Real-time intrusion detection system of network traffic data

Problem Definition :

Intrusion Detection Systems (IDS) have emerged as crucial components in computer and network security. While existing classification methods each have their strengths and weaknesses, there remains significant room for improvement in addressing their limitations. Recent years have seen a proliferation of proposed IDS models and extensive research comparing various classification techniques. We are building a real-time network intrusion detection system that uses trained anomaly detection to detect outliers that will be classified into one of the labels.

Dataset :

The CIC-IDS- 2017 dataset consists of generated network traffic data using CICF lowMeter, which resembles the true real-world data (PCAPs). We can monitor the network traffic in real-time and detect if the activity is suspicious or benign. The dataset can be found on this link. The malicious attacks that this dataset captures are Brute Force FTP, Brute Force SSH, DoS, Heartbleed, Web Attack, Infiltration, Botnet, and DDoS. There are 79 features, including connection-related attributes like timestamps, source and destination IPs, ports, protocols, etc.

Background

In paper [1], Sharafaldin et al. (2018) talk about how previous datasets to analyze network traffic are inefficient as they lack traffic diversity and volume or do not cover the variety of attacks to build robust anomaly detection and classification systems. The CIC-IDS2017 dataset produced by the authors of [1] contains realistic traffic patterns with benign and diverse label attack scenarios.

Paper [2] explores the AdaBoost machine-learning technique using the CIC-IDS2017 dataset. The authors use the Synthetic Minority Oversampling Technique (SMOTE) to solve class imbalance as the number of benign data exceeds the author data. Additionally, the authors use Principal component analysis and the SMOTE as preprocessing techniques to significantly enhance the AdaBoost classifier's accuracy. The paper uses metrics like accuracy, precision, recall, and F1 Score to evaluate the algorithm's performance.

Methodology

Data Preprocessing-

1. Synthetic Minority Oversampling Technique (SMOTE) :
    a. Addresses class imbalance in datasets, which can cause the model to perform poorly if it is not handled.
    b. Helps balance the dataset by interpolating minority instances. This helps the model learn patterns in the dataset better.
2. Standardization:
    a. Rescales features, ensuring that each contributes equally and no single feature influences the results.
3. Principal Component Analysis:
    a. The CIC-IDS2017 has 79 features. PCA will reduce the number of features and preserve as much variance.
4. Label Encoding:
    a. Label encoding converts categorical labels into numerical values, as many ML algorithms require numerical input.

Machine learning:

Supervised ML :
We will use **AdaBoost** classifier and **Random Forest** classifier to detect the nature of the network traffic and compare results using various metrics.

Unsupervised ML :
We will use **Isolation Forest** to detect outliers in network traffic data which will further use the classifier.

Results + Discussion

Project Goals-

Our project's goal is to use different ML strategies at different stages to detect real-time anomalies. For testing, we use multiple metrics at different stages of the process.

Testing-

1. **PCA:** We will use Cumulative Explained Variance Ratio

    a. Will indicate amount of information retained
    b. A higher value indicates that information is retained after dimensionality reduction- allows choosing optimal number of dimensions.
2. **Isolation Forest:** K-fold cross-validation
    a. Balances Precision and recall, ensures fewer false positives/ negatives.
    b. A balanced value will indicate a good separation of anomalies.
3. **Classification:** Similar to [2], we will use Precision, Recall, F1 score and accuracy
    a. **Precision**: *(True positives)/(Total positives)*- high value would mean lower number of false positives.
    b. **Recall**: *(True positives)/(Total positives)*- high values indicates model catches more anomalies
    c. **F1**: Harmonic mean of Precision and Recall
    d. **Accuracy**: *(Correctly Predicted Samples)/(Total Samples)*- gives a basic idea of model performance, but having the confusion matrix gives a comprehensive performance idea.

Gantt Chart

Contribution Table

Team Member	Proposal Contributions
Sahil Samantary	Ideation, Brainstorming, Introduction, Background, Lit Survey
Harshit Gupta	Ideation, Brainstorming,Methodology, Potential Results
Archit Vikas Shinde	Gantt Chart, Ideation, Brainstorming, Website
Aumkar Makarand Gadekar	Ideation, Brainstorming, PPT, Video

References

[1] Intrusion detection evaluation dataset (CIC-IDS2017)
[2] Yulianto, Arif, Parman Sukarno, and Novian Anggis Suwastika. "Improving adaboost-based intrusion detection system (IDS) performance on CIC IDS 2017 dataset." Journal of Physics: Conference Series. Vol. 1192. IOP Publishing, 2019.

[3] Towards Model Generalization for Intrusion Detection: Unsupervised Machine Learning Techniques