IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

REVIEW ON MACHINE LEARNING ALGORITHM FOR CYBER SECURITY

[1]S P Nayana, [2]Dr. Mohan Aradhya

[1]Student, [2]Assistant Professor
[1]Department of Computer Applications,
[1]RV College of Engineering, Bengaluru, India

Abstract: Machine learning methodology is being implemented in cybersecurity like never before. Starting with Classification of IP traffic, malicious intrusion detection filtering, ML is one of the promising reactions that may work against zero day attacks. New research is carried out using the techniques of statistical traffic and ML. This paper is a concentrated applying machine learning algorithm on cyber data set and its use in cyber analytics for intrusion detection, classification of network traffic and email filtering applications. Each method has been identified and summarized on the basis of relevance and the number of citations. Since ML approaches are essential, data sets include some well-known datasets. In addition, some suggestions are presented for using a certain algorithm. Various attacks were categorized using the ML-algorithms and each algorithm was eventually assessed for performance.

I. INTRODUCTION

This article highlights the use of information security machine learning and data mining techniques. With the application in the field of data security, few ML methods are specified. A set of criteria for comparing the ML method is given in the paper. A set of recommendations were made for the best method to use depending on the characteristics of the cyber security issues. This paper aims to help researchers willing to start working on ML and cyber security. In conjunction with this overview of the machine learning, leading works have been cited and helpful examples of how ML often addresses cyber problems have been presented. Several prominent ML research surveys were already described from the beginning of 2000.

And Nguyen.et.al[2] provides an extensive IP traffic grading study that does not rely on well-known port numbers or known payloads of packets. This paper discusses ML techniques along with traffic statistics used in the IP category and Nguyen. al. has examined 18 papers in the field of cyber security and is among the most valued researchers in the field of ML and related domains of any researchers.

Amomani et.al [4] has carried out an extensive survey of all major email filtering and ML techniques to classify and detect phishing emails. The state-of-the-art investigation into such attacks has been listed, and the techniques have been compared. and Tedero.[5] Presents an intrusion network mathematical method based on machine learning and knowledge. It focuses mainly on the detection of abnormalities and not the detection of signatures. Numerous cybersecurity staff have investigated whether traffic on fly is filtered or classified. Speroto and Speroto. to the. [3] NetFlow (Net Flow) used data and showed that it may not be possible to process packets at stream rate when the level of network traffic is limited.

II. ESSENTIAL DATASET IN CYBER SECURITY

Data are extremely important in ML approaches. Before conducting any analysis, one Machine Learning researcher must fully understand the data. In addition, raw data such as pcap, NetFlow and other network information can not be used to perform an ML analysis directly. In order to be used with popular ML instruments like WEKA[6], R [7] and RapidMiner[8], data needs to be processed beforehand. Researchers using ML custom system analysis therefore have to take into account methods for collecting data and how the data is pre-processed. This section contains few examples of small data sets and some of the common data collection tools for the network.

2.1 Data generated in Network

There are a number of internet protocols used by user-level programs (144, as per the Internet Engineering Task Force). The main mode of communication for these protocols are data packets. Network traffic is collected and stored by the appropriate and interface-transmitted (physical or wireless) packet capture (pcap) format. Libpcap and Wincap are common network tools for UNIX and Windows respectively. Furthermore, protocol analyzer and packet sniffer can include tools such as wireshark, tcpdump and network control.

The machine learning dataset has different characteristics and attributes. Such features describe the key features of each data set in the dataset. Thus when pcap is collected, the researcher needs to write a kind of script in order to separate the required attributes of the pcap to the ML tool. Fowler .al. [9] studied Weka 's Relationship attribute (arff) format was investigated and a method was developed for the conversion of any packet information markup language format (pdml) into a weka minable arff format. Pcap file can be converted to pdml tools with tshark.

2.2 Data from NetFlow

The Cisco network interface is monitored through Cisco's own NetFlow and IP traffic is collected on entering and leaving the interface. The data can be analyzed by a network administrator, such as source, destination traffic and service type. A traditional netFlow architecture has accumulated and transferred the traffic of the network to the collection system in three main aspects Flow Imports. Flow Collector collects and pre-processes data and eventually sets up the data. The existing network packets are stored in the compressed and preprocessed NetFlow files.

2.3 Additional Data Sets

DARPA has two sets of data, valuable for scientists of cyber security. The Agency for Defense Advanced Research (DARPA). Cyber Systems and Technology Group at Lincoln Laboratories Institute Massachusetts (MIT / LL) developed the 1998 and 1999 DARPA dataset. KDD 1999 is another well-established data collection used mainly by cyber-security experts. Another important SCADA data package was created by the Critical Infrastructure Protection Center of Mississippi University [1]. This data set will be analyzed for the accuracy of the SCADA algorithms in the following sections. This package tracks data from a gas pipeline simulated and separately documents 35 attacks by SCADA.

