International Journal of Scientific Research and Management (IJSRM)
||Volume||10||Issue||04||Pages||EC-2022-832-839||2022||
Website: www.ijsrm.in ISSN (e): 2321-3418
DOI: 10.18535/ijsrm/v10i4.ec05

A comparison of the NSL-KDD dataset and its predecessor the KDD Cup '99 dataset

Yassine SAHLI

Information Science and Technology, Nanjing University, Nanjing, china.

Abstract:
This study examines three datasets, notably the KDD Cup '99 and the NSL-KDD datasets, which are commonly used in intrusion detection research in computer networks. The KDD Cup '99 dataset contains five million records, each with 41 attributes that may be used to categorize malicious assaults into four categories: Probe, DoS, U2R, and R2L. Because it was developed by simulation over a virtual computer network, the KDD Cup '99 dataset cannot reflect real traffic statistics. Duplicate and redundant records from the KDD Cup '99 dataset are eliminated from the training and test sets, respectively, in the NSL-KDD dataset.

Key words: intrusion detection, KDD Cup '99, NSL-KDD.

Introduction:
Intrusion can be understood as an attempt to violate information protection, data integrity and resource accessibility [1]. The most popular way to protect a computer network from various malicious activities is to detect intrusion by using an intrusion detection system (IDS). The IDS consists of software applications and/or hardware devices that constantly monitor computer network for suspicious activities, and trigger intrusion alarms if unknown or malicious activities are detected. There are typically two kinds of IDSs. A host-based IDS detects and identifies any system changes by analyzing system or server log files and comparing them against database of common signatures for known attacks. To defend the system against network-based threats, a network-based IDS observes network traffic and looks for unusual activity by evaluating the content and header information of all packets. For monitoring, evaluating, and detecting network security violations, there are two well-known systems. Pattern recognition is used by misuse-based systems, which retain a database of indications (signatures) collected from prior assaults. Anomaly-based systems create statistical models of typical network data and look for anomalies to identify what's abnormal.

Many academics have been exploiting these datasets to create anomaly-based IDSs and develop other solutions for computer network security defense during the last few years. The KDD Cup '99 dataset is made up of data that was moved from a virtual environment. The learning competition's goal was to develop a prediction model (classifier) that could discriminate between genuine and malicious connections in a computer network [2].The KDD Cup '99 dataset is a subset of the 1998 DARPA dataset acquired by simulating the functioning of a typical US Air Force LAN with numerous assaults divided into four categories: probing, denial of service, user to root, and remote to local. The 41 characteristics in the KDD Cup '99 dataset records are divided into four categories: basic, traffic, content, and host-related aspects [3].

Because the KDD Cup '99 dataset is a network traffic simulation, there are a lot of duplicate records in the training set and the test set, making it difficult to categorize the non-redundant records. A new NSL-KDD dataset [4] was proposed to address these difficulties. The NSL-KDD dataset contains chosen characteristics from the KDD Cup '99 dataset, however there are no duplicates in the training set and no redundant entries in the test set. In addition, the training and test sets have a reasonable amount of records.

Datasets:
The KDD Cup '99:

The KDD Cup '99 dataset is the most well-known and commonly used dataset for anomaly detection studies in computer networks. The KDD Cup '99 dataset is a collection of data transfers from a virtual environment for the Third Knowledge Discovery and Data Mining Tools Competition [2]. It's as mentioned earlier a subset of the 1998 DARPA dataset that was gathered by simulating the operation of a typical US Air Force LAN with numerous assaults and acquiring nine weeks of TCP dump data, as we described before. At the Massachusetts Institute of Technology (MIT) Lincoln Laboratory, the data was gathered and dispersed.

The KDD Cup '99 intrusion detection test is made up of three parts: the whole KDD Cup '99 dataset, which comprises instances of assaults and regular connections, a 10% KDD dataset for training classifiers, and a KDD test dataset for testing [5]. The whole KDD Cup '99 dataset comprises 4,898,431 single connection records, each with 41 normal or attack attributes (Table 1).

