International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

Prediction of the Network Attacks by Finding the Best Accuracy using Supervised Machine Learning Algorithm

A. Sharfudeen

HOD, MCA, Department of MCA, Roever Engineering College, Elambalur, 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 dataset 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 machine learning-based method to accurately predict the DOS, R2L, UU2R, 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 dataset 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 F1 Score.***

Keywords: dataset, Machine learning-Classification method, python, Prediction of Accuracy result

1. 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.

Scope:
The scope of this study 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.

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 ie 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.

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

2. 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.

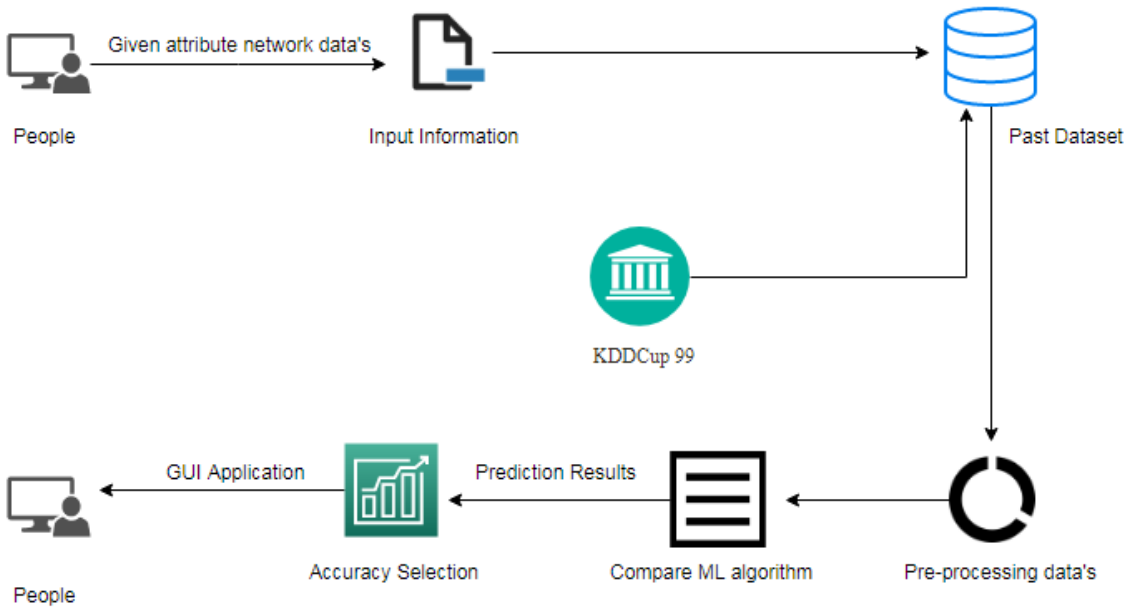
Data collection

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.

Design architecture:



3. Future Work

1) Network sector want to automate the detecting the attacks of packet transfers from eligibility process (real time) based on the connection detail.
2) To automate this process by show the prediction result in web application or desktop application.
3) To optimize the work to implement in Artificial Intelligence environment.