MISSISSIPPI STATE UNIVERSITY
PROJECT REPORT - SCADA ANOMALY DETECTION

Project Summary

Project Title: SCADA Anomaly Detection

Project Date: February 17th, 2014 – April 30, 2014

Project Team: Jeff Hsu - Computer Engineer - jvh52@msstate.edu
David Mudd - Computer Engineer - dbm157@msstate.edu
Zach Thornton - Computer Engineer - jzt3@msstate.edu

Key Words: **SCADA** - Supervisory Control And Data Acquisition
Anomaly - A deviation from the common rule[1]
ICS - Industrial Control System
PLC - Programmable Logic Controller
Algorithm - A step-by-step procedure for solving a problem or accomplishing some end, especially by a computer[1]
Machine Learning Algorithm - An algorithm designed for the construction and study of systems that can learn from data[3]
Weka - A software implementation of a collection of machine learning algorithm for data mining tasks developed, by the University of Waikato in New Zealand[2]

Project Description: This project aims to examine existing machine learning algorithms, develop criterion for algorithm selection, and to examine the effectiveness of machine learning algorithms in detecting anomalous SCADA transactions.

Task Delegation: Jeff Hsu - Electric Power Transmission Data Analysis
David Mudd - Water Storage Tower Data Analysis
Zach Thornton - Gas Pipeline Data Analysis

Motivation

Viruses that are designed to attack SCADA systems, such as Stuxnet, have given rise to many doubts about the level of security from cyber attacks in critical infrastructure. Society is highly dependent upon the standard operations of critical infrastructure. The security of these SCADA systems is paramount. Discerning between normal transactions and anomalies is of utmost importance.

There are many proposed algorithms available for use with anomaly detection based intrusion detection systems. Choosing an appropriate algorithm for use with SCADA systems requires multiple steps. First, criteria for algorithm selection system must be developed. Second, algorithms which meet minimum selection criteria should be compared. Finally, the strengths and weaknesses of each algorithm should be discussed.

This project aimed develop criterion for algorithm selection, and examine the effectiveness of several machine learning algorithms in detecting anomalous SCADA transactions. In addition to examining the algorithms, this project examined the training datasets to determine their usefulness in SCADA anomaly detection. Training datasets which include normal and cyber attack data logs were used from a laboratory scale gas pipeline, water storage tank, and electric transmission protection system.

Methods

1. Algorithm Criterion & Algorithm Selection

The first step of this project was to develop criterion for algorithm selection. In order to do this, a list of potential algorithms had to be developed, an implementation scheme chosen, and datasets chosen.

Because MSU's Dr. Thomas Morris had ready available datasets, and due to the untested nature of these datasets, 3 datasets were chosen for use in this examination. These included a dataset from a laboratory scale gas pipeline, a lab scale water tower, and a lab scale electric transmission system. All 3 of these datasets contained preprocessed network transaction data, preprocessed to strip lower layer transmission data(TCP, MAC, etc). The number of entries in the datasets ranged from 100,000 for the gas data, to 200,000 for the water data, to 5,000,000 for the electric data. The datasets included 24 unique parameters for the water data, 27 for the gas data, and 132 for the electric data.

The result parameter for the water and gas datasets categorized each entry into 1 of 7 attack vectors. The full list of parameters and attack vectors for the water and gas datasets is shown below in Tables 1 and 2.

Attack Name	**Abbreviation**
Normal	Normal(0)
Naïve Malicious Reponse Injection	NMRI(1)
Complex Malicious Response Injection	CMRI(2)
Malicous State Command Injection	MSCI(3)
Malicous Parameter Command Injection	MPCI(4)
Malicous Function Code Injection	MFCI(5)
Denial Of Service	DOS(6)

Reconnaisance	Recon(7)

Table 1: Attacks

Gas Parameters	Water Parameters
command address	command address
response address	response address
command memory	command memory
response memory	response memory
command_memory_count	command_memory_count
response_memory_count	response_memory_count
comm_read_function	comm_read_function
comm_write_fun	comm_write_fun
resp_read_fun	resp_read_fun
resp_write_fun	resp_write_fun
sub_function	sub_function
command_length	command_length
resp_length	resp_length
gain	HH
resest	HH
deadband	L
cycletime	LL
rate	control_mode
setpoint	control_scheme
control_mode	pump
control_scheme	crc_rate
pump	measurement
solenoid	time
crc_rate	result
measurement	
time	
result	

Table 2: Gas and Water Parameters

The "marker" parameter for the electric dataset identifies each entry as belonging to one of 40 scenarios, 26 of which are attacks. These in turn belong to 7 essential categories of behavior. The behavior categories and parameters are shown in Tables 3 and 4, respectively.

