Int. J. Advance Soft Compu. Appl, Vol. 13, No. 2, July 2021
Print ISSN: 2710-1274, Online ISSN: 2074-8523

Difficulties Faced and Applications of Machine Learning in Cyber-Security

Radwan M. Batyha[1], Tariq Khaled Aburashed[2], Bandar Rahil Alshammari[3]

[1] Department of Computer Science, Irbid National University, 2600 Irbid, Jordan
e-mail: rbatiha@inu.edu.jo
[2]Department of Computer Engineering, Saudi Electronic University, Riyadh, Saudi Arabia
e-mail: abuzead2001@yahoo.co.uk
[3]Department of Computer Science, Imam Muhammad ibn Saud Islamic University, Riyadh, Saudi Arabia
e-mail: Bandar.Alremali@hotmail.com

Abstract

Over time, internet technologies like social networks, cloud computing and web technologies have shown remarkable progress therefore the need for securing these platforms has also intensified. Machine learning has displayed eminence over the conventional rule-based algorithms. This paper is a literature survey of scrutinizing security attributes of protocols and replacing the first level of security analysis with superior machine learning algorithms. Although Machine Learning is one of the most promising answers against cyber-crimes and contemporary research is being undertaken by the use of machine learning techniques, the effectiveness of machine learning in cybersecurity must be evaluated. This paper is an attempt to identify the hindrances faced in applying these machine learning techniques practically.

Keywords: Machine Learning, Cybersecurity Network, Security Algorithms.

1 Introduction

Usability of Machine Learning is, in every field intelligible, ever-growing and extant techniques being continuously enhanced and have evolved to the point where they can address real-world scenarios with high precision therefore leading to the adoption of these ML techniques in several provinces. In some cases, usage of

Machine Learning is prioritized over traditional rule based methods and up to an extent, even over humans. This trend is also influencing the field of Cyber-Security where upgradation of detection systems (like intrusion detection systems etc.) is being done with ML modules [1]. Old-fashioned software requires a lot of individual time and energy to identify hazards. This onerous course can be made more proficient by applying Machine Learning algorithms and techniques. As an outcome, a number of researchers have probed various Machine Learning practices to identify attacks more time-effectively and dependably. Our paper is based on rigorous literature review and theoretical findings. Other academic papers compare ML methods for cyber-security by considering one explicit application (e.g.: [1], [2], [3]) and are stereotypically more concerned with Artificial Intelligence(AI) methods rather than with security operations.

This study is aimed at researchers looking for a head start in the field of Machine Learning in cybersecurity. It provides an insight of machine learning applications in Cybersecurity and how Cybersecurity issues are handled by Machine Learning by referencing noticeable works and citing examples from the works of prominent researchers. We point out an overall underestimation of intricacy of supervising Machine Learning architecture in Cybersecurity caused by deficiency of data available for training.

With due recognition interest of the researchers and experience, we decided to conduct a literary survey on “Problems Faced and Applications of Machine Learning in Cyber-Security.”.

2 Machine Learning Models

A. Decision Tree

Decision Trees are a hierarchical Model composed of internal decision nodes and terminal leaves. C4.5 and ID3 are few well favored technologies for initiating decision trees spontaneously. Intrusion Detection Systems (IDS’s) are built to fortify network security and to detect malicious activities. One of the cores of IDS is Pattern Matching. Misuse detection can make IDS have a passive detection of known attacks. An open source network intrusion detection system (NIDS) SNORT monitors network traffic in real time rules with new traffic is time consuming since there are an enormous number of signatures. Toth and Kruegel et al [4] substituted 150 SNORT rules by utilizing a variant of ID3 algorithm later to be replaced by Decision Tree as an efficacious measure to elevate the speed of processing. This model then attained a best case speed of 105% and a worst case speed of 5%. After further experiments on this model the maximum speed was increased from 150 to 1581.

B. Bayesian Network

A Network of a turbulent group of variables and their subject dependencies via a Directed Acyclic Graph (DAG) has evolved. Child nodes rely on the parent nodes and every single node commemorates the states of the probability configuration.

