MACHINE LEARNING IN CYBER SECURITY

[1]GEETANJALI R, [2]RAGNARE CHARLES

[1]MCA-RIT, Bangalore

[2]Arcurus Technologies

Abstract - The convention of Machine Learning procedure in cyber security has risen over the year. Preliminary from IP traffic classification, finding vulnerable systems, risk identification, intrusion detection, Machine Learning promises responses that can be operative in delivering a secure network. Lately, research shows use of statistical traffic characteristics and Machine Learning techniques. This paper is a engrossed literature survey of machine learning and its convention to cyber analytics for intrusion detection, traffic classification and email filter application. On the basis of the relevant and the quantity of citation each methods were recognized and summarized because datasets are the significant portion of the Machine Learning tactics few well know datasets are also stated. Some references are also as extensive as on when to use a certain algorithm. Several attacks have been classified using the Machine Learning algorithms and in conclusion the performance of each algorithm have been calculated.

Keywords - Machine Learning; Cyber Security; Cyber Analytics; Algorithm

I. INTRODUCTION

Some of the Machine Learning approaches are defined along with their convention in the theme of cyber security. A series of contrast standards for Machine Learning technique is seen in this paper and a series of proposals on the superlative method to use was made dependent on the possessions of the cyber security glitches. Usage of Unified Host and Network Dataset [1] is used to figure out the comparisons of algorithms. Receiver operating characteristic is often used to select optimal model and shed suboptimal autonomously from the cost content or the class distribution. Therefore, a Receiver operating characteristic curve is plotted to measure the functioning of the binary classifier. Nguyen et. al. [2] states a comprehensive analysis of IP traffic classification technique that does not bank on meticulous port numbers or recognised packet payloads. Procedures including Machine Learning with statistical traffic characteristics used in IP classification is studied in this paper. Amomani et.al [4] states an broad survey on all the key e-mail filtering and Machine Learning techniques that could be utilised to categorize and distinguish phishing emails from normal emails. The state of the art research on these attacks is numbered and a proportional study of the all those techniques have been carried out. Tedero et. al. [5] portrays a statistical, machine learning and knowledge centred tactics for network intrusion detection. It emphases primarily in the area of anomaly detection and not on signature based detection. Filtering or categorizing traffic on the fly is an significant characteristic that is researched by a lot of cyber security investigators. Speroto et. al. [3] spent Network Flow data and evidenced that the packet processing might not be conceivable at streaming speed if the quantity of network traffic is pastprecise threshold limit. These are substantial works that defines the existing works associated to Machine Learning and it's usage w.r.t cyber security. The role of the paper is to detect cyber security datasets that can be used by investigators and to aim at the algorithms that can be harnessed to cyber specific glitches. The precision of each Machine Learning algorithm in the seclusion of malicious traffic which has been analysed.

II.CYBER SECURITY DATASET FOR MACHINE LEARNING

Data is of extremely vital in Machine Learning methods. An investigator of machine learning has to fathom the data set scrupulously before doing any kind of scrutiny. Next, the raw data like packet capture , Network Flow and other network data is not unswervingly usable in the Machine Learning investigation, it needs to be pre-processed to turn it into usable in some Machine Learning tools like WEKA [6], R [7] and Rapid Miner [8]. Therefore, investigators using Machine Learning scrutiny on custom system has to understand the data collection methodology and the methods that are used in pre-processing the data. This part will lists few low level minutiae on the data sets, and few popular tools used in capturing the data from the network.

2.1 Network Packet Data

There are 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. Libpcap for UNIX and Win cap for Windows correspondingly are standard network tools. Few tools liketcpdump, wires hark are protocol analyser, packet sniffer and network monitor. The

dataset of Machine Learning has distinct characteristics and attributes. These characteristics outlines the key characteristics of each series of data in the dataset. Fowler et.al. [9] has examined the attribute relation file format of Weka and built a tool that can be used to alter any Packet Details Markup Language format to Weka minable attribute relation file format format. To convert a packet capture file to Packet Details Markup Language tools like tshark can be used.

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

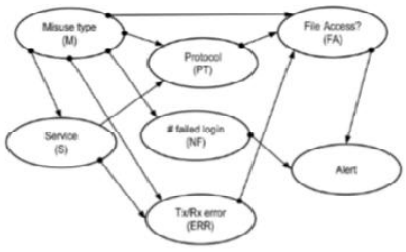
Where the input file is the .pcap file and the output file is the name of the .pdml file. Next the Fowler's tool accomplishes the final conversion, pdml2arff.py <Input file>here the input file is the name of the pdml file. This will generate an attribute relation file format format file called <Input file>.arff Fowler's concludes that for normal tcp traffic the tool performs well and successfully converts the raw data into weka usable format. A comprehensive list of packet headers of cyber security dataset of few important protocol have been enumerated by Buczak et. al [10].

III. MACHINE LEARNING TECHNIQUES FOR CYBER-SECURITY

Some popular Machine Learning techniques are defined in this section. For each method the application to cyber security have been identified.

