MACHINE LEARNING AND CYBER SECURITY

Gupta Kajal Santosh Sanju

Under the Guidance: Mrs. Hina Mehmood

Master of Science in Information Technology Part-I

Rizvi College of Arts, Science and Commerce

University Of Mumbai

ABSTRACT

The application of machine learning (ML) technique in cyber- security is increasing than ever before. Starting from IP traffic classification, filtering malicious traffic for intrusion detection, ML is the one of the promising answers that can be effective against zero day threats. New research is being done by use of statistical traffic characteristics and ML techniques. This paper isa focused literature survey of machine learning and its application to cyber analytics for intrusion detection, traffic classification and applications such as email filtering. Based on the relevance and the number of citation each methods were identified and summarized. Because datasets are an important part of the ML approaches some well know datasets are also mentioned. Some recommendations are also provided on when to use a given algorithm. An evaluation of four ML algorithms has been performed on MODBUS data collected from a gas pipeline. Various attacks have been classified using the ML algorithms and finally the performance of each algorithm have been assessed.

KEYWORDS

Machine learning; Data mining; cyber security

INTRODUCTION

This papers is a focused literature survey of machine learning and data mining methods for cyber security applications. Few ML methods are described along with their application in the field of cyber security. A set of comparison criteria for ML method is provided in the paper and a set of recommendations on the best method to use was made depending on the properties of the cyber security problems. Secondly, a MODBUS data set has been used to compare the effectiveness of five different algorithmswhen applied to ICS networks. Receiver operating characteristic (ROC) is often used to choose optimal models and to discard sub- optimal one independently from the cost content or the class distribution. Hence, a ROC curve has been plotted to assess the performance of the binary classifier used with the data set under study.

This paper is intended for researchers willing to start their work inthe field of ML and cyber security. Along with the description of the machine learning some references to prominent works have been cited and some valuable examples are put forth how cyber problems are often tackled by ML. From early 2000 several prominent surveys on the ML research has already been described. Nguyen et. al. puts forth a comprehensive study of IP traffic classification technique that does not rely on well- known port numbers or known packet payloads. The contribution of the paper is to identify cyber security datasets that can be used by researchers and to point out the algorithms that can be applied to cyber specific problems. A set of machine learning algorithm have been evaluated in the later part of the paper on collected ICS dataset to identify various attacks while analyzing remote terminal unit (RTU) in a gas pipeline. The data set used has 35 different types of simulated attacks against ICS. The accuracy of each ML algorithm in the segregation of malicious traffic have been analyzed.

RESEARCH METHODOLGY

The SLR aims to identify, evaluate and interpret all the available research in the area of interest to identify potential research gaps and highlight the frontiers of knowledge. It provides a high-quality, transparent and replicable review to summarize the large number of research studies. This study follows an SLR methodology for the following reasons: (i) AI for cybersecurity is a diverse field with a large quantity of literature; (ii) this study aims to answer specific research questions; (iii) the rigour and replicability it provides leads to an unbiased scientific study. The procedure if the SLR is describe in Selection of bibliometric database, Search strategy, Inclusion and Exclusion criteria and Selection of primary studies.

IMPORTANT CYBER SECURITY

DATASET FOR MACHINE LEARNING

Data is of utmost importance in ML approaches. A researcher of machine learning has to understand the data set thoroughly before doing any kind of analysis. Secondly, raw data like packet capture(pcap), NetFlow and other network data is not directly usable in the ML analysis. The data has to be pre-processed to make it usable in popular ML tools like WEKA, R and RapidMiner. Hence researchers using ML analysis on custom system has tounderstand the data collection methodology and the methods that are used in preprocessing the data. This section will enlist few lowlevel details on the data sets, and some popular tools used in capturing the data from the network.

1.1 Network Packet Data

There are a lot (144 as per Internet Engineering Task Force) of internet protocols that are used by programs running on the user levels. These protocols uses data packets as the main mode of communication. The network traffic in the form of packet received and transmitted at the interfaces (physical and wireless) can be captured and stored in

the form of packet capture (pcap). Libpcap and Wincap for UNIX and Windows respectively are very popular network tools. Some tools like wireshark, tcpdump can also be used as protocol analyzer, packet sniffer and network monitor. The dataset of machine learning has distinct features and attributes. These features defines the prime characteristics of each set of data in the dataset. To convert a pcap file to pdml tools like tshark can be used.

