DOI: http://dx.doi.org/10.26483/ijarcs.v10i2.6395

ISSN No. 0976-5697

Volume 10, No. 2, March-April 2019

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

# A REVIEW ON KDD CUP99 AND NSL-KDD DATASET

Ritu Bala
Research Scholar, GJU&ST,
Hisar, India

Dr. Ritu Nagpal
Associate Professor, GJU&ST
Hisar, India

***Abstract:*** Continues use of network services for information and resource sharing makes our work easier. But sometime the extensive use of network services leads many problems in the form of attacks or intrusions which demolish not only the privacy but also the integrity and accessibility of data. Detection of attacks or intrusions on the network is a serious issue of concern for the researchers. Intrusion Detection System solves the purpose of detecting intrusion on the network. Huge amount of data is required to simulate the powerful Intrusion Detection System (IDS) model as well as to train and testing the model. This paper, presents the review of datasets DARPA, KDDCup99 and NSL_KDD which are most widely used by researchers to detect the intrusion in computer network.

***Keywords:*** IDS, Dataset, DARPA, NSL_KDD, KDDCup99.

## INTRODUCTION

Now days, the use of network is growing due to the increased use of handheld devices. So, network security is the major issue. Our network suffers from various types of attacks like viruses, Trojan horse, worms. To identify and stop these attacks, a security management system is required. Confidentiality, Integrity and availability of data are the major objectives of security. An Intrusion Detection System serves this purpose by automatically alert the administrator when someone trying to violets the security policies. The role of intrusion detection system is to assemble the information from the network. Then after supervising and investigating this information, it separates them into normal & malicious behaviour and brings this result to system administrator [2].

Encryption, firewalls, virtual private network etc. are the conventional approaches which were used in early days. But they were not able to protect the network completely. Thus to increase the network security, an Intrusion Detection System is introduced. It is divided into two main categories as Signature and Anomaly based . Another name of signature based IDS is misuse based IDS. It identifies only the familiar attacks. But anomaly based IDS can identify known as well as novel attacks.

To figure out the conduct of Network Intrusion Detection System, various datasets are available. To provide the security to computer network, many researchers have suggested three most widely used datasets like DARPA 98/99, KDD99 and NSL- KDD. DARPA is the first dataset for the evaluation of intrusion detection system and was attempted in MIT Lincoln Laboratory in 1998. KDD CUP99 is the subset of DARPA98 dataset. It has 41 features. NSL-KDD dataset is derived from KDDCup99 by removing the redundant and duplicate records from training and testing datasets respectively so it is the revised version of the original KDDCup99 dataset. Each dataset has its advantages and shortcomings. It is very challenging to select a suitable dataset itself. Due to the increased use of network, the behaviour and the pattern changes and dependency on a particular dataset is not trust-worthy. So, there is need to update the dataset periodically.

In this paper the review of DARPA, KDD Cup99 and NSL_KDD datasets are made using various attributes.

## DATASETS

KDDCup99 dataset (Knowledge Discovery in Databases):

The KDDCup99 is the mostly and commonly used dataset for the identification of intrusion in computer network. Simulation of US Air Force LAN was done in order to get the subset of DARPA 1998 dataset by inducing different types of attacks. A nine week of TCP dump data was used for this purpose at MIT Lincoln Laboratory. KDDCup dataset contains about 4,900,000 single instances which are described by 41 features [1]. They are classified as either normal or an intrusion.

Table1. Features in the KDD Cup99 Dataset

| Sr. No. | Characteristics | Description |
|---|---|---|
| 1 | Kind of Protocol | Description of the protocol used |
| 2 | Time intervals | Time interval of the connection |

