Application of Machine Learning and Deep Learning in Cybercrime Prevention – A Study

Maya Kumari M
MSc CS, Indian Academy Degree College, Bangalore, India

Abstract—Threat of cyber security has drastically transformed over the past few years due to large amount of data produced every second and stored in various forms and in different platforms, which makes it necessary to develop techniques to curbe attacks and theft of critical information. This gives rise to Machine learning and Deep learning techniques to shine and show their power to escape cyber attacks. In this paper, the various attacks and techniques are discussed to prevent cyber crimes.

Keywords—*AI, Machine Lachine learning , Cyber Security, Cyber Crime.*

I. INTRODUCTION

Cybercrime refers to a crime that is put in effect with the involvement of computer systems and networks. The computer system can either be used as an attacker or a target.

Cybercrimes can be defined as: "Offences that are committed against individuals or groups of individuals with a criminal motive to intentionally harm the reputation of the victim or cause physical or mental harm, or loss, to the victim directly or indirectly, using modern telecommunication networks such as Internet (networks including chat rooms, emails, notice boards and groups) and mobile phones (Bluetooth/SMS/MMS)".[1]

There are several examples of Cybercrime , such as : online transaction theft, identity theft , unauthorized system access , hacking , virus diffusion , malware distribution , DoS (Denial of Service), software piracy , Email bombing and Spamming. Cyberterrorism remains the most consequential cybercrime of significant concern.

Collectively, these types of criminal activities are considered as major interest, and combating cybercrime has become a matter of great importance for both corporations and Governments.

Financial data, health records, personal identifiable information (PII), intellectual property and basically any valuable data are subject to attacks. Cybercriminals employ highly profitable strategies like disrupting business operations via DDoS attacks or monetizing data access by advanced ransomware techniques.[4]

The few most popular ways of stealing data consists of:

- **Phishing** - an attempt to obtain sensitive information such as credit card details , usernames , passwords etc , from a user by disguising as a trustworthy entity or by prompting the user to enter details at a fake website .
- **Skimming** - the stealing of payment information or hiding tax details from the tax authorities.
- **Malware** - computer viruses designed and introduced to intentionally cause harm to a target computer user, which takes forms of Trojan horses, spyware, adware, worms etc.

Cybercrime bounds an extensive range of affairs, these can be broadly classified as follows:

- Computer devices or networks targeted crimes. Example : viruses , malwares and Denial-of-Service (DoS)
- Crimes enabled with computer network aid to boost other criminal tasks. Example: Phishing, identity theft, CyberHunting..

II. USE OF MACHINE LEARNING IN CYBER CRIME PREVENTION

Machine Learning is a part of Data Science where the machines are designed in such a way that they learn and improve themselves from their previous experiences and the availability of huge amounts of data. This reduces human involvement exponentially. Machines can improve by learning from the enormous and extensive analytics performed over datasets .Hence, more the data, more intelligent the machine. This ability to compute large volumes of data leads to exponential growth in the applications of Artificial Intelligence (AI) and Machine Learning (ML), Algorithms are generally categorized as Supervised Learning, Unsupervised Learning and Reinforcement Learning. Supervised Learning is further divided into Classification and Regression. The goal of supervised learning is to train the computer to learn to predict a value or classify an input instance accurately. Unsupervised learning is used when a labeled dataset is not available.

Clustering is an unsupervised learning technique which results in grouping similar instances in clusters. Clustering is used to discover patterns in data. In some cases, clustering is performed to classify an unlabeled dataset and using the resulting classified dataset for supervised learning [2]. Commonly used ML algorithms are: Linear Regression, Decision Tree, Support Vector Machine, K-Nearest Neighbor, k-means, Logistic Regression, Ransom Forest.

Machine learning can be leveraged in various domains of cyber security to provide analytical based approaches for attack detection and response. It can also enhance security processes by automating routine tasks and making it easy for security analysts to quickly work with semi-automated tasks.[3]

The following section focuses on various applications of ML in Cyber Crime prevention:

A. Phishing Detection

The (Phishing aims at acquiring sensitive and personal information in a fraudulent manner. There are three principal groups of anti-phishing methods that are formed based on the research as follows: Detective, Preventive and Corrective.

Federations can detect these risks by acquiring descriptive data from emails, without trading off users' confidentiality. Machine learning algorithms can learn to analyze patterns that divulge malignant senders' emails by surveying the email headers and a part of body data. By acquiring and categorizing patterns, Machine Learning models can be trained to identify if a phishing attack is attempted.

