CMPE 255 Project Proposal - Group1

Project Title: Computer Network Intrusion Detection

Team Information:

- Yi Hu (015239913) yi.hu@sjsu.edu
- Arun Hiremath (014527877) arungangayya.hiremath@sjsu.edu
- Xin Miao (3086569) xin.miao@sjsu.edu

Project Description:

The objective of this project is to build some classifiers to distinguish computer network intrusions from normal connections.

Given the basic features of TCP connections, content features, and traffic features, we need to predict if the connection is a normal or a malicious one.

Dataset:

The dataset for the project is KDD Cup 1999. It is a simulation of the military computer network.

The training dataset contains 42 features and 4,898,431 records. The testing dataset contains 42 features (labeled) and 311,029 records. Each record stands for a network connection, and is labeled with either "normal" or "attack" with a specified attack type. The intrusion can be mainly categorized into 4 types: Denial of Service (DOS), unauthorized access from remote machines (R2L), unauthorized access to local root privileges (U2R), and Probing.

The training dataset contains 22 types of intrusions. The test dataset contains additional 14 types of intrusions, which makes the project more realistic.

Filename	Description
training_attack_types	A list of intrusion types (22)
kddcup.names	A list of features (42)
kddcup.data	The entire dataset (4,898,431 rows, 42 columns)
kddcup.data_10_percent	A 10% subset of the entire dataset (494,020 rows, 42 columns)
corrected	Test dataset with labels (311,029 rows, 42 columns)
kddcup.testdata.unlabeled	Test dataset without labels(2,984,154 rows, 41

	columns)
kddcup.testdata.unlabeled_10_percent	A 10% subset of the test dataset without labels (311,029 rows, 41 columns)
kddcup.newtestdata_10_percent_unlabeled	A 10% subset of a new test dataset without labels (311,079 rows, 41 columns)

Proposed Methodology:

In this project, we plan to apply data preprocessing, data analysis and cleaning, and multiple classification methods such as decision trees, k-means clustering, k nearest neighbor, multi-level perceptron, random forest classifier, support vector machine, naive bayes, logistic regression.

We plan to implement at least 3 of these algorithms and compare the prediction performance. We will implement more if time permits.

For the evaluation indicators, we plan to use accuracy, precision, recall, f1 score to evaluate the prediction performance.

References:

1. https://www.kaggle.com/galaxyh/kdd-cup-1999-data?select=kddcup.names
2. http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html