IJREAT International Journal of Research in Engineering & Advanced Technology, Volume 10, Issue 1, Feb – March 2022

ISSN: 2320 – 8791 (Impact Factor: 2.317)

www.ijreat.org

Cybersecurity Data Science: Machine Learning Algorithms

K. Uma Rani [1], V.S. Madhumala[2] A. Mounika[3]

#[1,2,3] Assistant Professor, Department of CSE, Sree Datta Groups Of institutions, Hyderabad Telangana.

Abstract:

Organizations all over the world have recognised the true intrinsic value of their results, resulting in a boom in demand for data scientists. It's becoming more common to set up business intelligence departments and make data-driven decisions. In a computing context, cybersecurity is undergoing massive shifts in technology and operations in recent days, and data science is driving the change. Extracting security incident patterns or insights from cybersecurity data and building a corresponding data-driven model is the key to making a security system automated and intelligent. To understand and analyses the actual phenomena with data, various scientific methods, machine learning techniques, processes, and systems are used, which is commonly known as data science. Over the last decade, a number of machine learning algorithms have been developed to make the data classification and knowledge extraction processes as easy as possible. In this paper, we explain some of the fundamental machine learning algorithms that every data science enthusiast should be familiar with.

Keywords: *Data Science, Machine Learning, Decision making, Cyberattack, Security modeling*

1. Introduction

Due to the increasing dependency on digitalization and Internet-of-Things (IoT) [1], various security incidents such as unauthorized access [2], malware attack [3], zeroday attack [4], data breach [5], denial of service (DoS) [2], social engineering or phishing [6] etc. have grown at an exponential rate in recent years. For instance, in 2010, there were less than 50 million unique malware executables known to the security community. By 2012, they were double around 100 million, and in 2019, there are more than 900 million malicious executables known to the security community, and this number is likely to grow, according to the statistics of AV-TEST institute in Germany [7]. Cybercrime and attacks can cause devastating financial losses and affect organizations and individuals as well. It's estimated that, a data breach costs 8.19 million USD for the United States and 3.9 million USD on an average [8], and the annual cost to the global economy from cybercrime is 400 billion USD [9]. According to Juniper Research [10], the number of records breached each year to nearly triple over the next 5 years.

Cybersecurity is a set of technologies and processes designed to protect computers, networks, programs and data from attack, damage, or unauthorized access. In recent days, cybersecurity is undergoing massive shifts in technology and its operations in the context of computing, and data science (DS) is driving the change, where machine learning (ML), a core part of "Artificial Intelligence" (AI) can play a vital role to discover the insights from data. Machine learning can significantly change the cybersecurity landscape and data science is leading a new scientific paradigm.

The ultimate goal of cybersecurity data science is data-driven intelligent decision making from security data for smart cybersecurity solutions. CDS represents a partial paradigm shift from traditional well-known security solutions such as firewalls, user authentication and access control, cryptography systems etc. that might not be effective according to today’s need in cyber industry. Thus, based on the concept of data-driven decision making, we aim to focus on cybersecurity data science, where the data is being gathered from relevant cybersecurity sources such as network activity, database activity, application activity, or user activity, and the analytics complement the latest data-driven patterns for providing corresponding security solutions.

2. Related work

Data science has gained importance since available data and hardware facilities have been ubiquitous. Algorithms to process a huge amount of data and extract information were developed decades ago. However, due to the lack of high-capacity computers, it was not possible to use them on real-life data and problems. Today, from finance to medicine data science plays an important role to solve problems.

Suffice it to say, machine learning algorithms are the core of this new phenomenon besides data itself. Artificial neural networks, deep learning, Support Vector Machines, Decision Tree Learning Models, and related algorithms have been used successfully and yielded very important results recently. On the other hand, text data have also gained importance being the fuel of machine learning in data science. Especially the emergence of social media and communication technology contributed to the popularity of texts in data science. In [11], the authors summarized the introductions about the most popular and also successful machine learning algorithms. This work will be helpful for those readers who do not have enough information about machine learning and its algorithms.

To address the challenging problems in current biomedical data science, we proposed several novel large-scale machine learning models for multi-dimensional data integration, heterogeneous multi-task learning, longitudinal feature learning, etc. Meanwhile, to deal with the big data computations. In [12], the authors proposed new asynchronous distributed stochastic gradient and coordinate descent methods for efficiently solving convex and non-convex problems, and also parallelized the deep learning optimization algorithms with layer-wise model parallelism. They applied our new large-scale machine learning models to analyze the multi-modal and longitudinal Electronic Medical Records (EMR) for predicting the heart failure patients' readmission and drug side effects, integrate the neuroimaging and genome-wide array data to recognize the phenotypic and genotypic biomarkers, and detect the histopathological image markers and the multi-dimensional cancer genomic biomarkers in precision medicine studies.

In [13], the authors presented the potential of machine learning and its contribution to the Internet of Things, whose goal is to provide the ability to automatically solve a wide range of complex decision-making and analytic tasks, and then we give an architecture that shows the relationship between ML and IoT. Additionally, we present an overview of different machine learning techniques, as well as studying classification algorithms, which is frequently used for intelligent data analysis IoT. Then, finally, we present an evaluation with an iris database to measure the performance of these algorithms.

