IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

An Efficient Model to defend cyber-attack using Data Science

1Sumanta Sharma, 2Dr S Anupama Kumar

1VI Semester Student, 2Associate Professor

1Department of MCA, RV College of Engineering,

2Department of MCA RV College of Engineering

Abstract

The recent rapid growth in big data, networking, and machine learning is due to exponential advances in processing, storage, and network technologies. As the world becomes more digitalized, there is a greater need for comprehensive and sophisticated security technologies and strategies to address the increasingly complex nature of cyber-attacks. This paper examines how machine learning, big data is being used in cyber security, both in defence and offense, with a focus on cyber-attacks against machine learning models. Machine learning can be used to carry out cyber-attacks, such as smart botnets, sophisticated spear fishing, and evasive malware. In the field of defence, big data analytics refers to the ability to collect large volumes of digital data in order to analyse, visualize, and derive knowledge that can help predict and prevent cyber-attacks. It gives us a stronger cyber defence stance when combined with security technologies. They allow businesses to identify patterns of behaviour that indicate network threats. Machine learning is used in cyber security for threat identification and prevention, malware detection and classification, and network risk rating, among other things.

Keywords

Cyber Security; Machine Learning; Malware; Threats Detection and Classification; Network Risk Scoring, Big Data Analytics

Introduction

Cyber space in this technology driven world is getting complex every day and due to this hike in technology data protection and privacy management is became the most important thing in the cyber security. There are thousands of approaches to deal with this problem, but this approach may not be sustainable when volume of data grows exponentially. System will be requiring a data driven decision making system to protect from potential cyber-attack or crime. In this paper, it is discussed that how data science driven approach can help us to prepare and build a shield for upcoming potential cyber-attack.

Big Data with Machine learning is an interdisciplinary discipline that employs computational techniques, procedures, algorithms, and systems to derive information and insights from structured and unstructured data, as well as to extend that knowledge and actionable insights to a variety of application domains. Data processing, machine learning, and big data are also linked to data science. Big Data is a massive compilation of data that continues to expand exponentially over time. It is a data set that is so massive and dynamic that no conventional data processing systems can effectively contain or process it. Big data is similar to regular data, but it is much larger.

In comparison to conventional cybersecurity programming processes, the principle of cybersecurity data science makes for more actionable and intelligent computing. Following that, it go over and outline a selection of related research topics and future directions. In addition, it have a multi-layered machine learning-based architecture for cybersecurity modelling. Overall, the aim is to concentrate on the applicability of data-driven informed decision making for defending networks against cyber-attacks, rather than only discussing cybersecurity data science and related approaches.

Literature survey

Machine learning with Big data provides us a better cyber defence stance when combined with security technology and volume of data. They allow businesses to spot trends of behaviour that indicate network risks. The goal of big data will be used to discover hidden insights, improve security related [1].Machine learning is an important platform for automating dynamic cybersecurity protection and offense operations. As a result, with cybercriminals using machine learning in their arsenal of cyber weapons, we should expect to see more advanced and large-scale AI-powered attacks [2].

A modern architecture, we will combine Big Data, Machine Learning, and Cyber Security. The goal is to secure any device from a data breach by reviewing existing data sets and using any previous attack history that might exist [3]. The paper [4] focuses on analysing attacks and gathering intruder footprints using live streaming evidence. Since data is increasing in both scale and variety, there is a need to protect it from unauthorized access. It's past time for a machine to be smart enough to do the analysis and provide valuable results. The end aim is to fully simplify the detection of cybercrime.

Big Data plays a significant role in major decisions, whether they are business-related or include the adoption of emerging technologies. We've been experimenting with it, but not extensively in Cyber Security. With its assistance, live streaming data can be used to track and evaluate major attacks, enabling us to identify hackers [5].

In comparison to conventional cybersecurity programming processes, the principle of cybersecurity data science makes for more actionable and intelligent computing. Following that, we go over and outline a selection of related research topics and future directions. In addition, we have a multi-layered machine learning-based architecture for cybersecurity modelling. Overall, our aim is to concentrate on the applicability of data-driven informed decision making for defending networks against cyber-attacks, rather than only discussing cybersecurity data science and related approaches [1].

