Cyber Attacks Prediction using Data Science

E.Sankar

PhD,Assistant Professor
Department of Computer Science and Engineering
Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya Enathur, Kanchipuram 631 502, Tamil Nadu, India

Mogillikunta Nikhil , Gangavarapu Srinivasulu Reddy

Student, IV Year B.E. Department of Computer Science and Engineering
Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya Enathur, Kanchipuram 631 502, Tamil Nadu, India

ABSTRACT

Cyber-attack, via cyberspace, targeting an enterprise's use of cyberspace for the purpose of disrupting, disabling, destroying, or maliciously controlling a computing environment/infrastructure; or destroying the integrity of the data or stealing controlled information. The state of the cyberspace portends uncertainty for the future Internet and its accelerated number of users. New paradigms add more concerns with big data collected through device sensors divulging large amounts of information, which can be used for targeted attacks. Though a plethora of extant approaches, models and algorithms have provided the basis for cyber-attack predictions, there is the need to consider new models and algorithms, which are based on data representations other than task-specific techniques. However, its non-linear information processing architecture can be adapted towards learning the different data representations of network traffic to classify type of network attack. In this paper, we model cyber-attack prediction as a classification problem, Networking sectors have to predict the type of Network attack from given dataset using machine learning techniques. The analysis of dataset by supervised machine learning technique(SMLT) to capture several information’s like, variable identification, uni-variate analysis, bi-variate and multi-variate analysis, missing value treatments etc. A comparative study between machine learning algorithms had been carried out in order to determine which algorithm is the most accurate in predicting the type cyber Attacks. We classify four types of attacks are DOS Attack, R2L Attack, U2R Attack, Probe attack. The results show that the effectiveness of the proposed machine learning algorithm technique can be compared with best accuracy with entropy calculation, precision, Recall, F1 Score, Sensitivity, Specificity and Entropy.

Keywords

Cyber-attack, Cyberspace, DOS Attack, R2L Attack, U2R Attack, Probe attack

INTRODUCTION:

The process of prediction analysis is a process of using some method or technology to explore or stimulate some unknown, undiscovered or complicated intermediate processes based on previous and present states and then speculated the results. In an early warning system, accurate prediction of DoS attacks is the prime aim in the network offence and defense task. Detection based on abnormity is effective to detect DoS attacks. A various studies focused on DoS attacks from different respects. However, these methods required a priori knowledge being a necessity and were difficult to discriminate between normal burst traffics and flux of DoS attacks. Moreover, they also required a large number of history records and cannot make the prediction for such attacks efficiently. Based on data from flux inspecting and intrusion detection, it proposed a prediction model of DOS attack’s distribution discrete probability based on clustering method of genetic algorithm and Bayesian method and the clustering problem first, and then utilizes the genetic algorithm to implement the optimization of clustering methods. Based on the optimized clustering on the sample data, we get various categories of the relation between traffics and attack amounts, and then builds up several prediction sub-models about DoS attack. Furthermore, according to the Bayesian method and deduce discrete probability calculation about each sub-model and then get the distribution discrete probability prediction model for DoS attack. This paper begins with the relation exists between network traffic data and the amount of DoS attack, and then proposes a clustering method based on the genetic optimization algorithm to implement the classification of DoS attack data. This method first gets the proper partition of the relation between the network traffic and the amount of DoS attack based on the optimized clustering and builds the prediction sub-models of DoS attack. Meanwhile, with the Bayesian method, the calculation of the output probability corresponding to each sub-model is deduced and then the distribution of the amount of DoS attack in some range in future is obtained.

1. SYSTEM DESCRIPTION

This analysis aims to observe which features are most helpful in predicting the network attacks of DOS, R2L, U2R, Probe and combination of attacks or not and to see the general trends that may help us in model selection and hyper parameter selection. To achieve used machine learning classification methods to fit a function that can predict the discrete class of new input. Apply the fundamental concepts of machine learning from an available dataset and Evaluate and interpret my results and justify my interpretation based on observed dataset. Create notebooks that serve as computational records and document my thought process and investigate the network connection whether attacked or not to analyses the data set. Evaluate and analyses statistical and visualized results, which find the standard patterns for all regiments.

2. IMPLEMENTATION

An implementation is a realization of a technical specification or algorithm as program, software elements, or other computer system though computer programming and deployment. Numerous implementations may exist for specifications or norms.
Implementation literally means to put into product or to carry out.