Category	Type	Num Scenarios
Primary protection properly working	Normal	6
Fault replay	Attack	6
Line maintenance	Normal	2
Command injection (one relay)	Attack	4
Command injection (two relays)	Attack	2

Primary protection disabled (one relay)	Attack	14
Primary protection disabled (two relay)	Attack	6

Table 3: Categories of Behavior for Electric Data

Network/Other	Relay 1	Relay 2	Relay 3	Relay 4
Date	R1-PA1:VH	R2-PA1:VH	R3-PA1:VH	R4-PA1:VH
Timestamp	R1-PM1:V	R2-PM1:V	R3-PM1:V	R4-PM1:V
control_panel_log1	R1-PA2:VH	R2-PA2:VH	R3-PA2:VH	R4-PA2:VH
control_panel_log2	R1-PM2:V	R2-PM2:V	R3-PM2:V	R4-PM2:V
control_panel_log3	R1-PA3:VH	R2-PA3:VH	R3-PA3:VH	R4-PA3:VH
control_panel_log4	R1-PM3:V	R2-PM3:V	R3-PM3:V	R4-PM3:V
relay1_log	R1-PA4:IH	R2-PA4:IH	R3-PA4:IH	R4-PA4:IH
relay2_log	R1-PM4:I	R2-PM4:I	R3-PM4:I	R4-PM4:I
relay3_log	R1-PA5:IH	R2-PA5:IH	R3-PA5:IH	R4-PA5:IH
relay4_log	R1-PM5:I	R2-PM5:I	R3-PM5:I	R4-PM5:I
snort_log1	R1-PA6:IH	R2-PA6:IH	R3-PA6:IH	R4-PA6:IH
snort_log2	R1-PM6:I	R2-PM6:I	R3-PM6:I	R4-PM6:I
snort_log3	R1-PA7:VH	R2-PA7:VH	R3-PA7:VH	R4-PA7:VH
snort_log4	R1-PM7:V	R2-PM7:V	R3-PM7:V	R4-PM7:V
marker	R1-PA8:VH	R2-PA8:VH	R3-PA8:VH	R4-PA8:VH
fault_loc	R1-PM8:V	R2-PM8:V	R3-PM8:V	R4-PM8:V
load_con	R1-PA9:VH	R2-PA9:VH	R3-PA9:VH	R4-PA9:VH
	R1-PM9:V	R2-PM9:V	R3-PM9:V	R4-PM9:V
	R1-PA10:IH	R2-PA10:IH	R3-PA10:IH	R4-PA10:IH
	R1-PM10:I	R2-PM10:I	R3-PM10:I	R4-PM10:I
	R1-PA11:IH	R2-PA11:IH	R3-PA11:IH	R4-PA11:IH
	R1-PM11:I	R2-PM11:I	R3-PM11:I	R4-PM11:I
	R1-PA12:IH	R2-PA12:IH	R3-PA12:IH	R4-PA12:IH
	R1-PM12:I	R2-PM12:I	R3-PM12:I	R4-PM12:I
	R1:F	R2:F	R3:F	R4:F
	R1:DF	R2:DF	R3:DF	R4:DF
	R1-PA:Z	R2-PA:Z	R3-PA:Z	R4-PA:Z
	R1-PA:ZH	R2-PA:ZH	R3-PA:ZH	R4-PA:ZH
	R1:S	R2:S	R3:S	R4:S

Table 4: Parameters for Electric Data

Note that the electric data differs fundamentally in that it is sequential in nature - in addition to including network transaction data, it consists largely of sensor measurements which have been sampled at a rate of 120 times per second. Thus, each instance of a scenario may be represented by thousands of data entries rather than a single entry as in the case of the water and gas data sets. As a result, the approach used for analysis differs somewhat from that used for the water and gas data sets, as shall be reported.

The implementation scheme chosen was the University of Waikato's WEKA software. This software includes 96 different machine learning algorithms, implemented with a graphical user interface for selecting the algorithm, the input data, the parameters to be used, the results, and other useful information. This software was chosen because of its ease of use, availability, and easily accessible literature and documentation.

The algorithms chosen for initial analysis were chosen based on research into similar applications using machine learning algorithms(see previous work), as well as their availability in WEKA. The full list of 35 algorithms is given as below Table 5.

Algorithm	**Category**
Best First Decision Tree(BFTree)	Decision Tree
Decision Stump	Decision Tree
FaultTree(FT)	Decision Tree
J48 Decision Tree	Decision Tree
J48Graft Decision Tree	Decision Tree
Logiboost Alternating Decision Tree(LADTree)	Decision Tree
Logistic Model Tree(LMT)	Decision Tree
RandomErrorPruning Tree(REPTree)	Decision Tree
RandomForrest	Decision Tree
RandomTree	Decision Tree
SimpleCart	Decision Tree
Naïve Bayes Tree(NBTree)	Decision Tree
Radial Basis Function Network(RBFNetwork)	Nerual Network
Multilayer Perceptron	Nerual Network
Logistic Regression	Regression
SimpleLogistic	Regression
Sequential Minimal Optimization(SMO)	Support Vector Machine
ConjunctiveRule	Rule Based
DecisionTable	Rule Based
DTNB	Rule Based
Jrip	Rule Based
Nnge	Rule Based
OneR	Rule Based
PART	Rule Based
Ridor	Rule Based
ZeroR	Rule Based
BayesNet	Bayes
ComplementNaiveBayes	Bayes
DMNBtext	Bayes
NaiveBayes	Bayes
NavieBayesMultinomial	Bayes

NaiveBayesMultinomialUpdateable	Bayes
NaiveBayesSimple	Bayes
NaiveBayesUpdateable	Bayes

Table 5: Initial Algorithm List

In order to determine the viability of each of these algorithms, each algorithm was run in WEKA with a 10% subset of the training data. This method was chosen because for most of the 35 algorithms, tests with the full dataset was time prohibitive, whereas running with a 10% subset was much more time efficient. In addition, due to ignorance of the algorithms specific workings, this method seemed to reveal the effectiveness of the algorithms for the MSU datasets, without requiring detailed knowledge of the algorithms.