Vitally Ford et al [6] emphasizes that the researchers compared six machine learning classifiers, using 1,171 raw phishing emails and 1,718 legitimate emails, – "Logistic Regression (LR), Classification and Regression Trees (CART), Bayesian Additive Regression Trees (BART), Support Vector Machines (SVM), Random Forests (RF), and Neural Networks (NNets)". For experimentation purpose, emails were parsed using text indexing techniques. Email headers were considered along with the html tags from the body part. Subsequently, all the peripheral words were discarded by exercising a stemming algorithm. Eventually, based on their frequency in emails, all items were classified. In conclusion, it can be said that LR is a more preferred option among users due to low false positive .Also, LR has the highest precision and relatively high recall in comparison with other classifiers under contemplation. The comparison of precision, recall, and F-measure is given in Table 1.

Table 1: Comparison of precision, recall and F1 [7]

Classifier	Precision	Recall	F1
LR	95.11%	82.96%	88.59%
CART	92.32%	87.07%	89.59%
SVM	92.08%	82.74%	87.07%
NNet	94.15%	78.28%	85.45%
BART	94.18%	81.08%	87.09%
RF	91.91%	88.88%	90.24%

B. Network Intrusion Detection

Network Intrusion Detection systems are used to recognize malignant or harmful network activities that activate violation of Confidentiality, Integrity or Availability of the systems in a network. Machine Learning techniques are strong grounds for the development of many Network Intrusion Detection (NID) systems due to their variability to mysterious attacks.

In [6, 8], the Researcher points out a unified effective solution for improving Genetic Network Programming (GNP) for the misuse and anomaly detection. Identical degree and genetic algorithm were combined in order to eliminate unnecessary axioms and process necessary ones. The researcher indicates that the system was tested on KDDcup99 [6, 9] data to indicate its coherence. The suggested pruning algorithm does not stand in need for a prior knowledge from experience. The axiom or the Rule is eliminated if the average matching degree is less than some standard.

Table 2: Comparison of NID systems [6,8]

NID	Detectio n Rate	ACC	FP	FN
Unified detection (with two-stage rule pruning)	97.75%	94.91%	2.01%	2.05%
Unified detection (without two-stage rule pruning)	95.79%	90.17%	4.41%	3.75%
GNP-based anomaly detection	86.89%	------	18.4%	0.75%
GNP-based misuse detection	94.71%	------	3.95%	8.54%
Genetic programming	90.83%	------	0.68%	------
Decision trees	------	89.70%	------	------
Support Vector Machines	95.5%	------	1.0%	------

On the training step, 8,068 randomly chosen connections were fed into their system (4,116 were normal, 3,952 – smurf and neptune attacks). After training the system, the proposed solution was tested on 4,068 normal connections and 4,000 intrusion connections. The accuracy (ACC) is reported to be 94.91%, false positive rate (FP) is 2.01%, and false negative rate (FN) is 2.05%. The below table displays the performance comparison of different algorithms including the proposed one.

C. Social Network Spam Detection

Conventional way of spam detection is usage of rules known as knowledge engineering [6, 11]. Ford and Siraj [6] discuss the observations of K.Lee et al [10] that spammers misuse social systems in order to employ phishing attacks, distributing software intended to harm the system such as malware, spyware, etc., and encouraging associate websites. To detect spammers in social systems like Twitter and Facebook, a social Honeypot was developed. The proposed solution is based on Support Vector Machine (SVM) and has a high precision as well as low false positive rate. A social honeypot is a mechanism for computer security that detects attempt at unauthorized use of information systems. A honeypot consists of data that appears to be an authorized part of the site but, in point of fact, is secluded and observed, and that seems to contain information of importance to attackers, who are then blocked. These honeypots collect valid and spam profiles and feeds that data into the SVM (Support Vector Machine) classifier.

MySpace and Twitter were examined by the researchers [10] in order to evaluate the performance of the proposed system. Certain valid user accounts were created in social networks so as to collect data over a few months. Ambiguous spam profiles, like click traps, friend infiltrators, duplicate spammers, promoters, and phishers were manually singled out into several groups. The SVM classifier was fed with the data (for MySpace: 388 legitimate profiles and 627 ambiguous spam profiles; for Twitter: 104 legitimate profiles, 61 spammers' and 107 promoters' profiles). Results demonstrate spam precision to be 70% for MySpace and 82% for Twitter.

