Machine Learning for Information Assurance in SCADA Systems

Steve H. Kim, Peter Bayiokos, Constanza Cabrera-Mendoza, Sabrin Kaur Guron,
Wildenski Osias, Charles C. Tappert and Avery Leider
Seidenberg School of Computer Science and Information Systems, Pace University
Pleasantville, NY 10570, USA
Email: {sk48149w, pb10842p, cc09237p, sg58867n, wo39632n, ctappert, aleider}@pace.edu

***Abstract*—With the evolution of technology and connectivity, SCADA systems have been used to manage and monitor the most critical infrastructures since the 1950s. As technology and connectivity have evolved so has SCADA systems. The connectivity of SCADA systems has presented an inherent risk of malicious security activities. With the remote monitoring of SCADA systems they are vulnerable to exposure over the internet as well as hackers and malware. If malicious security events were to unfold onto these systems it would be catastrophic for infrastructures such as power, water, telecommunication, and gas systems. One of the best ways of securing SCADA systems is through intrusion detection systems (IDSs). Machine learning techniques have been developed to vastly improve the utility of IDSs and in turn improve the security of SCADA systems when implemented. This paper will analyze three different machine learning techniques that are commonly used for intrusion detection systems. At the conclusion of our analysis we will make a recommendation on the best IDS using the most efficient and effective machine learning technique to properly secure SCADA systems.**

***Index Terms*—TBA.**

I. Introduction

Intrusion Detection Systems, (IDS), are software applications or hardware appliances that actively monitor traffic in networks and throughout technology systems to identify suspicious activity and threats. Systems are configured to send alerts to IT personnel if a network intrusion is suspected to be taking place. These systems can also be programmed to analyze traffic and identify patterns in said traffic to indicate cyber attacks.

There are two main types of IDSs: host-based and network-based. These mainly speak for where the IDS sensors have been placed in the system, (host/endpoint vs. the network). Experts in IDS further categorize them, including but not limited to, perimeter IDS, VM-based IDS, stack-based IDS, signature based IDS, and anomaly-based IDS [10].

A. Overview of Machine Learning

Coined by Arthur Samuel, a pioneer in the field of artificial intelligence, Machine Learning can be defined as giving computers the ability to act, through the use of statistical techniques and data, without them being explicitly programmed. Otherwise stated, Machine Learning "is concerned with the question of how to construct computer programs that automatically improve with experience... A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E" [11] as stated by Mitchell in his book Machine Learning. Machine Learning is subdivided in many categories but the three major ones are: Supervised Learning, Unsupervised Learning and Reinforcement Learning. Please see Figure 1 for a diagram describing Machine Learning.

- **Supervised Learning**: Simply put, supervised learning is a process in which we train the machine by using data that is well labeled—data that is tagged with the correct answer. For example, a group of pictures of apples with the tag apple on them would help the computer come up with the rules to classify pictures of apples. Supervised Learning can be, in turn, sub-categorized into classification and regression.
- **Unsupervised Learning**: Unsupervised learning refers to training the machine using non-tagged, non-labeled or non-classified data, thereby allowing the algorithm to act on the data without directions. For example, asking the computer to identify and group items that are frequently bought together on an e-commerce website involves unsupervised machine learning. Unsupervised Learning can be grouped into clustering and association.
- **Reinforcement Learning**: This type of learning focuses on taking appropriate action for the purpose of maximizing reward in a specific circumstance. Reinforcement Learning is mostly used in the search for best possible behavior or path in a certain situation. For example, a robot for accomplishing a certain task as requested.

From web search engines to photo tagging and spam detectors, Machine Learning is being used in many products and services and has become increasingly impactful in the technological realm. As Machine Learning gets closer to our everyday lives, researchers and tech enthusiasts search to better understand the opportunities and the challenges that are associated with it. In this paper, we focus on one of the many opportunities associated with Machine Learning, which is the use of "machine learning algorithms to detect malicious network traffic in SCADA systems and for other

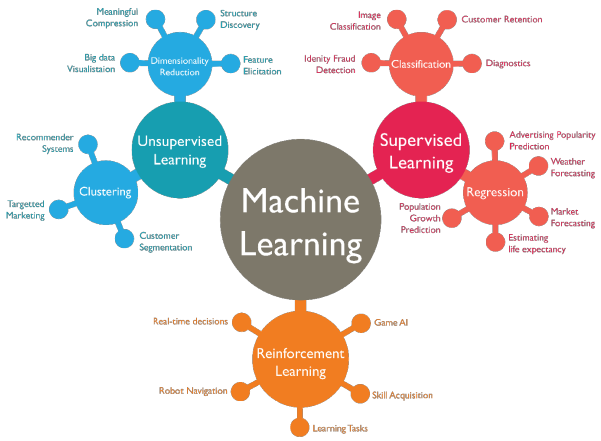

Fig. 1. Machine Learning Overview Diagram.

