Study of NSL KDD Dataset

Dharaneish V C
Department of Computer Science and Engineering
Amrita School of Engineering, Coimbatore
Amrita Vishwa Vidyapeetham, India
Coimbatore - 641112, India
dharaneish@gmail.com

I. Study of NSL KDD dataset

A. *Study of NSL KDD Dataset*

The NSL-KDD dataset was introduced as part of The Third International Knowledge Discovery and Data Mining Tools Competition, which was held alongside KDD-99, The Fifth International Conference on Knowledge Discovery and Data Mining. The primary objective of the competition was to develop a network intrusion detection system capable of accurately distinguishing between "bad" connections, known as intrusions or attacks, and "good" normal connections. The dataset comprises a comprehensive collection of auditable data, including a diverse array of simulated intrusions encountered in a military network environment. It has since become a widely used benchmark dataset in the field of network security and intrusion detection, facilitating the development and evaluation of new and improved models and algorithms.[34]

The dataset can be downloaded from the site https://www.unb.ca/cic/datasets/nsl.html .

B. *Dataset Splits*

This data set is comprised of four sub data sets: KDDTest+, KDDTest-21, KDDTrain+, KDDTrain+_20Percent, although KDDTest-21 and KDDTrain+_20Percent are subsets of the KDDTrain+ and KDDTest+.

KDDTrain+ is simply referred to as train and KDDTest+ is referred to as test. The KDDTest-21 is a subset of test, without the most difficult traffic records (Score of 21), and the KDDTrain+_20Percent is a subset of train, whose record count makes up 20% of the entire train dataset. That being said, the traffic records that exist in the KDDTest-21 and KDDTrain+_20Percent are already in test and train respectively and aren't new records held out of either dataset.

C. *Features*

The dataset contains 4,94,021 tuples and 43 features per record, with 41 referring to the traffic input itself [independent] and the last two being labels (whether the traffic input is normal or attack) and Score (the severity of the traffic input itself) [dependent].

There are 4 different classes of attacks: Denial of Service (DoS), Probe, User to Root(U2R), and Remote to Local (R2L). A brief description of each attack is presented in table 3.

TABLE I. Classes of Attacks in NSL KDD Dataset

Attribute	Attribute type	Purpose
DoS	Explicit	shut down traffic flow from the target system. (IDS is flooded with an abnormal amount of traffic) Eg: online retailer getting flooded with online orders on a day with a big sale
Probe	Implicit	get information from a network act like a thief and steal important information
U2R	Implicit	Exploit the vulnerabilities to gain root privileges (starts off with a normal user account and tries to gain access to the system or network, as a super-root user)
R2L	Implicit	gain local access to a remote machine (kinda hacking)

Here Important to note is - DoS acts differently from the other three attacks, where DoS attempts to shut down a system to stop traffic flow altogether, whereas the other three attempts to quietly infiltrate the system undetected.

Break- down of sub classes of each attack is presented in the table 4.

TABLE II. Sub-Classes of Attacks in NSL KDD Dataset

Classes	Sub-Classes	Total Count
DoS	apache2, back, land, Neptune, mailbomb, pod, processtable, smurf, teardrop, udpstorm, worm	11
Proble	Ipsweep, mscan, nmap, portsweep, saint, satan	6
U2R	Buffer_overflow, loadmodule, perl, ps, rootkit, sqlattack, xterm	7
R2L	ftp_write, guess_passwd, httptunnel, imap, multihop, named, phf, sendmail, Snmpgetattack, spy, snmpguess, warezclient, warezmaster, xlock, xsnoop	15

Essentially, more than half of the records that exist in each data set are normal traffic, and the distribution of U2R and R2L are extremely low. Although this is low, this is an accurate representation of the distribution of modern-day internet traffic attacks, where the most common attack is DoS and U2R and R2L are hardly ever seen.

The distribution of the Normal and Abnormal labels in the dataset was found to be equally distributed with 77,054 rows

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

of the normal class and 71,463 rows of the attack class. A pie chart of the distribution of the Normal and attack classes is shown in Fig 1, which indicates that the dataset is well balanced between the two classes.

