Leveraging Data Science and Machine Learning for Enhanced Cyber security

Rising Threats and Evolving Tactics

Today, the world we live in is advancing in each field, especially in terms of technological development. Any work can be done through smart devices. The biggest boon for our century is the internet. With that, we can know what is happening around us and in every part of the world. With the increase in the technological aspects and flexibility of doing things online, the risk of cyber-attacks is also growing exponentially.

Data science is one of the top emerging technologies that has proven to aid many organizations, and cyber security is not an exception for that. Data science includes many machine learning algorithms that are used for predicting cyber-attacks in advance. Existing Cyber security techniques could be helpful in preventing the attacks or threats as they occur, but with data science, these attacks could be predicted in advance and take necessary steps to avert them completely. This paper gives a general introduction to what Data science is and how that can be used in Cyber security, along with some basic machine learning algorithms.

Machine learning is one of the topmost emerging technologies that is currently booming the business of any type of organization. Generally, data science is the field of extracting and analyzing the data. This process involves extracting useful information from either structured or unstructured data that has been collected from various sources. As the data is collected from multiple sources, it would be generally unformatted data. For the analysis process, data should be in an organized manner. Then by using machine learning algorithms and tools, the information can be analyzed to make predictions about unexpected events.

Data Science helps in predicting the actions that might occur by analyzing past data. Most organizations are implementing data science in their business activities in order to predict possible activities. The task of data analysis is performed by the data scientists. They establish a working relationship with the stakeholders in order to know which information is to be analyzed so that they can find the algorithms that need to be used to run the data models that might help in business growth.

On the other hand, Cybersecurity is another big thing happening in the world. Security is needed everywhere and for everyone, especially when we are dealing with devices online. Cyber security deals with threat management issues related to any type of organization. It provides the procedures and methods that can be used to prevent cyber attackers. Also, they aid in recovering from the attacks. Generally, hackers and attackers with malicious intentions try to steal confidential data from devices through various attacks.

A cyber-attack involves stealing, modifying, or deleting sensitive data from a system or accessing another's system without their knowledge. There are innumerable cases where many industries have lost millions of dollars because of cyber threats. Hackers can target either an individual or a group of people in order to gain access to the systems. These attacks come in different forms. Some of the attacks that are frequently used by the attackers are phishing, a man-in-the-middle attack, denial-of-service attacks, viruses, malware, etc., and Social engineering attacks, which are the most commonly used attacks on individuals and on employees of any company. In this method, the hacktivist's main goal is to make the target believe that they are legitimate and trust them.

Relationship between Machine learning and Cyber Security:

Data analysts use machine learning tools in order to conduct a thorough analysis of the collected data to reveal trends and patterns. For example, based on the analysis, future occurring attacks can be predicted so that necessary preventive measures can be taken by the organization. Cyber security uses a wide range of tools and intrusion styles in order to monitor the activities on the devices and to stop dangerous activities. They come into action only when any unusual things happen. But data science can be used for both enhancing and simplifying cyber security tools. By using past and present data as input to machine learning algorithms in data science, the possibility of the occurrence of future attacks can be estimated.

Another biggest concern about a cyber-attack is losing valuable information. Cyber security uses encryption algorithms in order to prevent the loss of data from an organization's database. But, by using data science, inaccessible protocols can be developed. For instance, by analyzing past data, spending a huge amount on the detection and response phases. But no organization is sure of the results because if they find new countermeasures for the attacks, then the attackers are also taking new forms to exploit the systems.

This is where data science comes in. Most companies these days have a team of data scientists, but they do not work in security. Data scientists working with the security team can inform what the data needs to be focused on. on As the organization starts to look to gain continuous visibility to risk and security performance, there are three critical questions that need to be answered.

They are:

i. What are the available data and the quality of the available data?
ii. What does that mean for the insight we can get in?
iii. What is the plan to follow and to improve data sources to answer the questions that matter most?

The Impact of Data Science on Cybersecurity:

Generally, data scientists use machine learning tools and algorithms to predict cyber-attacks and to stop them. So that the data scientists can identify the risks based on past attacks, machine learning algorithms can find the attacks that might take place by analyzing the past data. This is very useful to organizations that are prone to attacks. Machine learning tools can also be used to make repetitive security tasks occur automatically.