Tshark –T pdml –r <input file> < output file>

Where the input file is the .pcap file and the output file is the name of the pdml file. Secondly the Fowler's tool "pdml2arff.py" (available in GitHub) can be used to do perform the final conversion.

pdml2arff.py <Input file>

Where the input file is the name of the pdml file. This will generate an arff file called <Input file>.arff

1.2 Data from NetFlow

Cisco has its own feature called NetFlow to monitor the network interface and collect IP network traffic as it enters and exits the interface. A network administrator can determine things such as the source and the destination traffic and class of service by analyzing the data provided. A typical NetFlow architecture has three main components Flow exporter- accumulates the network traffic and exports the flow towards flow collectors, Flow collector- receives and preprocesses the data and finally stores the data, Analysis application- Segregates the flowing data and profiles it on the basis of need.

1.3 Other Data Sets

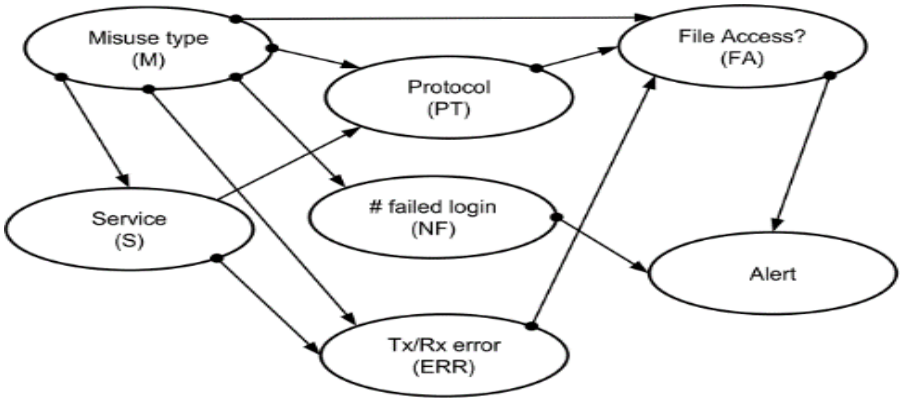
The DARPA (Defense Advanced Research Project Agency) has two datasets that are invaluable for cyber security researchers. The DARPA 1998 and 1999 dataset was developed by Cyber Systems and Technology group of the Massachusetts Institute of Technology Lincoln Laboratory (MIT/LL). KDD 1999 is another famous data set that is predominantly used by cyber security researchers. Another prominent data set involving SCADA protocol was generated by the Mississippi State University's critical Infrastructure protection center. This dataset will be analyzed in the later sections to evaluate the accuracy of ML algorithms on the SCADA protocols. This data set records the data from a simulated gas pipeline and documents 35 distinct attacks on the SCADA system.

MACHINE LEARNING TECHNIQUES FOR CYBER-SECURITY

Few popular ML techniques are described in this section. For each method the application to cyber security have been identified.

2.1 Bayesian Network

The network is developed as a random set of variable and their conditional dependencies via a directed acyclic graph. The nodes representing the child are dependent on the parent nodes and each node maintains the states of the conditional probability form and the random variable. Fig 1 shows the attack signature detection using Bayesian network. Each state is an input to the underlying state with varying state values. The calculated probability tables are calculated and shown in the figure. Bayesian networks can also be used to infer unobserved variables.



File Access state input variables and values	P(FA = True)	P(FA = False)
M=R2H, PT=NSF, ERR=0	0.95	0.05
M=R2H, PT=FTP, ERR=0	0.99	0.01
M=Probe, PT=none, ERR=50%	0.80	0.20
M=Probe, PT=PING, ERR=0	0.50	0.50
M=DoS, PT=POP, ERR=100%	0.80	0.20
M= DoS, PT=HTTP, ERR=50%	0.90	0.10

Fig 1. Attack signature detection using Bayesian network**.**

2.2 DECISION TREES

The decision tree is very much analogous to a tree. The trees have leaves which represents the various classifications and the branches are the links or features that in-turn provides the path to the classifications. ID3 and C4.5 are few popular algorithms for generating decision trees automatically. The comparing process of the SNORT rules with

the incoming traffic is slow because of the large number of signatures. Kruegel and Toth et al. replaced 150 SNORT rules by using a variant of ID3 algorithm. Their aim was to replace these algorithm by a decision tree model. This would be effective in increasing the speed of processing. Rule clustering was used to replace the Snort rules. This minimizes the number of necessary comparisons. This also allows parallel evaluation hence speeds up the comparison procedure. The clustered rule was applied to DARPA 1999 dataset.