N°	name	Description
1	duration	Length of connection
2	protocol type	Type of protocol (TCP, UDP...)
3	service	Destination service (ftp, telnet...)
4	flag	Status of connection
5	source bytes	No. of B from source to destination
6	destination bytes	No. of B from destination to source
7	land	If the source and destination address are the same land=1/ifnot, then 0
8	wrong fragments	No. of wrong fragments
9	urgent	No. of urgent packets
10	hot	No. of hot indicators
11	failed logins	No. of unsuccessful attempts at login
12	logged in	If logged in=1/if login failed 0
13	# compromised	No. of compromised states
14	root shell	If a command interpreter with a root account is running rootshell=1/if not, then 0
15	su attempted	If an su command was attempted su attempted=1/if not, then 0 (temporary login to the system with other user credentials)
16	# root	No. of root accesses
17	# file creations	No. of operations that create new files
18	# shells	No. of active command interpreters
19	# access files	No. of file creation operations
20	# outbound cmds	No. of outbound commands in an ftp session
21	is hot login	is host login=1 if the login is on the host login list/if not, then0
22	is guest login	If a guest is logged into the system, is guest login=1/if not,then 0
23	count	No. of connections to the same host as the current connection at a given interval
24	srv count	No. of connections to the same service as the currentconnection at a given interval
25	serror rate	% of connections with SYN errors
26	srv error rate	% of connections with SYN errors
27	rerror rate	% of connections with REJ errors
28	srv rerror rate	% of connections with REJ errors
29	same srv rate	% of connections to the same service

30	diff srv rate	% of connections to different services
31	srv diff host rate	% of connections to different hosts
32	dst host count	No. of connections to the same destination
33	dst host srv count	No. of connections to the same destination that use thesame service
34	dst host same srcrate	% of connections to the same destination that use the sameservice
35	dst host srv rate	% of connections to different hosts on the same system
36	dst host same srvport rate	% of connections to a system with the same source port
37	dst host srv diffhost rate	% of connections to the same service coming from differenthosts
38	dst host serror rate	% of connections to a host with an S0 error
39	dst host srv serrorrate	% of connections to a host and specified service with an S0error
40	dst host serror rate	% of connections to a host with an RST error
41	dst host srv serrorrate	% of connections to a host and specified service with anRST error

Table1: the KDD Cup '99 dataset properties[5]

The qualities that describe the linkages may be divided into four groups:

The packet header is used to extract basic information without having to examine the information about the packet (duration, protocol type, service, flag and the number of bytes sent from the source to the destination and the other way around).

The content characteristics of a TCP packet are determined by evaluating the packet's content (number of unsuccessful attempts to login to the system).
The length of a connection from a source IP address to a target IP address is determined by time characteristics. The connection is made up of a series of data packets that begin and stop at predetermined times.

The traffic characteristics are based on a window with a set number of connections in it. This is appropriate for describing assaults that persist longer than the specified time interval.

All assaults in the KDD Cup '99 dataset are categorized into one of four groups (Table 2) [6].

Attack category	Attack name
Probe	ipsweep, nmap, portsweep, satan
DoS (Denial of Service)	back, land, neptune, pod, smurf, teardrop
U2R (User to Root)	buffer_overflow, loadmodule, perl, rootkit
R2L (Remote to Local)	ftp_write, guesspasswd, imap, multihop, phf, spy,warezlient, warezmaster

Table2: Attacks classification

Probe: By scanning a machine or a networking device for weaknesses or vulnerabilities that may subsequently be exploited in order to compromise the system, the attacker gathers information about the system or computer network in order to identify (known) flaws.

DoS: The attacker denies legitimate users access to computational resources or overburdens them to the point that requests aren't completed in real time. This attack causes resource unavailability, which means that resources are too busy or full to handle valid networking requests, preventing people access to a system.

U2R: In order to get root privileges, the attacker exploits vulnerabilities. The attacker logs in as a regular user and searches the system for weaknesses in order to achieve superuser capabilities [7].

R2L: Because the attacker lacks a user account on the victim's PC, he attempts to get access to the remote system without one [5].

Tables 3,4,5 shows the number of instances in the entire dataset, the 10% training set (that consists of 10% of the total number of instances), and the test set, which consists of 311,029 instances, organized by categories and datasets, as well as the percentage of the total share of each category within each dataset.