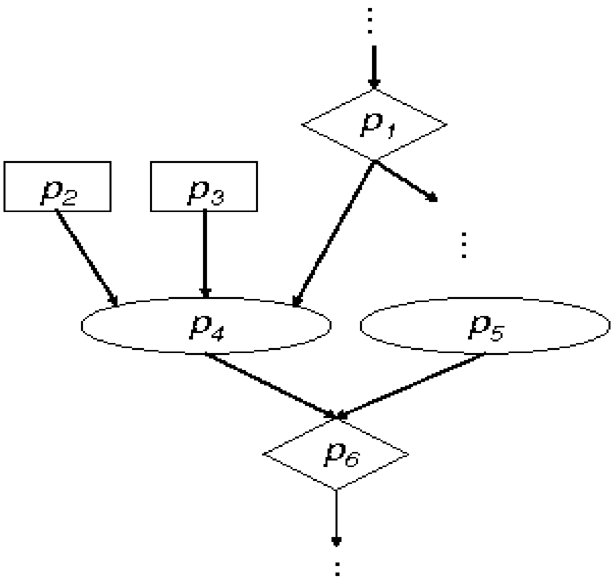

Fig.1: shows the relationship of child and parent node

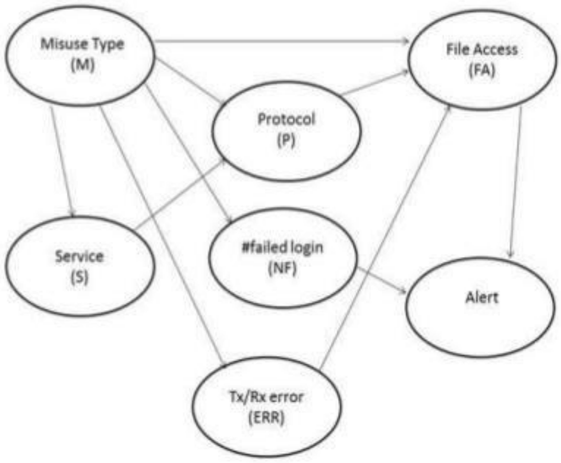

Fig. 2: shows different types of attack detection with the help of Bayesian Network.

C. Clustering

Clustering methods allows learning from mixture parameters of data. Being an unsupervised Machine Learning technique, clustering algorithms can create models from unlabeled data therefore the requirement of a description of the data is nullified. There are various clustering procedures: K-means Clustering, Spectral Clustering, Hierarchical Clustering etc.

Hendry et.al [5] illustrated one experiments based on Clustering real time signature detection. Where traffic generation (both anomalous and normal) was done using Simple Log-file Clustering Tool (SLCT) which is a clustering scheme based on the density. There were two clustering methods used: First, for differentiating between the normal and attack scenarios and the second is used to identify and cluster the normal data-flow through the network. The model uses M as a framework in this model to annotate the attributes contained in the cluster. A 15% FAR in attack data is obtained by setting the value of M to 97-98% with. The KDD set was used to endorse the model. Using this model for anonymous attacks, a 70-80% accuracy was attained which if not practically feasible, is still considerably stunning.

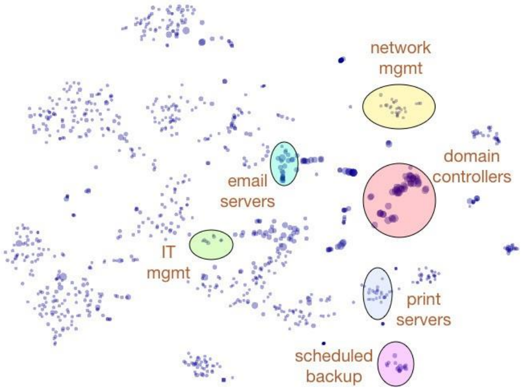

Fig. 3: Visualization of model-based clustering when trained on network connection relationships [6].

D. Artificial Neural Networks (ANN)

Artificial Neural networks consist of millions of artificial cells that mainly behave like human brain cells. Every node has a unique activation level and to activate these nodes each must uphold some set of axioms.

Artificial neural networks have two operating domains: when data is provided to the ANN to develop its experience and learn from the data, the ANN is in learning mode and when the experience gained from the learning part is applied to unlabeled data, the ANN is said to be in operating mode. ANN's utilize back-propagation which is a method to provide feedback to the previous nodes to tune the output towards intended results. This usage of experience and feedback by ANN's is analogous with humans using past experiences for decision making.

Neural networks are often used for security purposes. For instance, a bank processing thousands of credit card transactions may need an automated method of recognizing unethical transactions [7]. With enough pre-programmed unethical activity inputs and clues, an artificial neural network would be able to distinguish any dubious activity swiftly.

