## PREDICTION OF NETWORK ATTACKS BY FINDING THE BEST ACCURACY USING SUPERVISED MACHINE LEARNINGALGORITHM

**Dr.Prasanna Srinivasan**

Associate Professor, Department of Information Technology,

R.M.D Engineering College, Tamil Nadu, India.

**A.Anjana, K.Arumuga Kumari, T.C.Keerthana**

Student, Department of Information Technology,

R.M.D Engineering College, Tamil Nadu, India.

## ABSTRACT

*Generally, to create data for the Intrusion Detection System (IDS), it is necessary to set the real working environment to explore all the possibilities of attacks, which is expensive. Software to detect network intrusions protects a computer network from unauthorized users, including perhaps insiders. The intrusion detector learning task is to build a predictive model (i.e. a classifier) capable of distinguishing between “bad” connections, called intrusions or attacks, and “good” normal connections. To prevent this problem in network sectors have to predict whether the connection is attacked or not from KDDCup99 data set using machine learning techniques. The aim is to investigate machine learning based techniques for better packet connection transfers forecasting by prediction results in best accuracy. To propose a machine learning-based method to accurately predict the DOS, R2L,U2R, Probe and overall attacks by prediction results in the form of best accuracy from comparing supervise classification machine learning algorithms. Additionally, to compare and discuss the performance of various machine learning algorithms from the given data set with evaluation classification report, identify the confusion matrix and to categorizing data from priority and the result shows that the effectiveness of the proposed machine learning algorithm technique can be compared with best accuracy with precision, Recall and F1Score.*

**Key words:** Dataset, Machine learning-Classification method, Python, Prediction of Accuracy.

# 1.INTRODUCTION

Machine learning (ML) is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed. Machine learning focuses on the development of Computer Programs that can change when exposed to new data and the basics of Machine Learning, Process of training and prediction involves use of specialized algorithms. It feed the training data to an algorithm, and the algorithm uses this training data to give predictions on a new test data. Machine learning can be roughly separated in to three categories. There are supervised learning, unsupervised learning and reinforcement learning. Supervised learning program is both given the input data and the corresponding labeling to learn data has to be labeled by a human being beforehand. It provided to the learning algorithm. This algorithm has to figure out the clustering of the input data. Finally, Reinforcement learning dynamically interacts with its environment and it receives positive or negative feedback to improve its performance. At a high level, these different algorithms can be classified into two groups based on the way they “learn” about data to make predictions: supervised and unsupervised learning.

**2.PROBLEM DESCRIPTION**

Lately, an internet network company in Japan has been facing huge losses due to malicious server attacks. They’ve encountered breach in data security, reduced data transfer speed and intermittent breakdowns in user-user & user-network connections. When asked, a company official said, “There’s a significant dip in the number of active users on our network. The company is looking are some predictive analytics solution to help them understand, detect and counter the attacks and make their network connection secure. Think of a connection as a sequence of TCP packets starting and ending at some well-defined times, between which data flows to and from a source IP address to a target IP address under some well-defined protocol. In total, there are 3 major type of attacks to which their network is vulnerable to. But, 3 of them cause the maximum damage. In this challenge, you are given an anonymised sample dataset of server connections. You have to predict the type of attack(s) like Dos, R2L, U2R,Probe.

**3. LITERATURE SURVEY**

**TITLE**: Distributed Secure Cooperative Control under Denial-of-Service Attacks From Multiple Adversaries

**AUTHOR**: Wenying Xu , Guoqiang Hu

**YEAR:**2019

**DESCRIPTION:** This paper has investigated the distributed secure control of multiagent systems under DoS attacks. We focus on the investigation of a jointly adverse impact of distributed DoS attacks from multiple adversaries. In this scenario, two kinds of communication schemes, that is, sample-data and event-triggered communication schemes have been discussed and, then, a fully distributed control protocol has been developed to guarantee satisfactory asymptotic consensus. Note that this protocol has strong robustness and high scalability. Its design does not involve any global information, and its efficiency has been proved. For the event-triggered case, two effective dynamical event conditions have been designed and implemented in a fully distributed way, and both of them have excluded Zeno behaviour. Finally, a simulation example has been provided to verify the effectiveness of theoretical analysis. Our future research topics focus on fully distributed event/self-triggered control for linear/nonlinear multiagent systems to gain a better understanding of fully distributed control.

