CHAPTER 3

INTRUSION DATA ANALYSIS

In order to create data for the IDS, it is necessary to set the real working environment to explore all the possibilities of attacks, which is expensive. Data analysis phase systematically identifies the patterns in the gathered information, and narrates them to the defined issue. It is a process of examining, transforming and modeling of data and deciding how to organize, classify, interrelate, compare and display it. Data quality focuses the correctness and reliability of information gathered and utilized in an evaluation. Data quantity deals with the quantity of information gathered for the evaluation.

The research task requires various ground truth databases in its region and the experimentation would be completed effectively if the quality and features of data for the specific region are good. Image processing, web site analysis, medical applications, remote sensing, etc. have standard and authorized ground truth databases for analysis. Likewise, most of the computer network intrusion detection systems use the KDD Cup99 for the classification analysis of network traffic. This chapter explains the formation of KDD Cup99 dataset and its features.

3.1 KDD CUP99 DATA

The Association for Computing Machinery (ACM) has a special interest group on Knowledge Discovery and Data mining (KDD)

(http://www.sigkdd.org/kddcup) which is the most popular professional organization of data miners. The KDD organized the annual Data Mining and Knowledge Discovery competition called KDD Cup in different areas. The various focused areas of KDD and its period have been tabulated in Table 3.1.

Table 3.1 KDD-CUP center of attention

Year	Focused Area
KDD-CUP 1997	Direct marketing for lift curve optimization
KDD-CUP 1998	Direct marketing for profit optimization
KDD-CUP 1999	Computer network intrusion detection
KDD-CUP 2000	Online retailer website click stream analysis
KDD-CUP 2001	Molecular bioactivity and Protein locale prediction
KDD-CUP 2002	Bio Medical document and Gene role classification
KDD-CUP 2003	Network mining and usage log analysis
KDD-CUP 2004	Particle physics; plus Protein homology prediction
KDD-CUP 2005	Internet user search query categorization
KDD-CUP 2006	Pulmonary embolisms detection from image data
KDD-CUP 2007	Consumer recommendations
KDD-CUP 2008	Breast cancer
KDD-CUP 2009	Fast scoring on a large database

Under the sponsorship of DARPA and Air Force Research Laboratory, the Lincoln Laboratory at Massachusetts Institute of Technology generated standard network traffic data for evaluation of computer network intrusion detection systems. The evaluation efforts were conducted between 1998 and 1999. The intention was to review and assess research activities in intrusion detection.

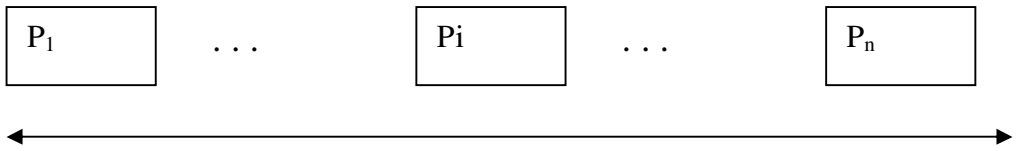
A standard set of data to be audited was provided which consists of a wide variety of intrusions simulated in a military network environment. It created an environment to acquire raw TCP/IP dump data for a network by simulating a typical US Air Force LAN. The LAN was focused like a real environment and blasted with multiple attacks. A connection is a sequence of TCP packets starting and ending at some time duration between which data flows to and from a source IP address to a target IP address under some well defined protocol. Also, each connection is labeled as either normal or as an attack with exactly one specific attack type. Each connection record consists of about 100 bytes. The raw training data is about four gigabytes of compressed binary TCP dump data obtained from seven weeks of network traffic. Finally, the completed process generated around five million connection records. Likewise, the two weeks of test data gives around two million connection records. For each TCP/IP connection, 41 various quantitative and qualitative features are obtained with normal and attack data. In 1999, the KDD acknowledged and approved DARPA data as the conventional benchmark data base for IDS called KDD Cup99 which is available in http://www.kdd.ics.uci.edu/databases/kddcuip99/task.html. There are 41 features used to represent each group of packets.