The results of these tests from the gas and water datasets are given as Table 9 in the Results section. From this, 7 algorithms were chosen for further study. Additionally, the tests from the 10% subset were repeated with the full dataset for the 7 selected algorithms. These results were compared to determine whether using a 10% subset is a legitimate method for algorithm criterion.

As the electric dataset is an order of magnitude larger than the water and gas datasets, running the entire dataset in WEKA was not possible, due to program memory constraints. Thus, a 10% subset of one of the 10 constituent datasets was used for preliminary analysis. This subset was run with 19 algorithms, the results of which are given in Table 10 in the Results section. From this, 3 algorithms were chosen for further study.

2. Algorithm and Dataset Analysis

2.1 Gas and Water Data

After the 7 algorithms were chosen for further study, the effectiveness of the algorithms was called into question. In order to determine which parameters were the most useful to the algorithms in determining an attack, the number of parameters was reduced from the full parameter set to a minimal parameter set with the same results as the full dataset.

This was done first by removing non-changing parameters. From here, each remaining parameter was removed 1 at a time, and a test performed without that parameter. This was done to determine the effect of that parameter on the algorithm's performance. After removing all parameters whose effect was negligible, a reduced parameter set was determined. Tables 6 and 7 below give the reduced parameter set, and which attack vector detection requires that parameter for the gas pipeline.

Parameter	**Abbreviation**
command_address	CA
resp_address	RA
resp_length	RL
com_read_fun	CRF
resp_read_fun	RRF
subfunction	SF

setpoint	SP
control_mode	CM
control_scheme	CS
Measurement	M

Table 6: Reduced Parameter Set

Algorithm	CA	RA	RL	CRF	RRF	SF	SP	CM	CS	M
J48Graft Decision Tree	DOS	N/A	N/A	DOS	N/A	MFCI	MPCI	MSCI	MSCI	NMRI, CMRI, Recon
Logistic Regression	DOS	Recon	N/A	DOS	CMRI	MFCI	MPCI	Normal	MSCI	N/A
Multilayer Perceptron	DOS	N/A	Recon	DOS	CMRI	MFCI	MPCI	MFCI	MSCI	N/A
RandomErrorPruningTree (REPTree)	DOS	N/A	N/A	DOS	N/A	MFCI	MPCI	MSCI	MSCI	NMRI, CMRI, Recon

Table 7: Reduced Parameter Set Attack Vectors

After this reduced parameter set was discovered, an investigation began into the relationship between each parameter and the result parameter. The result of this investigation for the gas pipeline is given as Table 11 in the Results section below.

2.2 Electric Data

After the 3 algorithms were chosen for further study, a similar approach of parameter reduction was taken. To establish some leads on what to select, the parameters were analyzed using the InfoGain evaluator within WEKA. The results of this evaluation is given in Table 12 in the Results section. Tests on the 3 algorithms were then run with certain parameters stripped based on information gain. The results of these tests are given as Table 13 in the Results section. It was then decided that a separate dataset should be generated by isolating two scenarios (a normal fault and fault replay) and extracting every instance from all 10 constituent datasets. Tests with one of the selected algorithms were then performed. The results of this are contained in Table 14 in the Results section.

Results

1. Algorithm Criterion & Algorithm Selection

1.1 Gas and Water Data

The first noteworthy result is that the effectiveness of the algorithms with the 10% subset did prove a legitimate criterion for algorithm selection. After running 7 of the algorithms with both the 10% subset and the full dataset, the results proved very similar. A comparison of the performance of both for 4 of the algorithms using the gas data is given as Table 8 below.

Algorithm	Normal(0)	NMRI(1)	CMRI(2)	MSCI(3)	MPCI(4)	MFCI(5)	DOS(6)	Recon(7)
J48Graft Decision Tree(100%)	100%	94%	100%	95%	98%	96%	97%	100%
J48Graft Decision Tree (10%)	100%	95%	100%	89%	99%	68%	88%	100%
Logistic Regression(100%)	98%	1%	99%	95%	98%	96%	71%	100%
Logistic Regression(10%)	100%	4%	99%	93%	99%	95%	67%	100%
Multilayer Perceptron(100%)	98%	3%	99%	95%	98%	96%	77%	100%
Multilayer Perceptron(10%)	98%	2%	99%	93%	99%	95%	68%	100%
RandomErrorPruning Tree(100%)	100%	98%	100%	95%	98%	96%	97%	100%
RandomErrorPruning Tree(10%)	100%	95%	100%	90%	99%	95%	95%	100%

Table 8: 10% Vs. 100%

The second noteworthy result is that almost all of the algorithms examined performed very well, even with the 10% subset. The results of these tests from the gas and water datasets are given as Table 9 below