Machine Learning in Cyber Security:

Generally, any machine learning algorithm that is used in cyber security contains two phases which are the training phase and the protection phase.

Training Phase: In this phase, both positively labelled and negatively labelled features are given as input to a system which makes a predictive model out of that.

Protection Phase: In this phase, the predictive model identifies whether the incoming feature is benign or harmful.

Clustering Algorithm:

Clustering is a technique of separating the data points which are of the same kind. That means all the data points in a cluster contain similar features, and those features are different from the data points of another cluster.

The above figure represents the clustering algorithm in a two-dimensional space. The x-axis and y-axis represent two different features, and the input data is represented in the form of data points in that space. The algorithm involves several steps.

They are:

Step 1: Firstly, the input data is represented in data points in two-dimensional space, then two random points are selected from all the data points.

Step 2: In the second step, the distance from each selected point to the rest of the points in that space is calculated.

Step 3: The points which are nearest to the selected points are formed into clusters. As there are two selected data points, there will be two clusters.

Step 4: After the clusters are formed then, the mean point of each cluster is determined. The mean is calculated as the sum of all the data points divided by the number of data points. As there are two clusters, two means are determined in our example.

Step 5: In this step distance from each mean point to all other points is measured, and the nearest points to the mean points are again formed as clusters.

Step 6: Steps 4 and 5 are repeated until we get the same mean points consecutively. Then the clusters formed are considered the final ones.

Initially, all the incoming executables are sent into the algorithm. Then based on the number of clusters required, the initial clusters are formed. Then as explained in Figure 1, after several iterations, final clusters are formed. In this way, the clustering algorithm is used to identify the malicious objects in a network.

Challenges:

By using a clustering algorithm in the field of cyber security, any unusual activities on the network can be identified immediately, and the chances of the occurrence of attacks can be predicted. But there are a few challenges that a data scientist is facing. Those include:

1. Number of Clusters:

Identifying the count of the clusters that need to be formed is a difficult task because without knowing the exact clusters, the analysis may not be appropriate to consider. Also, there is no perfect way to determine the number of clusters.

2. Distance Measuring:

The distance can be calculated by using Manhattan, Euclidean, or the maximum distance measure. Finding the correct method of measuring the distance based on the labels is difficult.

3. Choosing Initial Data Points:

Choosing the initial data points is very crucial as the rest of the functions depends on that.

Conclusion:

Cyber security and machine learning can help many organizations in identifying and predicting several cyber-attacks in advance. Companies are losing millions of dollars as a result of cyber-attacks each year. Machine Learning algorithms are used to predict the possibility of the occurrence of data breaches in advance based on past data. Clustering is one such algorithm that is used for identifying malicious objects in a network. The importance of data science in cyber security and functionality and one example of a clustering algorithm are explained in this paper.

References

1. Howe, S. (Jul 18, 2018). The Value of Data Science in Security. Retrieved from: https://www.csoonline.com/article/3500646/the-value-of-data-science-in-security.html
2. James, M. (August 27, 2019). How To Improve Cybersecurity With Data Science. Retrieved from: https://www.smartdatacollective.com/how-to-improve-cybersecurity-with-data-science/
3. Tianfield, H. (2017). Data Mining Based Cyber-Attack Detection. Retrieved from: https://www.researchgate.net/publication/321491605_Data_Mining_Based_Cyber-Attack_Detection
4. Drinkwater, D. (December 12, 2017). 5 Top Machine Learning Use Cases for Security. Retrieved from: https://www.csoonline.com/article/3240925/5-top-machine-learning-use-cases-for-security.html
5. Raghupathi, K. (2018). 10 Interesting Use Cases for the K-Means Algorithm. Retrieved from: https://dzone.com/articles/10-interesting-use-cases-for-the-k-means-algorithm
6. Agarwal, P., Alam, M. A., & Biswas, R. (2011). Issues and Tools of Clustering Algorithms. Retrieved from: https://www.semanticscholar.org/paper/Issues%2CChallenges-and-Tools-of-Clustering-Agarwal–Alam/7b49bd891f632ca6e86e5ccccdc3761ceb3fd277