Sardar Patel Institute of Technology,Mumbai

Department of Electronics and Telecommunication Engineering

T.E. Sem-V (2018-2019)

ETL54-Statistical Computational Laboratory

Lab-6:Analysis of Various Classifiers and to detect computer network intrusion

Name:Manish Dsilva **Roll No.15**

UID 2017120013

Objective: To detect computer network intrusion using machine learning techniques.

Outcomes:

1) T load dataset in R or Python

2)To process and prepare data

3)Build a model for intrusion detection.

4) Use machine learning algorithms to detect intrusion.

5) Calculate accuracy and confusion matrix and prediction time.

to be completed by each Batch on their respective Lab daySystem Requirements:

Meta-Data for the Database Used:

This database(KDD Cup 1999 Data) contains a standard set of data to be audited, which includes a wide variety of intrusions simulated in a military network environment.

Attacks fall into four main categories:

DOS | denial-of-service | syn flood |

R2L | unauthorized access from a remote machine | guessing password |

U2R | unauthorized access to local superuser privileges | ‘’buffer overflow’’ attacks |

probing | surveillance and other probing | port scanning

Analysis:

We have taken 1% of the original (around 50,000 train and 40,000 test)(Certain Models i.e SVM were taking more than 30mins while doing for 10%)

Algorithm	Time to Train	Time to Predict	Accuracy	Precision	Recall	F1-score
LDA	0.64s	0.0378s	95.63%	0.99	0.96	0.97
Naïve-Bayes	0.33s	0.67s	86.88%	0.99	0.87	0.91
SVM	68.75s	14.54s	98.75%	0.98	0.99	0.99
Random Forest	0.436s	0.0371s	99.76%	1	1	1
k-NN	7.91s	27.59s	99.63%	1	1	1

Conclusion:

- The lesser the time taken to train the model the better it is for the user and hence we can say that the model is computationally less expensive.
- Often the model is trained once at the server and we need to predict multiple times this means time to predict is also very crucial in choosing a model
- Accuracy, Precision, Recall are 3 more parameters which tells us how our model has performed on unseen data
- For the above Problem we observe that Random Forest has performed better than other algorithms on most parameters. It is an example of ensemble learning.
- Scikit-Learn has provided us a user-friendly environment for using all the above classification algorithms. As a result to change between algorithms we didn't have to change more than 2 lines