3.2 ATTRIBUTES IN KDD CUP99

The data (referring to a group of packets over a time duration of 2 seconds, also named as packet data in upcoming discussion) set in KDD Cup99 have 41 features. Among the 41 features, 1-9 are used to represent the basic features of a packet, 10-22 employ the content features, 23-31 are used for traffic features with two seconds of time window and 32-41 for host based features (Wenke Lee et al 1999). They are basically grouped into three categories: basic features of individual connection, content features within a connection, and traffic features which are computed using a two seconds time

window. Figure 3.1 shows the schematic diagram depicting the observation of the traffics (packet p_i) over a time window to construct the 41 features.

Also, the KDD Cup99 data comprises normal and 22 different types of attacks (Chi-Ho Tsang et al 2007). The description of all the 41 features is shown in Table 3.2. The features are labeled as v1, v2… v41 for the convenient representation which will be utilized in the proposed method. The variable type 'C' and 'D' represents continuous feature and categorical features (Which are discrete inform) respectively.



41 features collected from the packet data observed over the time window of 2 seconds

Figure 3.1 Extraction of features from the network connections for detecting intrusion

Some of the terminologies associated with the data set are:

- The term 'same host' refers to the connections in the past two seconds that have the same destination host as the current connection, `and is attached to the features like protocol activities, service etc.
- The term 'same service' refers to the connections in the past two seconds that have the same service as the current connection.
- The features based on 'same host' and 'same service' are collectively called as time-based traffic features of the connection records.

Table 3.2 Description of IDS features

Feature Name	*Variable type	**Type	Label	Description
duration	C	1	v1	Number of seconds of the connection
protocol_type	D	1	v2	Type of the protocol, e.g., TCP, UDP, etc.
service	D	1	v3	Network service on the destination, e.g., http, telnet, etc.
flag	D	1	v4	Normal or error status of the connection
src_bytes	C	1	v5	Number of data bytes from source to destination
dst_bytes	C	1	v6	Number of data bytes from destination to source
land	D	1	v7	1-connection is from/to the same host/port; 0-otherwise
wrong_fragment	C	1	v8	Number of 'wrong' fragments
urgent	C	1	v9	Number of urgent packets
hot	C	2	v10	The count of access to system directories, creation and execution of programs
num_failed_logins	C	2	v11	Number of failed login attempts
logged_in	D	2	v12	1 - successfully logged in; 0 - otherwise
num_compromised	C	2	v13	Number of "compromised" conditions
root_shell	C	2	v14	1 - root shell is obtained; 0 - otherwise
su_attempted	C	2	v15	1 – 'su root' command attempted; 0 - otherwise
num_root	C	2	v16	number of 'root' accesses
num_file_creations	C	2	v17	Number of file creation operations

Table 3.2 (Continued)

Feature Name	*Variable type	**Type	Label	Description
num_shells	C	2	v18	Number of shell prompts
num_access_files	C	2	v19	Number of write, delete, and create operations on access control files
num_outbound_cmds	C	2	v20	Number of outbound Commands in a ftp session
is_hot_login	D	2	v21	1 - the login belongs to the 'hot' list (e.g., root, adm, etc.) ; 0 – otherwise
is_guest_login	D	2	v22	1 - the login is a 'guest' login (e.g., guest, anonymous, etc.) ; 0 – otherwise
count	C	3	v23	Number of connections to the same host as the current connection in the past 2 seconds
srv_count	C	3	v24	Number of connections to the same service as the current connection in the past 2 seconds
serror_rate	C	3	v25	% of connections that have 'SYN' errors to the same host
srv_serror_rate	C	3	v26	% of connections that have 'SYN' errors to the same service
rerror_rate	C	3	v27	% of connections that have 'REJ' errors to the same host
srv_rerror_rate	C	3	v28	% of connections that have 'REJ' errors to the same service
same_srv_rate	C	3	v29	% of connections to the same service and to the same host
diff_srv_rate	C	3	v30	% of connections to different services and to the same host
srv_diff_host_rate	C	3	v31	% of connections to the same service and to different hosts

Table 3.2 (Continued)

Feature Name	*Variable type	**Type	Label	Description
dst_host_count	C	3	v32	Number of connections to the same host to the destination host as the current connection in the past 2 seconds
dst_host_srv_count	C	3	v33	Number of connections from the same service to the destination host as the current connection in the past 2 seconds
dst_host_same_srv _rate	C	3	v34	% of connections from the same service to the destination host
dst_host_diff_srv _rate	C	3	v35	% of connections from the different services to the destination host
dst_host_same_src _port_rate	C	3	v36	% of connections from the port services to the destination host
dst_host_srv_diff_ host_rate	C	3	v37	% of connections from the different hosts from the same service to destination host
dst_host_serror _rate	C	3	v38	% of connections that have 'SYN' errors to same host to the destination host
dst_host_srv_ serror_rate	C	3	v39	% of connections that have 'SYN' errors from same service to the destination host
dst_host_rerror _rate	C	3	v40	% of connections that have 'REJ' errors from the same host to the destination host
dst_host_srv_ rerror_rate	C	3	v41	% of connections that have 'REJ' errors from the same service to the destination host

