Cybersecurity in the Machine Learning Lens: Challenges and Solutions

Alex Vazquez
University of British Columbia
Dept. of Electrical and Computer Engineering

Abstract

Machine learning has become an industry standard for making sense of big data through correlation, however the efficiency of a machine learning model is influenced by how the data is consumed, it's labels and availability.. Throughout the pandemic we saw a rapid increases of an already alarming number of cyber-attacks. This paper will report, summarize and explore the findings of the papers "Machine Learning in Cyber-Security Threats" and "Machine Learning in Cyber-Security - Problems, Challenges and Data Sets".

1 Introduction

Cyber-Security plays a crucial role in today's world, as our world transforms into a more digital world, our information, privacy and newly digitized data is now exposed to anyone with the right tools and ill intentions. According to Stouffer et al.[3], there is a cybersecurity attack every 39 seconds and within the last year we saw a 65 percent increase in security breaches between 2014 and 2019 Stouffer et al.[3]. With this rapid increase of cyber-attacks and high number of daily attacks, Machine Learning becomes of high importance due to its ability to deal with big numbers.

Machine Learning is efficient in the detection of Cyber-Security risks compared to traditional software associated methodologies, thus reduces challenges faced by security analyst. However due to the dynamic nature of cyber-attacks it is important to acknowledge that ML in Cyber-Security involves "challenges that require elegant methodological and theoretical handling"Idan et al. [2]. Moreover, Idan et al. [2] emphasize that ML in cybersecurity lack data sets and other challenges which we will summarize in this paper.

The scope of this paper is a merge between both papers which complement one another. First we will begin by exploring the challenges and solutions that arise from using Machine Learning for Cyber-Security, specially data-sets validation and creation then we will elaborate on a proposed methodology and algorithm for ML applications in Cyber-Security and lastly we will present new challenges and results for future considerations.

2 Methods

2.1 Malware Classification and Detection

Malware can be defined as a program or file which is harmful to a computer system. The objective of the malware is to enter the computer system undetected using common evasion techniques such as metamorphism, impersonation, compression and obfuscation.You et al. [4].

Current methods for labeling and classifying malware rely on using the signatures(SHA1, MD5) of the malware to label it, then using Naive Bayes Classifier, SVM and/or Random Forest to clean

data. However, the problem relies of the nature of malware and what the attackers do to overpass such labeling. Polymorphism, for example, is the easiest way to create duplicates in a data-set, with a complete different signature. Since malware classification relies on signatures, if the malware is slightly modified, its signature is changed completely. This leads to labeling errors which become crucial when it comes to ML in Cyber-Security.

1. Identifying Malicious Programs
   Currently, malware classification relies purely on signatures and domain experts. However as explained above, it's easy for attackers to overpass signature labeling. Since singature based classification can lead to errors and domain experts are time constricted to the number of malware they can analyze. Idan et al. [2] propose labeling malware via operator domain pivoting, by grouping malware based on its contact with malicious site's. Labeling is performed in the following way.

   Let m_1, m_2 be malware.
   Let $OperatorDomains(m)$ be the unique domains with which the malware communicates.
   If $(m_1) \cap OperatorDomains(m_2) \neq \emptyset$ then m_1, m_2 belong to the same operator.
   Hence, given data set D of malware and a $OperatorDomains(m)$ they label

   $$\forall m_1, m_2 \in D, (m_1, m_2, OperatorDomains(m_1) \cap OperatorDomains(m_2) \neq \emptyset)$$

   This method results in a larger number of labels, agnostic to the content. Which provides a more robust and efficient way of classifying malware, preventing polymorphism from generating labeling errors.

2. Identifying Malicious Sources
   A threat actor will often use the same hosts with the same service on them, since they serve the same need. Using the same Operator Domain Pivot technique, Idan et al. [2] proposes that the nature of malicious sources allows to "deduce on one host using similar hosts"Idan et al. [2]. If there is threat intelligence of malicious activity in one host, it is most likely that its associated hosts are also malicious. Idan et al. [2] propose the following labeling method for sources:

   For each domain d, it is possible to resolve the ip address on which it is being hosted (nslookup), once resolved we mark the function $resolve(ip)$ and its signature as $signature(ip)$.
   Given a data set D of malware and the functions $OperatorDomains(m)$, $resolve(ip)$, $signature(ip)$ we label

   $$\forall m_1, m_2 \in D, \forall d \in OperatorDomains(m_1) \cap OperatorDomains(m_2),$$

   $$\forall ip_1, ip_2 \in resolve(d)(signature(ip_1), signature(ip_2), True)$$

2.2 Proposed Methodology

The paper Mathew et al. [1], explores the difficulties of obtaining reliable data-sets and proposes intelligent data-mining for labeling. For the scope of this paper we assume we performed the methods for labeling proposed by Idan et al. [2] as well as the provided data-sets to implement with Mathew's ML algorithms for cybersecurity.