**TITLE** : Cyber Attacks Prediction Model Based on Bayesian Network

**AUTHOR**: Jinyu W1, Lihua Yin and Yunchuan Guo

**YEAR:** 2012

**DESCRIPTION:**The prediction results reflect the security situation of the target network in the future, and security administrators can take corresponding measures to enhance network security according to the results. To quantitatively predict the possible attack of the network in the future, attack probability plays a significant role. It can be used to indicate the possibility of invasion by intruders. As an important kind of network security quantitative evaluation measure, attack probability and its computing methods has been studied for a long time. Many models have been proposed for performing evaluation of network security. Graphical models such as attack graphs become the main-stream approach. Attack graphs which capture the relationships among vulnerabilities and exploits show us all the possible attack paths that an attacker can take to intrude all the targets in the network. The traffics to different hosts or servers may differ from each other. The hosts or servers with big traffic may be more risky since they are often important hosts or servers, and intruders may have more contacts and understanding with them. In our cyber-attacks prediction model, they used attack graph to capture the vulnerabilities in the network. In addition we consider 3 environment factors that are the major impact factors of the cyber-attacks in the future. They are the value of assets in the network, the usage condition of the network and the attack history of the network. Cyber-attacks prediction is an important part of

risk management. Existing cyber-attacks prediction methods did not fully consider the specific environment factors of the target network, which may make the results deviate from the true situation. In this paper, we propose a cyber-attacks prediction model based on Bayesian network. We use attack graphs to represent all the vulnerabilities and possible attack paths. Then we capture the using environment factors using Bayesian network model. Cyber-attacks predictions are performed on the constructed Bayesian network.

**TITLE** : New Attack Scenario Prediction Methodology

**AUTHOR**: Seraj Fayyad, Cristoph Meinel

**YEAR :** 2013

**DESCRIPTION:** Intrusion detection systems (IDS) are used to detect the occurrence of malicious activities against IT system. Through monitoring and analyzing of IT system activities the malicious activities will be detected. In ideal case IDS generate alert(s) for each detected malicious activity and store it in IDS database. Some of stored alerts in IDS database are related. Alerts relations are differentiated from duplication relation to same attack scenario relation. Duplication relation means that the two alerts generated as a result of same malicious activity. Where same attack scenario relation means that the two related alert are generated as a result of related malicious activities. Attack scenario or multi-step attack is a set of related malicious activities run by same attacker to reach specific goal. Normal relation between malicious activities belong to same attack scenario is causal relation. Causal relation means that current malicious activity output is pre-condition to run the next malicious activity. Possible multi-step attack against a network start with information gathering about network and the information gathering is done through network Reconnaissance and fingerprinting process. Through reconnaissance network configuration and running services are identified. Through fingerprint process Operating system type and version are identified. propose a real time prediction methodology for predicting most possible attack steps and attack scenarios. Proposed methodology benefits from attacks history against network and from attack graph source data. it comes without considerable computation overload such as checking of attack plans library. It provides parallel prediction for parallel attack scenarios. Possible third attack step is to identify attack plan based on the modeled attack graph in the past step. The attack plan usually will include the exploiting of a sequence of founded vulnerabilities. Mostly this sequence is distributed over a set of network nodes. This sequence of nodes vulnerabilities is related through causal relation and connectivity. Lastly Attacker start orderly exploits the attack scenario sequences till reaching his/her goal. Attack plan consist of many correlated malicious activities end up with attacking goal.

**4. EXISTING SYSTEM**

The goal of link prediction was to estimate the link likelihood of two unconnected nodes based on available network data and analysis tools, such as machine learning and network theory. Various link prediction methods has used and they are divided into four categories i.e. structural similarity based methods, maximum likelihood based methods ,stochastic probabilistic models and information theory based methods. The established link prediction methods are widely used in online product recommendation, bionetwork reconstruction and evolution process prediction of infrastructure networks. Real-world networks suffer from random failures and targeted attacks. Many scale-free networks, such as the Internet, are vulnerable to degree based targeted attacks. A small initial attack can trigger a large-scale cascading failure which is one of the main security issues in power networks. Novel attack strategies has been proposed, such as the edge attacks and path attacks these random or intentional attacks significantly affect the structure and dynamics of real-world networks. The predictability of real-world networks keeps changing as the network attack continues.

**Drawbacks:** The proposed method did not predict the specific or particular attack. Algorithm prediction results by best accuracy of algorithms with classification report of precision, recall and f1-score and additionally, to categorized other attacks of network.

