NAME: **Nicole L. Huddleston**

PARENTS' NAMES: **Marta Burbano and Richard Huddleston**

HOMETOWN: **La Porte, Indiana**

CAREER OBJECTIVE**: I would like to become a Project Engineer focusing on children's toys.**

BIOGRAPHY: **I am currently about to completing my bachelor's degree in mechanical engineering technology. I enjoy helping people and I have over 400 hours of community service and volunteering. My aspiration in life is to be happy and successful enough to be able to help others do the same. My passions are soldering, reading, cooking, fixing things, and keeping a clean environment. I am happy, extraverted, and energetic with many supportive people in my life. I love animals and have a miniature schnauzer, Jojo, who could be described as a grumpy old man that I couldn't live without. Together, we enjoy going on long walks on the pier near her mother's home, going to Starbucks to get a shaken espresso and a pup cup, as well as having snuggle time. I enjoy working with my hands-on electronic sets and woodworking, as well as CAD modeling and coding. I am currently watching the *Queens Gambit* on Netflix, but my favorite series is *The Good Place*. I care about the environment and equal rights for all. I play many instruments, but my favorite is the alto saxophone. I spent 13 years playing in various jazz bands and prefer to play with others, instead of solo.**

FACULTY LSAMP SPONSOR: **Dr. Hany Abdel-Khalik, Associate Professor of Nuclear Engineering**

GOAL OF THE WORK: **During this experience, I was aiming to provide a proof-of-principle review for a new idea that can be used to bolster the cyber defenses of a wide range of industrial control systems.**

PERSONAL STATEMENT ABOUT THE LESSONS LEARNED FROM THIS EXPERIENCE: **During my experience working under Dr. Hany Abdel-Khalik for LSAMP, I learned about how important cyber security is. Without strong cyber security, anybody that learned basic hacking techniques could turn a nuclear power plant into a time bomb. I also learned that there are so many simulations that can be run to help test theories using MATLAB. Before this, I had only ever graphed basic bell-curves, and now I understand how to input big data and find discrepancies in that data. Finally, I got to learn about how nuclear power plants stay up and running, as well as the programs used to keep them safe for us, and keep them doubly-safe from outsiders with ill intent. Overall, I think working with Dr. Abdel-Khalik and LSAMP was a great experience, and I would highly recommend for students following me to consider doing the same.**

Data Mining to Save the World

By Nicole Huddleston

Abstract

Data analysts look for patterns and characteristics in the large sets of data being collected under modern electronic conditions. When unusual circumstances occur, patterns within the data can alert oversight personnel to the incident before other traditional warning signs. This work utilized the standard deviation of single signal outputs for analysis. This data mining application was specifically applied to safety within the nuclear power industry.

Keywords

data mining; hacking defense; nuclear plant safety; principal component analysis; statistical analysis

Introduction

This paper provides the author's experiences during a faculty-directed research program through the Louis Stokes Alliance for Minority Participation (LSAMP) project under the supervision of Dr. Hany Abdel-Khalik of Nuclear Engineering in the Center for Education and Research in Information Assurance and Security (CERIAS) investigating the potential of data mining to foresee certain predictive patterns within the data. Many colleges have started data science programs, because they believe in its importance for the future. Data science is creating meaning from data or can be considered the practice of learning from the data. Surprisingly enough, these programs generally start-off with statistics classes. Statistics is a set of tools for collecting data and analyzing it. It allows the prediction of the probability of certain things occurring, the distribution of outcomes, and how much those outcomes will vary from the center of the distribution. However, some statisticians are more interested in data analysis, which falls under the statistical sciences, rather than pure mathematics. Instead of creating a whole new field for that, some statisticians want to broaden the scope of what they can do to include data analysis. However, statistics already covers so much that it's near impossible to tack on another burgeoning field. Statisticians primarily work on processing data that already exists in a way that makes it helpful and easier to understand. Data

analysts create data, and then they use it to learn about how it works to apply it to different applications. Statisticians work within the present, but data analysts predict the future.

Background

Data analysis is complex and takes a long time to understand. Data science can offer more than statistics can, but it's becoming quite apparent that it takes much more study and effort to understand it. Statistics is not a large part of data science, but data scientists must have a firm grasp of it. According to Donoho (2017), "Data scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician." With this combination of skills, a data scientist can offer more than a statistician. Unfortunately, current data science degrees do not fully cover all that a data scientist should know. Becoming a data scientist requires acquiring computing and database skills, as well as experience in industry and training with the technological tools they will use. In addition, a statistics background is necessary for data analysts to be effective. As stated by Donoho (2017), true data science is "facing essential questions of a lasting nature using scientific techniques". Without the proper training, experience, and tools, it is not possible for data scientists to live up to their potential.

Modern methods of collecting data create data sets overwhelm other investigative methodologies by many orders of magnitude. In fact, these data sets are so large that they are often referred to as "Big Data". The only feasible way to work with these enormous piles of data is to create computational methods to analyze them. Computers process data much faster and in much higher quantities than humans can. These computational methods are called data mining. Data mining is a set of mathematical algorithms used to process very large amounts of data in order to derive insight from the associated system of information. The differences between data mining and statistics can be confusing, because they are both means to analyze data. Big data can be useful in both data science and statistics. Statisticians can use very large data sets for their own purposes, but data scientists use specific methods of extracting information from huge data sets to learn from the data. Effectively, data mining consists of collecting vast amounts of data and processing it in order to parse new information from the data set.