D. Malware Detection

Malware refers to malignant software viruses that infect computers or an entire computer network. It abuses target system's susceptibilities, such as a bug in authorized software that can be hijacked. Over the last few years, traditional anti- malware companies have stiff competition from new generation of endpoint security vendors that major on machine learning as a method of threat detection [11, 12]. With the help of Machine Learning, machines are taught to detect threats. With this knowledge, machines can analyze and predict future threats.

E. Lateral Movement

Lateral Movement attack vectors represent an attacker's movement across a network looking for susceptibilities and when found, misusing them. An attacker's transformation from survey to data retrieval is specifically suggestive of threat growth along the kill chain. Especially, when the attacker moves from low-level users' machines to more valuable data personnel.

Network traffic input records tell about users' interactions with a website. Machine learning-informed examining of this data can offer a dynamic view of normal traffic data. With a better understanding of typical traffic flow, algorithms can perform change-point detection (i.e., they can identify instances when the probability distribution of a given traffic pattern changes and becomes unlikely based on "normal" traffic activity) to detect potential threats.

F. Ransomware

Ransomware is a kind of malignant application from the cryptovirology that jeopardizes to broadcast the prey's details or hinder approach to it unless a payment is made to the perpetuator. Simple ransomware techniques are effortlessly reversible, whereas, there are some advanced malwares such as cryptoviral blackmail under which victim's data is encrypted and is made unobtainable and claims for a payment to decrypt it.

Appropriately executed extortion attack results in an unmanageable issue where reclaiming the data without decryption key is difficult and digital currencies are accustomed for the ransoms, causing tracing and charging the agents difficult.

Ford et al [6] concludes that Machine Learning is a constructive tool that can be utilized in many fields of cyber security. There exist some robust anti-phishing algorithms and network intrusion detection systems. The machine learning classifiers themselves are liable to harmful attacks despite the fact that machine learning helps keeping various systems secure. There are many opportunities in information security to apply machine learning to address various challenges in such complex domain.

According to [13] , the safety of computer systems from high-tech cyber-attacks is one of the major concerns for national and international security. Artificial intelligence and Machine Learning play a noteworthy role in protection of computer systems. Numerous researches have been organized using certain datasets. Yavanoglu et al [13] observes comprehensive class of various datasets together with their advantages and disadvantages.

The researcher [3] remarks that machine learning is a robust tool that can be used for automating complex defense cyber activities and threats. The researcher holds the position for the basis of future research that can focus on scrutinizing existing security solutions and the various challenges influencing machine learning to develop and deploy extensible cybersecurity systems in production environments.

[14] Suggests that exclusive or intimate facts of an individual along with the crucial framework is liable to numerous cybercrimes. Likewise situations, Machine Learning is serving mankind efficiently in focusing on the issues of cyber security considering its intelligent kind and adaptability. Scholarly assets have exhibited that there are several applications of Machine learning which assist the human to defend against rigorous cyber-attacks, furthermore constituting the phenomenon that plenty more is yet to be researched in the pot of opportunities of Machine learning.

The researcher [14] precisely conferred the progress formed in exercising numerous procedures of Artificial intelligence, their present position and the area of subsequent work.

III. MACHINE LEARNING IN APPLICATION SECURITY

Sangani and HarootZarger [15] comments on the high dependence of a country's safety and monetary progress on a protected cyber space.

According to them, the attackers keep an eye on the susceptible loopholes in order to sneak crucial information, data, currency, having facilities interrupted etc. In this new age, many web applications are abused by the new innovations of the hackers. But the war between the owners and the attackers has compelled the company owners to evolve and automate the identification techniques to forecast the attempts and breaches in their application.

Taking into account the prediction of attacks, [15] remarks that Machine learning (ML) has greatly evolved in the entire cyber space. With time, Machine learning has proved itself as a simple problem solver for any kind of problems in the space where ML follows a simple rule of prediction, and is adapted in many industries based on its ability to provide security.

The capability to identify breaches and prevent attacks, based on the analysis of the use case patterns, has led to the integration of ML in many web applications.

The researchers [15] concluded that ML has been successful in identifying the attacks and furthermore, research has been carried out for the same. And also adds that the security of the web applications lies in the hands of ML in the coming days where huge data are being stored every millisecond and hackers are on the lookout for a loophole.

A peculiar technique to Machine learning (ML) is Deep learning. This is a Neural Network based subfield of ML. Artificial Neural Network (ANN) has been successful in the recent times due to its composition of many layers that are stimulated on biological neural systems, and are used to assess functions that can rely on enormous inputs. ANN is enabled to adapt and learn from its previous data.