Mathew et al. [1] outlines his proposed methodology for data pre-processing in order to convert it for relevant data for the ML. The block diagram shown in Figure one is a conceptual framework which reveals the methodical approach applied in the paper. Assume the data set in the first block from Idan et al. [2] samples an ANN using SVM applicable for Cyber-Security management with the use of Gaussian Naive Bayes and Random Forest algorithms. Idan et al. [2] emphasizes on checking missing values as an "analytical approach to ensure focus on evidence-based data for resourceful data evaluation".

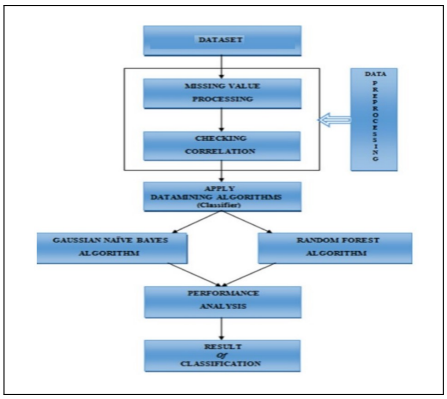


Figure 1: Block Diagram

Mathew et al. [1] states that the ML algorithms which are most applicable to Cyber-Security form a tree-like classification due to the numerous classification available and broadness as shown in Figure 2. For the purpose of the paper, Mathew et al. [1] focuses on Supervised Deep Learning, due to the nature of Cyber-Security data. It explains that Deep Learning algorithms use Deep Neural Networks (DNN) which are large neural systems organized in different layers and ways to promote autonomous depiction learning. Supervised DL algorithms, such as random forest, relate to Feed-forward Deep Neural Networks (FNN). FNN does not assume much involvement of data and offers general purpose resolution for categorization with increased computational cost, while Convolutional Feed-forward Deep Neural networks CNN receive an extra subset of neurons from subsequent layers. CNN has better cost-benefits compared to FNN. Lastly, Recurrent Deep Neural Networks (RNN), send output of previous alters with excellent sequence generators which work for long-term memory applications.

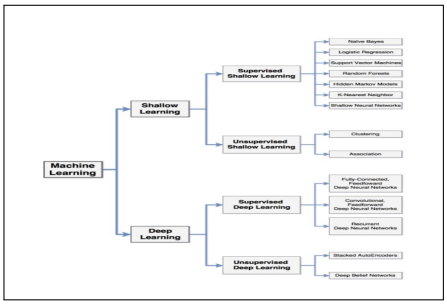


Figure 2: Machine Learning Algorithms

The Naive Bayes Model is listed on top due to its facilitation of data assessment and compatibility with various data-set. Mathew et al. [1] shows a flow-chart which illustrates in summary the proposed methodology to apply ML for cybersecurity and search flow in Figure 3. A performance matrix indicates an average class-wise accuracy with test accuracy, which shows testing accuracy for all classes, which creates an indicator for performance criterion.

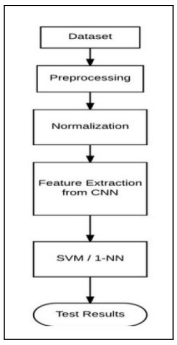
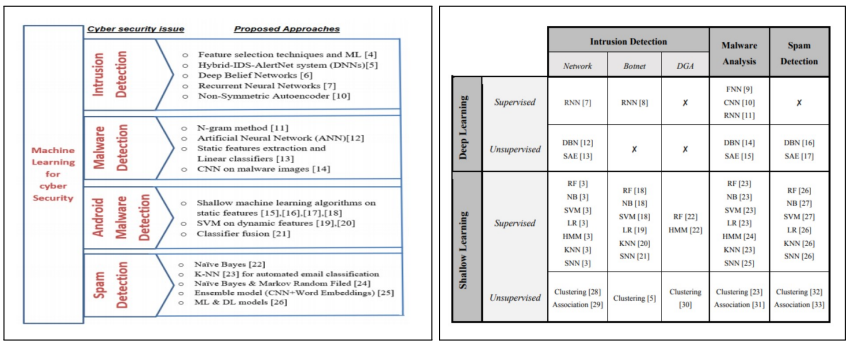


Figure 3: Flow Chart Summary

3 Discussion

3.1 Result Analysis for ML

Mathew et al. [1] results were emphasised on four major atack types of cybersecurity: malware analysis, spam detection, android malware detection and intrusion detection. The study was conducted without the labeling of Mathew et al. [1], thus identification and detection of malware were labeled by anomaly or signatures.Spam detection and intrusion were achieved by network traffic monitoring and determined by anomalies. Although the study covered all ML algorithms, data mining and pre-processing restricted the tested ML algorithms due to the difficulty of modeling such algorithms with the given data. Figure 4 shows the results found by the study, and the recommended approaches for four cybersecurity attack types.



