Large-Scale GPU-Accelerated Machine Learning Project

Reema Aldabass

KDD Cup 99 Dataset

Introduction to the KDD Cup 99 Dataset

The KDD Cup 99 dataset is a benchmark dataset widely used for evaluating intrusion detection systems (IDS). It consists of simulated network traffic records labeled as either normal or various types of attacks, with each record described by 41 features.

Dataset Overview:

- Source: KDD Cup 99 dataset
- Size: 3,000,000 rows × 41 columns
- Download Speed: ~0.3s at 69.0 MB/s
- GPU Data Loading Time: 0.93 seconds

Data Preprocessing:

- Dropped Columns: num_outbound_cmds, is_host_login, su_attempted, src_bytes
- Label Encoding: Applied on categorical columns protocol_type, service, flag
- Missing Values: No missing values detected
- Feature Scaling: StandardScaler (GPU-based)
- GPU Preprocessing Time: 8.66 seconds

Data Imbalance & Oversampling:

- Original Class Distribution:
  - Class 0.00 (Normal): 2,924,875 samples
  - Class 1.00: 35,693 samples
  - 99 other rare attack types
- Oversampling:
  - Minority classes upsampled using sklearn.utils.resample on CPU
  - Final training set rebalanced and moved back to GPU memory

Model Training :

- GPU Acceleration drastically reduced both loading and training times (especially for Random Forest).
- High Accuracy (>98%) achieved across all models after rebalancing the dataset.
- KNN is computationally expensive at prediction stage—less suitable for large-scale datasets.

```
Logistic Regression Accuracy: 98.81%
Logistic Regression Prediction time: 0.05 seconds

Random Forest Accuracy: 98.81%
Random Forest Prediction time: 0.35 seconds

KNN Accuracy: 98.46%
KNN Prediction time: 168.36 seconds
```

```
 Training Logistic Regression...
Logistic Regression training time: 91.05 seconds

 Training Random Forest...
Random Forest training time: 25.41 seconds

 Training KNN...
KNN training time: 0.13 seconds
```