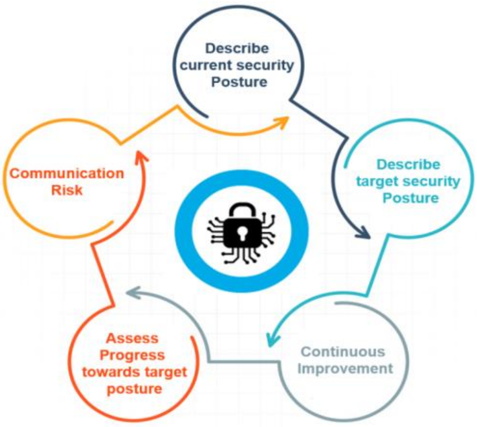
Describe current security Posture
Describe target security Posture
Continuous Improvement
Assess Progress towards target posture
Communication Risk

Fig. 4

E. Hidden Markov Model (HMM)

Hidden Markov Model along with a set of orders is a statistical representation which are reticulated with the help of transition probabilities that stimulates the analysis situs of the prototype. The model consists of unnoticed parameters. The probability distribution varies in individual order, is capable of reconstructing states work time and is adept to exemplifying transient flow [8]. Multiplicity of approaches have been scrutinized to spot bizarre phenomenon from standard ones along with statistical representation and Hidden Markov Model. The multi-layer preposition initiated the advancement and affirmation to overcome the typical flaws in the utilization of Hidden Markov Model (HMM) to IDS (Intrusion Detection System) is often quoted as the "curse of dimensionality". Lack of prosaic technique for the rephrasing of the observed network packet data into significant Markov Model is one of the major challenges faced while applying the Markov Technique to IDS. [9].

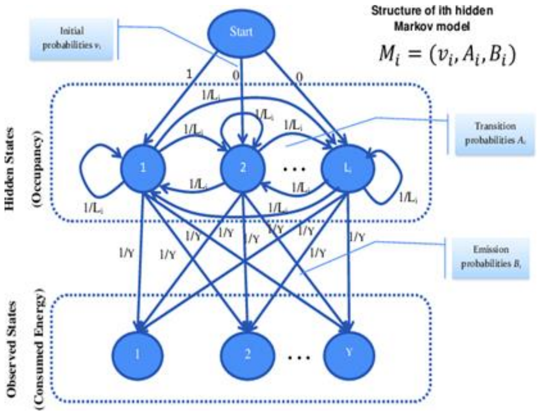
Structure of ith hidden Markov model
$M_i = (v_i, A_i, B_i)$
Initial probabilities v_i
Start
1
0
0
$1/L_i$
Transition probabilities A_i
Hidden States (Occupancy)
1
2
...
L_i
Emission probabilities B_i
1/Y
Observed States (Consumed Energy)
1
2
...
Y

Fig. 5: Hidden Markov Model components and internal structure after initialization (before training) [9].

F. Cryptography

The process of transmitting and receiving sensitive and classified data in an encrypted form such that only scepter person can access the protected data.

The constituents of Cryptography are:

- Plain text
- Cipher text
- Encryption and Decryption Algorithms
- Keys

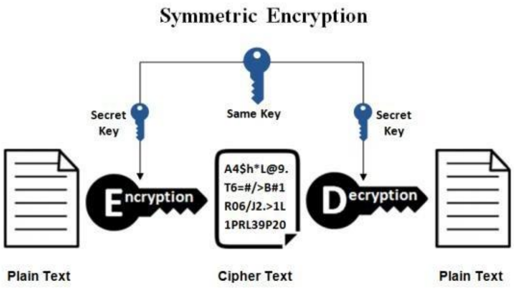

Fig. 6: Various components of Cryptography.

To surpass the incursion on networks, different methods are often used promptly. Some of them are [10]: -

• Authentication: - Authenticity and originality of all data and documents must be checked in case the source is not trusted and potentially malicious.

• Antivirus: - Antivirus software must be frequently updated.

• Firewalls: - Firewalls help in keeping a record of incoming and outgoing traffic of the system and works as an adviser whether the access and usage is safe or not.

• Access Control: - Authenticity via username and password so that only legitimate users may log-in and access data.

• Cryptography: - Process of transmitting and receiving classified data in an encrypted form such that only scepter people can access the data.

3 Problem Description

Provided the abundance of the aforementioned and other machine learning techniques to tackle cybersecurity issues currently faced, applying machine learning solutions although seems like a trivial task, is actually an intricate one. There exist if not a plethora of but still crucial obstacles in the way of successfully integrating Machine Learning solutions and network security.