| | | |
|---|---|---|
| 3 | Src_bytes | number of bytes to be transferred from source |
| 4 | Service | Service on the destination |
| 5 | Urgent | statistics of important packets |
| 6 | Flag | Gives the description about the status whether normal or error |
| 7 | Dst_bytes | number of bytes to be transferred from destination |
| 8 | Wrong_fragmant | Number of wrong segments |
| 9 | Land | When IP address of sending end, receiving end and port number are equal then set 1 otherwise 0. |
| 10 | Num_file_creation | Total operation to create a file |
| 11 | Root_shell | Whenever root account is running then set 1otherwise 0 |
| 13 | Hot | statistics of hot indicators |
| 14 | Su_attempted | Whenever su command is used then set 1 otherwise 0 |
| 15 | Logged_in | If login successfully then set to 1 otherwise 0 |
| 16 | Num_root | Statistics of the performed operations as a root |
| 11 | Num_failed_logins | Total of failed login attempt |
| 13 | Num_compromised | Total of Compromised situations |
| 14 | Num_access_files | Total actions required to access the control files |
| 15 | Num_outbound_cmds | In the ftp period, number of outgoing commands |
| 18 | Num_shell | Total shell stimulate |
| 21 | Is_hot_login | Whenever signed through hot list set to one else zero |
| 22 | Count | In past two seconds, the total of connection to the similar destination node as the ongoing connection |
| 23 | Is_guest_login | If signed through guest then set 1 otherwise 0 |
| 24 | Serror_rate | % of SYN fault in connections |
| 25 | Srv_count | Describe the total connections in past two second to the similar service (port number) as the ongoing Connection |
| 27 | Srv _rerror _rate | % of REJ fault in connections |
| 26 | Srv_ error_rate | % of SYN fault in connections |
| 27 | Rerror_ rate | % of REJ fault in connections |
| 28 | Diff_ srv_ rate | % of variation in connections |
| 29 | Same_ srv _rate | % of connections to the similar service |
| 32 | Dst_ host | Count number of connections to the same destination |
| 31 | Srv_ diff_ host_ rate | % of connections to different hosts |
| 33 | Dst _host_ srv _count | Count the connection with same destination and also use the same service |
| 34 | Dst _host _same _srv_rate | % of connections to the same service and destination |
| 35 | Dst_host_srv_serror_rate | % of connections to a host as well as specified service with an S0 error |
| 36 | Dst_host_same _src_port_rate | % of connections on the same source port |
| 37 | Dst_host_srv_ diff_host_rate | % of connections on different destination |
| 38 | Dst_host_diff_ srv_rate | % of connections to various services |
| 39 | Dst_host_serror_rate | % of S0 fault in a host |

| | | |
|---|---|---|
| 40 | Dst_host_srv_rerror_rate | Connections in % that triggered the flag (4) REJ |
| 41 | Dst_host_rerror_rate | % of connections that triggered the flag (4) REJ |

Features of KDD dataset are classified into four different classes as:

**Basic:** This class consist of all the features of TCP connection.

**Content**: This class consist of all the features given by domain knowledge within a connection.

**Traffic:** this class consist of features which are computed within a time frame of two seconds.

**Host:** This class consist of features which last for more than two seconds.

All the attack are divided into one of the following categories[3]:

**Denial of Service Attack (DoS)**: In this attack our system resources knowingly occupied by some unnecessary or unwanted processes in order to make the server too busy to handle the other important requests which result in rejection of legitimate request.

**User to Root Attack (U2R):** In this type of attack, the intruder tries to gain the access of an authorized user account in the system and exploit some vulnerability to gain super user privilege.

**Remote to Local Attack (R2L)**: A person who doesn't have an account on a machine but yet sending packets to the same machine on a network to personify the legal user for gaining the local access to the machine.

**Probing Attack**: These attacks gather the network activity information for the supposed objective to bypass its security controls.

**Advantages of KDDCup99:**

KDDCup99 dataset has some improvement over DARPA 1998 dataset:

- Conversion the network traffic from TCP dump file into relational structure is not required.
- The dataset contains direct and derived features which are readily available.
- The memory and processing power is less required.
- To optimize accuracy and detection rate over KDD99 dataset many machine learning algorithms are used. Most frequent algorithms used with KDD99 are decision tree derivatives and support vector machines.

**Problems of KDD99 Data Set:**

- The synthesized data is not matching to real traffic of network data.
- Training and test sets are too large which make it very complex.
- Detection accuracy is very low.
- Cannot be detect dropped packets.
- Unreliable for building real NIDS.
- It has redundant and duplicate records
- Machine learning algorithm cannot be applied to R2L and U2R.
- Large gap between the number of instances of normal traffic and number of instances of attack.

