Hands-on Lab Description

CS-ML-00101 – Understanding NSL-KDD Dataset

Category:

```
CS-ML: Machine Learning
```

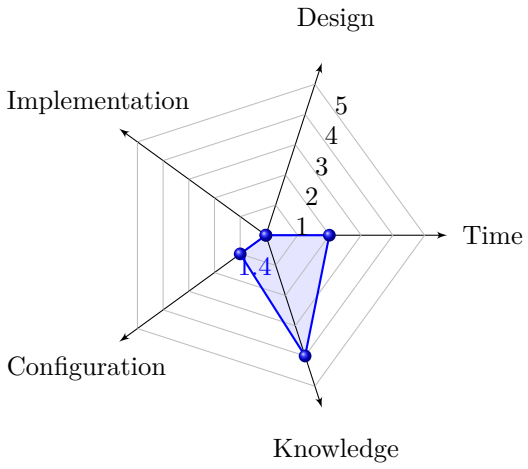
Objectives:

```
Learn NSL-KDD dataset: data distribution, features, and flags
```

Estimated Lab Duration:

```
Expert: 20 minutes
Novice: 100 minutes
```

Difficulty Diagram:



Difficulty Table.

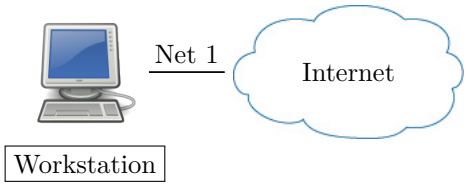
Measurements	Values (0-5)
Time	2
Design	0
Implementation	0
Configuration	1
Knowledge	4
Score (Average)	1.4

Required OS:

```
Linux: Ubuntu 18.04 LTS
```

Lab Running Environment:

```
VirtualBox https://www.virtualbox.org/ (Reference Labs: CS-SYS-00101)
```




```
1 Server: Linux (Ubuntu 18.04 LTS)
2 Network Setup: connected to the Internet
```

Lab Preparations:

```
Initial setup: basic Ubuntu 18.04 LTS is required for this lab
Basic Linux knowledge and operations. Reference Lab: CS-SYS-00001.
Python and Anaconda software packages installed on the Linux VM. Reference Lab:
    CS-ML-00001
```

In this lab, you will use NSL-KDD dataset, which is a refined version of KDD'99 dataset. NSL-KDD dataset is now considered as one of most common dataset for network traffic and attacks, and it is a benchmark for modern-day internet traffic.

Task 1 Understand NSL-KDD datasets

NSL-KDD is comprised of four sub datasets: KDDTest+, KDDTest-21, KDDTrain+, KDDTrain+_20Percent, although KDDTest-21 and KDDTrain+_20Percent are subsets of the KDDTrain+ and KDDTest+.

Usually, KDDTrain+ is used for training purpose and KDDTest+ is used for testing purpose. 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.

These data sets contain the records of the internet traffic seen by a simple intrusion detection network and are the *ghosts* of the traffic encountered by a real IDS and just the traces of its existence remains. The data set contains 43 features per record, with 41 of the features referring to the traffic input itself and the last two are labels, where one tells whether it is a normal or attack and another scores the severity of the traffic input itself.

Within the data set exists 4 different classes of attacks: Denial of Service (DoS), Probe, User to Root(U2R), and Remote to Local (R2L). They are summarized in Table CS-ML-00101.1.

Table CS-ML-00101.1
Classification of NSL-KDD Dataset Attacks

Classes	Dos	Probe	U2R	R2L
Sub-classes	apache2 back land neptune mailbomb pod processtable smurf teardrop udpstorm worm	ipsweep mscan nmap portsweep saint satan	buffer_overflow loadmodule perl ps rootkit sqlattack xterm	ftp_write guess_passwd httptunnel imap multihop named phf sendmail snmpgetattack spy snmpguess warezclient warezserver xlock xsnoop
Total	11	6	7	15

- DoS is an attack that tries to shut down traffic flow to and from the target system. The IDS is flooded with an abnormal amount of traffic, which the system cannot handle, and shuts down to protect itself. This prevents normal traffic from visiting a network. An example of this could be an online retailer getting flooded with online orders on a day with a big sale, and because the network cannot handle all the requests, it will shut down preventing paying customers to purchase anything. This is the most common attack in the data set.
- Probe or surveillance is an attack that tries to get information from a network. The goal here is to act like a thief and steal important information, whether it be personal information about clients or banking information.
- U2R is an attack that starts off with a normal user account and tries to gain access to the system or

network, as a super-user (root). The attacker attempts to exploit the vulnerabilities in a system to gain root privileges/access.