Algorithm	Normal(0)	NMRI(1)	CMRI(2)	MSCI(3)	MPCI(4)	MFCI(5)	DOS(6)	Recon(7)
Best First Decision Tree(BFTree)	100%	97%	99%	87%	99%	95%	95%	100%
Decision Stump	99%	0%	0%	0%	0%	0%	0%	100%
FaultTree(FT)	100%	94%	100%	93%	99%	95%	96%	100%
J48 Decision Tree	100%	95%	100%	90%	99%	73%	91%	100%
J48Graft Decision Tree	100%	95%	100%	89%	99%	68%	88%	100%
Logiboost Alternating Decision Tree(LADTree)	100%	94%	99%	93%	99%	0%	73%	100%
Logistic Model Tree(LMT)	100%	86%	100%	93%	99%	95%	93%	100%
Logistic Regression	100%	4%	99%	93%	99%	95%	67%	100%
Multilayer Perceptron	98%	2%	99%	93%	99%	95%	68%	100%
Naïve Bayes Tree(NBTree)	100%	96%	99%	93%	98%	95%	95%	100%
Radial Basis Function Network(RBFNetwork)	98%	1%	99%	93%	99%	95%	88%	100%
RandomErrorPruning Tree(REPTree)	100%	95%	100%	90%	99%	95%	95%	100%
RandomForrest	100%	96%	100%	90%	99%	93%	93%	100%
RandomTree	99%	96%	100%	90%	98%	81%	91%	100%
SimpleCart	100%	96%	100%	88%	99%	95%	94%	100%
SimpleLogistic	98%	36%	99%	93%	99%	95%	68%	100%
Sequential Minimal Optimization(SMO)	98%	1%	99%	93%	99%	73%	44%	100%
BayesNet	98%	98%	95%	97%	100%	100%	99%	100%
ComplementNaiveBayes	100%	0%	0%	0%	29%	100%	0%	100%
DMNBtext	100%	0%	0%	0%	44%	100%	0%	100%
NaiveBayes	43%	0%	99%	97%	99%	100%	96%	100%
NavieBayesMultinomial	100%	0%	0%	97%	81%	100%	39%	100%
NaiveBayesMultinomialUpdateable	100%	0%	0%	97%	81%	100%	39%	100%

NaiveBayesSimple	0%	95%	98%	56%	62%	0%	0%	2%
NaiveBayesUpdateable	43%	0%	99%	97%	99%	100%	96%	100%
ConjunctiveRule	100%	0%	0%	0%	0%	0%	0%	100%
DecisionTable	98%	98%	95%	94%	98%	95%	91%	100%
DTNB	98%	98%	95%	97%	99%	100%	100%	100%
Jrip	99%	98%	94%	97%	99%	100%	100%	100%
Nnge	97%	97%	75%	97%	99%	100%	97%	100%
OneR	97%	98%	95%	0%	0%	0%	0%	100%
PART	99%	98%	95%	97%	99%	100%	100%	100%
Ridor	99%	98%	94%	97%	99%	100%	100%	100%
ZeroR	0%	0%	0%	0%	0%	0%	0%	0%

Table 9: Results From 10% Test of Water & Gas Datasets

1.2 Electric Data

Using the 10% dataset within WEKA resulted in suspiciously high performance for many algorithms. The test results are given as Table 10 below.

Algorithm	Category	Overall Accuracy	Cross-Validation
BayesNet	Bayes	99.2%	10
DMNBtext	Bayes	75.9%	10
NaïveBayes	Bayes	98.5%	10
NaiveBayesUpdateable	Bayes	98.5%	10
Logistic	Regression	100.0%	10
Multilayer Perceptron	NeuralNet	100.0%	2
RBFNetwork	NeuralNet	98.7%	10
Conjunctive Rule	Rule Based	23.4%	10
Jrip	Rule Based	99.9%	10
OneR	Rule Based	100.0%	10
PART	Rule Based	100.0%	10
ZeroR	Rule Based	12.6%	10
DecisionStump	Decision Trees	23.4%	10
J48	Decision Trees	100.0%	10
J48graft	Decision Trees	100.0%	10
RandomForest	Decision Trees	100.0%	10
RandomTree	Decision Trees	99.9%	10
REPTree	Decision Trees	100.0%	10
SimpleCart	Decision Trees	100.0%	10

Table 10: Results from 10% Test of Electric Dataset

Note: The default 10-fold cross-validation was used for all algorithms except Multilayer Perceptron, which used 2-fold cross-validation due to time constraints.

2. Algorithm and Dataset Analysis

2.1 Gas and Water Data

The effectiveness of the algorithms in determining an attack, as shown in Tables 4 and 5, gave rise to questions about these numbers. It was not expected that so many of these algorithms should perform so well. This was not expected because many of the attacks, such as Complex Malicious Response Injection, should resemble Normal network traffic in most ways. However, the detection of CMRI attacks ranged from 94% to 99% in most of the examined algorithms.

It was in response to these questions that the usefulness of the dataset itself was called into question. If obvious trends could be found in the dataset, and these trends could be shown to be avoidable and due to human error in their creation, the datasets could be shown to be not useful for IDS research.