**NSL-KDD Dataset**: This dataset designed by Tavallaee et al. [5]. It is developed after the removal of redundant and duplicate records from training and test data of KDDCup . it contains only selected and necessary records from. There are total 37 attacks out of which 27 attacks are used by testing dataset and 23 attacks are used by training dataset for experiments[8]. The number of feature in NSL-KDD dataset has same as that of in KDDCup. This dataset contains 41 features and 5 attack classes. There is one normal class and other 4 are different types of attack. These different attacks are grouped into four categories: Probe attack, Denial of service attack(DoS), User to Root (U2R) and Remote to Local (R2L). The above dataset holds a binary class attribute as well as reasonable number of training and test instances [6]. This dataset is publically available for researchers.

**Advantages of NSL-KDD:**

- Removal of redundant records helps the classifier to produce unbiased result.
- Since not even a single record found identical in proposed test set; therefore, learner's performance is not biased by the methods having better detection rates on the frequent records.
- Detection rate is high as compared to KDD Cup.
- The record counts in the train and test sets are reduced. Therefore selection of a chunk of data is not required randomly and all the experiments can be done on the entire set. As well as it gives consistent result of different research work.

There are total 21 different types of attacks which are present in training dataset. While test dataset contain 16 additional attacks. Major attacks are categorised as Probe, DoS, U2R and R2L[7].

Table II. Categories of Attacks for Training And Testing Datasets

| DOS | Probe | R2L | U2R |
|---|---|---|---|
| Back | Ipsweep | Spy | bufferoverfl ow |
| Land | Mscan | Warezclient | Loadmodule |
| Mailbom b | Nmap | ftp_write | Perl |
| Neptune | Portswe ep | Guesspassw d | Ps |
| Pod | Saint | Httptunnel | Rootkit |

| | | | |
|---|---|---|---|
| Processable | Satan | Imap | Snmpguess |
| smurf | | Multihop | Sqlattack |
| Teardrop | | Named | Worm |
| udpstorm | | Phf | xterm |
| | | Sendmail | |
| | | Snmpgetatt ack | |
| | | Warezmaste r | |
| | | Xlock | |
| | | Xsnoop | |

## CONCLUSION

In order to develop new tools and in the research area of IDS KDD Cup99 is most known dataset for the protection of computer network against malicious activities. This dataset also have many limitations like redundant and duplication of records, imbalance between normal traffic and number of attacks and many more listed above. The solution of this is NSL-KDD which has removed unnecessary and same records in both training and test sets. Continuous use of computer network and information system has become the vital source for large number of attacks. Now a days , in all over the world, many researcher are developing new datasets by taking the help from KDD Cup, NSL- KDD and DARPA datasets depending upon the issues in problem solving and purpose of IDS.

## REFERENCES

1. L. Dhanabal, Dr. S.P. Shantharajah "A Study on NSL-KDD Dataset for Intrusion Detection system Based on Classification Algorithms", in International Journal of Advanced Research in Computer and Communication engineering, vol. 4, pp. 446-452, 2015.
2. Rung-Ching Chen, Kai-Fan Cheng and Chia-Fen Hsieh, "Using Rough Set and Support Vector Machine for Network Intrusion Detection", in International Journal of Network Security & Its Applications (IJNSA), vol 1, pp. 1-13 2009.
3. Al-Dhafian, B., Ahmad, I. & Al-Ghamid, A. An Overview of the Current Classification Techniques" in International Conference on Security and Management, pp.82-88, 2015.
4. Alzobaidy, L. "Anomaly network intrusion detection system based on distributed time-delay neural network (DTDNN)", Journal of Engineering Science and Technology (JESTEC), vol.5, pp. 457-471, 2010.
5. Tavallaee, M.; Bagheri, E.; Wei Lu; and Ghorbani, A. " A detailed analysis of the KDD CUP 99 data set" IEEE Symposium on Computational Intelligence in Security and Defense Applications (CISDA 2009),pp. 1-6, 2009.
6. Shaheen, A. "A comparative analysis of intelligent techniques for detecting anomalous internet traffic", MSc. Thesis, King Fahd University, 2010.
7. Danijela D. Protić "Review of KDD Cup '99, NSL-KDD and Kyoto 2006+ Datasets", vol. 66, pp. 580-596, 2018.Dr.K.Arunesh, M. Manoj Kumar, "A Comparative Study Of Classification Techniques For Intrusion Detection Using Nsl-Kdd Data Sets", in International Conference on Recent Trends in Engineering Science, Humanity and Management, pp. 288-295, 2017.