- R2L is an attack that tries to gain local access to a remote machine. An attacker does not have local access to the system/network, and tries to "hack" their way into the network.

Task 2 NSL-KDD Data Distribution and Features

Although these attacks exist in the data set, the distribution is heavily skewed. A breakdown of the record distribution can be seen in the Table CS-ML-00101.2. 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.

Table CS-ML-00101.2
Statistics of NSL-KDD Dataset Attacks

Dataset	Number of Records					
	Total	Normal	DoS	Prob	U2R	R2L
KDDTrain+20%	25192	13449 (53%)	9234 (37%)	2289 (9.16%)	11 (0.04%)	209 (0.8%)
KDDTrain+	125973	67343 (53%)	45927 (37%)	11656 (9.11%)	52 (0.04%)	995 (0.85%)
KDDTest+	22544	9711 (43%)	7458 (33%)	2421 (11%)	200 (0.9%)	2654 (12.1%)

The features in a traffic record provide the information about the encounter with the traffic input by the IDS and can be broken down into four categories: Intrinsic, Content, Host-based, and Time-based. Below is a description of the different categories of features:

- Intrinsic features can be derived from the header of the packet without looking into the payload itself, and hold the basic information about the packet. This category contains features 1–9.
- Content features 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.
- Time-based feature hold the analysis of the traffic input over a two-second window and contains 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. This category contains features 23–31.
- Host-based 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. This category contains features 32–41.

The feature types in this data set can be broken down into 4 types:

- 4 Categorical (Features: 2, 3, 4, 42)
- 6 Binary (Features: 7, 12, 14, 20, 21, 22)
- 23 Discrete (Features: 8, 9, 15, 23–41, 43)
- 10 Continuous (Features: 1, 5, 6, 10, 11, 13, 16, 17, 18, 19)

A breakdown of the possible values for the categorical features can be seen in the Table CS-ML-00101.3. There are 3 possible Protocol Type values, 60 possible Service values, and 11 possible Flag values.

Table CS-ML-00101.3
Category features of NSL-KDD Dataset

Protocol Type (2)	Service (60)				Flag (11)
icmp	other	urh_i	time	private	OTH
tcp	link	ssh	hostnames	http_2784	S1
udp	netbios_ssn	http_8001	name	echo	S2
	smtp	iso_tsap	ecr_i	http	RSTO
	netstat	aol	bgp	ldap	RSTRs
	ctf	sql_net	telnet	tim_i	RSTOS0
	ntp_u	shell	domain	netbios_dgm	SF
	harvest	supdup	ftp_data	uucp	SH
	efs	auth	nnsp	eco_i	REJ
	klogin	whois	courier	Remote_job	S0
	systat	discard	finger	IRC	S3
	exec	sunrpc	uucp_path	http_443	
	nntp	urp_i	X11	red_i	
	pop_3	Rje	imap4	Z39_50	
	printer	ftp	mtp	Pop_2	
	vmnet	daytime	login	gopher	
	netbios_ns	domain_u	tftp_u	Csnet_ns	
		pm_dump	kshel		

Task 3 NSL-KDD Dataset Flags

Unlike *Protocol Type* and *Service* whose values are self-explanatory (these values describe the connection), Flag is not very easy to understand. The Flag feature describes the status of the connection, and whether a flag was raised or not. Each value in Flag represents a status a connection had and the explanations of each value can be found in Table CS-ML-00101.4.

Table CS-ML-00101.4
Illustration of Flags of NSL-KDD Dataset

Flag	Description	Flag	Description
SF	Normal establishment and termination. Note that this is the same symbol as for state S1. You can tell the two apart because for S1 there will not be any byte counts in the summary, while for SF there will be	RSTO	Connection reset by the originator
REJ	Connection attempt rejected	RSTR	Connection reset by the responder
S0	Connection attempt seen, no reply	OTH	No SYN seen, just midstream traffic (a "partial connection" that was not later closed)
S1	Connection established, not terminated	RSTOS0	Originator sent a SYN followed by a RST, we never saw a SYN-ACK from the responder
S2	Connection established and close attempt by originator seen (but no reply from responder)	SH	Originator sent a SYN followed by a FIN, we never saw a SYN ACK from the responder (hence the connection was "half' open)
S3	Connection established and close attempt by responder seen (but no reply from originator)	SHR	Responder sent a SYN ACK followed by a FIN, we never saw a SYN from the originator. (Not in NSL-KDD but still a flag)

Related Information and Resource

```
NSL-KDD dataset:
https://www.unb.ca/cic/datasets/nsl.html

KDD CUP ARCHIVES
https://www.kdd.org/kdd-cup
```