2.3 CLUSTERING

This is an unsupervised learning method where similarity measure is used to group data together. Clustering algorithms can learn from audit data and explicit description of different attack classes by the system administrator is not necessary. Hendry et. al. demonstrates the application of real-time signature detection using clustering algorithm. The normal and anomalous network traffic was created by a density based clustering scheme known as Simple Logfile Clustering Tool (SLCT). Two clustering schemes are used: Firstly, for detection of normal and attack scenarios, secondly the other scheme can be used to determine the normal traffic in a supervised manner. In this model parameter M is used to define the feature that is contained in the cluster.

2.4 ARTIFICIAL NEURAL NETWORK (ANN)

The ANN behaves mainly like human brain. The neural network has a layer layout. The input from the data actuates the neuron the second layer of the network. Which in turn outputs to the next layer of the hierarchy. This carries on and finally the output is produced by the last layer of the network. The internal network which plays an important part in the neural network are black boxed from the environment and is known as hidden layers. One major drawback of neural network is the huge learning time due to the occurrence of local minima. This approach was prevalent in mid-nineties but due to the advent of support vector machines (SVMs) ANN started to fade away. With the introduction of convolution NN the popularity of neural network is on the rise again.

2.5 GENETIC ALGORITHM AND GENETIC PROGRAMMING

Two of the most popular computation method based on the principle of survival of the fittest is- GA and GP. These algorithms functions on the population of the chromosomes that evolve based on certain operators. The three basic operator used is selection, crossover and mutation. The algorithm is started with a randomly generated population, a fitness value is computed for each individual. This signifies the ability of the each individual to solve the current problem and individuals with higher probability have higher chance of being chosen in the mating pool. Two capable individual will perform the next step called crossover and finally each will undergo mutation. Among the two mutated individual the highest fit chromosome will be rallied over to the next generation.

2.6 HIDDEN MARKOV MODELS (HMM)

This is a statistical Markov model with a set of states which are interconnected using transition probabilities that determines the topology of the model. The system is assumed to be a Markov process with unobserved parameters. This model provides a forward- backward correlation which can be used to determine the hidden parameters from the observable parameters. Since the probability distribution in each state is different the system can change states overtime and is capable of representing non- stationary sequences. Joshi et. al. made use of HMM to develop an intrusion detection system. Five definite states are used each having six observation symbol per state. The interconnection between the states are developed in such a way that any state can transition into any different state.

2.7 INDUCTIVE LEARNING

The inference of certain information from a dataset is known as deduction. On the other hand the other approach of moving from specific observation to develop theories and patterns is known as inductive learning. These are the two primary methods used for the inference of information from the data. Inductive analysis develops some general patterns and which are used to develop some hypothetical conclusions. Two prime approaches of distribution based anomaly generation and the filtered artificial anomalies was used to generate these random anomaly.

ML RECOMMENDATIONS FORANOMALY DETECTION

Machine learning is used in cyber-security in three main areas: IDS, Anomaly detection module and misuse detection. Anomaly detection is specifically aimed at segregating abnormal traffic from normal one while misuse detection classifies attack signature comparing it with known ones. Clustering algorithm (Density based like DBSCAN) works the best with anomaly detection. Apart from the high processing speed clustering algorithms are easy to implement and the parameters to configure are also less in number. SVM also performs considerably well for anomaly detection. For misuse detection the classifiers has to have the capability to generate signatures. Branch feature in a decision tree or chromosomes in genetic algorithm generates signatures that are apt for such task. Hence algorithms like ANN and SVMs which has hidden nodes are not well suited.

EVALUATION OF ML ALGORITHMS ON MODBUS DATA

The main aim of this evaluation is to test the applicability of certain ML algorithms to detect cyber-attacks on MODBUS data. Tenfold cross validation was used to develop the ML models. This analysis was performed in Weka. In 10 fold cross validation Weka produces 10 different models for the data set provided. Then the weighted

average of these models are calculated which is showed as the final result. The data set used was labeled telemetry data from gas pipeline developed by the Critical Infrastructure Protection Centre of Mississippi state university.

