previously addressed problem.

While the authors provide significant detail as to the nature and intent of the attacks, they do not discuss their subtly (or lack thereof). Many protection systems can detect attacks wherein the evidence of the attack is clear to anyone looking. The values on the Y axis of the authors' figures (2, 3, 4) are seemingly large numbers (especially as compared to the values in Table II) and the presense of the attack should be obvious to any observer. Of more interest would be the system's ability to detect attacks that barely reached the thresholds established in Table II.

While it is more opinion than anything else, I believe that their figures (2, 3, 4) could have been re-organized to more clearly communicate their attacks and the relations between the values being displayed. I would have preferred to see the various sensors stacked vertically such that it was easier to understand the time series relationship between each sensor's values. Further, providing figures that focused on a single attack (rather than many) may have stretched the time axis (X) such that the variance in the normative (non-attack) case could be clearly seen. As presented, it appears that unless the system is under attack, there exists no variance which trivializes the task of detection.

Finally, I find their claim of novelty/contribution regarding their system's ability to detect *which sensor* is being attacked a bit hollow. Their system develops sensor-specific models/evaluation routines which, by definition, will allow you to detect which sensor is providing anomalous data. Their claim here would be more interesting if their system provided a unified evaluation/model yet still had the ability to ascertain individual sensor misbehavior.

Compare/Contrast

One aspect of my proposed research is to take a grey-box approach to assessing the "slack" or "slew" available in the anomaly detection systems protecting cyber-physical systems. In the authors' description of their experimentation setup, they state that for each sensor measurement, they defined an upper and lower bounds beyond which things would be in a "warning" or "error" state (ref: Table II). As such, my work would seek to determine these boundaries emperically. These limits, however, are not simple scalars based on measurement readings, but rather the CUSUM value representing the built-up distance, in time-sequence, of the difference of the actual value from that which was predicted. This makes the task of ascertaining the boundaries far more complex than a simple parameter sweep. Once these boundaries are obtained (or, rather, estimated), I will work to adjust the standard series of attacks to assess if they (the attacks) can be successfully executed within the confines of those boundaries. It is expected that, in some cases, the answer will be "yes", however the attack will need to be deliberately slowed or otherwise modified. I will then generate a composite score describing the overall risk to the system given the current configuration, the ability to successfully attack, and the time duration required for the attack. *NOTE: It is quite possible that under some configurations an attack may be theoretically possible yet impractical due to excessive duration (e.g. brute-forcing a password in 18.28 centuries).*

One of the testbeds that I will be working on is a microgrid[6] - a local engery grid capable of independent operation ("islanding") including localized generation and distribution as well connecting (and disconnecting) from a larger energy grid. Within this environment there are many deployed sensors and controls used to maintain proper and safe operation of the facility. A key aspect of this (and similar electrical grids) are their protection systems such as current and voltage-based circuit breakers ("protection" here is used as protecting from physical phenomena that may damage the equipment or present a saftey issue rather than protection from cyber-based attacks). In this enviornment, you can imagine an anomaly detection system such as described in the authors' paper being utilized to monitor the sensor readings from each of the current, voltage, and temperature sensors within the microgrid with the purpose of tripping the protection systems regardless

of the "official" logic when an anomaly of sufficient import is detected. In this scenario, the anomaly detection system would serve as a secondary, or out-of-band (OOB) protection system.

In my proposed work, I would begin by demonstrating an attack on the system such that, for example, the control system recieves a high current and voltage reading (falsified) resulting in a triggering of the protection activitiy (opens the circuit) and creates an outage (denial of service) when no such protection was needed. I would then have the anomaly detection system deployed, configured, and trained on the system. Running the same attack in the same fashion should trigger the anomaly detection system (based, perhaps, on the increased volume of data points submitted to the system in order to spoof or override the actual sensor). The anomaly would be detected thereby deferring the circuit opening and service would not be interrupted. I would then attempt to ascertain the boundaries of the configuration, modify the attack accordingly and attempt to reproduce the affect without being detected as anomalous. Specifically, I will work to adjust the attacks in a manner that the rate of change (a driver in the CUSUM approach) is such that the delta from the predicted value is not *so* high (or constant) that it builds up to the point of triggering detection.

> *NOTE: It may be fairly observed that the scenario described above is unrealistic as the designers of the systems would generally err on the side of safety (opening the circuit) and would never want the circuit to remain closed when the sensors' readings say they should be opened. The provided scenario could just as easily be inverted: control system sees low/normal values when, in reality, the values are high causing the circuit to remain closed and thereby damaging equipment. However, it is easier (safer, less expensive) to test the non-destructive scenario and the one establishes the validity of both.*

The methodology I propose differs from what the authors of the paper presented in that, at least according to the included figures, their attacks were hard/fast/noisy (smash-and-grab vs. low-and-slow). One of the key objectives of my work is to ascertain the susceptibility of the system to a patient attacker or one who is willing to demonstrate finesse.

References

[1] J. Goh, S. Adepu, M. Tan, and Z. S. Lee, "Anomaly detection in cyber physical systems using recurrent neural networks," in *2017 ieee 18th international symposium on high assurance systems engineering (hase)*, 2017, pp. 140–145.

[2] L. Cheng, K. Tian, D. Yao, L. Sha, and R. A. Beyah, "Checking is believing: Event-aware program anomaly detection in cyber-physical systems," *CoRR*, vol. abs/1805.00074, 2018.

[3] iTrust | Signapore University of Technology and D. (SUTD), "Secure water treatment." [Online]. Available: https://itrust.sutd.edu.sg/testbeds/secure-water-treatment-swat/.

[4] E. S. PAGE, "CONTINUOUS inspection schemes," *Biometrika*, vol. 41, nos. 1-2, pp. 100–115, 1954.

[5] P. Malhotra, L. Vig, G. Shroff, and P. Agarwal, "Long Short Term Memory Networks for Anomaly Detection in Time Series," *European Symposium on Artificial Neural Networks*, pp. 22–24, 2015.

[6] A. for the U. D. of E. Lantero, "How microgrids work." [Online]. Available: https://www.energy.gov/articles/how-microgrids-work.