3. Types of machine learning

Machine learning is used to predict, categorize, classify, finding polarity, etc. from the given datasets and concerned with minimizing the error. It uses training data for artificial intelligence. Since there are many algorithms like SVM Algorithm in Python, Bayes algorithm, logistic regression, etc. which will use training data to match with input data and then it will provide a conclusion with maximum accuracy as shown in Figure 1.

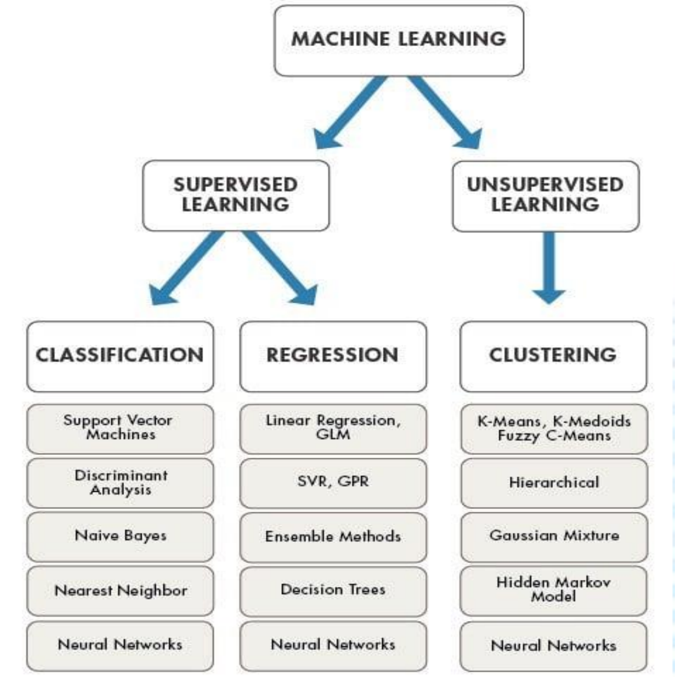

Figure 1: Overview of Machine Learning algorithms

The critical element of data science is Machine Learning algorithms, which are a process of a set of rules to solve a certain problem. Some of the important data science algorithms include regression, classification and clustering techniques, decision trees and random forests, machine learning techniques like supervised, unsupervised and reinforcement learning. In addition to these, there are many algorithms that organizations develop to serve their unique needs.

Supervised learning:

It is used for the structured dataset. It analyzes the training data and generates a function that will be used for other datasets.

Unsupervised learning:

It is used for raw datasets. Its main task is to convert raw data to structured data. In today’s world, there is a huge amount of raw data in every field. Even the computer generates log files which are in the form of raw data. Therefore, it’s the most important part of machine learning.

a) Multilayer Perceptions

A neural network-based classifier, called Multilayer perception (MLP), is used to classify the handwritten digits. Multilayer perceptron consists of three different layers, input layer, hidden layer and output layer. Each of the layers can have certain number of nodes also called neurons and each node in a layer is connected to all other nodes to the next layer. For this reason, it is also known as feed forward network. The number of nodes in the input layer depends upon the number of attributes present in the dataset. The number of nodes in the output layer relies on the number of apparent classes exist in the dataset. The convenient number of hidden layers or the convenient number of nodes in a hidden layer for a specific problem is hard to determine. But in general, these numbers are selected experimentally. In multilayer perceptron, the connection between two nodes consists of a weight. During training process, it basically learns the accurate weight adjustment which is corresponds to each connection. For the learning purpose, it uses a supervised learning technique named as Back propagation algorithm.

b) Support Vector Machine

SVM or Support Vector Machine is a specific type of supervised ML method that intents to classify the data points by maximizing the margin among classes in a high-dimensional space [14]. SVM is a representation of examples as points in space, mapped due to the examples of the separate classes are divided by a fair gap that is as extensive as possible. After that new example are mapped into that same space and anticipated to reside to a category based on which side of the gap they fall on [15]. The optimum algorithm is developed through a "training" phase in which training data are adopted to develop an algorithm capable to discriminate between groups earlier defined by the operator (e.g. patients vs. controls), and the "testing" phase in which the algorithm is adopted to blind-predict the group to which a new perception belongs. It also provides a very accurate classification performance over the training records and produces enough search space for the accurate classification of future data parameters. Hence it always ensures a series of parameter combinations no less than on a sensible subset of the data. In SVM it's better to scale the data always; because it will extremely improve the results. Therefore, be cautious with big dataset, as it may lead to the increase in the training time

c) Random Forest Algorithm

Random forest as is an ensemble of un-pruned regression or classification trees, activated from bootstrap samples of the training data, adopting random feature selection in the tree imitation process. The prediction is made by accumulating the predictions of the ensemble by superiority voting for classification. It returns generalization error rate and is more potent to noise. Still, similar to most classifiers, RF may also suffer from the curse of learning from an intensely imbalanced training data set. Since it is constructed to mitigate the overall error rate, it will tend to focus more on the prediction efficiency of the majority class, which repeatedly results in poor accuracy for the minority class.