Few standard classifier was considered for the evaluation. The methods used were:-

1. Naïve Bayes- Bayes' theorem based probabilistic classifier.
2. Random Forest- A Ml based on decision tree algorithms.
3. OneR- Each feature of the rule set is evaluated andfinally the optimum or the best one is chosen.
4. J48- A basic implementation of C4.5 decson treealgorithm

3.1 Information about the Dataset

The dataset used was in Weka minable arff format. It had 20 total attributes. The Table I below enlists all the features present in the dataset.

Features	
address	reset rate
control scheme	command response
function	deadband
pump	time
length	cycle time
solenoid	binary result
setpoint	rate
pressure measurement	categorized result
gain	system mode
crc rate	specific result

Morris et. Al. Gives a comprehensive overview about each features of the dataset and why each aspect is important from the perspective of cyber security and intrusion detection. In thisdataset a total of 35 attack was performed. These attack can be broadly classified into 7 categories: Naïve Malicious Response Injection (NMRI), Complex Malicious Response Injection (CMRI), Malicious State Command Injection (MSCI), Malicious Parameter Command Injection (MPCI), Malicious Function Code Injection (MFCI), Denial of Service (DoS) and Reconnaissance. The Final class in the arff data set has these 7 attack catagories along with normal traffic data. 97019 Instances were recorded in the dataset. The Distribution of the final class is enlisted below.

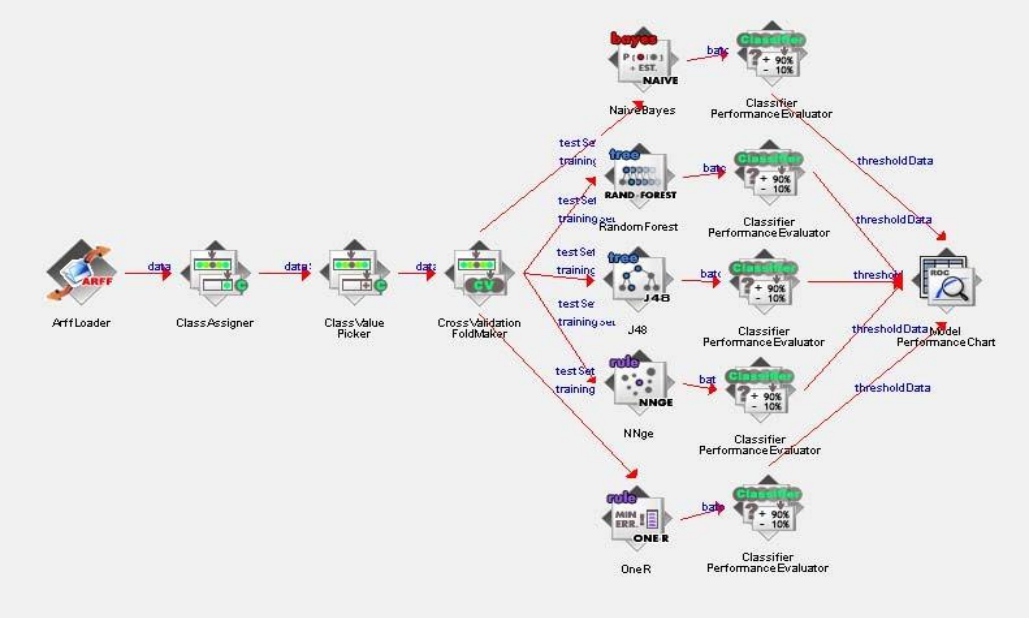
Class Label	Count
Normal	61156
Naïve Malicious Response Injection (NMRI)	2763
Complex Malicious Response Injection (CMRI)	15466
Malicious State Command Injection (MSCI)	782
Malicious Parameter Command Injection (MPCI)	7637
Malicious Function Code Injection (MFCI)	573
Denial of Service (DoS)	1837
Reconnaissance	6805