Proposed work or implementation

Big Data refers to data sets that are very large or complex, and with which conventional data set analysis framework software is insufficient or unable to cope. The number, velocity, and variance between traditional and big data are the most significant differences. Volume denotes the sum of data generated; velocity denotes the rate at which the data is generated; and variance denotes the different types of structured and unstructured data. Predictive and Pattern Discovery are two types of machine learning algorithms. There is often a goal variable in supervised learning, the meaning of which the machine learning model learns to simulate using various learning algorithms. For example, a machine learning model would predict whether a given IP address was part of a botnet based on the location of the IP address, the frequency of Web requests, and the times of the requests, this helps to protect from DDoS (distributed Denial of service) attack.

Cybercrime and surveillance, which has many use cases for machine learning, such as ransomware and log processing, is one area where machine learning is seeing widespread acceptance. Cybercriminals and defence researchers both use machine learning to their advantage. Now we'll look at how machine learning is being used in both cybercrime and cyber defence.

Big Data Analytics and Machine Learning

Because of the rapid growth of data, it is difficult to store, manage, and use it effectively to identify a training dataset and build a predictive algorithm to automatically classify future events. Data mining is used to find correlations and derive main associations from data variables, while machine learning is used to refine predictive algorithms. As a result, the more data there is, the more computation is needed, and the more effective the applications would be, necessitating Big Data Analysis in any domain. These activities take up a lot of time and need a lot of thought.

Some Big Data analysis techniques are:-

1. Sentiment analysis - as the name implies—researches sentiments and produces sentiment-related outcomes, i.e., a set of target audiences' sentiments are reported based on a given situation, and relevant outcomes are obtained.
2. Optimisation is used by genetic algorithms, which are inspired by evolution processes.
3. Social network research- Since social networking is still on, it plays an essential role in the corporate world.

Some of the big data tools which can be used for the above cases are:-

1. Hadoop is a free and open source data processing platform. Apache provides it to process and interpret large amounts of data.
2. Pig is a tool for dealing with large amounts of data that was created by Yahoo.
3. Hive is a SQL-like query language for handling large amounts of data in a warehouse environment.
4. Map Reduce jobs can be run on HBase, which is used as a datastore.
5. Bash Reduce - Map Reduce implementation for UNIX standard commands.

Machine Learning is a collection of algorithms and techniques, the majority of which are based on statistics and probability that infer an approximate model from enough observations of the target system. For different fields, learning from these large data sets is supposed to offer tremendous opportunities and disruptive potential. Deep learning, also known as Machine Learning, and Big Data, are buzzwords in the field and in recent computer security developments. Learning can be described as the capacity to learn information and use it to achieve results. The Learning Process: In machine learning, measurement, programming, function selection or prediction, and model learning analysis outcomes are all critical.

The security aspects of these technology

"In general, Information Security is a domain challenge rather than a domain solution, because it looks for solutions in other areas or realms."

Basically security aspects cover these points in general:-

- Secrecy is achieved by the use of cryptography, or the practice of cracking codes.
- Hash functions are used to maintain the consistency of a device.

When it comes to firewalls, which are a must in today's world, they keep us secure, but Big Data and Firewalls slow down transmission because the more private your data is, the better firewall you use, slowing down the system because IP addresses, routers, and other factors get in the way! However, we would finally need a firewall.

Threats can be:-

The aim of traditional attacks is to infect a large number of computers and steal their energy. This can be monitored using Big Data analysis so they are easier to spot than Advanced Persistent Threats.

In Advanced Persistent Threats, the attacker can stay in the environment as long as its target isn't met, it can deal with standard defences, and it has the ability to steal data. This looks a lot like a zombie apocalypse. Since the attacks are one-of-a-kind and have no prior experience, they are able to get through the defence mechanisms to complete their mission without being identified!

