Intrusion Detection Using Machine Learning

Adnan Athar Janwari V00959422
MEng, Electrical And Computer Engineering

Supervisor: Dr. Fayez Gebali
Intrusion Detection System?

An Intrusion Detection System (IDS) is software or hardware that detects any suspicious behavior and notifies the network administrator.
Growing cyberattacks and increasing vulnerabilities in network security

Detecting cyberattacks that haven't happened yet
METHODOLOGY

- CSE-CiC-IDS2008 Dataset
- Data Preprocessing
- Feature Reduction
- Model Building and Testing
- Result Analysis
METHODOLOGY (continued)

- CSE-CIC-IDS dataset (Canadian Institute for Cybersecurity and the Communications Security Establishment)
- 14 network attacks and benign traffic
- 80 features
- Weka (Waikato Environment for Knowledge Analysis) is used to implement ML algorithms
METHODOLOGY (continued)

Steps involves in the data preprocessing

• Removal of Zero Value Features
• Removal of Duplicate Value
• Linear Discriminant Analysis (LDA) is used to reduced to dimension of the data
• 10 highly ranked features are selected
METHODOLOGY (continued)

Three classifiers are used.

- Naive Bayes (NB)
- Random Forest (RF)
- Decision Tree (DT)
SMOTE and Resample are used on the dataset to balance it.

- SMOTE is an oversampling technique that replicate the instances of minority classes.
- Resample is an undersampling technique that removes the instances of the majority classes.
## METHODOLOGY (continued)

### Dataset Instances

| Dataset                        | Total Number of Instances | Benign Instances | Attack Instances |
|-------------------------------|---------------------------|------------------|------------------|
| Without under and over sampling | 1,714,800                 | 381,936          | 1,332,864        |
| With under and over sampling  | 1,893,238                 | 381,936          | 1,511,302        |
PERFORMANCE EVALUATION METRICS

- Accuracy
- Precision
- Recall
- F-Measure
- Execution Time
# EXPERIMENT ONE RESULT

Result without Oversampling and Undersampling

| Classifier     | Accuracy | Precision | Recall | F-Measure | Execution Time (s) |
|----------------|----------|-----------|--------|-----------|--------------------|
| Naive Bays     | 76.10    | 82.8      | 76.1   | 75        | 14.17              |
| Random Forest  | 92.99    | 92.5      | 93.0   | 92.7      | 4165.05            |
| Decision Tree  | 93.50    | 93.2      | 93.5   | 92.6      | 351.64             |
# EXPERIMENT TWO RESULT

Result with Oversampling and Undersampling

| Classifier       | Accuracy | Precision | Recall | F-Measure | Execution Time (s) |
|------------------|----------|-----------|--------|-----------|--------------------|
| Naive Bays       | 84.14    | 82.9      | 84.1   | 82.7      | 4.19               |
| Random Forest    | 93.28    | 92.8      | 93.3   | 92.8      | 2151.78            |
| Decision Tree    | 93.39    | 93.2      | 93.4   | 92.5      | 94.66              |
RESULT COMPARISON

Result comparison in terms of accuracy with and without Under sampling and Oversampling

| Algorithm        | Accuracy Without U & O sampling | Accuracy with U & O sampling |
|------------------|---------------------------------|-----------------------------|
| NAIVE BAYS       | 76.1                            | 84.14                       |
| RANDOM FOREST    | 92.99                           | 93.28                       |
| DECISION TREE    | 93.39                           | 93.5                        |
RESULT COMPARISON

Result comparison in terms of execution time with and without Under sampling and Oversampling

| METHOD          | Execution Time (s) Without U & O sampling | Execution Time (s) with U & O sampling |
|-----------------|------------------------------------------|----------------------------------------|
| NAIVE BAYES     | 14.17                                     | 14.17                                   |
| RANDOM FOREST    | 4165.05                                   | 2151.78                                 |
| DECISION TREE    | 351.64                                    | 94.66                                   |
CONCLUSION

• In terms of accuracy Decision Tree performed the best followed Random Forest and Naive Bays
• Naive Bays execution time was fastest
FUTURE WORK

- Other ML classifiers can be used in future
- WEKA supports ScikitLearnCLassifier written in Python can be considered
Thank You
Question?