**5. PROPOSED SYSTEM**

**Exploratory Data Analysis**

This analysis is not meant to be providing a final conclusion on the reasons leading to network sector as it doesn’t involve using any inferential statistics techniques/machine learning algorithms. Machine learning supervised classification algorithms will be used to give the network connection dataset and extract patterns, which would help in predicting the likely patient affected or not, thereby helping the attack of avoids for making better decisions in the future. Multiple datasets from different sources would be combined to form a generalized dataset, and then different machine learning algorithms would be applied to extract patterns and to obtain results with maximum accuracy.

**Data Wrangling**

In this section of the report will load in the data, check for cleanliness, and then trim and clean given dataset for analysis. Make sure that the document steps carefully and justify for cleaning decisions.

**Data collection**

The data set collected for predicting the network attacks is split into Training set and Test set. Generally, 7:3 ratios are applied to split the Training set and Test set. The Data Model which was created using Random Forest, logistic, Decision tree algorithms, K-Nearest Neighbor (KNN) and Support vector classifier (SVC) are applied on the Training set and based on the test result accuracy, Test set prediction is done.

**Preprocessing**

The data which was collected might contain missing values that may lead to inconsistency. To gain better results data need to be preprocessed so as to improve the efficiency of the algorithm. The outliers have to be removed and also variable conversion need to be done. The correlation among attributes can be identified using plot diagram in data visualization process. Data preprocessing is the most time consuming phase of a data mining process. Data cleaning of connections, data removed several attributes that has no significance about the behavior of a packet transfers. Data integration, data reduction and data transformation are also to be applicable for network connections dataset. For easy analysis, the data is reduced to some minimum amount of records. Initially the Attributes which are critical to make a loan credibility prediction is identified with information gain as the attribute-evaluator and Ranker as the search-method.

**6. MODULE DESCRIPTION**

* + - Data validation process by each attack
    - Performance measurements of DoS attack
    - Performance measurements of R2L attack
    - Performance measurements of U2R attack
    - Performance measurements of Probe attack
    - Performance measurements of overall network attacks
    - GUI based prediction results of Network attack

**Description:**

**Data Validation Process by Each Attack**

Validation techniques in machine learning are used to get the error rate of the Machine Learning (ML) model, which can be considered as close to the true error rate of the dataset. Importing the library packages with loading given dataset. To analyzing the variable identification by data shape, data type and evaluating the missing values, duplicate values.The primary goal of data cleaning is to detect and remove errors and anomalies to increase the value of data in analytics and decision making. Data visualization can be helpful when exploring and getting to know a dataset and can help with identifying patterns, corrupt data, outliers, and much more.

**Performance Measurements of DoS Attack**

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. An application layer DoS attack is a form of DoS attack where attackers target application-layer processes. The attack over-exercises specific functions or features of a website with the intention to disable those functions or features. This application-layer attack is different from an entire network attack, and is often used against financial institutions to distract IT and security personnel from security breaches.

**Performance Measurements of R2L Attack**

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.

**Performance Measurements of U2R Attack**

These attacks are exploitations in which the hacker starts off on the system with a normal user account and attempts to abuse vulnerabilities in the system in order to gain super user privileges e.g. perl, xterm.

**Performance Measurements of Probing Attack**

Probing is an attack in which the hacker scans a machine or a networking device in order to determine weaknesses or vulnerabilities that may later be exploited so as to compromise the system. This technique is commonly used in data mining e.g. saint, portsweep, mscan, nmap etc.