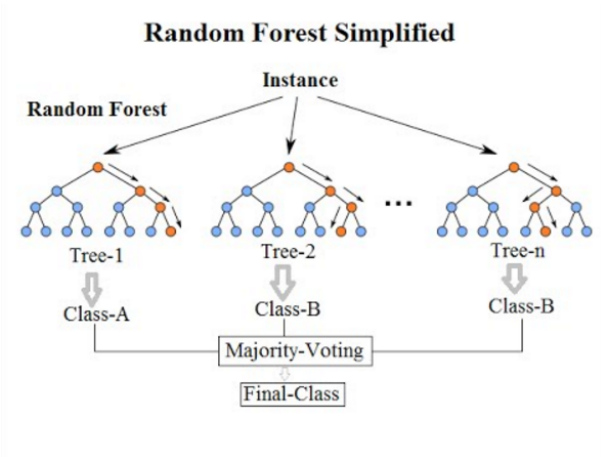

Fig. 2. Random Forest Algorithm Diagram.

intrusion and anomaly detection purposes" [12]. As a subdivision of Supervised Machine Learning, we mention the term classification. Machine Learning Classification is defined as "a process of categorizing a given set of data into classes, It can be performed on both structured or unstructured data" [13]. Classification plays a crucial in the effort to protect data from unauthorized access in that it helps classify data. Given the large amount of redundant data included in network traffic it is pivotal "to develop robust model that can classify the data with high accuracy" [14].

B. Overview of Random Forest Algorithm (RF)

Three of the best machine learning algorithms for intrusion detection systems are Random Forest (RF), Decision Table (DT), Naive Bayes (NB).A random forest algorithm is comprised of a large number of trees that all operate as a forest. Each tree splits into its own set branches of prediction. The power of the RF algorithm comes from the ability to classify different sets of data simultaneously. Each tree and it's branches are uncorrelated which isolates themselves from their own errors. For example the forest ingests a single large data set. The data set is then split into three different trees for classification and regression. One tree could be right, while the other three could be wrong. See an example of the Random Forest algorithm in the diagram below (Fig. 2).

C. Overview of Decision Table Algorithm (DT)

A decision table algorithm is a classification model used for predictions and actions. The table itself is hierarchical so that the values in the first table are broken down and used in tables below it. When you have a data set with multiple different values it's beneficial to use a decision table so that the system can understand each value and the corresponding decision. The predictions/actions that a decision table makes are conventionally based on a set of conditional values. For example the table ingests a data set, the algorithm then asks if X is being attacked, it can then offer a solution for yes or no. If yes then the table suggests additional measures, if not the table could suggest additional conditional questions [9]. See a decision table broken down in the diagram below (Fig. 3).

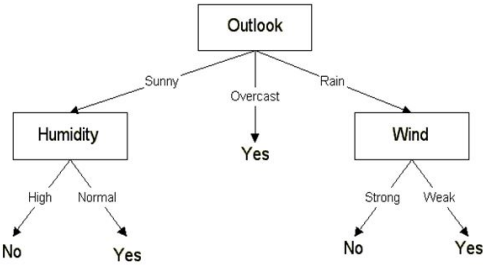

Fig. 3. Decision Table Algorithm Diagram.

D. Overview of Naive Bayes Algorithm (NB)

The Naive Bayes algorithm in machine learning is used primarily for classification and prediction. NB is based off of Bayes' Theorem. Bayes' Theorem lets you gain the probability of an event based on prior knowledge of any event that is related to the former event. For example, the probability that the price of house A is high. However, we can make a better assessment of house A if we know the neighborhood around it. Then we take the assessment of the house that was made without the knowledge of the neighborhood around it. With those three inputs we have the probability. Naive Bayes algorithm "dumbs down" Bayes Theorem, hence the name naive. In NB you have X, Y and Z. All three variables are independent only if the probability governing X is independent of the value Y given Z. The variables basically should not provide any information on the likelihood of them occurring [4]. Please see the equation written out below (Fig. 4).

