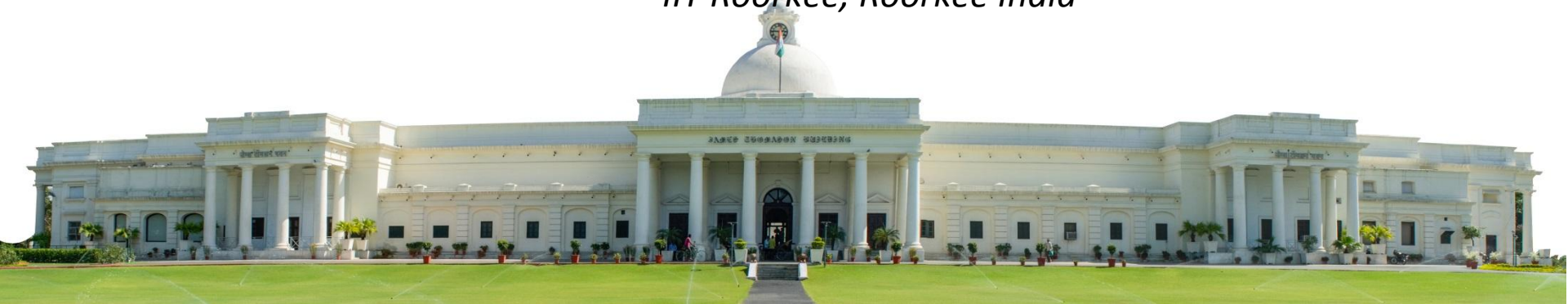
KDD99 Dataset

Under The Guidance
of
DR.Partha Pratim Roy

Presented by
Tofik Ali
and
Shrey Singh

Department of Computer Science and Engg.
IIT Roorkee, Roorkee India



Introduction

- Although KDD99 dataset is more than 15 years old, it is still widely used in academic research.

- To investigate wide usage of this dataset in Machine Learning Research (MLR) and Intrusion Detection Systems (IDS).

Fig1: The relation between main and extracted datasets. KDD99 is created from DARPA, NSL-KDD is created from KDD99.

Challenges

- To design a software to detect network intrusions protects a computer network from unauthorized users, including perhaps insiders.

- The intrusion detector learning task is to build a predictive model (i.e. a classifier) capable of distinguishing between ”bad” connections, called intrusions or attacks, and “good” normal connections.

Dataset Description

- A connection is a sequence of TCP packets starting and ending at some well defined times, between which data flows to and from a source IP address to a target IP address under some well defined protocol.

- 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 datasets contain a total of 24 training attack types, with an additional 14 types in the test data only.

Dataset Description Conti...

- **Feature Description**

Table 1: Basic features of individual TCP connections.

feature name	*description*	*type*
duration	length (number of seconds) of the connection	continuous
protocol_type	type of the protocol, e.g. tcp, udp, etc.	discrete
service	network service on the destination, e.g., http, telnet, etc.	discrete
src_bytes	number of data bytes from source to destination	continuous
dst_bytes	number of data bytes from destination to source	continuous
flag	normal or error status of the connection	discrete
land	1 if connection is from/to the same host/port; 0 otherwise	discrete
wrong_fragment	number of “wrong” fragments	continuous
urgent	number of urgent packets	continuous

Dataset Description Conti..

Table 2: Content features within a connection suggested by domain knowledge.

feature name	*description*	*type*
hot	number of “hot” indicators	continuous
num_failed_logins	number of failed login attempts	continuous
logged_in	1 if successfully logged in; 0 otherwise	discrete
num_compromised	number of “compromised” conditions	continuous
root_shell	1 if root shell is obtained; 0 otherwise	discrete
su_attempted	1 if “su root” command attempted; 0 otherwise	discrete
num_root	number of “root” accesses	continuous
num_file_creations	number of file creation operations	continuous
num_shells	number of shell prompts	continuous
num_access_files	number of operations on access control files	continuous
num_outbound_cmds	number of outbound commands in an ftp session	continuous
is_hot_login	1 if the login belongs to the “hot" list; 0 otherwise	discrete
is_guest_login	1 if the login is a “guest”login; 0 otherwise	discrete

Dataset Description Conti..

Table 3: Traffic features computed using a two-second time window..

feature name	*description*	*type*
count	length (number of seconds) of the connection	continuous
	Note: The following features refer to these same-host connections.	
serror_rate	% of connections that have “SYN” errors	continuous
rerror_rate	% of connections that have “REJ”" errors	continuous
same_srv_rate	% of connections to the same service	continuous
diff_srv_rate	% of connections to different services	continuous
srv_count	number of connections to the same service as the current connection in the past two seconds	continuous
	Note: The following features refer to these same-service connections.	
srv_serror_rate	% of connections that have “SYN" errors	continuous
srv_rerror_rate	% of connections that have ``REJ" errors	continuous
srv_diff_host_rate	% of connections to different hosts	continuous

References

- KDD Cup 1999 (http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html) .
- A review of KDD99 dataset usage in intrusion detection and machine learning between 2010 and 2015.

Thanks