Machine learning can be used as following:-

Machine learning algorithms can be used in software to detect and respond to cyber-attacks before they happen. This is typically accomplished by the use of a model built through the analysis of large data sets of security incidents and the identification of malicious behaviour patterns. As a consequence, when related behaviours are discovered, they are dealt with immediately. The training dataset for the models is usually composed of previously defined and registered Indicators of Compromise (IOC), which are then used to create models and structures that can monitor, recognize, and react to threats in real time.

With artificial intelligence's rapid development, we're seeing a growing amount of activities being automated. This AI driven tools are being developed by modelling the correct data and insights that can be derived or extracted from the utilization of Machine learning with Big data technologies. It's enticing to believe that artificial intelligence can accelerate automation and that robots will take over those functions actually done by humans.

Studies have been conducted on the use of machine learning algorithms, such as genetic algorithms and decision tries, to develop applications that produce rules for classifying network connections in order to improve security analysts' activities.

Solution to the cybercrime problem using Machine Learning and Big Data

With the growing challenge of cyber-attacks, researchers are concentrating their efforts on machine learning and its large range of tools and techniques for detecting, stopping, and responding to sophisticated cyber-attacks. Machine learning can be used to provide analytical-based methods for attack detection and response in a variety of cyber security domains.

Detection and assessment of threats: Machine learning algorithms can be used in software to detect and respond to cyber-attacks before they happen. This is typically accomplished by the use of a model built through the analysis of large data sets of security incidents and the identification of malicious behaviour patterns. As a consequence, related events are automatically dealt with when they are detected. The training dataset for the models is usually composed of previously defined and registered Indicators of Compromise (IOC), which are then used to create models and structures that can monitor, recognize, and react to threats in real time.

Network Risk Scoring: This applies to the use of predictive metrics to allocate risk scores to different parts of a network, allowing businesses to prioritize their cyber defence services based on risk scores. Through reviewing historic cyber-attack databases and evaluating which areas of networks were most often involved in those forms of attacks, machine learning may be used to simplify this process. The use of machine learning has the advantage that the resulting scores will not only be dependent on domain knowledge of the networks, but they will also be data driven. This score will help organisations measure the probability and effect of an attack in relation to a specific network location, lowering the chance of becoming a target of an attack.

Automation & Human Driven Learning: During security operations, machine learning can be used to simplify routine tasks performed by security analysts. This can be accomplished by reviewing records/reports of previous activities taken by security experts to effectively diagnose and react to specific threats, and then using the information to create a model that can detect and respond to similar attacks without the need for human interference. Though it is impossible to fully automate the security process and replace the human security researcher, machine learning can automate certain parts of the investigation, such as malware detection and network log analysis.

We will use Big Data to gather data from people or organizations that have been targeted in the past, and then do an investigation to determine the nature of attack, the attack's target, and the attacker's identity! The attacker attacks the machine and gets around standard encryption, but it's incredibly difficult for him to grasp what's in the massive dataset if he isn't a data analyst.

Furthermore, APT attacks harvest limited information, which does not function in Big Data because it is useless until all of the data is connected, and repeating the APT attack would mean a long delay, as extracting 1 TB of data would take years. As a result, it is essential to anticipate and react to all security risks, as well as to thoroughly investigate them. We will secure our device in real time by analysing streaming data.

These steps can be utilized to get the big data on the go tool for security analysis detection.

- Suspicion is aroused by the discovery of suspicious domains that match public DNS registry records.
- Using Big Data on email to identify trends.
- Determine the number of redirects and goals.
- Visualization and effects study of advanced threats.
- Finding out something from the hacker.
- Identification of the hacker using forensic documents