In order to further examine this, the set of input parameters was reduced in order to find a minimal parameter set for the gas and water data. After a minimal parameter set was found, each of the parameters in this minimal set was examined and a strong correlation was found between each of these parameters and an attack to be predicted. All of these correlations were due to human error and were avoidable. These correlations are shown in below Table 11.

command_address
Always 4, unless DOS attack
response_address
only 0 when Recon attack
response_length
always 19 unless Recon attack
comm_read_function
always 3 unless DOS attack
resp_read_fun
only 1 when normal or CMRI attack
subfunction
always 0 unless MFCI attack
setpoint
always 20 unless MPCI attack
control_mode
only 1 when MSCI
control scheme
only 0 when MSCI
Measurement
All CMRIs in range 6-11
all NMRIs grossly out of bounds

Table 11: Reduced Parameter Set Vs. Result

2.2 Electric Data

As shown in Table 10, many of the algorithms have exceptionally high accuracies. Given the wide array of scenarios and the discrete nature of some of the attacks, such high accuracies are not expected. After selecting 3 algorithms from the set of 19, focus was shifted to the reduction of parameters. Parameter selection was guided by the results of an attribute information gain evaluation (InfoGain within WEKA), which are shown in Table 12.

Parameter	Ranking	Parameter	Ranking	Parameter	Ranking	Parameter	Ranking
load_con	3.811	R2-PA10:IH	3.152	R3-PM3:V	2.090	R1:DF	0.066
Timestamp	3.811	R1-PA4:IH	3.149	R1-PA:Z	2.081	relay1_log	0.057
Date	3.811	R4-PM5:I	3.146	R1-PM1:V	1.759	R3:DF	0.057
R4-PA1:VH	3.473	R2-PA4:IH	3.134	R1-PM3:V	1.749	R1-PM9:V	0.052
R4-PA7:VH	3.469	R3-PM5:I	3.056	R1-PM7:V	1.744	R3-PA9:VH	0.052
R4-PA2:VH	3.464	R2-PM5:I	3.011	R1-PM2:V	1.568	R2-PA9:VH	0.052
R4-PA3:VH	3.461	R1-PM5:I	2.883	R4-PA:ZH	1.242	R1-PM8:V	0.052
R2-PA1:VH	3.423	R4-PM10:I	2.820	R3-PA11:IH	1.224	R1-PA9:VH	0.052
R2-PA3:VH	3.421	R2-PM10:I	2.726	R2-PA11:IH	1.223	R1-PA8:VH	0.050
R2-PA2:VH	3.421	R4-PM4:I	2.719	R1-PA11:IH	1.221	R2-PA8:VH	0.050
R2-PA7:VH	3.420	R3-PM10:I	2.664	R4-PA11:IH	1.218	R3-PA8:VH	0.050
R3-PA3:VH	3.419	R1-PM10:I	2.661	R4-PA12:IH	1.191	R2-PM9:V	0.044
R3-PA7:VH	3.417	R2-PM4:I	2.636	R3-PA12:IH	1.170	R3-PM9:V	0.043
R3-PA1:VH	3.416	R4-PM6:I	2.602	R2-PA12:IH	1.153	R4-PA9:VH	0.043
R3-PA2:VH	3.416	R2-PM6:I	2.585	R2-PA:ZH	1.128	R4-PA8:VH	0.043
R1-PA1:VH	3.398	R3-PM4:I	2.571	R1-PA12:IH	1.082	R2-PM8:V	0.042
R1-PA7:VH	3.396	R1-PM4:I	2.553	R1-PA:ZH	1.056	R3-PM8:V	0.042
R1-PA2:VH	3.393	R3-PM6:I	2.518	R3-PA:ZH	1.035	R4-PM8:V	0.042
R1-PA3:VH	3.392	R1-PM6:I	2.511	R4-PM12:I	0.519	R4-PM9:V	0.040
R4-PA5:IH	3.387	R2-PM1:V	2.469	R2-PM12:I	0.509	relay4_log	0.038
R3-PA5:IH	3.384	R2-PM7:V	2.451	R2-PM11:I	0.497	R1:S	0.038
R4-PA6:IH	3.310	R2-PM3:V	2.424	R4-PM11:I	0.495	R4:S	0.036
R4-PA10:IH	3.305	R2-PM2:V	2.414	R2:F	0.450	relay3_log	0.036
R4-PA4:IH	3.286	R4-PM7:V	2.311	R4:F	0.445	R3:S	0.019
R3-PA10:IH	3.273	R4-PM1:V	2.299	R4:DF	0.240	snort_log4	0.000
R3-PA6:IH	3.268	R4-PM3:V	2.257	R2:DF	0.221	snort_log1	0.000
R3-PA4:IH	3.239	R4-PM2:V	2.236	R3-PM12:I	0.170	snort_log2	0.000
R2-PA5:IH	3.232	R4-PA:Z	2.212	R1-PM11:I	0.161	control_panel_log4	0.000
R1-PA5:IH	3.224	R3-PM2:V	2.207	R3-PM11:I	0.157	control_panel_log3	0.000
fault_loc	3.174	R3-PM1:V	2.192	R1-PM12:I	0.138	R2:S	0.000
R1-PA10:IH	3.166	R3-PA:Z	2.186	R3:F	0.091	control_panel_log1	0.000
R1-PA6:IH	3.159	R3-PM7:V	2.176	R1:F	0.078	control_panel_log2	0.000
R2-PA6:IH	3.152	R2-PA:Z	2.147	relay2_log	0.066	snort_log3	0.000

Table 12: Information gain evaluation for 10% electric data

From the above results, it is clear that there are an abundance of parameters which are extremely revealing to the algorithms. To help confirm this, some parameters were stripped and a test was performed with the selected algorithms, with the expectation that a minor reduction in parameters would have minor impact on the accuracies (i.e. the high results would remain). This expectation was met as shown in Table 13.