Machine learning is a buzzword in today's technology, and it's gaining momentum at a breakneck speed. Even if we aren't aware of it, we use machine learning in our everyday lives through Google Maps, Google Assistant, Alexa, and other similar services. The following are some of the most common real-world Machine Learning applications:

Image Recognition: One of the most popular applications of machine learning is image recognition. It's used to recognize things like people, locations, and digital images. Automatic friend tagging recommendation is a common application of image recognition and face detection.

Speech Recognition: When we use Google, we have the option to "Search by voice," which falls under the category of speech recognition and is a common machine learning application. Speech recognition, also known as "Speech to text" or "Computer speech recognition," is the method of translating voice commands into text.

Self-driving vehicles: Self-driving cars are one of the most exciting applications of machine learning. In self-driving vehicles, machine learning plays a significant role. Tesla, the most well-known automobile manufacturer, is developing a self-driving vehicle. It trains car models to detect people and objects when driving using an unsupervised learning process.

Tools used in Machine Learning

Microsoft Azure Machine Learning: Azure Machine Learning is a cloud platform for building, training, and deploying AI models. Microsoft is still upgrading and enhancing its machine learning software, and it recently announced improvements to Azure Machine Learning, including the retirement of the Azure Machine Learning Workbench.

IBM Watson: Watson Machine Learning is a cloud service from IBM that uses data to deploy machine learning and deep learning models. Users will use this machine learning method to perform two basic machine learning operations: training and scoring. Remember that IBM Watson is best suited for developing machine learning applications via API connections.

Google Tensor Flow: TensorFlow is a Google product. TensorFlow is an open-source software library for dataflow programming that Google uses for research and development. TensorFlow is, at its heart, a machine learning system. This machine learning tool is new to the market and is rapidly developing. The ease with which TensorFlow enables developers to visualize neural networks is likely the most appealing feature.

5. Conclusion

Motivated by the growing significance of cybersecurity and data science, and machine learning technologies, in this paper, we have discussed how cybersecurity data science applies to data-driven intelligent decision making in smart cybersecurity systems and services. We also have discussed how it can impact security data, both in terms of extracting insight of security incidents and the dataset itself. We aimed to work on cybersecurity data science by discussing the state of the art concerning security incidents data and corresponding security services. We also discussed how machine learning techniques can impact in the domain of cybersecurity, and examine the security challenges that remain. In terms of existing research, much focus has been provided on traditional security solutions, with less available work in machine learning technique-based security systems.

References

[1] Li S, Da Xu L, Zhao S. The internet of things: a survey. Inform Syst Front. 2015;17(2):243–59.

[2] Sun N, Zhang J, Rimba P, Gao S, Zhang LY, Xiang Y. Data-driven cybersecurity incident prediction: a survey. IEEE Commun Surv Tutor. 2018;21(2):1744–72.

[3] McIntosh T, Jang-Jaccard J, Watters P, Susnjak T. The inadequacy of entropy-based ransomware detection. In: International conference on neural information processing. New York: Springer; 2019. p. 181–189

[4] Alazab M, Venkatraman S, Watters P, Alazab M, et al. Zero-day malware detection based on supervised learning algorithms of API call signatures (2010)

[5] Shaw A. Data breach: from notification to prevention using PCI DSS. Colum Soc Probs. 2009;43:517.

[6] Gupta BB, Tewari A, Jain AK, Agrawal DP. Fighting against phishing attacks: state of the art and future challenges. Neural Comput Appl. 2017;28(12):3629–54.

[7] Av-test institute, Germany, https://www.av-test.org/en/statistics/malware/. Accessed 20 Oct 2019.

[8] IBM security report, https://www.ibm.com/security/data-breach. Accessed on 20 Oct 2019.

[9] Fischer EA. Cybersecurity issues and challenges: In brief. Congressional Research Service (2014)

[10] Juniper research. https://www.juniperresearch.com/. Accessed on 20 Oct 2019

[11] Silahtaroğlu, Gökhan & Dincer, Hasan & Yüksel, Serhat. (2021). Introduction to Data Science and Machine Learning Algorithms. 10.1007/978-3-030-74176-1_1.

[12] Huang, Heng. (2019). Large-Scale Machine Learning Algorithms for Biomedical Data Science. BCB '19: Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics. 4-4. 10.1145/3307339.3342130.

[13] Bouhlal, Meriem & Aarika, Kaoutar & Ait Abdelouahid, Rachida & SANAA, ELFILALI & Benlahmar, EL Habib. (2020). Machine learning algorithms for data classification. International Journal of Web-Based Learning and Teaching Technologies.

[14] Li, Jing & Peng, Jiangtao. (2009). A support vector machine learning algorithm. Journal of Hubei University. Natural Science Edition. 21.

[15] Gammermann, A. (2000). Support vector machine learning algorithm and transduction. Computational Statistics - COMPUTATION STAT. 15. 31-39. 10.1007/s001800050034.