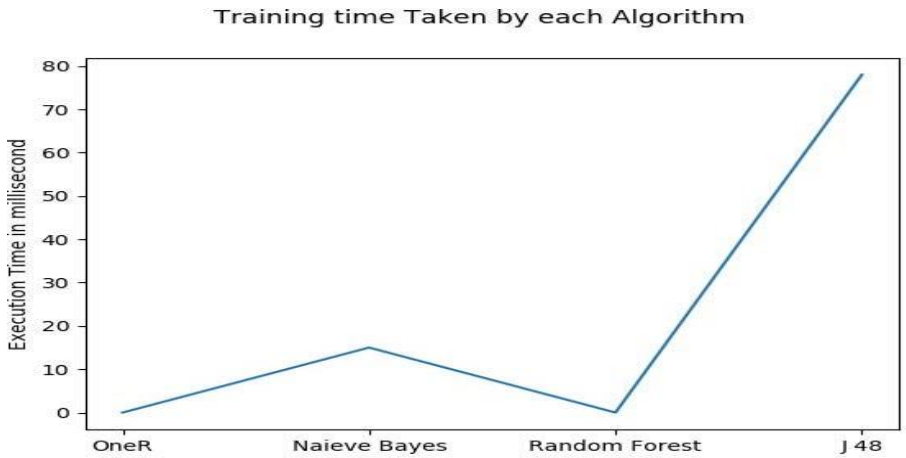
3.2 Accuracy and ROC curve for MLAlgorithm Evaluation

Beaver et. al. have already used the ICS dataset for ML algorithm analysis. But ROC curve was not plotted for any algorithm and hence it is very hard to make out the overall performance of the algorithms. The receiver operating characteristics (ROC) curve is the plot of false positive rate (FAR)in the x-axis versus the plot of test sensitivity in the y-axis. The area under the curve of the ROC is an important parameter. It is used to measure the sensitivity and the specificity. Where the sensitivity is the number of true positive decisions and the specificity is known as the number of true negative decisions.

The Weka Knowledge flow model for the current analysis was developed. It is shown in Figure 2. The Roc curve generated for the four ML algorithms are show in Figure 5. From the ROC curve it is evident that j48 algorithm produces the most optimized results in general overall classification for the power system dataset. The AUC of the four algorithms is shown in the table below.

Algorithm	Area under the Curve (AUC)	Precision	Recall
OneR	0.887	0.862	0.894
Naïve Bayes	0.967	0.947	0.936
Random Forest	0.989	0.988	0.988
J 48	0.995	0.992	0.992





Hence in this 9 bus IEEE system even though J48 might outperform Random forest in terms of accuracy a little compromise on the accuracy can yield better real-time performance when the algorithms are implemented as a core part of the intrusion detection system.

FUTURE RESEARCH

Through specific procedures, such as the Automated Validation of Internet Security Protocols and Applications (AVISPA), the security of the newly constructed protocol can be verified formally. This tests the security of the protocol against replay and man-in-the-middle attacks. In addition, the Burrows-Abadi-Needham (BAN) logic test can be used to determine whether there is a chance for "safe mutual authentication among the communicating organizations." In addition to these, we can analyze the formal security of a security protocol using the Real-or-

Random (ROR) model implementation, which highlights the potential for an attack on the planned authentication, access control, or key management protocol that involves unauthorized session key computation. In this approach, the security of the designed protocol may be assessed and examined. Therefore, new security protocols are needed that have more security and functionality characteristics and can withstand zero day vulnerabilities as well.

CONCLUSION

In this paper an elaborate survey was performed to enlist few popular datasets then few ML algorithms were discussed along with their application in cyber-security. Finally few recommendations were made regarding the choice of ML. In the later part of the paper a brief analysis was performed with an ICS data set and performance of a few ML algorithm was evaluated. Although J48 algorithm performs better than other algorithms in the scope of analysis, more analysis needs to be performed to ascertain the performance of the algorithms because the performance of algorithms tends to skewed depending upon the dataset on which it is being applied on. Secondly, Random forest might be more suitable as a core IDS algorithm for its optimal real-time performance in the current scenario being considered.

REFERENCE

- Buczak, A., & Guven, E. (2015). A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection. IEEE Communications Surveys & Tutorials, 1–1. http://doi.org/10.1109/COMST.2015.2494502
- F. Jemili, M. Zaghdoud, and A. Ben, “A framework for an adaptive intrusion detection system using Bayesian network,” Intelligence and Security Informatics, IEEE, 2007
- Abraham, C. Grosan, and C. Martin-Vide, “Evolutionary design of intrusion detection programs,” International Journal of Networks Security, 4, 2007, pp. 328–339