A. Polymorphic Software

Malwares have long been playing a crucial role in the cybersecurity era and are often the precipitants of major cyber-attacks. Smuggling malware into the victim computers is a demanding task at the malicious attacker’s end, therefore leading to the evolution of the attackers’ approach and the development of polymorphic malwares capable of changing characteristics so as to bypass detection.

Traditional malware detection systems and threat intelligence repositories are based on malware signatures (SHA1, MD5 etc.) [11], therefore it is fair to infer that a malware capable of changing its signature is automatically capable of hoodwinking the malware detector system [12].

B. Similarity Between Hosts and Identifying True Sources

Hosts in cybersecurity jargon are the sources of the attacks. One crucial task in countering attacks on a network is the identification of the host so that the nature and motivation behind the attack can be traced back to its source and future attacks of similar kind are averted. When searching for hosts, it can sometimes be uncovered that several potential host machines have the same or almost the same configurations. It can also be concluded that a malicious attacker using several hosts for attacks will have the same services on all of the machines since the need is the same. Therefore, if one host machine is detected, then all other machines in the search domain having the same services automatically become potential hosts. Attributing of hosts has already been done [13] and also done using the agnostic label [11]. labelling of hosts via operator domain pivoting is yet another method to label similar hosts.

Similar to the problem of identifying similar hosts present in a domain is identifying malicious attackers using Virtual Private Network (VPN) systems which allow users to route their traffic through virtual servers located offsite, even off country for the sake of the users' privacy and security. Though the VPN service providers cannot be blamed for malicious use of their services, it is still extant as a major host tracing problem in the current network-security scenario.

C. The Game

The events of network-security can also be inferred as a game of cat and dog where both the attacker and the victim have to play both roles depending on what the circumstances demand. Both parties are involved in a power-struggle over dominance over the network-security space by continually advancing and improving the attack and defense mechanisms respectively and although it is also common sense that improvements in the mechanisms of one faction will lead to its triumph over the cyber-security domain, the opposite is as a matter of fact true, meaning advancements in one faction leads to the game advancing to another level with a higher difficulty setting. Citing Slick Willie Sutton: "I rob banks because that's where the money is" [11]; this cat and dog game is perennial.

The game seems to be commensurate for both the parties (the attacker and the defender) from an observer point of view however the on-field facts debunk the belief. There exist dissimilarities in the scope of the mechanisms of the factions. The defenders have to have a prescience of the attackers' modus operandi and a system has to be deployed to protect against potential tactics used by the attackers. The attackers can gain advantage by acquainting the defense system of the victim and prepare the steps of the attack likewise. The attacker also has the edge of choosing crucial parameters of the attack like the time and date of the attack, which, if calculated carefully enough provide enough incentive to the attacker to catch the defense system at its most vulnerable moment. Citing the asymmetry in this case, the defender has the advantage of deploying honeypots [14] for the attacker to fall right into.

The attacker might as well be a malicious insider having access to the logic of the defense mechanisms but it can also prove counterproductive for the attacker if the defender has honeypots deployed of which the insiders are oblivious to. The existence of asymmetries allows the application of Game Theory to network-security [15], [16].

The defending community has the incentive of sharing details about any and all potential malwares to cyber-security forums and threat intelligence repositories like www.virustotal.com [11].

D. Data Set Imbalance

Since Machine Learning is the key concern of this paper, the mention of datasets is impending. A major challenge faced by researchers and network-security experts in developing and training models for practical applications is the absence of labelled datasets. The data required although need not be completely labelled, does need a significant percentage of labelled data to be considered balanced.

An imbalanced dataset is when the ratios between the majority and minority sets are large [17]. The Machine Learning community considers an imbalance ratio of 1 to 10 for a dataset to be called imbalanced however it is uncommon for cyber security datasets to have imbalance ratios of 1 to 10,000.

The terms relative imbalance and absolute imbalance are used to explain the chasm between the two types of imbalances found commonly in cyber-security datasets where relative imbalance refers to datasets having large ratios but enough samples in each set and absolute imbalance refers to datasets having large imbalance ratios and not enough samples in each set.

E. Domain Diversity

Clearly the cyber world is omnipresent and there is no domain in the physical world which remains unscathed from its influence. This expanse has led to the presence of diverse cultures, scales and domains and the cyber world looks different in each of them. A scenario where the test environment and the training environment differ is referred as domain adaptation [18] and developing Machine Learning models which can perform in multiple domains without losing efficiency is another challenge yet to be overcome.