III. METHODOLOGIES OF MACHINE LEARNING FOR CYBER-SECURITY

There are few popular ML techniques in this section. Each method has been identified for the cyber security application.

3.1 Network of Bayesian

The system is built via a directed acyclic graph as a random variable and its dependency. The child's nodes depend on the parent nodes, the conditional probability status, and the random variable for each node are maintained. For the underlying state each state has different values. The determined probability tables are estimated and shown in the figure. Including unobserved variables, Bayesian network for anomaly detection and knowledge Bayesian network can be utilized

The attack and pattern signature can also be compared to known attack streaming data. Jemili and Jemili. al.[14] developed a Bayesian network intrusion detection system. The KDD of 1999 was applied to models nine attributes of the device. A score of 88% and 89% was accomplished in standard and attack scenarios. For sample, scan, DOS and R2L the detection rate for the model was 99%, 21% 89% and 7%. The model's accuracy has been greatly affected by the reality that the number of training cases in R2L can also be very beneficial.

3.2 Decision trees

The tree of decisions is just like a forest. The leaves of the trees reflect different gradations and the branches are the relations or characteristics on which the grade is centered. ID3 and C4.5 are just some of the common decision-making algorithms.

The relation between the SNORT rules and incoming traffic is sluggish due to the large number of signatures. Kruegel and Toth et al. [15] substituted 150 SNORT rules by using an ID3 algorithm variant. Its goal was a model of the decision tree to replace this algorithm. This would increase processing speed effectively. The Snort rules have been replaced by rule clustering. This reduces the number of comparisons needed. Parallel evaluation also allows the comparison process to be speeded up. DARPA 1999 dataset was subject to the clustered law. The software model was equivalent to its processing speed and performance with the snort study. The model's average speed was 105% and the minimum speed was 5%. For more study, the number of substituted laws was increased from 150 to 1581. The number of alternate laws was increased from 150 to 1581 for further research. While Toth has yet to detect a deep rate through the decision tree procedure, the second is a significant decrease in the treatment time and not all quantitative data are present.

3.3 Applying Clustering Algorithms

This is an unregulated method of learning used with similarity in group data. Audit data can be used as clustering algorithms, and it is unnecessary for the system manager to describe different attack classes specifically.

Hendry.al. [16] illustrates the need for real-time signature detection using clustering algorithms. The Simple Log Clustering Tool (SLCT) clustering framework generated regular and abnormal traffic on the network. Two rating schemes are employed: first, for the identification of regular traffic and attack situations, the second can be used in monitoring. This model parameter M specifies the cluster function. When M is set to 97%, 98% threat data with the FAR value is detected. Samples in the high-density clusters produce these signatures. This model was tested using the KDD dataset. In order to enhance model precision, the integrity of the cluster was used as performance indicators. For unknown attacks, an accuracy of 70 % to 80% was achieved.

3.4 Applying Artificial Neural Networks (ANN) algorithm

The ANN works primarily like the human mind. The layer of the neural network is structured. The input drives the neuron to the network's second layer. In addition, the next layer results hierarchically. Tests are carried out and final results are generated in the last network layer. In the neural network the internal network plays a key role in black-boxing. Because of local minimum learning time, the neural network has a major downside. In the mid-1990s, this was common, but ANN began to decrease as support vector machines developed. With the implementation of convolution NN, the success of the neural network is increasing again. ANN that uses a multi category abnormal classifier is described by Canady [17]. The Real Secure network controller has been used for data generation. The attack signatures have been included in the program. About 3000 attacks have been simulated by software such as Satan or Internet Scanner from the 10 000 reported attacks. In addition to data preprocessing, nine selected features were included: ICMP code, type of ICMP, source, destination address, protocol Id, source port, destination port, raw longitude and raw data sort. The analysis then used normal data to train the ANN and assault data. During the training and test scenario, report an error rate of 0.058 and 0.070. Therefore, the normal accuracy of a 0.070 RMS for the test phase corresponds to 93 percent. The information here is classified as traffic that is normal or malicious.

