Assignment 5 - Network Intrusion Detection using ML Techniques

Group 13
Satvik Padhiyar
Nisha M

PART- A (Datasets)

Q1. What are the different objectives of generating benchmark datasets for intrusion detection and explain each objective in no more than two sentences? Refer this

Realistic network and traffic

Analyzing a dataset should be as accurate as possible, with no artificial post-capture trace insertion. This includes both normal and anomalous traffic, as well as normal and non-anomalous traffic.

Labeled dataset

Analyzing data in a controlled and deterministic environment allows for the separation of anomalous activity from normal traffic while also eliminating the inefficient process of manually labeling individual pieces of data."A labeled dataset is critical in the evaluation of various detection mechanisms."

Total interaction captures

All network interactions, whether inside or between internal LANs, must be included in a dataset. This allows for the detection of abnormal behavior and is required for post-evaluation and interpretation of the data.

Complete capture

Network security researchers have struggled to get real-world network traces. The goal of this study is to create network traces in a regulated testbed environment,

eliminating the requirement for sanitization and keeping the naturalness of the generated dataset. Data suppliers are hesitant to give this type of information.

Diverse intrusion scenarios

Attack scenarios are designed to perform a diverse set of multi-stage attacks. Such attacks can cause far more serious disruptions than traditional brute force attempts and require more in-depth insight into IP services and applications for their detection. The aim of this research is to identify emerging trends in security threats.

Q2. What are the drawbacks of the existing **KDDCup'99** dataset that led to the formation of its refined version **NSL-KDD**? Why are KDPCUP'99 and NSL-KDD datasets considered unreliable to validate novel intrusion detection algorithms of late? Refer: https://ieeexplore.ieee.org/document/8586840

1) Size And Redundancy

KDD's training dataset comprises 4,898,431 data points, as specified. However, due to considerable duplication (78%), only 1,074,992 unique data points exist. Similarly, the test dataset is 89.5 percent redundant; it was reduced from 2,984,154 to 311,029 patterns. In this study, we look at these reduced datasets.

2) Features

As with KDD-99, certain parameters were found unnecessary. A reduced set of 20 features found by Mean Decrease Impurity was used in this paper.

3) Skewness

The dataset's biased character is obvious. 98.61 per cent of the data falls into one of two categories: normal or dos. This impairs classifier performance in the remaining classes.

4) Non-Stationarity

The KDD-99 dataset contains non-stationary examples rather than those that are trained on a stationary partition. The training set has 23% of its data as DoS examples versus 73.9% in the test set. It has been demonstrated that such divergence negatively affects performance.

Because these datasets are more than two decades old, they do not reflect modern-day scenarios such as 5G traffic. It also has fewer classification labels when compared to datasets like UNSW-NB15 etc.

Q3. Sketch a table specifying the different properties of the following datasets KDD CUP'99, NSL-KDD, CICIDS 2017, CICIDS 2018, UNSW-NB15

Properties - Year of public availability of dataset, Number of features, Number of different class labels, Names of different types of attacks.

	Year of public availability of dataset.	**Number of features**	**Number of different class labels**	**Names of different types of attacks**
KDD CUP'99	October 28, 1999	41	5 such as DOS, Probe, R2L, U2R, Normal.	back dos buffer_overflow u2r ftp_write r2l guess_passwd r2l imap r2l ipsweep probe land dos loadmodule u2r multihop r2l neptune dos nmap probe perl u2r phf r2l pod dos portsweep probe rootkit u2r satan probe smurf dos spy r2l teardrop dos warezclient r2l warezmaster r2l
NSL-KDD	Modified version of KDD 99	43	5 such as DOS, Probe, R2L, U2R, Normal.	back dos buffer_overflow u2r ftp_write r2l

				guess_passwd r2l imap r2l ipsweep probe land dos loadmodule u2r multihop r2l neptune dos nmap probe perl u2r phf r2l pod dos portsweep probe rootkit u2r satan probe smurf dos spy r2l teardrop dos warezclient r2l warezmaster r2l
CICIDS 2017	2017	83	7	DDoS DoS slowloris Slowhttptest Hulk GoldenEye Heartbleed PortScan Bot Brute-Force FTP-Patator SSH-Patator Web AttackWeb Attack-Brute Force Attack-XSS Attack-SQL Injection Infiltration
CICIDS 2018	2018	83	4 types Benign, DDoS, Port scan, Bot	DDoS DoS slowloris Slowhttptest Hulk GoldenEye Heartbleed PortScan Bot Brute-Force FTP-Patator SSH-Patator Web AttackWeb Attack-Brute Force Attack-XSS Attack-SQL Injection Infiltration
UNSW-NB15	2015	49	10	Fuzzers, Analysis, Backdoors, DoS, Exploits, Generic, Reconnaissance, Shellcode Worms

PLAGIARISM STATEMENT

We certify that this assignment/report is our own work, based on our personal study and/or research and that we have acknowledged all material and sources used in its preparation, whether they be books, articles, packages, datasets, reports, lecture notes, and any other kind of document, electronic or personal communication. We also certify that this assignment/report has not previously been submitted for assessment/project in any other course lab, except where specific permission has been granted from all course instructors involved, or at any other time in this course, and that we have not copied in part or whole or otherwise plagiarized the work of other students and/or persons. We pledge to uphold the principles of honesty and responsibility at CSE@IITH. In addition, We understand my responsibility to report honor violations by other students if we become aware of it.

Names: Nisha M , Satvik Padhiyar
Date: 21/03/2022
Signature: Nisha M , Satvik Padhiyar