Assignment-5: Network Intrusion Detection using ML Techniques

Group Size: 2 (refer the appendix to know your group details)

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: The data set should represent data about realistic network traffic and not any artificial capture traces. Artificial captures would negatively affect the data and introduce inconsistencies in the dataset. When the captures are realistic, we can clearly understand the effects of attacks over the network and how the corresponding workstations respond to it.

· Labeled dataset: A labeled dataset is extremely important for evaluation of various models. So, it is created in a controlled and deterministic environment which eliminates the impractical task of manual labeling.

· Total interaction capture: A good dataset must include all the network interactions both intraLAN and interLAN. Because the more the data, the better the analysis we can do.

· Complete Capture: The traces should be completely captured including the payload for accurately evaluating and comparing systems. The limited captures prove as a massive disadvantage to researchers.

· Diverse Intrusion Scenarios: The complexity of attacks has increased exponentially over the years. So, the researchers need to do an in depth analysis of possible scenarios and recreate situations to identify and resolve all sorts of 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

Drawbacks of KDDCup'99:

· Too much data, so it becomes computationally intensive to use for training models

· Many features/parameters were found unnecessary

· KD99 is a skewed dataset towards Normal and Dos categories.

· Dataset is non-stationary which causes divergence which in turn negatively impacts performance

Although NSL KDD is a refined version of the KDDCup'99 dataset, it is still irrelevant in modern times w.r.t many aspects. Machine learning techniques optimized on KDD-99 and NSL-KDD may be exposed to minority class attacks while claiming a higher efficacy. They are still relatively a lot skewed in comparison with the latest datasets such as UNBW and CICIDIS. The newer datasets are less redundant and less skewed and in relevance to today's attacks and latest technologies available. The KDD datasets are more than a decade old which is less relevant.

Q3. Sketch a table specifying the different properties of the following datasets

KDD CUP'99, NSL-KDD, CICIDS 2017, CICIDS 2018, UNSW-NB15

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

Properties -> Names	Year of availability	Number of features	Number of class labels	Names of different attacks
KDD CUP'99	1999	41	23 subclasses in 5 classes	Normal,Dos,Probe U2R, R2L
NSL-KDD	2009	43	39 Subclasses in 5 classes	Normal,Dos,Probe U2R, R2L are the names of the attack classes.
CICIDS 2017	2017	83	15 classes	Benign,DoS Hulk, PortScan, DDoS attack, DoS GoldenEye, FTP-Predator, SSH-Patator,DoS slowloris, DoS Slowhttptest, Bot, BruteForce Web Attack XSS,Web Attack SQL Injection

CICIDS 2018	2018	83	7 classes	Normal,Bruteforce attack, DoS attack, Web attack, Infiltration attack, Botnet attack, DDoS+PortScan
UNSW-NB 15	2015	49	10 classes	Normal,Fuzzers,Analysis,Backdoors, Dos,Exploits,Generic, Reconnaissance,ShellCode,Worms

PART-B (Anomaly Detection)

For the following experiments, we will use the **CICIDS2017 and UNSW-NB15 partial datasets.**

CICIDS2017:
Friday-WorkingHours-Afternoon-DDos.pcap_ISCX.csv,Friday-WorkingHours-Afternoon-PortScan.pcap_ISCX.csv and Friday-WorkingHours-Morning.pcap_ISCX.csv

UNSW-NB15:
UNSW-NB15_4.csv, UNSW-NB15_3.csv

Required **CSV** files can be downloaded from here

This is a process of sanitizing our data. We need to ensure that our data is free from ambiguities, outliers, and biases. The first step towards realizing it would be to remove source IP, destination IP, Source port, Destination port, and Timestamp from CSV files which may add bias to the detection algorithm. Now concatenate each individual CSV file of corresponding datasets to form a unified single dataset per each CICIDS2017 and UNSW-NB15. Replace all the attack labels with label '1' and benign/no-attack sample labels with label '0'.

At this point, we translated the dataset which is suitable for anomaly detection. Replace all the categorical feature values with label encoding (don't use one-hot encoding), which inherently doesn't increase the dimensionality of the data. Drop the columns which have more than 25% NaN or NULL values, and replace NaN/NULL values in other columns with the average(or most repetitive) value of the corresponding column. Next, remove data duplication by eliminating all redundant rows and also multi-label rows from each dataset. Eventually, normalize the data using a min-max scaler

Feature Selection (Subtask-2)

This is an important component to select the most representative features on which the predictor variable has a higher dependency. In other words, it also reduces the computation time of the learning algorithm. We use the following feature selection technique(s)

Univariate feature selection: use the SelectKBest method with chi-square as score function. Try with at least three different 'k'-values (Refer here)

At this juncture, we will have three variants of a refined version of datasets corresponding to different 'k'-value per each CICIDS and UNSW-NB15 (In total we will have six datasets).

