Botnet Detection Report

Ahmed Yasser Ibrahim

222110758

SE495

Introduction

Bots are groups of computer devices that are compromised by a hacker without their owner's knowledge. They can be used for various purposes, such as launching cyberattacks or spamming.
In this lab, our task is to identify these bots by studying a set of features, including total forwarded packets, total acknowledgment flags, rate of flow of bytes, etc.
This study will identify the most important features that can help us detect those bots and mitigate their damage.

Problem Statement

The use of bots by hackers is rapidly increasing, and there is a need for models that can classify devices connected to the network as bots or legitimate devices to help mitigate the damage that bots cause when connected to the internet.

Data Analysis

To conduct this study, we used the CTU-13 dataset, which is a well-known dataset that is considered a benchmark for measuring the performance of botnet detection tools. This dataset comprises over 90,000 entries and 33 attributes (split between the normal and attack CTU CSV files). This dataset is labeled, allowing us to train and test this model with relative ease.

We start by reading both CSV files and concatenating them. In the data cleaning phase, we check for null values and check the statistical description of the data. The dataset is already clean given that it is widely used and carefully constructed by the provider.

We constructed the correlational matrix, which showed some correlations between features but was challenging to interpret given the sheer number of features included. Therefore, we decided to truncate some of the features in the dataset to reduce the computing load and facilitate the study. The features retained are predicted to have the most effect on the classification results based on domain knowledge. Further testing can be done to optimize the results.

We constructed some plots and histograms to visualize the data and added some additional features through calculations that combined certain features. We then scaled the data using standard scaling to ensure that features contribute to the model equally and achieve consistent results. Then, we split the data between training and testing and started training the models. The results are shown below:

Feature Engineering

We performed feature engineering to enhance the quality of our dataset. This included:

- Cleaning and standardizing numerical features.
- Extracting new features through mathematical transformations.
- Removing redundant or highly correlated features.
- Standardizing feature values to ensure equal contribution to the models.

Results and Discussion

Logistic Regression:

```
precision    recall  f1-score   support

      0       0.79      0.89      0.84      2750
      1       0.84      0.72      0.77      2237

accuracy                           0.81      4987
macro avg       0.82      0.80      0.81      4987
weighted avg       0.82      0.81      0.81      4987

False Positive Rate (FPR): 0.1102
```

SVM Report:

```
precision    recall  f1-score   support

       0       0.91      0.87      0.89      2750
       1       0.85      0.90      0.87      2237

accuracy                           0.88      4987
macro avg       0.88      0.89      0.88      4987
weighted avg       0.89      0.88      0.88      4987

False Positive Rate (FPR): 0.1276
```

Decision Tree Report:

```
 precision    recall  f1-score   support

       0       0.98      0.99      0.98      2750
       1       0.98      0.98      0.98      2237

 accuracy                           0.98      4987
 macro avg       0.98      0.98      0.98      4987
 weighted avg       0.98      0.98      0.98      4987

 False Positive Rate (FPR): 0.0142
```

Based on the results, logistic regression and SVM yielded similar results, with the accuracy being approximately 88%. However, decision trees yielded an accuracy of 98%, showing significantly better performance than the other two models with a low FPR, indicating that the model is not too lenient either.

Feature Importance Ranking for Botnet Detection:

Feature Importance

Packet_Size_Variance 0.183262

SYN Flag Cnt 0.173465

ACK Flag Cnt 0.158713

Bwd IAT Mean 0.103290

Tot Fwd Pkts 0.071716

TotLen Fwd Pkts 0.069952

Fwd IAT Mean 0.062624

BytesPerPacket 0.059462

Flow Pkts/s 0.039510

Flow Duration 0.025270

FlowRate 0.018832

PacketRate 0.017998

FIN Flag Cnt 0.005908

ByteRatio 0.005221

PacketToByteRatio 0.004778

This shows that most features were important in achieving an accurate classification, but the most important features were packet size variance, count of SYN/ACK flags, and inter-arrival times.

Conclusion

The study successfully identified key features for botnet detection and evaluated three classification models: Logistic Regression, SVM, and Decision Trees. While Logistic Regression and SVM performed adequately, the Decision Tree model significantly outperformed them with 98% accuracy and the lowest false positive rate. This suggests that decision trees provide an effective approach for botnet classification in the CTU-13 dataset.

Future work can focus on optimizing hyperparameters, experimenting with ensemble methods, and testing on other datasets to enhance generalization. Additionally, real-time detection implementation should be explored to ensure proactive botnet mitigation in live network environments.