Project Proposal Report

Abstract

The objective of this project is to use various classification and clustering algorithms to classify cyber-attacks in network traffic. With the growing number of cyber-attacks, it has become increasingly important to detect and classify them in order to prevent and mitigate the damage they can cause. In this project, different algorithms will be applied to classify and cluster network traffic data to determine if an attack has occurred. The results will be analyzed and compared to identify the best algorithm for this task.

Introduction

Cyber-attacks can cause significant damage, including financial losses, reputation damage, and the compromise of sensitive information. With the increasing reliance on computer networks and the Internet, the risk of cyber-attacks has become even greater. In order to prevent and mitigate the damage caused by cyber-attacks, it is important to detect and classify them in a timely and accurate manner. In this project, we aim to apply various classification and clustering algorithms to the problem of classifying cyber-attacks in network traffic.

Problem Formulation

The problem formulation for this project involves identifying the most effective classification and clustering algorithms for detecting and classifying cyber-attacks in network traffic data. The project will involve collecting network traffic data and preprocessing it for analysis. Moreover, it includes determining which features of network traffic data are most useful for detecting cyber-attacks.

Objectives

1. Collect and preprocess network traffic data
2. Identify relevant features for classification
3. Apply different classification algorithms
4. Compare and analyze the results
5. Draw conclusions and make recommendations

Dataset Description

- The "Dataset.txt" file in the project folder contains the complete dataset for the classification of cyber-attacks in network traffic. Each row in the dataset represents a single network traffic instance, and each column specifies the attributes of that instance. These attributes may include packet size, protocol type, source and destination IP addresses, and others.
- The "Attack_types.txt" file summarizes the possible attack types that may be present in the network traffic data. Each line in the file corresponds to a unique attack type, and includes a description of the attack. This file is useful for understanding the different types of attacks that may be present in the network traffic data, and can be used for labeling the instances in the dataset accordingly.

Data Preprocessing

1. Data cleaning
2. Feature selection
3. Data normalization
4. Handling categorical variables
5. Splitting the data
6. Balancing the dataset

Classification and Clustering Algorithms

- Support Vector Machines (SVM): SVM is a powerful classification algorithm that can be used for both binary and multi-class classification. It works by finding the hyperplane that maximizes the margin between the two classes.
- K-Nearest Neighbors (KNN): KNN is a simple and intuitive classification algorithm that assigns a new data point to the class that is most common among its K nearest neighbors in the training data.
- K-Means: K-Means is a clustering algorithm that partitions the data into K clusters based on the similarity between the data points. It works by minimizing the sum of squared distances between the data points and their assigned cluster centers.

Comparison and Performance Evaluation

- Accuracy: The percentage of correctly classified instances.
- Precision: The proportion of correctly classified positive instances out of all instances that were classified as positive.
- Recall: The proportion of correctly classified positive instances out of all actual positive instances.
- Confusion matrix: A table that shows the number of correctly and incorrectly classified instances for each class.

Conclusions

Based on the evaluation, we can conclude that the performance of the algorithms varies depending on the type and complexity of the cyber-attacks present in the network traffic dataset. Therefore, it is essential to consider various metrics and factors such as computational efficiency, ease of implementation, and interpretability while selecting the most suitable algorithm for a particular scenario.