3.1 Bayesian Network

The network is picked as a random set of variable and their qualified dependencies through a directed acyclic graph. The nodes represents the child are dependent on the parent nodes and each node holds the condition of the conditional probability form and the random variable. Fig 1 [10] displays the attack signature detection using Bayesian network. Each condition is an input to the principal condition with wavering conditional values. The assessed probability tables are gauged and depicted in the figure, it is also used to gather 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 [10]

Bayesian network is utilised for anomaly detection and recognized attack signature and patterns which could be judged with the streaming data for recognized attacks. Jemili et. al. [14] created an intrusion detection system using the Bayesian network. The KDD 1999 was utilised with 9 of its attributes to model the system where a performance of 88% was attained in normal state and 89% attack state. The model presented detection rate of 99% for probe, 21% for scan, 89% for DOS and 7% for R2L. As the number of training cases were highly truncated in case of R2L the precision of the model bore significantly.

3.2 Decision trees

The decision tree is highly similar to a tree. The trees have leaves which signifies the several classifications and the branches are the links or features which in-turn delivers the path to the classifications where ID3 and C4.5 are some standard algorithms for making decision trees spontaneous. The evaluating procedure of the SNORT rule with the incoming traffic is slow since, the enormous number of signatures. Kruegel and Toth et al. [15] substituted 150 SNORT rule with a usage of modified ID3 algorithm. Their objective here was to substitute these algorithm by a decision tree model, which would be efficient in rising the speed of processing. Rule clustering was used to supplant the Snort rules hence, lessens the number of essential comparisons, which also lets parallel evaluation hence, speeds up the assessment method. The clustered rule was harnessed to DARPA 1999 dataset. The processing speed and efficacy of the model was judged with the snort analysis. The model attained a highest speed up of 105% and the lowest speed up was of 5%. For further investigation the amount of rules replaced was intensified from 150 to 1581. Though Toth does not give any sort of quantitative figures but the study detected a intense speed up by means of the decision tree technique, secondly the processing time was shrunk drastically.

3.3 Clustering

It is an unsupervised learning method which is of same measure which is used to group data composed. Clustering algorithms is used study from audit data and explicit description of numerous attack classes by the system administrator which is not essential. Hendry et. al. [16] exhibits the purpose of real-time signature detection using clustering algorithm. The normal and anomalous network traffic was crafted by a density based clustering scheme known as Simple Log file Clustering Tool . Two clustering schemes are used: First for finding the normal and attack scenarios, then the other scheme could be used to gauge the normal traffic in a supervised method. This model uses the parameter M to identify the feature that is comprised in the cluster. When parameter M

was set to 97%, 98% attack data is detected with a 15% FAR. The signatures are created from the samples of the high density clusters of the model. The KDD dataset was used to validate the generated model. Cluster integrity was used as the performance metrics to improve the accuracy of the model. An precision of 70 % to 80 % was attained for unknown attacks. Considering the indefinite nature of the attacks this level of accuracy is highly impressive.

IV. MACHINE LEARNING RECOMMENDATIONS FOR ANOMALY DETECTION

Cyber-security requires Machine Learning in three main areas: IDS, Anomaly detection module and misuse detection. Anomaly detection is explicitly intended at separating abnormal traffic from normal one whereas misuse detection categorizes attack signature assessing it with known ones. Clustering algorithm is the best while working with anomaly detection. Though the high processing speed clustering algorithms are simple to apply and the parameters to configure are also very less. Support Vector Machine also achieves significantly well for anomaly detection. For misuse detection the classifiers has to have the competence to create signatures. Branch feature in a decision tree or chromosomes in genetic algorithm creates signatures that are appropriate for such task. Therefore algorithms like ANN and SVMs which has hidden nodes are inappropriate.

V. EVALUATION OF MACHINE LEARNING ALGORITHMS

Primary objective of this assessment is to test the applicability of few Machine Learning algorithms to detect cyber-attacks on dataset. Tenfold cross validation was applied to develop the Machine Learning models. This investigation was achieved in Weka [6]. In 10 fold cross validation Weka generates 10 different models for the dataset defined. Then the weighted average of these models are computed which is presented as the final result. Couple of standard classifier was studied for the evaluation. The approaches used were:-

1) Naive Bayes- Bayes' theorem based probabilistic classifier.
2) Random Forest-A Machine Learning based on decision tree algorithms.
3) One R- Each feature of the rule set is evaluated and finally the optimum or the best one is chosen.
4) J48- A basic implementation of C4.

5.1 Information about the Dataset

The dataset used was in Weka minable arff format. It had 31 total attributes. The features present in the dataset are Time, V1-V28, Amount and Class. Morris et. al.[1] States a complete overview about each aspects of the dataset and why each elements is significant from the standpoint of cyber security and intrusion detection. In this dataset a total of 284807 there were 35 attack was performed.

Here the attack can be roughly classified into 7 types:

1. Naive Malicious Response Injection (NMRI).
2. Complex Malicious Response Injection (CMRI)
3. Malicious State Command Injection (MSCI),
4. Malicious Parameter Command Injection (MPCI)
5. Malicious Function Code Injection (MFCI).
6. Denial of Service (DoS).
7. Reconnaissance.