Deep learning is efficient compared to other ML techniques because of its advantages of starting up with the raw data leading to the elimination of the most time consuming part of ML - feature engineering. Large number of training data is required to be fed in the classifier in order to detect patterns in the decrypted malware. The stability of Deep learning depends on the huge data that helps in the training of the classifier. Other drawback of Deep learning is the highly overpriced systems that are hard to train. On the other hand, it takes a lot of time to learn from the fed data and forms clusters. Deep learning models are slow at adapting to new malwares that leaves the user susceptible to different malwares.

IV. HOW CAN ML CHANGE THE STATE OF CYBERSECURITY

Considering the condition of the implementation of ML systems can be a real game changer. Certainly, 52% of cyber professionals presume that current systems are not precise enough. However, these systems are benefited with abilities that will empower cybersecurity professionals with abundant possibilities to defend against cyber-attacks and secure their company. Today, techniques, like machine learning and deep learning, are probable because of more effective algorithms and the large amounts of available data.

CONCLUSION

In this paper, we have seen various attacks threatening the cyber space, and different techniques to deal with the same. Machine learning and Deep learning techniques have given enterprises the capability to ensure the safety of its data. Furthermore, these techniques are themselves vulnerable to various malignant attacks.

References

[1] Halder, D., &Jaishankar, K. (2011) Cybercrime and the Victimization of Women: Laws, Rights, and Regulations. Hershey, PA, USA: IGI

Global. ISBN 978-1-60960-830-9

[2] Lambert, Glenn M. II, "Security Analytics: Using Deep Learning to Detect Cyber Attacks" (2017). UNF Graduate Theses and Dissertations. 728, https://digitalcommons.unf.edu/etd/728

[3] ManjeetRege& Raymond Blanch K. Mbah, Machine Learning for Cyber Defense and Attack , DATA ANALYTICS 2018 : The Seventh International Conference on Data Analytics, Copyright (c) IARIA, 2018. ISBN: 978-1-61208-681-1 , pp.73–78.

[4] Dmitri Koteshov, How Can Ai Change The State Of Cybersecurity, March 7, 2018, https://www.elinext.com/industries/financial/trends/ai-and-security/

[5] Anti-Phishing Working Group, “Phishing and Fraud solutions”. [Online], [Accesses: March 18, 2019] http://www.antiphishing.org/

[6] S. Abu-Nimeh, D. Nappa, X. Wang, and S. Nair, “A Comparison of Machine Learning Techniques for Phishing Detection”, APWG eCrime Researchers Summit, October 4-5, 2007, Pittsburg, PA

[7] N. Lu, S. Mabu, T. Wang, and K. Hirasawa, “An Efficient Class Association Rule-Pruning Method for Unified Intrusion Detection System using Genetic Algorithm”, in IEEJ Transactions on Electrical and Electronic Engineering, Vol. 8, Issue 2, pp. 164 – 172, January 2, 2013.

[8] Knowledge Discovery and Data Mining group, “KDD cup 1999” [Online], [Accessed:March 18, 2019], http://www.kdd.org/kddcup/index.php

[9] K. Lee, J. Caverlee, and S. Webb, “Uncovering social spammers: social honeypots + machine learning”, SIGIR’10, July 19-23, 2010, Geneva, Switzerland.

[10] Nilaykumar Kiran Sangani & HarootZarger, Machine Learning in Application Security, [Accessed: March 18, 2019] http://dx.doi.org/10.5772/intechopen.68796

[11] Security Week Network. Symantec Adds Machine Learning to Endpoint Security Lineup [Internet]. 2016. Available from: http://www.securityweek.com/symantec-adds-machine-learning-endpoint-security-lineup

[12] Ozlem Yavanoglu & Murat Aydos, A Review on Cyber Security Datasets for Machine Learning Algorithms, 11-14 Dec. 2017, 2017 IEEE International Conference on Big Data (Big Data), INSPEC Accession Number: 17504859

[13] Md. Zeeshan Siddiqui &Sonali Yadav, Application Of Artificial Intelligence In Fighting Against Cyber Crimes: A Review, International Journal of Advanced Research in Computer Science April 2018 , (ISSN: 0976-5697), ISBN: 978-93-5311-643-9, page[118-121]

[14] Nilaykumar Kiran Sangani&HarootZarger, Machine Learning in Application Security, [Accessed: March 18, 2019] http://dx.doi.org/10.5772/intechopen.68796

[15] Dmitri Koteshov, How Can Ai Change The State Of Cybersecurity, March 7, 2018, https://www.elinext.com/industries/financial/trends/ai-and-security/