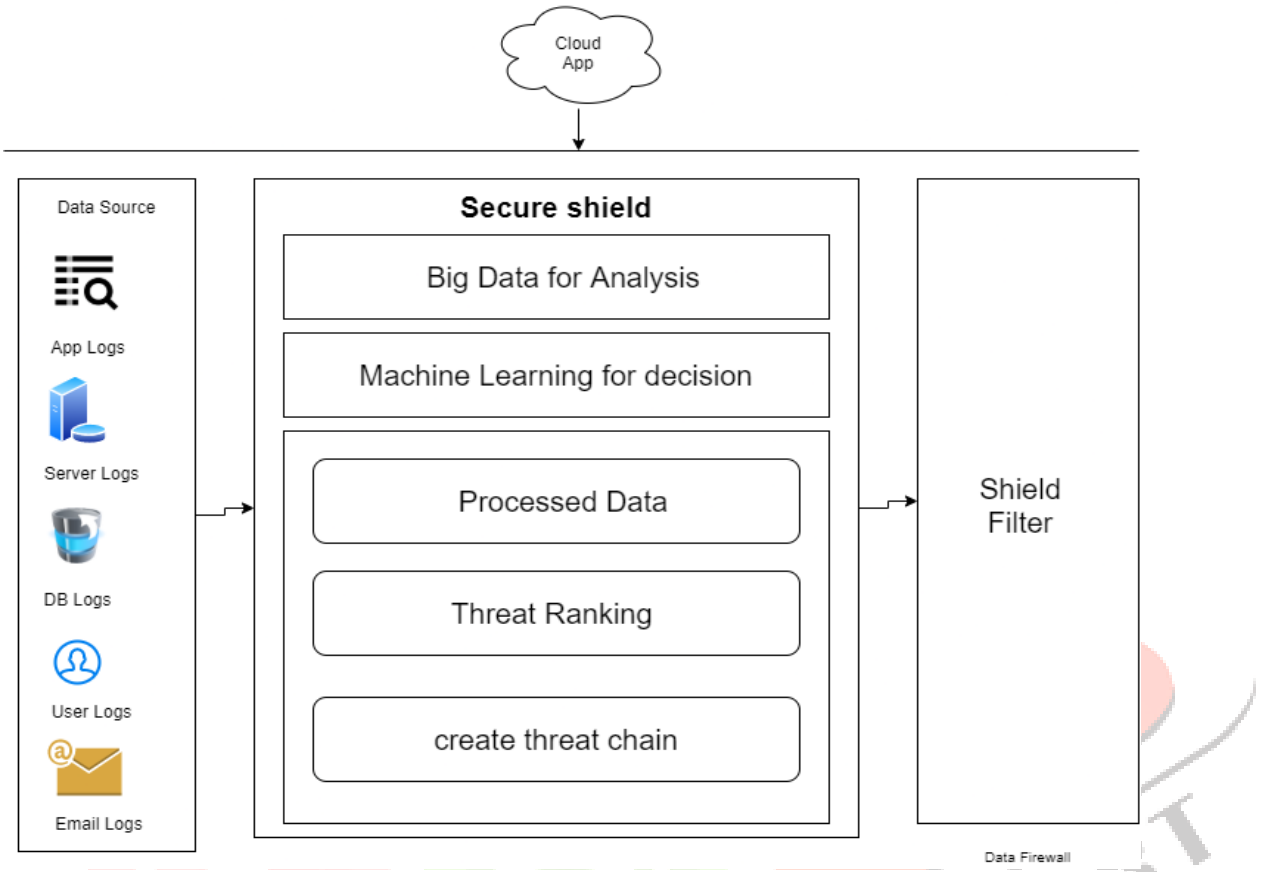

Fig: Working Mechanism for the proposed system.

Conclusion

Machine learning, as can be observed, is a versatile technology that can be used to automate complicated cyber security and offense operations. As a result, with cybercriminals using machine learning in their arsenal of cyber weapons, we should expect to see more advanced and large-scale AI-powered attacks.

Big Data plays a significant role in major decisions, whether they are business-related or include the adoption of emerging technologies. We've been experimenting with it, but not extensively in Cyber Security. With its assistance, live streaming data can be used to track and evaluate major attacks, helping us to identify the hackers. With the help of Big Data, Machine Learning, and Cyber Security, we can create a system that can collect data, interpret it, and provide us with the information of the individual or entity that attempted to hack into our system or gain access to our data by unethical means. To summarize, automated processes are needed, or, to put it another way, we want all material to be automated.