4 Conclusion

This paper contains the findings of an extensive literature survey kindred to the applications and usability of machine learning in the cyber-security domain and the snags experienced in real world applications of proposed Machine Learning solutions given the recent advancements in Machine Learning. A list of Machine

Learning models apt for use in cyber-security were discussed and in the latter part, an in depth analysis of the problems faced was discussed inclusive of the expanse of the domain in which the problem exists and a major problem recognized was the absence of datasets convenient for training Machine Learning models so as to render them applicable. Also it is noteworthy how the incessant battle between the malicious attacker and the defender has always led to the development of improved infiltrating and defensive mechanisms which serves as a reminder for potential victims to continually upgrade their defenses.

References

[1] Buczak, A. L., & Guven, E. (2015). A survey of data mining and machine learning methods for cyber security intrusion detection. IEEE Communications surveys & tutorials, 18(2), 1153-1176.

[2] Blanzieri, E., & Bryl, A. (2008). A survey of learning-based techniques of email spam filtering. Artificial Intelligence Review, 29(1), 63-92.

[3] Gardiner, J., & Nagaraja, S. (2016). On the security of machine learning in malware c&c detection: A survey. ACM Computing Surveys (CSUR), 49(3), 1-39.

[4] Kruegel, C., & Toth, T. (2003, September). Using decision trees to improve signature-based intrusion detection. In International Workshop on Recent Advances in Intrusion Detection (pp. 173-191). Springer, Berlin, Heidelberg.

[5] Hendry, G. R., & Yang, S. J. (2008, March). Intrusion signature creation via clustering anomalies. In Data Mining, Intrusion Detection, Information Assurance, and Data Networks Security 2008 (Vol. 6973, p. 69730C). International Society for Optics and Photonics.

[6] Brain Jones. (2016, March). ”Threat Research”, Available at: https://www.fireeye.com/blog/threat-research/2016/03/relational_learning.html.

[7] An illustration of ANN in Cyber security, Available at: https://www.forcepoint.com/cyber-edu/neural-network

[8] Das, R., & Morris, T. H. (2017, December). Machine learning and cyber security. In 2017 International Conference on Computer, Electrical & Communication Engineering (ICCECE) (pp. 1-7). IEEE.

[9] Ullah, I., Ahmad, R., & Kim, D. (2018). A prediction mechanism of energy consumption in residential buildings using hidden markov model. Energies, 11(2), 358.

[10] Shree, D., & Ahlawat, S. (2017). A Review on Cryptography, Attacks and Cyber Security. International Journal of Advanced Research in Computer Science, 8(5).

[11] Amit, I., Matherly, J., Hewlett, W., Xu, Z., Meshi, Y., & Weinberger, Y. (2018). Machine learning in cyber-security-problems, challenges and data sets. arXiv preprint arXiv:1812.07858.

[12] You, I., & Yim, K. (2010, November). Malware obfuscation techniques: A brief survey. In 2010 International conference on broadband, wireless computing, communication and applications (pp. 297-300). IEEE. [13] Starov, O., Zhou, Y., Zhang, X., Miramirkhani, N., & Nikiforakis, N. (2018, April). Betrayed by your dashboard: Discovering malicious campaigns via web analytics. In Proceedings of the 2018 World Wide Web Conference (pp. 227-236).

[14] Spitzner, L. (2003, December). Honeypots: Catching the insider threat. In 19th Annual Computer Security Applications Conference, 2003. Proceedings. (pp. 170-179). IEEE.

[15] Manshaei, M. H., Zhu, Q., Alpcan, T., Bacşar, T., & Hubaux, J. P. (2013). Game theory meets network security and privacy. ACM Computing Surveys (CSUR), 45(3), 1-39.

[16] Roy, S., Ellis, C., Shiva, S., Dasgupta, D., Shandilya, V., & Wu, Q. (2010, January). A survey of game theory as applied to network security. In 2010 43rd Hawaii International Conference on System Sciences (pp. 1-10). IEEE.

[17] Chawla, N. V., Japkowicz, N., & Kotcz, A. (2004). Special issue on learning from imbalanced data sets. ACM SIGKDD explorations newsletter, 6(1), 1-6.

[18] Daumé III, H. (2009). Frustratingly easy domain adaptation. arXiv preprint arXiv:0907.1815.