**Performance measurements of overall network attacks**

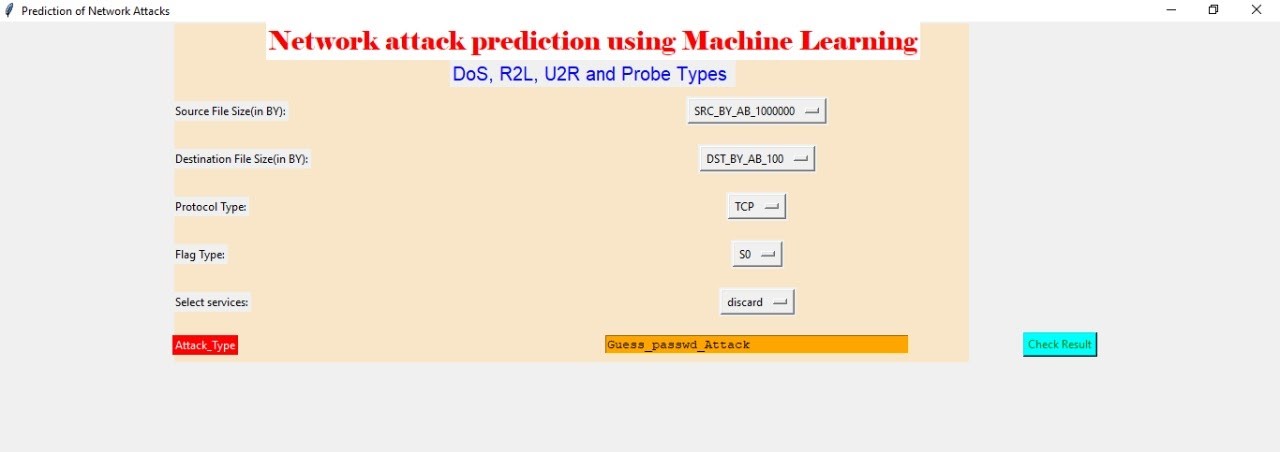
Comparing Algorithm with prediction in the form of best accuracy from the below algorithms,

* Logistic Regression
* Decision Tree
* Navies Bayes
* K- Nearest Neighbor
* Random Tree

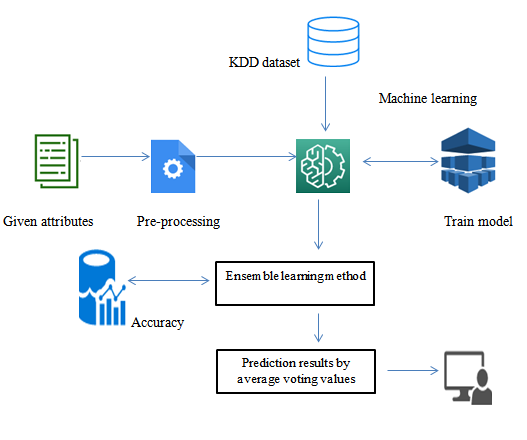
It is important to compare the performance of multiple different machine learning algorithms consistently and it will discover to create a test harness to compare multiple different machine learning algorithms in Python with scikit-learn. It can use this test harness as a template on your own machine learning problems and add more and different algorithms to compare.

**GUI based prediction results of Network attack**

Tkinter is a python library for developing GUI (Graphical User Interfaces). We use the tkinter library for creating an application of UI (User Interface), to create windows and all other graphical user interface and Tkinter will come with Python as a standard package, it can be used for security purpose of each users or accountants.



**7. SYSTEM ARCHITECTURE**



**8. FUTURE ENHANCEMENT**

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.

**9. CONCLUSION**

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.

**REFERENCES**

[1] R. Olfati-Saber, J. A. Fax, and R. M. Murray, “Consensus and cooperation in networked multi-agent systems,” Proc. IEEE, vol. 95, no. 1,pp. 215–233, Jan. 2007.

[2] Y. Su and J. Huang, “Cooperative output regulation of linear multi-agent systems,” IEEE Trans. Autom. Control, vol. 57, no. 4, pp. 1062–1066, Apr. 2012.

 [3] Z. Feng, G. Hu, W. Ren, W. E. Dixon, and J. Mei, “Distributed coordination of multiple unknown Euler-Lagrange systems,” IEEE Trans. Control Netw. Syst., vol. 5, no. 1, pp. 55–66, Mar. 2018.

 [4] Z. Feng, C. Sun, and G. Hu, “Robust connectivity preserving rendezvous of multirobot systems under unknown dynamics and disturbances,” IEEE Trans. Control Netw. Syst., vol. 4, no. 1, pp. 725–735, Dec. 2017.

[5] X. Dong and G. Hu, “Time-varying formation control for general linear multi-agent systems with switching directed topologies,” Automatica,vol. 73, pp. 47–55,Nov.2016.

 [6] Z. Li, G. Wen, Z. Duan, and W. Ren, “Designing fully distributed consensus protocols for linear multi-agent systems with directed graphs,” IEEE Trans. Autom. Control, vol. 60, no. 4, pp. 1152–1157, Apr. 2015.

 [7] I. Shames, A. M. H. Teixeira, H. Sandberg, and K. H. Johansson, “Distributed fault detection for interconnected second-order systems,”Automatica, vol. 47, no. 12, pp. 2757–2764, 2011.