References

[1] qbal H. Sarker1,2* , A. S. M. Kayes3 , Shahriar Badsha4 , Hamed Alqahtani5 , Paul Watters3 and Alex Ng3, "Cybersecurity data science: an overview from machine learning perspective" [Sarker et al. J Big Data, 2020]
[2] About. (n.d.). Retrieved: November 03, 2018, from http://www.palantir.com/.
[3] H. Weiwei., and Y. Tan. Generating Adversarial Malware Examples for Black-Box Attacks Based on GAN. 2017, February 20, Retrieved: November 03, 2018, from https://arxiv.org/abs/1702.05983v1.
[4] S. Dolev and S. Lodha, "Cyber Security Cryptography and Machine Learning", In Proceedings of the First International Conference, CSCML 2017, Beer-Sheva, Israel, June 29-30, 2017.
[5] S. Dolev and S. Lodha, "Cyber Security Cryptography and Machine Learning", In Proceedings of the First International Conference, CSCML 2017, Beer-Sheva, Israel, June 29-30, 2017.
[6] G. A. Wang, M. Chau, and H. Chen. Intelligence and Security Informatics: 12th Pacific Asia Workshop, PAISI 2017, Jeju Island, South Korea, May 23, 2017, Proceedings. Cham, Switzerland:Springer.
[7] M. P. Stoecklin. DeepLocker: How AI Can Power a Stealthy New of Malware. 2018, August 13. Retrieved: September 20, 2018, from https://securityintelligence.com/deeplocker-how-ai-can-powera-stealthy-new-breed-of-malware/
[8]D. Gershgorn. Microsoft's AI is learning to write code by itself, not steal it., 2017, March 1. Retrieved: November 03, 2018, from https://qz.com/920468/artificial-intelligence-created-by microsoftand-university-of-cambridge-is-learning-to-write-code-by-itself-notsteal-it/
[9] Hong-Mei Chen, Rick Kazman and Serge Haziyev ,"Agile Big Data Analytics for Web-Based Systems: An Architecture-Centric Approach", IEEE TRANSACTIONS ON BIG DATA, VOL. 2, NO. 3, JULY-SEPTEMBER 2016.
[10] Christopher Phethean, Elena Simperl, ThanassisTiropanis, RamineTinati, and Wendy Hall, University of Southampton, "The Role of Data Science in Web Science", IEEE Co*mputer Society, May-June 2016.
[11] A.A. Cardenas, P.K. Manadhata., and S.P. Rajan, "Big Data Analytics for Security," In IEEE Security & Privacy, Vol. 11, No. 6, pp.74-76, 2013.
[12] A. Cuthbertson. Ransomware attacks have risen 250 percent in 2017, hitting the U.S. hardest. 2017, May 28. Retrieved: September 21, 2018, from http://www.newsweek.com/ransomware-attacks-rise-250- 2017-us-wannacry-614034.
[13] B. M. Cooper. Resiliency and Recovery Offset Cybersecurity Detection Limits. 2015, January 16. Retrieved: September 21, 2018, from https://www.afcea.org/content/resiliency-and-recovery-offsetcybersecurity-detection-limits.
[14] K. Rieck, P. Trinius, C. Willems, and T. Holz. Automatic analysis of malware behavior using machine learning. Journal of ComputerSecurity, 19(4), 639-668, 2011.
[15] T. Nguyen, and G. Armitage. A survey of techniques for Internet traffic classification using machine learning. IEEE Communications Surveys and Tutorials, 10(4), 56-76. 2008.
[16] Youssef Gahi, MouhcineGuennoun, Hussein T. Mouftah, "Big Data Analytics: Security and Privacy Challenges"[School of Electrical Engineering and Computer Science, University of Ottawa, 800 King Edward Ave., Ottawa, ON, Canada].