* C- Continuous, D- Discrete **1-Intrinsic, 2-Content, 3-Traffic

The protocol_type, service, flag, land, logged_in, is_hot_login, and is_guest_login are labeled as discrete or categorical features and other 34 features are labeled as continuous features. The categorical features protocol_type, service and flag have different values which are listed in Table 3.3.

Table 3.3 Different values of protocol, services and flag

Protocol_ type (v2)	**Label**	**Service (v3)**	**Label**	**Service (v3)**	**Label**	**Service (v3)**	**Label**
icmp	1	netbios_dgm	1	Z39_50	25	time	49
tcp	2	netbios_ssn	2	gopher	26	echo	50
udp	3	netbios_ns	3	domain	27	ldap	51
Flag (v4)	**Label**	remote_job	4	finger	28	link	52
		http_8001	5	klogin	29	http	53
RSTOS0	1	hostnames	6	kshell	30	smtp	54
RSTR	2	uucp_path	7	supdup	31	uucp	55
RSTO	3	http_2784	8	systat	32	auth	56
OTH	4	iso_tsap	9	telnet	33	nnsp	57
REJ	5	csnet_ns	10	shell	34	nntp	58
S0	6	domain_u	11	imap4	35	name	59
S1	7	ftp_data	12	eco_i	36	exec	60
S2	8	http_443	13	ecr_i	37	aol	61
S3	9	daytime	14	red_i	38	IRC	62
SF	10	harvest	15	pop_2	39	X11	63
SH	11	discard	16	pop_3	40	bgp	64
		netstat	17	login	41	ctf	65
		courier	18	tim_i	42	mtp	66
		pm_dump	19	urh_i	43	rje	67
		printer	20	urp_i	44	ssh	68
		private	21	ntp_u	45	efs	69
		sql_net	22	vmnet	46	ftp	70
		tftp_u	23	other	47		
		sunrpc	24	whois	48		

The feature ‘protocol_type’ has 3 different values of icmp, tcp and udp. Likewise, the feature ‘service’ has 70 different values and ‘flag’ feature

has 11 different values. The description of the different 'flag' values are listed in Table 3.4. These 3 features and their different values acquire significant position to construct grammars in the proposed method.

Table 3.4 Description of flag values

Flag	Description
RSTOS0	Originator sent a SYN followed by a RST, never see a SYN ACK from the responder
RSTR	Established, responder aborted
RSTO	Connection established, originator aborted (sent a RST)
OTH	No SYN seen, just midstream traffic (a "partial connection" that was not later closed)
REJ	Connection attempt rejected
S0	Connection attempt seen, no reply
S1	Connection established, not terminated
S2	Connection established and close attempt by originator seen (but no reply from responder)
S3	Connection established and close attempt by responder seen (but no reply from originator)
SF	Normal establishment and termination
SH	Originator sent a SYN followed by a FIN (finish 'flag') , never saw a SYN ACK from the responder (hence the connection was "half" open)

3.3 CLASSIFICATION OF ATTACKS

The data set in KDD Cup99 have normal and 22 attack type data with 41 features and Table 3.5 shows few data set. All generated traffic patterns end with a label either as 'normal' or any type of 'attack' for upcoming analysis.

Table 3.5 Sample packet data

Feature Name	Packet-1 (normal)	Packet-2 (neptune)
duration	0	0
protocol_type	tcp	tcp
service	http	private
flag	SF	REJ
src_bytes	327	0
dst_bytes	467	0
land	0	0
wrong_fragment	0	0
urgent	0	0
hot	0	0
num_failed_logins	0	0
logged_in	1	0
num_compromised	0	0
root_shell	0	0
su_attempted	0	0
num_root	0	0
num_file_creations	0	0
num_shells	0	0
num_access_files	0	0
num_outbound_cmds	0	0
is_hot_login	0	0
is_guest_login	0	0
count	33	136
srv_count	47	1
serror_rate	0.00	0.00
srv_serror_rate	0.00	0.00
rerror_rate	0.00	1.00
srv_rerror_rate	0.00	1.00
same_srv_rate	1.00	0.01