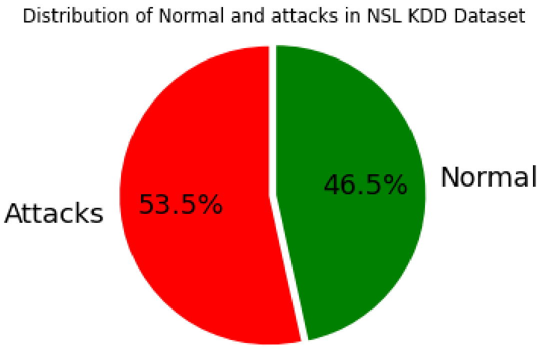

Fig. 1. Pie Chart distribution of Normal and attacks in dataset

D. *Class level Details*

The features can be broken down into four categories: Intrinsic, Content, Host-based, and Time-based. The description of it is presented in Table 5.

TABLE III. CLASSIFICATION OF FEATURES IN NSL KDD DATASET

Category	Description	Features
Intrinsic features	These can be derived from the header of the packet without looking into the payload itself, and hold the basic information about the packet.	Features 1-9
Content features	These hold information about the original packets, as they are sent in multiple pieces rather than one. With this information, the system can access the payload. This category contains features 10–22.	Features 10-22
Time-based features	These features hold the analysis of the traffic input over a two-second window and contain information like how many connections it attempted to make to the same host. These features are mostly counts and rates rather than information about the content of the traffic input	Features 23-31
Host-based features	These features are similar to Time-based features, except instead of analyzing over a 2-second window, it analyzes over a series of connections made (how many requests made to the same host over x-number of connections). These features are designed to access attacks, which span longer than a two-second window time-span.	Features 32-41

E. *Feature Types*

These features types can be broken down into Categorical, Binary, Discrete and Continuous

a) 4 Categorical (Features: 2, 3, 4, 42)

b) 6 Binary (Features: 7, 12, 14, 20, 21, 22)

c) 23 Discrete (Features: 8, 9, 15, 23 –41, 43)

d) 10 Continuous (Features: 1, 5, 6, 10, 11, 13, 16, 17, 18, 19)

The detailed description about each feature in dataset is presented in Table IV.

TABLE IV. DESCRIPTION OF FEATURES IN NSL KDD DATASET

#	Feature Name	Description	Type	Value Type	Ranges
1	Duration	Length of time duration of the connection	Continuous	Integers	0 - 54451
2	Protocol Type	Protocol used in the connection	Categorical	Strings	
3	Service	Destination network service used	Categorical	Strings	
4	Flag	Status of the connection – Normal or Error	Categorical	Strings	
5	Src Bytes	Number of data bytes transferred from source to destination in single connection	Continuous	Integers	0 - 1379963888
6	Dst Bytes	Number of data bytes transferred from destination to source in single connection	Continuous	Integers	0 - 309937401
7	Land	If source and destination IP addresses and port numbers are equal then, this variable takes value 1 else 0	Binary	Integers	{0, 1}
8	Wrong Fragment	Total number of wrong fragments in this connection	Discrete	Integers	{0,1,3}
9	Urgent	Number of urgent packets in this connection. Urgent packets are packets with the urgent bit activated	Discrete	Integers	0 - 3
10	Hot	Number of "hot" indicators in the content such as: entering a system directory, creating programs and executing programs	Continuous	Integers	0 - 101
11	Num Failed Logins	Count of failed login attempts	Continuous	Integers	0 - 4
12	Logged In	Login Status: 1 if successfully	Binary	Integers	{0, 1}