Algorithm	AttrRem	2	3	6	12	15	16	19	22	23	27	28	30	37	38	wavg
Bayes Net	2	0.99	0.99	0.99	0.99	1.00	0.99	0.99	0.98	0.99	0.98	0.99	0.99	1.00	0.99	0.99
	3	0.99	0.99	0.99	0.99	1.00	0.99	0.99	0.98	0.99	0.97	0.99	0.99	0.99	0.99	0.99
JRip	2	1.00	1.00	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	3	0.99	0.99	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
RandomForest	2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table 13: Results of parameter reduction on 10% electric data

Through visualization of the features within WEKA, it did not seem that there is anything inherent to the data itself causing excessive information gain. To investigate further, a separate dataset was generated by selecting a scenario "pair" - one scenario of normal behavior and an attack scenario meant to resemble the normal behavior - and extracting all instances of the two scenarios into a new dataset. The benefit of this was that there were many more instances of a given scenario (albeit only two scenarios total) compared to the 10% dataset, as well as a higher degree of variability within the features. This dataset was run with the selected algorithms. As shown in Table 14, this still produced exceptionally high accuracies as previously.

Algorithm	Category	Overall Acc	Cross-Validation	TP Rate Scenario 1	TP Rate Scenario 7
BayesNet	Bayes	95.28%	10	0.959	0.946
Jrip	Rules	99.98%	10	1	1
RandomForest	Trees	99.96%	10	1	0.999

Table 14: Results of two-scenario constructed data set

From the above results, it is strongly implied that there is a fundamental skewing of results when running the sequential-type electric data in WEKA with little or no advanced preprocessing performed prior. In particular, the issue lies with WEKA considering data entries to be individual entities to be classified, whereas the electric data has many entries corresponding to a single event to be classified, despite each entry having a marker parameter associating it with a member of the classifier. Thus, WEKA considers orders of magnitude more instances than are genuinely represented in the electric data, and the accuracies reported by the tool are not indicative of the true classification of sequential events.

Furthermore, what may in actuality be a change in a parameter during a single instance of a scenario is instead seen by WEKA as individual instances of a scenario with differing parameter values. This has a particularly strong impact on the networking parameters (control panel, relay, and snort logs), which tend to be characterized by bursts but are otherwise zeroed. Instead of treating such transactions as a meaningful sequence within a single event, the tool considers each in isolation and thus tends to classify

based on the more prevalent, normal, zeroed state. This effect seems to be supported by the extremely low rankings of information gain given by WEKA to all networking parameters.

Open Problems and Future Work

Gas and Water Data

The primary open problem for the gas pipeline and water storage tower datasets, is that they are unsuitable in their current form for use in IDS research. In each case of a correlation between a parameter and an attack, the correlation could have been avoided.

1. Gas Pipeline Dataset

1.1 command_address

If the command_address is something other than 4, Weka classifies it as a DOS attack. Some of this is good, because the MODBUS address of the device sending commands is 4, so something else could easily be an attacker. However, a few man in the middle attacks impersonating device 4 would add randomness.

1.2 response_address

This value is only ever 4 or 0 and only 0 when it's a recon attack.
This is because 0 is the MODBUS address of a broadcast message and the reconnaissance attacks send a broadcast message to determine the address of a device that responds. Some of this is legitimate because broadcast messages are not common in an established ICS system. However, broadcast messages are a legitimate MODBUS function, so adding legitimate broadcast messages would help to add randomness.

1.3 response_Length

This is always only 19 unless it's a reconnaissance attack. Then it's 123.
Again, this is because the device response to a broadcast message is of size 123. so adding legitimate broadcast messages would help to add randomness.

1.4 comm_read_function

This value is almost always 3, except in case of a DOS attack. This because 3 is the MODBUS read registers function code. I don't think parameter is needed, because there is no particular "read" field in the MODBUS data, just a function code. This field should be combined with resp_read_fun and subfunction to give just the MODBUS function code.

1.5 resp_read_fun

This value is only ever 3 or 1. CMRI only happens when it is 1. I don't think

parameter is needed. This field should be combined with comm_read_function and subfunction to give just the MODBUS function code.

1.6 subfunction

There are only three values of subfunction: 0, 1.5, and 4. It's always 0 unless it's an MFCI attack. . I don't think parameter is needed. This field should be combined with comm_read_function and resp_read_fun to give just the MODBUS function code.

1.7 setpoint

Setpoint only has unique 4 values: 20, 70, 80, and 90. Anytime the setpoint is not 20, it's an MPCI attack. More randomness could easily be added by modifying the setpoint legitimately to a wide range of pressures.

1.8 control_mode

control_mode is only ever either 0, 1, or 2. If it's ever 1, it's certainly MSCI. 1 is for manual mode of the pipeline. The system should be run in all 3 modes(manual, automatic, and off) legitimately to add randomness

. 1.9 control_scheme

control_mode is only ever 0 or 1. If it's 0, it's an MSCI attack. control_mode indicates whether the system is in "pump" control or "solenoid" control. The system should be run in both modes legitimately to add randomness.