Table 3.5 (Continued)

Feature Name	Packet-1 (normal)	Packet-2 (neptune)
diff_srv_rate	0.00	0.06
srv_diff_host_rate	0.04	0.00
dst_host_count	151	255
dst_host_srv_count	255	1
dst_host_same_srv_rate	1.00	0.00
dst_host_diff_srv _rate	0.00	0.06
dst_host_same_src_port_rate	0.01	0.00
dst_host_srv_diff_host_rate	0.03	0.00
dst_host_serror_rate	0.00	0.00
dst_host_srv_serror_rate	0.00	0.00
dst_host_rerror_rate	0.00	1.00
dst_host_srv_rerror_rate	0.00	1.00

There are varieties of attacks which are entering into the network over a period of time and the attacks are classified into the following four main classes.

- Denial of Service (DoS)
- User to Root (U2R)
- Remote to User (R2L)
- Probing

3.3.1 Denial of Service

Denial of Service is a class of attacks where an attacker makes some computing or memory resource too busy or too full to handle legitimate requests, denying legitimate users access to a machine. The different ways to launch a DoS attack are

- by abusing the computer's legitimate features
- by targeting the implementation bugs
- by exploiting the misconfiguration of the systems

DoS attacks are classified based on the services that an attacker renders unavailable to legitimate users.

3.3.2 User to Root

In User to Root attack, an attacker starts with access to a normal user account on the system and gains root access. Regular programming mistakes and environment assumption give an attacker the opportunity to exploit the vulnerability of root access.

3.3.3 Remote to User

In Remote to User attack, an attacker sends packets to a machine over a network that exploits the machine's vulnerability to gain local access as a user illegally. There are different types of R2L attacks and the most common attack in this class is done by using social engineering.

3.3.4 Probing

Probing is a class of attacks where an attacker scans a network to gather information in order to find known vulnerabilities. An attacker with a map of machines and services that are available on a network can manipulate the information to look for exploits. There are different types of probes: some of them abuse the computer's legitimate features and some of them use social engineering techniques. This class of attacks is the most common because it requires very little technical expertise.

Table 3.6 Description of attacks

Name of the attack	Type	Mechanism	Effect of the attack
back	DoS	Abuse/Bug	Slows down server response
land	DoS	Bug	Slows down server response
neptune	DoS	Abuse	Slows down server response
smurf	DoS	Abuse	Slows down the network
pod	DoS	Abuse	Slows down server response
teardrop	DoS	Bug	Reboots the machine
loadmodule	U2R	Poor environment sanitation	Gains root shell
buffer_overflow	U2R	Abuse	Gains root shell
rootkit	U2R	Abuse	Gains root shell
perl	U2R	Poor environment sanitation	Gains root shell
phf	R2L	Bug	Executes commands as root
guess_passwd	R2L	Login misconfiguration	Gains user access
warezmaster	R2L	Abuse	Gains user access
imap	R2L	Bug	Gains root access
multihop	R2L	Abuse	Gains root access
ftp_write	R2L	Misconfiguration	Gains user access
spy	R2L	Abuse	Gains user access
warezclient	R2L	Abuse	Gains user access
satan	Probe	Abuse of feature	Looks for known vulnerabilities
nmap	Probe	Abuse of feature	Identifies active ports on a machine
portsweep	Probe	Abuse of feature	Identifies active ports on a machine
ipsweep	Probe	Abuse of feature	Identifies active machines

The different types of attack, their mechanism and consequences (Kristopher Kendall 1999) are listed in Table 3.6. The unauthorized persons mostly abuse the network or system in various manners and gain the network access or slow down the response.

Although many irregularities are present in KDD Cup99 data set (Terry Brugger 2007), research activities in IDS are still using the KDD Cup 99 dataset for analyzing and exploring new approaches for better IDS. Hence, the proposed method has been experimented and analyzed with KDD Cup99.

3.4 SUMMARY

This chapter outlines the structure of the dataset used in the proposed work. The various kinds of features such as discrete and continuous features are studied with a focus on their role in the attack. The attacks are classified with a brief introduction to each. The next chapter discusses the clustering and classification of the data with a direction to learning by machine.