		Intrusion Detection			Malware Analysis	Spam Detection
		Network	*Botnet*	*DGA*		
Deep Learning	*Supervised*	RNN [7]	RNN [8]	✗	FNN [9] CNN [10] RNN [11]	✗
	Unsupervised	DBN [12] SAE [13]	✗	✗	DBN [14] SAE [15]	DBN [16] SAE [17]
Shallow Learning	*Supervised*	RF [3] NB [3] SVM [3] LR [3] HMM [3] KNN [3] SNN [3]	RF [18] NB [18] SVM [18] LR [19] KNN [20] SNN [21]	RF [22] HMM [22]	RF [23] NB [23] SVM [23] LR [23] HMM [24] KNN [23] SNN [25]	RF [26] NB [27] SVM [27] LR [26] KNN [26] SNN [26]
	Unsupervised	Clustering [28] Association [29]	Clustering [5]	Clustering [30]	Clustering [23] Association [31]	Clustering [32] Association [33]

Figure 4: ML algorithms for Cyber-Security a & b

Figure A and the study's results show that DL models with features such as RNNs are most suitable for intrusion detection. ANNs and CNNs are most suitable for malware detection, it is worth noting that such samples were converted to pictures before CNNs applications.
Then shallow ML algorithms addressed Android malware detection better. Lastly, both DL models such as CNN and shallow models such as K-NN and Naive Bayes models, are the most efficient for spam detection.

Figure B summarizes the study's findings and shows how effective each ML method is for each major cybersecurity attack type in comparison to other ML methods. It suggest the most effective techniques to tackle such attack types.

3.2 Challenges

Some challenges discussed in the papers are expressed below, this challenges were summarized and only those of major importance were listed below:

1. Lack of Labeled Samples and Certainty in Ground Truth
   In Cyber-Security most tasks are supervised learning tasks, which requires cyber-security experts to determine whether the activity is malicious or benign. This can be problematic since often labels are missing for supervised learning. Manual labeling is tedious and restricted by time.
   A common approach is pooling community knowledge, such as global blacklist or white lists, however this only applies to popular entities and the information can be not trust-worthy or labeled differently. Algorithmic approaches might help such as the use of unsupervised learning specifically for anomaly detection, however with the lack of labeling and extensive plethora of anomalies, we might end up with a lot of anomalies, none which relate to the attacker. Idan et al. [2] emphasizes how promising research in active learning and semi-supervised learning will help in identifying most informative entities to label in small numbers. Given the small labeled data set, "one can apply semi supervised learning and by doing so be able to predict as in supervised learning",Idan et al. [2].
2. Imbalanced Data Sets
   In ML an imbalanced set is when in a data-set ratios between the majority set and the minority set's are large. According to Idan et al. [2] in ML a 1:10 ratio is considered imbalance, however in Cyber-Security the rarity of data sets leads to a ratio of 1:10,000 on average. This leads to an absolute imbalance in Cyber-Security data sets.
3. Data Sets
   Due to the sensitivity of information concerning cyber-attacks, most companies keep their data-sets and information confined. Although understandable Idan et al. [2] expresses the lack of suitable data sets for academic researchers. Idan et al. [2] proposes the use of anonymization, old data and removal of sensitive information in order to share without compromising data and sensitive information.

4 Conclusion

Mathew et al. [1] study findings prove that ML algorithms can be effective at classifying and predicting Cyber-Security issues. Although Mathew et al. [2 paper lacks depth of why each ML model is better at predicting specific issues than other models or showing the metrics used, the results show and even recommend specific ML models to use for each type of cyber-attack.
Idan et al. [2]. paper does a thorough analysis of addressing problems with ML in Cyber-Security and depicting incident labeling, which can be identified as a major limitation with ML in Cyber-Security. The use of Idan et al. [2] labeling via operating domain pivoting together with the proposed models and algorithms by Mathew et al. [1] could potentially form a more efficient and accurate model for applying ML in Cyber-Security, more specifically malware identification. A further study of the conjunction of both proposed methods would be of high interest due to it's potential to predict cyber-security related incidents more efficiently and potentially become a commonly adapted method for Cryber-threat prevention.

References

[1] Mathew, Alex Machine Learning in Cyber-Security Threats (January 19, 2021). ICICNIS 2020, Available at SSRN: https://ssrn.com/abstract=3769194 or http://dx.doi.org/10.2139/ssrn.3769194

[2] Amit, Idan & John Matherly & William Hewlett & Zhi Xu & Yinnon Meshi & Yigal Weinberger (2019) Machine Learning in Cyber-Security - Problems, Challenges and Data Sets.AAAI-19 Workshop on Engineering Dependable and Secure Machine Learning Systems, 2019. 1812.07858v3

[3] Stouffer, W. by C.İ15 cybersecurity statistics and trends you need to know in 2021. Norton. Retrieved March 31, 2022, from https://us.norton.com/internetsecurity-emerging-threats-cyberthreat-trends-cybersecurity-threat-review.html

[4] You, I., and Yim, K. 2010. Malware obfuscation techniques: A brief survey. In 2010 International Conference on Broadband, Wireless Computing, Communication and Applications, 297–300