2.1 Project Modules

The modules incorporated in this project are:

□Data validation process and Visualization (Module-01)

□DOS Attack Algorithm Comparison (Module-02)

□R2L Attack Algorithm Comparison (Module-03)

□U2R Attack Algorithm Comparison (Module-04)

□Probe Attack Algorithm Comparison (Module-05)

□Overall Attack Algorithm Comparison (Module-06)

2.2 Module Description

2.2.1 Data validation process and Visualization:

Pre-processing refers to the transformations applied to our data before feeding it to the algorithm. Data Preprocessing is a technique that is used to convert the raw data into a clean data set. In other words, whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis. To achieving better results from the applied model in Machine Learning method of the data has to be in a proper manner. Some specified Machine Learning model needs information in a specified format; for example, Random Forest algorithm does not support null values. Therefore, to execute random forest algorithm null values have to be managed from the original raw data set. And another aspect is that data set should be formatted in such a way that more than one Machine Learning and Deep Learning algorithms are executed in given dataset.

2.2.2 DOS Attack Algorithm Comparison:

In computing, a denial-of-service attack (DoS attack) is a cyber-attack in which the perpetrator seeks to make a machine or network resource unavailable to its intended users by temporarily or indefinitely disrupting services of a host connected to the Internet. Denial of service is typically accomplished by flooding the targeted machine or resource with superfluous requests in an attempt to overload systems and prevent some or all legitimate requests from being fulfilled. In a distributed denial-of-service attack (DDoS attack), the incoming traffic flooding the victim originates from many different sources. This effectively makes it impossible to stop the attack simply by blocking a single source. A DoS or DDoS attack is analogous to a group of people crowding the entry door of a shop, making it hard for legitimate customers to enter, disrupting trade.

2.2.3 R2L Attack Algorithm Comparison:

Now-a-days, it is very important to maintain a high level security to ensure safe and trusted communication of information between various organizations. But secured data communication over internet and any other network is always under threat of intrusions and misuses. To control these threats, recognition of attacks is critical matter. Probing, Denial of Service (DoS), Remote to User (R2L) attacks is some of the attacks which affect large number of computers in the world daily. Detection of these attacks and prevention of computers from it is a major research topic for researchers throughout the world.

2.2.4 U2R Attack Algorithm Comparison:

Remote to local attack (r2l) has been widely known to be launched by an attacker to gain unauthorized access to a victim machine in the entire network. Similarly user to root attack (u2r) is usually launched for illegally obtaining the root's privileges when legally accessing a local machine. Buffer overflow is the most common of U2R attacks. This class begins by gaining access to a normal user while sniffing around for passwords to gain access as a root user to a computer resource. Detection of these attacks and prevention of computers from it is a major research topic for researchers throughout the world.

2.2.5 Probe Attack Algorithm Comparison:

Probing attacks are an invasive method for bypassing security measures by observing the physical silicon implementation of a chip. As an invasive attack, one directly accesses the internal wires and connections of a targeted device and extracts sensitive information. A probe is an attack which is deliberately crafted so that its target detects and reports it with a recognizable “fingerprint” in the report. The attacker then uses the collaborative infrastructure to learn the detector's location and defensive capabilities from this report. This is an attack where the attacker attempts to gather information about the target machine or the network, to map out the network. Information about target may reveal useful information such as open ports, its IP address, hostname, and operating system. Network Probe is the ultimate network monitor and protocol analyzer to monitor network traffic in real-time, and will help you find the sources of any network slow-downs in a matter of seconds.

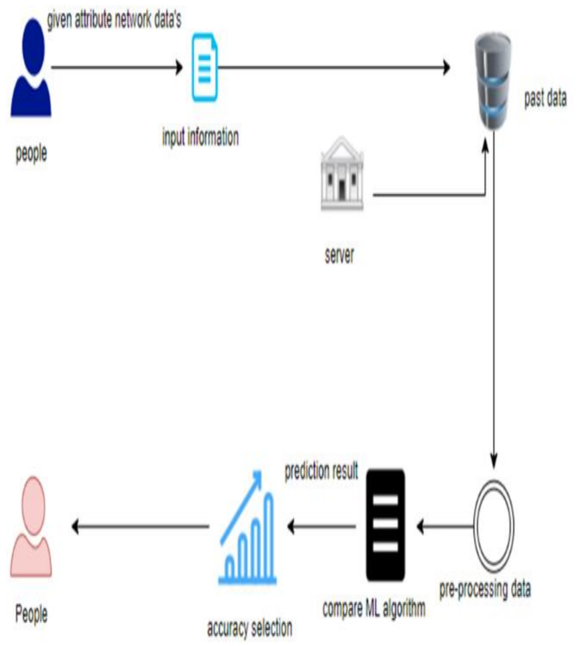
2.2.6 Overall Attack Algorithm Comparison:

Increasingly, attacks are executed in multiple steps, making them harder to detect. Such complex attacks require that defenders recognize the separate stages of an attack, possibly carried out over a longer period, as belonging to the same attack. Complex attacks can be divided into exploration and exploitation phases. Exploration involves identifying vulnerabilities and scanning and testing a system. It is how an attacker gathers information about the system. Exploitation involves gaining and maintaining access. At this stage, the attacker applies the know-how gathered during the exploration stage. An example of a complex attack that combines exploration and exploitation is a sequence of a phishing attack, followed by an exfiltration attack. First, attackers will attempt to collect information on the organization they intend to attack, e.g., names of key employees. Then, they will craft a targeted phishing attack. The phishing attack allows the attackers to gain access to the user’s system and install malware. The purpose of the malware could be to extract files from the user’s machine or to use the user’s machine as an attack vector to attack other machines in the organization’s network. A phishing attack is usually carried out by sending an email purporting to come from a trusted source and tricking its receiver to click on a URL that results in installing malware on the user’s system. This malware then creates a backdoor into the user’s system for staging a more complex attack.

2.3 Technologies Used:

Python is an interpreted high-level general-purpose programming language. Its design philosophy emphasizes code readability with its use of significant indentation. Its language constructs as well as its object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects. Python is dynamically-typed and garbage-collected. It supports multiple programming paradigms, including structured (particularly, procedural), object-oriented and functional programming. It is often described as a "batteries included" language due to its comprehensive standard library. Guido van Rossum began working on Python in the late 1980s, as a successor to the ABC programming language, and first released it in 1991 as Python 0.9.0. Python 2.0 was released in 2000 and introduced new features, such as list comprehensions and a garbage collection system using reference counting. Python 3.0 was released in 2008 and was a major revision of the language that is not completely backward-compatible. Python 2 was discontinued with version 2.7.18 in 2020. Python consistently ranks as one of the most popular programming languages.

2.4 Architecture



3. CONCLUSION

The analytical process started from data cleaning and processing, missing value, exploratory analysis and finally model building and evaluation. The best accuracy on public test set is higher accuracy score will be find out by comparing each algorithm with type of all network attacks for future prediction results by finding best connections. This brings some of the following insights about diagnose the network attack of each new connection. To presented a prediction model with the aid of artificial intelligence to improve over human accuracy and provide with the scope of early detection. It can be inferred from this model that, area analysis and use of machine learning technique is useful in developing prediction models that can helps to network sectors reduce the long process of diagnosis and eradicate any human error.

4. FUTURE SCOPE

The scope of this project is to investigate a dataset of network connection attacks for KDD records for medical sector using machine learning technique. To identifying network connection is attacked or not. Network sector want to automate the detecting the attacks of packet transfers from eligibility process (real time) based on the connection detail. To automate this process by show the prediction result in web application or desktop application. To optimize the work to implement in Artificial Intelligence environment.

5. REFERENCES

1. Wentao Zhao, Jianping Yin and Jun Long- a Prediction Model of DoS Attack's Distribution Discrete Probability, 2008

2. Apriori Viterbi Model for Prior Detection of Socio-Technical Attacks in a Social Network, Preetish Ranjan, Abhishek Vaish, 2014

3. New Attack Scenario Prediction Methodology, Seraj Fayyad, Cristoph Meinel, 2013

4. Cyber Attacks Prediction Model Based on Bayesian Network,Jinyu W1, Lihua Yin and Yunchuan Guo,2012

5. A Prediction Model of DoS Attack's Distribution Discrete Probability, Wentao Zhao, Jianping Yin, 2008

6. Adversarial Examples: Attacks and Defenses for Deep Learning, Xiaoyong Yuan, Pan He, Qile Zhu, and Xiaolin Li, 2019

7. Distributed Secure Cooperative Control Under Denial-of-Service Attacks From Multiple Adversaries, Wenying Xu, Guoqiang Hu , 2019.