E. Overview of SCADA Systems

Supervisory Control and Data Acquisition Systems, (SCADA Systems), are collections of hardware and software components used to supervise and control plants, either from

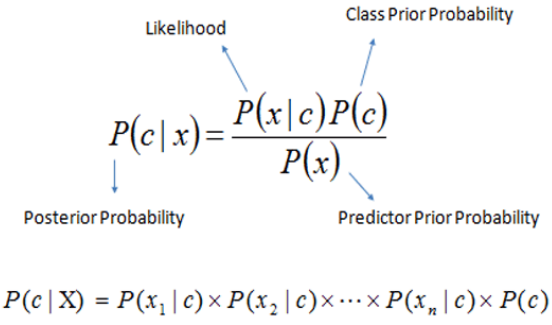

Fig. 4. Decision Table Algorithm Diagram.

a local or a remote location, through examination, collection, and process of data in real time. These SCADA Systems work hand in hand with Human Machine Interface software (HMI), to facilitate indirect user control of hardware in the plants. Remote Terminal Units (RTUs) and Programmable Logic Controllers (PLCs) further enhance the ability of the user to indirectly analyze and respond to plant events. SCADA was introduced to lower the need for onsite personnel at plants. Before, personnel were required on a 24 hr schedule in order to monitor and maintain plant machinery. SCADA eliminated this need by giving personnel the ability to remotely control and supervise plant processes. As industrial plants grew, a larger importance was placed on automation and reliable SCADA systems; focus and money was thrown into this field and it boomed into the subtle yet extremely important industry it is today.

The initial breaths of SCADA systems began back in the 1950s, where processors were first being used to automate control of machinery at plants. The 1960s provided telemetry, (the remote record and transmission of instrument readings), and further facilitated the rise of SCADA systems. SCADA was officially recognized in the 1970s, but was highly inefficient. The SCADA systems were comparable to standalone mainframes, large and cumbersome to manage and maintain. The '80s and '90s saw SCADA system development through the inclusion of Local Area Networking, but still didn't connect efficiently over longer distances because of needed wiring. These wire relying systems were named distributed SCADA systems.

The most impactful change these systems saw in the late '90s and early 2000s was the introduction of open system architectures, which let the systems become more easily accessible by vendors. These Networked SCADA systems almost became obsolete once SQL Databases were introduced. SQL Databases were not anticipated to be combined with SCADA systems, pushing them into a "past technology" category rather quickly. Once these SQL Databases were adapted and incorporated, along with web-based applications, SCADA Systems became what they are today: a staple of various, if not all, plant industries. Present day operator interactions now include real time facilitated responses to SCADA system queues based on field collected data and system analysis from almost anywhere in the world. This system-wide improvement also incorporated trend analysis, company mandated record keeping, plant process automation, and an immense decrease in required personnel. [3] Today, the most common applications of SCADA Systems include the following industries and plants: telecommunications, water and waste control, energy, oil and gas refining and transportation.

II. Literature Review

Information assurance has increasingly become much involved in the supervisory control and data acquisition (SCADA) systems. With the further use of this system, there brings a growing number of security threats and vulnerabilities. "Security Issues in SCADA Networks" highlights the alarming issues present in SCADA networks and the challenges that need to be tackled in order to improve these systems. This paper states that the reason behind the lack of security in organization's that implement SCADA networks is due to the use of commercial off-the-shelf (COTS) hardware and software to develop devices for operating in SCADA networks. With the use of this equipment, the further development of SCADA protocols are needed to create a more secure environment. Even though this paper highlights the security issues SCADA systems hold, there is an increasing effort placed on the information assurance of this system. Ensuring information assurance has become a center-piece to many network frameworks and it allows for sustainability in security for those organizations that incorporate SCADA systems. In addition, there is generally a low involvement in personnel in organizations that have SCADA systems, therefore, the rise in intrusion detection systems (IDS) are becoming more present in these organizations as a result [1].