We also offer additional Bonus points for those extending the feature selection strategy to Principal Component Analysis (PCA) with a total variance of the components of at least 90%

This completes the feature engineering task. Now split each dataset into Train, ~~validation~~, and Test subsets (80:20)

Experiment with different models (Subtask-3)

Modeling is typically a relationship assumption between the dependent and predictor variable(s). We approximate this relationship with the help of function estimators ranging from simple linear to complex non-linear models. In this sub-task, we will try to model anomaly detection with the following models

Gaussian Naive Bayes, Logistic Regression, SVM (RBF kernel), Decision Trees (ID3, C4.5), Random Forest, XGBoost (Refer), Voting classifier (Refer), AdaBoost Classifier (you are free to choose the parameter), MLP classifier (with one hidden layer of size 100).

~~Q1. For all the models, plot a Training loss Vs validation loss. Comment about the **overfit (if exist)** from the plot.~~
Q2. Create the confusion matrix for each model (with the test data), sketch a single table containing accuracy, precision, recall, f1-score, and running time(s) comparison among all the models
Q3. Plot the AUC-ROC curve for all the models

Dealing with Class Imbalance in Network Traces(Subtask-4)

Class imbalance is an inherent problem in anomaly detection/novelty detection. There are various ways to handle this issue i.e., data and algorithmic driven. In this subtask, we focus on data-driven approaches (a.k.a resampling techniques). We take the help of a well-known class imbalance library for our experiments.

Now for this task, we take the dataset after processing the subtask-2, and we will apply the following resampling techniques (free to choose the parameter)

1. Random OverSampling(ROS)
2. Random underSampling(RUS)
3. Synthetic Minority Oversampling TEchnique (SMOTE)

Now your latest datasets are partly balanced (class imbalance ratio will be reduced), **repeat the subtask-3 for these balanced datasets.**

Credits

Task	**Akash**	**Srivathsa**
Part A Q1	Proofread	Wrote Total Answer
Part A Q2	1. Written Drawbacks of KDD 2. KDPCUP'99 and NSL-KDD datasets are considered unreliable to validate novel intrusion (4/7)	3. KDPCUP'99 and NSL-KDD datasets are considered unreliable to validate novel intrusion (3/7)
Q3	Among 5 datasets I've written about CICIDS 2018 and UNSW-NB15	Among 5 datasets I've written about KDD,NSL-KDD,CICIDS 2017
PART B		
Sub task 1	1. Decided upon what features to be removed and dealing with missing values and Nans (discussion)	1. Data understanding and learning ML concepts for assignment and Code. 2. Decided upon what features to be removed and dealing with missing values and Nans (discussion)
Sub task 2	1. Written code for select K best and PCA	1. Discussion on how SelectKBest works
Sub task 3	1. Choosing models (discussion) 2. Written code	1. Choosing models(discussion)
Sub task 4	1. Discussion 2. Written code	1. Discussion 2. Helped searching resources online and discussed while writing code

Deliverables

A tar ball (prefix as your Group ID) with:

1. A report describing your answers for PART-A of the assignment
2. For PART-B, you need to submit the python-notebook containing
    a. All the plots, confusion matrix, and comparison tables
    b. It should be executable on Google's Colab environment
3. Credit Statement (1-pager): share an accurate and detailed description of each of the group member's contributions to the assignment in terms of coding, plots, report writing, bug fixes, analysis, etc.

PLAGIARISM STATEMENT <Include it in your report>

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: Akash Tadwai, Srivathsa L Rao
Date: 21/03/2022
Signature: Akash, Srivathsa

Late Policy:

10% cut in marks for each late day beyond buffer days

Appendix (Group Details)

Group1	Akash Tadwai	Srivathsa L Rao
Group2	Amit Kumar	Devang Deviprasad Dubey
Group3	Arkadeb Ghosh	Harinder Kaur
Group4	Ayan Kumar Pahari	K Shiv Kumar
Group5	Divya Pathak	Nilesh Shivanand Kale
Group6	Gurpreet Singh	Pallavi Saxena
Group7	Kamal Shrestha	Pradhumn Kanase
Group8	Kishan Nayanbhai Bhinde	Pratik Abhijeet Bendre
Group9	KOMMANA VIKAS	Priyansha Tiwari
Group10	Koustav Choudhury	REVANTH ROKKAM
Group11	Madhvendra Singh Chouhan	Rishabh Dongre
Group12	MUDAVATH SHATHANAND SAI	Sampath Kumar Sivampeta
Group13	NISHA M	Satvik Padhiyar
Group14	Pamuluri Ravi Sankar	Siddhesh Pratim Sovitkar
Group15	PELLURI SRIVARDHAN	Supriya Kumari
Group16	Piyush Madhukar Dadgal	Suranjan Daw
Group17	Raguru Sai Sandeep	TATIPELLY VAMSHI
Group18	Ravi Chandra Duvvuri	Unnati Dixit
Group19	Sai Varshittha Ponnam	Visakh K Vijayan
Group20	Srivathsa L Rao	Bharti Sahu
Group21	VINTA REETHU	Praveen Chandrahas
Group22	Yogesh Ahirwar	Anwesha Kar