The whole dataset (100%)	Attack category	Number of instances	(%)
	Normal	492,708	19.86%
	Probe	41,102	0.84%)
	DoS	3,883,370	79.30%
	U2R	52	0.00%
	R2L	1,126	0.02%

Table3: *The KDD Cup '99 whole dataset's* instances

Training set (10%)	Attack category	Number of instances	(%)
	Normal	97,278	19.69%
	Probe	4,107	0.83%
	DoS	391,458	79.24%
	U2R	52	0.01%
	R2L	1,126	0.23%

Table4: *The KDD Cup '99* training set's instances

Test set	Attack category	Number of instances	(%)
	Normal	60,593	19.48%
	Probe	4,166	1.34%
	DoS	229,853	73.94%
	U2R	70	0.02%
	R2L	16,347	5.26%

Table5: The KDD Cup '99 test set's instances

The KDD Cup '99 dataset has been criticized in a number of ways. The main objection is that the KDD Cup '99 dataset isn't a true representation of real-world network traffic. Authors also discuss the following issues [8], [9], and [10]:

– The complexity of the computations;
– The complexity of the computations;
– How duplicate data affects machine learning (ML) methods
– The number of attack instances is too excessive in comparison to the number of regular traffic instances;
– There is no practical link between particular attack types.
– As a result of converting data from the DARPA dataset to the KDD Cup '99 dataset, R2L instances of individual assaults are equivalent to normal traffic instances.
– Poor detection accuracy of attack distribution, etc.

For these reasons, alternate sets for training and testing might be created as follows:

– utilize only the training set,
– utilize only the training set,
– For training and testing, create a union of sections of the training and test sets
– Filter occurrences to ensure attack proportionality, and so forth.

The composition of alternative sets is determined by the IDS model's evaluation.

The NSL-KDD dataset:

The KDD Cup '99 dataset has a large number of duplicate and redundant entries (78%) that make it difficult to categorize the remaining records [11]. A new NSL-KDD dataset [4] has been suggested to address these concerns. The NSL-KDD dataset is made up of a small number of characteristics from the KDD Cup '99 dataset that aren't redundant in the training set or duplicates in the test set [12]. There are three compelling reasons to use the dataset in the experiments, given the dataset's design:

- removing duplicate data from the training set allows classifiers to be more unbiased when dealing with increasingly frequent records;

- By excluding duplicate records from the test set, the performance of a classifier will not be influenced by techniques that have higher decision rates on frequent records;

- Training and test sets contain a sufficient number of instances to allow for tests on the complete set without the requirement to select a small piece at random.

Out of the 37 assaults in the test dataset, the training dataset has 21 distinct attacks. The known attacks are those found in the training set, whereas the other 16 attacks can only be found in the test set (Table 4). Probe, DoS, U2R, and R2L are the different sorts of attacks [13].
There are a total of 126,620 occurrences in the regular traffic in the training set. The test set's regular traffic has 9,711 occurrences, bringing the total to 22,850.

Attack categories	Number of instancesin the training set	Number of instances in the test set
DoS	45,927	7,460
	back (956), land (18), neptune (41,214), pod (201), smurf (2,646), teardrop (892)	back (359), land (7), neptune (4,657), pod (41), smurf (665), teardrop (12), apache2 (737), udpstorm (2), processtable (685), worm (2), mailbomb (39)
	11,656	2,421

Probe	satan (3,633), ipsweep (3,599), nmap (1,493), portsweep (2,931)	satan (753), ipsweep (141), nmap (73), portsweep (157), mscan (996), saint (319)
R2L	1,642	3,191
	guess_passwd (53), ftp_write (6), imap (658), phf (4), multihop (7), warezmaster (20), warezclient (890), spy (2)	guess_passwd (1,231), ftp_write (3), imap (307), phf (2), multihop (18), warezmaster (944), xsnoop (4), xlock (9), snmpguess (331), snmpgetattack (178), httptunnel (133), sendmail (14), named (17)
U2R	52	67
	buffer_overflow (30), loadmodule (9), rootkit (10), perl (3)	buffer_overflow (20), loadmodule (2), rootkit (13), perl (2), xterm (13), sqlattack (2), ps (5)
Total	59,277	13,139

Table 6: The number of attack instances in the training and test sets

The comparison:

The author in [6]) compared five datasets: the KDD Cup '99, the NSL-KDD, and other datasets. Table 8 displays the results simple of the two datasets.

Dataset (year)	Features	Pros	Cons
KDD Cup'99 (1999)	41 features (32 numeric and 9 categorical)	– Used for evaluating anomaly detection systems. – Attack types in training set are distinctive from the testing set.	– Includes redundant and duplicate records. –Does not reflect the modern environment.
NSL-KDD (2009)	41 features (32 numeric and 9 categorical)	– Does not include redundant and duplicate records. – The selected records are inversely proportional to the percentage of records in the KDD Cup '99 dataset. – The number of records is reasonable.	– Not perfect for representing the existing real networks.