The paper "Sustainable Security for Infrastructure SCADA" emphasizes the various security structures set in place in SCADA systems and how these systems are built in order to keep the organization's information secure. The paper states there are three elements in sustainable security in SCADA systems: the first is to secure implementations of technology and procedures managed by effective security administration including enforcement and audit; the second is better security technology, including SCADA-specific capabilities; and finally, third party assessment of administration and implementation. SCADA systems influence all tiers of an organization and each level of an organization has some sort of effect on each other. In a SCADA system, the IT control framework has an influence on the security policy, the security policy affects the security plan, and finally the security plan affects the implementation guidance of the security plan. An emphasized security policy, creates for more opportunity to ensure the information assets in the organization are secure. SCADA and IDS systems are relying less on personnel and as a result, the growth in creating a reliable information assurance system is being incorporated into the organizations' security policies. The SCADA security policy needs to have a reliable control

objective for the organization to ensure a secure SCADA design, implementation, and operation [2].

III. Project Requirements

For this project we will be identifying and analyzing three different machine learning algorithms that are traditionally used in intrusion detection systems. These IDSs use machine learning techniques to adaptively secure SCADA systems with little to no computer training needed. Additionally we will analyze each machine learning theory individually to mark it's efficiency when learning new attack threats. The analysis will be done using python and Scikit-learn. Scikit-learn is one of the leading machine learning tools that is designed to interpret different algorithms using data sets and python.

The three machine learning algorithms that will be analyzed are as follows: Random Forest (RF), Decision Table (DT), and Naive Bayes (NB). These three algorithms have been proven to provide decisions based on data classification, probabilistic classifiers, and conditional decision making. Through this analysis we will make a determination what is the most efficient and effective intrusion detection system using machine learning techniques, that will assist in securing SCADA systems.

References

[1] V. M. Igure, S. A. Laughter, and R. D. Williams, "Security issues in scada networks," *computers & security*, vol. 25, no. 7, pp. 498–506, 2006.

[2] J. Stamp, P. Campbell, J. DePoy, J. Dillinger, and W. Young, "Sustainable security for infrastructure scada," *Sandia National Laboratories, Albuquerque, New Mexico (www. sandia. gov/scada/documents/SustainableSec urity. pdf)*, 2003.

[3] *What is SCADA?* YouTube, Jun 2019. [Online]. Available: https://www.youtube.com/watch?v=nlFM1q9QPJw

[4] P. Gupta, "Naive bayes in machine learning," Nov 2017. [Online]. Available: https://towardsdatascience.com/naive-bayes-in-machine-learning-f49cc8f831b4

[5] B. Miller and D. Rowe, "A survey scada of and critical infrastructure incidents," in *Proceedings of the 1st Annual conference on Research in information technology*, 2012, pp. 51–56.

[6] M. Hentea, "Improving security for scada control systems," *Interdisciplinary Journal of Information, Knowledge, and Management*, vol. 3, no. 1, pp. 73–86, 2008.

[7] E. J. Byres, M. Franz, and D. Miller, "The use of attack trees in assessing vulnerabilities in scada systems," in *Proceedings of the international infrastructure survivability workshop*. Citeseer, 2004, pp. 3–10.

[8] C. Davis, J. Tate, H. Okhravi, C. Grier, T. Overbye, and D. Nicol, "Scada cyber security testbed development," in *2006 38th North American Power Symposium*. IEEE, 2006, pp. 483–488.

[9] B. G. Becker, "Visualizing decision table classifiers," in *Proceedings IEEE Symposium on Information Visualization (Cat. No. 98TB100258)*. IEEE, 1998, pp. 102–105.

[10] "Guide to idps." [Online]. Available: https://www.esecurityplanet.com/products/top-intrusion-detection-prevention-systems.html

[11] T. M. Mitchell, *Machine Learning*. McGraw-Hill, 1997.

[12] A. AYODEJI, "Machine learning approach to industrial control system health monitoring and cyber security: Similarities, conflicts and limitations."

[13] "Classification in machine learning: Classification algorithms," Dec 2019. [Online]. Available: https://www.edureka.co/blog/classification-in-machine-learning/

[14] D. P. Mohapatra and S. Patnaik, *Intelligent Computing, Networking, and Informatics Proceedings of the International Conference on Advanced Computing, Networking, and Informatics, India, June 2013*. Springer India, 2014.