1.10. measurement

All the CMRI attacks are in exact same measurement range from about 6 to about 11. The CMRI attacks should be more spread out to add randomness.

All the NMRI attacks are all above 100 or below -1. This is acceptable for attacks like Negative Sensor Measurement, Sensor Measurement Grossly Out Of Bounds , or Random Sensor Measurement, as all of these attacks will produce measurements that are an anomaly.

2. Water Storage Tank

2.1 command_address

The *command_address* attribute is needed by all algorithms to classify DOS attacks. The command address used for normal transactions is 7. Any command address that is not 7 is classified as a DOS attack.

2.2 com_write_fun

This attribute is needed by the algorithms to classify a variety of attacks. The NaïveBayes and PART classifiers use it to identify DOS attacks. PART and Ridor need it to classify CMRI attacks. Ridor also needs it to identify NMRI and MSCI attacks.This attribute is only ever of value 0x10 or 0x11. The value 0x10 is the normal Modbus function code for writing multiple registers.

2.3 resp_write_fun

This attribute is only need by the NaïveBayes classifier to aide in identifying Normal transactions. This attribute is only ever of value 0x00 or 0x10. The value 0x10 is the normal Modbus function code for writing multiple registers. When the value is 0x00, the transaction is a Recon attack.

2.4 sub_function

This attribute is needed by the algorithms to aide in classifying various attacks. NaïveBayes and Ridor use this attribute to classify MFCI attacks. Ridor also uses this attribute to classify normal transactions as well as CMRI and MPCI attacks. PART uses this attribute to aide in DOS attack classification. Sub_function is only ever of value 0x00 or 0x10. If the value is 0x10, the transaction is an MFCI attack.

2.5 resp_length

This attribute is also used by the algorithms to classify various attacks. All the algorithms use this attribute to aide in normal transaction classification. PART needs this attribute to classify MSCI attacks. Ridor needs this attribute to classify NMRI and CMRI attacks.This attribute is only ever of value 21 or 123. When the value is 123, the transaction is a recon attack.

2.6 HH

The NaïveBayes classifier uses this attribute to identify MPCI attacks. Ridor uses this attribute to classify MSCI attacks. For normal operations, this attribute's value is 90. When it is anything else, the transaction is an MPCI attack. Whenever this attribute's value is changed, the attribute H is changed as well.

2.7 H

The NaïveBayes and Ridor classifiers use this attribute to identify MPCI attacks. Ridor also uses this to classify NMRI, CMRI, and MSCI attacks. For normal operations, this attribute's value is 80. When it is anything else, the transaction is an MPCI attack. Whenever this attribute's value is changed, the attribute HH is changed as well.

2.8 L

The NaïveBayes and Ridor classifiers use this attribute to identify MPCI attacks. Ridor also uses this to classify NMRI, CMRI, and MSCI attacks. For normal operations, this attribute's value is 20. When it is anything else, the transaction

is an MPCI attack. Whenever this attribute's value is changed, the attribute LL is changed as well.

2.9 LL

Ridor uses this attribute to classify CMRI attacks. For normal operations, this attribute's value is 10. When its value is not 10 but L is 20, the transaction is an MSCI attack.

2.10 control_mode

NaïveBayes and Ridor use this attribute to classify MPCI attacks. NaïveBayes also uses this attribute to classify CMRI attacks. Control_mode is only ever of value 0 or 2. There is no clear relation between the values and any transaction classification.

2.11 crc_rate

Ridor uses this attribute to identify MPCI attacks. This attribute's value is only ever 0 or 1. There is no clear correlation between this attribute's value and transaction classification.

2.12 measurement

All algorithms use this attribute to identify CMRI attacks. PART and Ridor also uses this attribute to identify NMRI attacks. NaïveBayes does not correctly classify any NMRI attacks.

2.13 time

Ridor uses this attribute to help classify CMRI and MSCI attacks. The value varies greatly; therefore, there is no clear correlation.

It is recommended that these datasets be recreated with a wider range of normal transactions and attacks that mimic more closely the behaviors of SCADA attackers, as described above.

3. Electric Dataset

The primary open problem with the electric transmission dataset is that without considerable preprocessing, they are unsuitable for use with the implementation of algorithms used by WEKA and thus results of analysis using the tool are likely not indicative of the true classification strength of the algorithms. The recommended solution is to attempt to preprocess the data in such a way that it remains representative of the original data, particularly the scope of behavioral scenarios, while being better suited to use with algorithms within WEKA. Given that high accuracies are achieved with such a technique, a test of downsampling could then be applied to determine whether or not the effective sampling rate of the data can be reduced while retaining similar value of information.

Conclusions

The primary conclusion of this project is that 2 of the datasets being used for analysis(the water and gas datasets in particular) are unsuitable for IDS research as they currently exist, due to the obvious correlations between particular parameters and the result to be predicted. These correlations are unrealistic in real SCADA transactions, which is what renders the datasets unsuitable in their current form.

The remaining dataset (electric power transmission) is not necessarily unsuitable for IDS, but rather is not well suited to the implementations of machine learning algorithms used for analysis due to its sequential nature. A significant amount of preprocessing on the electric transmission dataset is likely required in order to conduct appropriate research using given methods.