		logged in; 0 otherwise			
13	Num Compromised	Number of "compromised" conditions	Continuous	Integers	0 - 7479
14	Root Shell	1 if root shell is obtained; 0 otherwise	Binary	Integers	{0 , 1}
15	Su Attempted	1 if "su root" command attempted or used; 0 otherwise	Discrete		
16	(Dataset contains ‘2’ value)	Integers	0 - 2		
17	Num Root	Number of "root" accesses or number of operations performed as a root in the connection	Continuous	Integers	0 - 7468
18	Num File Creations	Number of file creation operations in the connection	Continuous	Integers	0 - 100
19	Num Shells	Number of shell prompts	Continuous	Integers	0 - 2
20	Num Access Files	Number of operations on access control files	Continuous	Integers	0 - 9
21	Num Outbound Cmds	Number of outbound commands in an ftp session	Continuous	Integers	{0}
22	Is Hot Logins	1 if the login belongs to the "hot" list i.e., root or admin; else 0	Binary	Integers	{0, 1}
23	Is Guest Login	1 if the login is a "guest" login; 0 otherwise	Binary	Integers	{0, 1}
24	Count	Number of connections to the same destination host as the current connection in the past two seconds	Discrete	Integers	0 - 511
25	Srv Count	Number of connections to the same service (port number) as the current connection in the past two seconds	Discrete	Integers	0 - 511
26	Serror Rate	The percentage of connections that have activated the flag (4) s0, s1, s2 or s3, among the connections aggregated in count (23)	Discrete	Floats (hundredths of a decimal)	0 - 1
27	Srv Serror Rate	The percentage of connections that have activated the flag (4) s0, s1, s2 or s3, among the connections aggregated in srv_count (24)	Discrete	Floats (hundredths of a decimal)	0 - 1
28	Rerror Rate	The percentage of connections that have activated the flag (4) REJ, among the connections aggregated in count (23)	Discrete	Floats (hundredths of a decimal)	0 - 1
29	Srv Rerror Rate	The percentage of connections that have activated the flag (4) REJ, among the connections aggregated in srv_count (24)	Discrete	Floats (hundredths of a decimal)	0 - 1
30	Same Srv Rate	The percentage of connections that were to the same service, among the connections aggregated in count (23)	Discrete	Floats (hundredths of a decimal)	0 - 1
31	Diff Srv Rate	The percentage of connections that were to different services, among the connections aggregated in count (23)	Discrete	Floats (hundredths of a decimal)	0 - 1
32	Srv Diff Host Rate	The percentage of connections that were to different destination machines among the connections aggregated in srv_count (24)	Discrete	Floats (hundredths of a decimal)	0 - 1
33	Dst Host Count	Number of connections having the same destination host IP address	Discrete	Integers	0 - 255
34	Dst Host Srv Count	Number of connections having the	Discrete	Integers	0 - 255

		same port number			
35	Dst Host Same Srv Rate	The percentage of connections that were to different services, among the connections aggregated in dst_host_count (32)	Discrete	Floats (hundredths of a decimal)	0 - 1
36	Dst Host Diff Srv Rate	The percentage of connections that were to different services, among the connections aggregated in dst_host_count (32)	Discrete	Floats (hundredths of a decimal)	0 - 1
37	Dst Host Srv Diff Host Rate	The percentage of connections that were to different destination machines, among the connections aggregated in dst_host_srv_count (33)	Discrete	Floats (hundredths of a decimal)	0 - 1
38	Dst Host Serror Rate	The percentage of connections that have activated the flag (4) s0, s1, s2 or s3, among the connections aggregated in dst_host_count (32)	Discrete	Floats (hundredths of a decimal)	0 - 1
39	Dst Host Srv Serror Rate	The percent of connections that have activated the flag (4) s0, s1, s2 or s3, among the connections aggregated in dst_host_srv_count (33)	Discrete	Floats (hundredths of a decimal)	0 - 1
40	Dst Host Rerror Rate	The percentage of connections that have activated the flag (4) REJ, among the connections aggregated in dst_host_count (32)	Discrete	Floats (hundredths of a decimal)	0 - 1
41	Dst Host Srv Rerror Rate	The percentage of connections that have activated the flag (4) REJ, among the connections aggregated in dst_host_srv_count (33)	Discrete	Floats (hundredths of a decimal)	0 - 1
42	Class	Classification of the traffic input	Categorical	Strings	
43	Difficulty Level	Difficulty level	Discrete	Integers	0 - 21
44	Dst Host Srv Diff Host Rate	The percentage of connections that were to different destination machines, among the connections aggregated in dst_host_srv_count (33)	Discrete	Floats (hundredths of a decimal)	0 - 1