Data Mining

The goal of this project is to apply data mining algorithms to support the creation of defenses for cyber-attacks against complex industrial systems, such as a nuclear reactor. It is believed that the information gained from these algorithms could be used to save a nuclear power plant from executing problematic instructions to the operating system during a damaging cyber-attack. Conventional security is typically based

on the concept of a perimeter defense, which is essentially like building a fence around what needs protection. However, recent attacks have shown that fences can generally be breached by sophisticated attackers. The current use of perimeter defense, such as passwords and two factor IDs, can be hacked with unlimited attempts and time. In this project, a new idea was explored that attempted to find unique patterns in the data collected from an industrial system to detect intrusion. Simplified models were employed to simulate the data generated by an industrial system, and basic data mining algorithms were applied to such data to find the resulting patterns. If successful, this research will provide a proof-of-concept evaluation for a new idea that can be used to bolster the cyber defenses of a wide range of industrial control systems.

Nuclear power plants are controlled mainly by computers, because the reactions occur quickly in a highly controlled environment, and the effects are so small that they are typically not visible to the naked eye. If a hacker decided to attack a minimally secured nuclear power station and succeeded in gaining control of the plant, then that person would be able to manipulate several components of this plant. The hacker might be able to compromise the entire reactor and could overheat it by simply closing a few coolant valves or turning a pump off. This could cause an explosion big enough to poses a fatal threat to people downwind and in the local area. Because of this threat, it is no surprise that protection from cyber-attacks is a serious problem. It is difficult to detect a breach in computer systems before any harm can be done, but there is a possible solution. Data mining is being researched to help detect the intrusion of hackers into critical control systems. Development of these defenses could improve cybersecurity for many different types of industries and data in the future.

Data science is being used to help secure industrial systems from cyber-attacks. The idea is to create specific signatures for a reactor's components using operating data. These signatures are patterns in the data that can be identified using data mining. They can be represented in graphical form, so that an operator can see how the reactor looks when it is functioning correctly. Then, if a hacker starts changing anything, it will change the typical signature due to the difference in the reactor's performance. This change will show-up on the graph, immediately signaling operators about the change.

Proposed Methodology

In order to understand how a data signature is created, a process's components must be understood. The simplest metric is standard deviation, which is related to the spread of data within a large data set. The standard deviation of a data set can be used to create a bell curve or Gaussian Distribution. The variables within the code are picked to form a perfect bell curve, and the data correlation coefficient tells us how closely data in a

scatterplot falls against the predicted line. A larger standard deviation of incoming data or a decreasing correlation coefficient could potentially mean something is destabilizing a process. For complex multi-dimensional data, Principal Component Analysis (PCA) allows one to rotate view point perspectives, by moving into new display planes to see if new hyperspatial data relationships conform or are independent of earlier patterns and relationships. This provides a relative strength for a principal component. The principal components are the outliers in these new planes that could cause changes on a higher order correlation graph. Data mining is used to trace these patterns.

The procedure starts by creating the data to be analyzed. This was done by graphing the equation:

$$\frac{dy}{dt} + Ly = e^{-Bt} \text{; using y(0) = 1.0} \quad (1)$$

where:

y is the variable of interest,
t is time, and
B and L represent randomly chosen empirical values.

The ODE45 solver, a MATLAB® tool, was used to complete this task, as well as a function to make the y values output in a matrix for each attempt. The algorithm then uses a Gaussian distribution to randomly pick numbers between 0 and 1 for B and between -1 and 1 for L. These values are placed into a matrix. Afterwards, the data was plugged into PCA to find higher order principle components or correlation strengths. This analysis found four principal components. The first was much larger than the other three, indicating its significance to the data. Using data distribution, correlation coefficients, and principal component analysis in MATLAB®, an analysis of the full set of data, graphs, and strengths for the given relationship in the provided data. The data retrieved from working with this equation in MATLAB® will now be compared to work done using another program that analyzes the data differently. Changes will then be made which will be accompanied by more tests to ensure that those changes can indeed be detected.

Conclusions

Although there is much more experimentation to be done, this method of future cybersecurity appears promising. For the future, this methodology needs more testing with MATLAB® and usage of other programs such as Java Script® and R®. There also needs to be an evaluation with more data to make sure that the procedure works well with other big data, in a more complex setting, such as an actual industrial control system. Assuming this system of using data mining to detect malicious traffic in a large nuclear power plant system is successful, it could be used to stop hackers in their tracks. Instead of allowing a hacker to turn-off a coolant pump, blow-up a plant, or turn entire cities or states to a nuclear

wasteland, this technique could be used to ensure that everyone will stay safe from these possible harmful outcomes. This is a form of data mining that could save the world.

References

Donoho, D. (2017) “50 Years of Data Science”, *Journal of Computational and Graphical Statistics 26:4*, pp 745-766, doi: 10.1080/10618600 .2017.1384734