3.5 Applying Genetic algorithm and genetic programming

GA and GP based on the most suitable survival principles are two of the most common calculation methods. These algorithms function with those operators for the chromosome population. The three primary operators used are range, mutation and crossover. The innovation is started by a random population with a fitness value for each person. This means that each person has the potential and the chances of solving the existing problem are better selected in the corresponding pool. The next move, known as a crossover, is taken by two able individuals and then mutated. The most fit chromosome will be united among the two mutated individuals to the next generation and Abhram.Al .[18] used an assault classifier using a simple GP template. The research included three common GP models, Linear Genetic Program (LGP) and Gene Expression Programming (GEP). The model used several mathematical operators' function sets. The 1998 DARPA data set is the primary dataset for validating the created model. The dataset consisted of four major attack types with a total of 24 scenarios (U2R, R2L, doS and test). Depending on the type of attack investigated, the false alarm rate of the above model is as small as 0 to 5 percent.

3.6 Applying Inductive Learning

Deduction is defined as the deduction of other data sets. The other approach is known as inductive learning, which is to switch from special insight to the creation of theories and models. Those are the two key methods used to evaluate results. Inductive analysis produces certain general patterns that are used to derive conclusions from hypotheses.

Fan et. al.[20] developed a random events and anomalous traffic creator for artificial anomalies. To produce this random anomaly, two primary approaches were used to establish distribution-based anomalies and the artificial anomalies filtered. These data were fused with the 1998 DARPA dataset randomly. Fan et. al. used these data to research inductive learning model output established. The detection rate of 94 was successful and the FAR was low of 2 percent.

3.7 SVM (support Vector Machine)

The use of an SVM will increase the accuracy of IDSs in the field of cyber security. This approach can be used to predict two possible effects (i.e. malicious or non-malicious network traffic) by the classifier generated. The analysis of ML techniques for intrusion detection tasks will allow us to identify a classification technique that could add to our anomaly bases IDS and, at the same time, make it possible to build a benchmark to compare the performance of our IDS with the hybrid detection line.

An SVM strives to find the optimal hyperplane separation, optimize the exercise margin and eliminate uncertainties and risk overlap. An SVM can be introduced easily, requires a small data set and can be analyzed on extremely large datasets. An SVM also takes very little time for the classification process, once the hyperplane of the optimal classification is built. The Matlab and LibSVM based classifier. The LibSVM is an integrated reliable software that includes multiple kernel functions to support vector classification and distribution estimation.

Using SVM to improve IDS output or as an alternative detection technique as an ML technique. It can be measured the output of unregulated, linear and nonlinear anomaly-based IDS against first- and second-class SVMs. A number of network traffic data sets, collected from real networks, consisting of various types of attacks have been used for analysis of SVM algorithms.

The two-class linear SVM achieves the best overall efficiency. With almost any sample, this methodology hits 100% DR and OSR. This SVM technique, however, requires previously defined, training data consisting of both data groups. On the other hand, the accuracy of the SVM linear one-class is comparable to the accuracy of the SVM linear two-class without the need for malicious data processing. Only in the case of Probing, the OSR reaches 81.67 percent. For the rest of the datasets, the OSR reaches 99.25 percent. DR reaches at least 93,51% for most datasets. Only tests produce false positive alarms. In WiFi datasets, IDS detects all malicious traffic. The accuracy of IDS, however, is reduced significantly when examining the Probing data collection. DR and FPr hit 18.82% and 15.99%, respectively, of this dataset. The size and non-homogeneity of the dataset will cause this. Therefore, it should be stressed that the findings are completely unmonitored without any more detail on the existence of the network traffic information. The use of ML techniques to improve detection accuracy may thus aid anomaly-based IDSThe use of linear SVM, both two- and one-class RBF-forms, may theoretically complement the output of IDS in particular in the analysis of non-homogeneous data.

IV. ML GUIDELINES FOR ANOMALY IDENTIFICATION

Cybersecurity machine learning is implemented in three major areas: the IDS, the detection module and the detection of abuse. The detection of abnormal traffic is specifically intended for the segregation from normal traffic whilst the detection of abuses classifies the attack signature in comparison.

The Density-Based Algorithm (DBSCAN) is best used to identify anomalies. Besides the high speed processing cluster algorithms, they are simple to implement and the design parameters are less numerical. SVM is also substantially effective in detecting anomalies. The classifiers must be able to generate signatures for misuse detection. Features of Brane produce signatures appropriate for the role of the decision tree or genetic algorithm chromosomes. Therefore, ANN and SVM algorithms with hidden nodes are not suitable.

V. CONCLUSION

Very some ML algorithms were discussed in this paper with their information security applications. Finally, there Although J48 is better than other analytical algorithms, further analyses are required to evaluate algorithm efficiency because the efficiency of an algorithm appears to be biased based on the data set it is used.

VI. REFERENCES

[1] Morris, T. H., Thornton, Z., & Turnipseed, I. (n.d.). Industrial Control System Simulation and Data Logging for Intrusion Detection System Research.