Previous Work

Jianmin Jiang and Lasith Yasakethu in their paper "Anomaly Detection via One Class SVM for Protection of SCADA Systems"[4] write about using Support Vector Machines(SVMs), a class of Machine Learning algorithms, in an intrusion detection system developed at The University of Surrey. While detailed in their analysis of the basic theory of SVMs, not much time is devoted to the particulars of the data used for analysis. It is our intention to include an analysis of the particular data being used, and how this affects the performance of the tested algorithms.

Maria Muntean et al. in their paper "*Data Mining Learning Models and Algorithms on a SCADA System Data Repository*"[5] perform a similar analysis of 3 Machine Learning Algorithms in the WEKA environment and their effectiveness at predicting inlet water temperature. While helpful in their baseline analysis of 3 algorithms, the data being analyzed has only 2 features and is therefore only minimally representative of a real SCADA control system. The data sets used for the proposed project are much more extensive and more representative of actual SCADA control systems.

Mohammad Al-Subaie and Mohammad Zulkernine in their paper "*Hidden Markov Models Over Neural Networks in Anomaly Intrusion Detection*"[6] state the importance of accounting for sequential relationships between events of patterns when analyzing system behavior. To this end, the authors investigate and compare the performance of two machine learning techniques: Hidden Markov Models (HMMs) and Multilayer Perceptron (MLP) neural network. While the paper does affirm the strength of sequential learning-based techniques for anomaly detection, it does not have specific focus on SCADA systems - our intention is to investigate and compare machine learning techniques in the particular context of SCADA systems.

In the paper *"Predicting Mine Dam Levels and Energy Consumption Using Artificial Intelligence Methods,"*[7] authors Ali Hasan, Bhekisipho Twala, et al. use four machine learning algorithms to determine the viability of the use of artificial intelligence in the mining industry to predict dam levels and energy consumption. The four algorithms used are: artificial neural networks, a naïve Baye's classifier, an SVM, and decision trees. Their results show that artificial neural networks worked the best in predicting both cases. While artificial neural networks may be the best of the four at predicting their cases, we intend to study the effectiveness of other algorithms in predicting our case.

In the paper *"A Log Mining Approach for Process Monitoring in SCADA,"*[8] authors Dina Hadžiosmanović, Damiano Bolzoni, et al. propose a method to "identify process-related threats in

SCADA" systems. Their goal is to prove that their proposed methodology can effectively detect anomalous behavior. Similar to our project, the authors of this paper also test their method on data obtained from a real SCADA system. While the goal of anomaly detection is similar to ours, the authors of this paper do not use machine learning techniques or algorithms. Also, the data sets we will use are produced from different SCADA systems than the one used in their research.

In the paper *"Neural Network Based Intrusion Detection System for Critical Infrastructures,"*[9] authors Ondrej Linda, Todd Vollmer, et al. propose the Intrusion Detection System using Neural Network based Modeling (IDS-NNM). Similarly to our project, the authors use real data from a SCADA system; however, the authors do not specify from which kind of system the data was obtained. The IDS-NNM uses two neural network algorithms. They are the Error-Back Propagation and the Levenberg-Marquardt algorithms. While the authors performed analysis with data from one SCADA system using neural network algorithms, we intend to analyze the effectiveness of three different algorithms on detailed data sets obtained from three different SCADA systems.

Appendix A: References

[1] Merriam Webster Dictionary - http://www.m-w.com

[2] WEKA - University Of Waikato - http://www.cs.waikato.ac.nz/ml/weka/

[3] Wikipedia - Machine Learning - http://en.wikipedia.org/wiki/Machine_learning

Appendix B: Previous Work References

[4] Jiang,Jianmin ,Yasakethu,Lasith, *Anomaly Detection via One Class SVM for Protection of SCADA Systems,* http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6685663&sortType%3Dasc_p_Sequence%26filter%3DAND(p_IS_Number:6685639)

[5] Muntean,Maria, Ilean,Ioan, Rotar,Corina, Risteiu,Mircea, *Data Mining Learning Models And Algorithms on a SCADA System Data Repository* http://brain.edusoft.ro/index.php/brain/article/view/106

[6] M. Al-Subaie, M. Zulkernine. *Hidden Markov Models Over Neural Networks in Anomaly Intrusion Detection*. In *Proceedings of the 30th Annual International Computer Software and Applications Conference*, pages 325-332, 2006.

[7] Hasan, A.N.; Twala, B.; Marwala, T., "Predicting mine dam levels and energy consumption using artificial intelligence methods," *Computational Intelligence for Engineering Solutions (CIES), 2013 IEEE Symposium on* , vol., no., pp.171,175, 16-19 April 2013.

[8]Hadžiosmanović D, Bolzoni D, Hartel P. "A log mining approach for process monitoring in SCADA," *International Journal Of Information Security* [serial online], vol. 11, pp. 231-251, August 2012.

[9] Linda, O.; Vollmer, T.; Manic, M., "Neural Network based Intrusion Detection System for critical infrastructures," *Neural Networks, 2009. IJCNN 2009. International Joint Conference on* , vol., no., pp.1827,1834, 14-19 June 2009.