The Final class in the data set has these 7 attack types along with normal traffic data. 97019 Instances were recorded in the dataset. The Distribution of the final class is enlisted below in the Table I.

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

Table I

5.2 Accuracy and ROC curve for ML Algorithm Evaluation

Beaver et. al. [21] has beforehand utilised the ICS dataset for ML algorithm analysis. Nevertheless ROC curve was not planned for any algorithm hence, it is arduous to compose the complete implementation of the algorithms. The receiver operating characteristics (ROC) curve is the plot of False Positive against the plot of Test Sensitivity in the x-axis and y-axis respectively. The area beneath the curve of the ROC is critical parameter. It's utilised to gauge1) the sensitivity is the amount of true positive decisions, 2) the specificity is known as the amount of true negative decisions. Hence, the area beneath the ROC curve is the linked measure of the sensitivity and specificity. As the area under the ROC curve is the measure of the overall performance of any test hence this parameter can be utilised to gauge the complete performance of the Machine Learning algorithms utilised in the classification of the data. Hence, an AUC analysis of the ROC curve for diverse Machine Learning algorithm will bare the classification performance of the algorithms. The Weka Knowledge flow model for the present analysis was established, which is displayed in Figure 2. The ROC curve produced for the four Machine Learning algorithms

are displayed in Figure 5. It shows from the ROC curve that it is obvious that j48 algorithm creates the maximum optimized results in general overall classification for the dataset. The AUC of the four algorithms is shown in the table below

Table III

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

Based on binary classification problem the analysis was performed but a different approach of multiclass classification could also be utilised. Then in the calculation involving the AUC, Precision and Recall shown in Table III, the weighted average of all 8 class is measured, each class can be distinctly examined which will deliver information about the model's capability to categorize each type of attack from each other and from the normal traffic. The results given in Table III shows the model's capability to categorize the traffic completely.

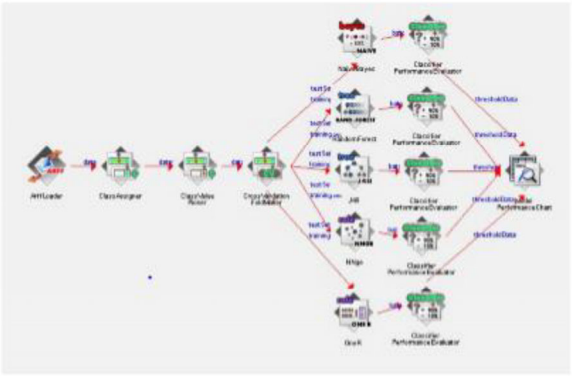


Fig 2. Weka Knowledge flow model for generating the ROC curves

From Fig. 3 it is obvious that the J48 achieves the best in the overall classification as the area beneath the curve value for the ROC curves is nearer to 1. The ROC graph produced from weka is shown in the appendix as figure 5. In ICS the implementation competence of the machine learning intrusion detection system made utilised is of maximum importance. Hence, the training must be best so that newly streamed data can be trained within a realistic amount of time and hence, the algorithm could still uphold its real-time data monitoring. Thus, when an algorithm is selected an optimal accuracy and training time pair is frequently chosen for each domain of the ICS. With the help of a python script was utilised to calculate the training time of the algorithm during the K-fold validation. The calculated training time is the time taken to authenticate all folds hence, realistically the training time can be assessed to be 1/Kth of the plotted time, which is in millisecond as portrayed in the figure 4.

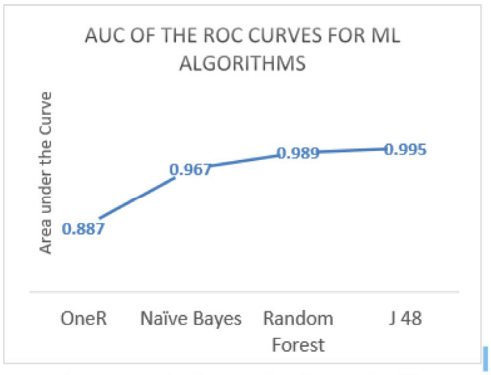


Fig. 3 Area under the curve for all 4 ML algorithm



Even though J48 could surpass Random forest in case of precision a slight negotiation on the precision can bring good real-time performance when the algorithms are applied as an integral part of the intrusion detection system.

VI. CONCLUSION

In this paper a detailed analysis was studied to point out a few standard datasets and a few Machine Learning algorithms were considered and how they can be applied in cyber-security. Few suggestions were also inferred regards the options of Machine Learning. Later part of the paper a concise study was accomplished with an ICS data set and execution of a few Machine Learning algorithm was assessed. Though J48 algorithm achieves better than other algorithms in the bounds of analysis, more analysis needs to be achieved to discover the performance of the algorithms. Random forest seems more appropriate as a principal IDS algorithm for its ideal real-time performance in the present scenario being measured.

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

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

[14] 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

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

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

[17] 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

[18] 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

[19] 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

[20] 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 Applications, 2, 54–59. http://doi.org/10.1109/ICMLA.2013.105

★★★