[2] Nguyen, T. T. T., & Armitage, G. (2008). A survey of techniques for internet traffic classification using machine learning. Communications Surveys & Tutorials, IEEE, 10(4), 56–76. http://doi.org/10.1109/SURV.2008.080406

[3] A. Sperotto, G. Schaffrath, R. Sadre, C. Morariu, A. Pras, and B. Stiller, "An overview of IP flow-based intrusion detection," IEEE Communications Surveys & Tutorials, 12(3), 2010, pp. 343–356

[4] Almomani, A., Gupta, B. B., Atawneh, S., Meulenberg, A., & Almomani, E. (2013). A survey of phishing email filtering techniques. IEEE Communications Surveys and Tutorials, 15(4), 2070–2090. http://doi.org/10.1109/SURV.2013.030713.00020

[5] P. Garcia-Teodoro, J. Diaz-Verdejo, G. Maciá-Fernández, and E. Vázquez, "Anomaly-based network intrusion detection: Techniques, systems and challenges," Computers & security 28, no. 1, 2009, pp. 18–28

[6] M. Hall, E. Frank, J. Holmes, B. Pfahringer, P. Reutemann, and I. Witten, "The WEKA data mining software: an update," ACM SIGKDD Explorations Newsletter, 11 (1), 2009, pp. 10–18

[7] R. Core Team, "R Language Definition," 2000

[8] M. Graczyk, T. Lasota, and B. Trawinski, "Comparative analysis of premises valuation models using KEEL, RapidMiner, and WEKA," Computational Collective Intelligence. Semantic Web, Social Networks and Multiagent Systems. Springer Berlin Heidelberg, 2009, pp. 800–812

[9] Fowler, C. A., & Hammel, R. J. (2014). Converting PCAPs into Weka mineable data. 2014 IEEE/ACIS 15th International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing SNPD 2014 Proceedings http://doi.org/10.1109/SNPD.2014.6888681

[10] Buczak, A., & Guven, E. (2015). A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection. IEEE Communications Surveys & Tutorials, (1), 1–1. http://doi.org/10.1109/COMST.2015.2494502

[11] R. Lippmann, J. Haines, D. Fried, J. Korba, and K. Das, "The 1999 DARPA offline intrusion detection evaluation," Computer Networks, 34, 2000, pp. 579–595

[12] R. Lippmann, D. Fried, I. Graf, J. Haines, K. Kendall, D. McClung, D. Weber, S. Webster, D. Wyschogrod, R. Cunningham, and M. Zissman," Evaluating Intrusion Detection Systems: the 1998 DARPA Offline Intrusion Detection Evaluation," Proceedings of the DARPA Information Survivability Conference and Exposition, Institute of Electrical and Electronics Engineers (IEEE) Computer Society Press, Los Alamitos, CA, 2000, pp. 12–26

[13] M. Tavallaee, E. Bagheri, W. Lu, and A. Ghorbani, "A detailed analysis of the KDD Cup 1999 data set," Proceedings of the Second IEEE Symposium on Computational Intelligence for Security and Defence Applications, 2009

[14] F. Jemili, M. Zaghdoud, and A. Ben, "A framework for an adaptive intrusion detection system using Bayesian network," Intelligence and Security Informatics, IEEE, 2007

[15] C. Kruegel and T. Toth, "Using decision trees to improve signature- based intrusion detection," Proceedings of the 6th International Workshop on the Recent Advances in Intrusion Detection, West Lafayette, IN, 2003, pp. 173–191

[16] R. Hendry and S. J. Yang, "Intrusion signature creation via clustering anomalies," SPIE Defense and Security Symposium, International Society for Optics and Photonics, 2008

[17] J. Cannady, "Artificial neural networks for misuse detection," Proceedings of the 1998 National Information Systems Security Conference, Arlington, VA, 1998, pp. 443– 456

[18] A. Abraham, C. Grosan, and C. Martin-Vide, "Evolutionary design of intrusion detection programs," International Journal of Networks Security, 4 (3), 2007, pp. 328–339

[19] S. S. Joshi and V. V. Phoha, "Investigating hidden Markov models capabilities in anomaly detection," Proceedings of the 43rd Annual Southeast Regional Conference, Vol. 1, ACM, 2005, pp. 98–103

[20] W. Fan, M. Miller, S. Stolfo, W. Lee, and P. Chan, "Using artificial anomalies to detect unknown and known network intrusions," Knowledge and Information Systems, 6 (5), 2004, pp. 507–527

[21] Beaver, J. M., Borges-Hink, R. C., & Buckner, M. a. (2013). An Evaluation of Machine Learning Methods to Detect Malicious SCADA Communications. 2013 12th International Conference on Machine Learning and