Table 7: KDD cup'99 and NSL-KDD datasets in IDSs comparison

The DARPA dataset was preprocessed to create the KDD Cup '99 dataset, which categorized records into 41 features. In both the training and testing sets, the dataset has a large number of entries, however it contains redundant and duplicate records and does not reflect real network traffic. The KDD Cup '99 dataset, on the other hand, is commonly utilized in the creation of new intrusion detection systems and tools for data protection to run tests on huge volumes of data or wherever repeatability is required.

Selected characteristics from the KDD Cup '99 dataset are included in the NSL-KDD dataset. It is intended to address issues like as redundant entries in the training set and redundant records in the test set, as well as to minimize data volume to a manageable amount.

Conclusion:

The KDD Cup is a competition in the field of machine learning and data mining held every year at the Data Mining and Knowledge Discovery conference. Competitors had to overcome the challenge of protecting computer networks against intrusions in 1999. The KDD Cup '99 dataset had been developed for competitive purposes. The whole dataset, 10% training set, and test set make up the KDD Cup '99 benchmark. Each record is made up of 41 characteristics that characterize the network activity of a computer network that has been simulated. The dataset includes information on the following attacks: Probe, DoS, U2R, and R2L.

The KDD Cup '99 dataset is frequently used as a reference for IDS research and the creation of new solutions to protect computer networks from different threats. Complexity, the influence of duplicates and redundant records, an imbalanced number of attacks compared to each other, and a mismatch between the number of attacks and regular traffic are all flaws that might damage the research. Using the NSL-KDD dataset, which does not contain duplicates the training and the test set, is one option to avoid these issues.

Researchers from all around the globe are developing new IDSs to defend computer networks from hackers by using known datasets and their pre- and post-processed versions, as the fast growth of computer networks and information systems has resulted in a huge number of complex assaults. The KDD Cup '99 and NSL-KDD datasets have been extensively utilized in studies to build a variety of technologies for defending against malicious assaults. The objective of a given IDS and the security goals in specific issue solution determine which of the bases is employed.

References:

1. Protić, D. 2016. Neural Cryptography. Vojnotehnički glasnik/Military Technical Courier, 64(2), pp.483-495. Available at:*https://doi.org/10.5937/vojtehg64-8877.*
2. SIGKDD - KDD Cup. KDD Cup 1999: Computer network intrusion detection. Available at: *https://www.kdd.org.*
3. Aggarwal, P. & Sharma, S.K. 2015. Analysis of KDD Dataset Attributes – Class Wise for Intrusions Detection. In: Procedia Computer Science, 57, pp.842- 851. . Available at: *https://doi.org/10.1016/j.procs.2015.07.490.*
4. Tavallaee, M., Bagheri, E., Lu, W. & Ghorbani Ali, A. 2009. A Detailed Analysis of the KDD CUP ‘99 Data Set. In: Proceedings of the 2009 IEEE Symposium on Computational Intelligence in Security and Defense Applications. Ottawa, ON, Canada. Available at: *https://doi.org/10.1109/CISDA.2009.5356528.*
5. Gifty Jeya, P., Ravichandran, M. & Ravichandran, C.S. 2012. Efficient Classifier for R2L and U2R Attacks. International Journal of Computer Applications, 45(21), pp.28-32.
6. Available at:*http://www.ijcaonline.org/archives/volume45/number21/7076-9751.*
7. Al-Dhafian, B., Ahmad, I. & Al-Ghamid, A. 2015. An Overview of the Current Classification Techniques. In: International Conference on Security and Management, Las Vegas, USA, pp.82-88.
8. Paliwal, S. & Gupta, R. 2012. Denial-of-Service, Probing & Remote to User (R2L) Attack Detection using Genetic Algorithm. International Journal of Computer Applications, 60(19), pp.57-62. Available at: *http://www.ijcaonline.org/archives/volume60/number19/9813-4306.*
9. Kolez, A., Chowdhury, A. & Alspector, J. 2003.Workshop on Learning from Imbalanced Data Sets (II), Whashington.
10. Maček, N. & Milosavljević, M. 2013. Critical Analysis of the KDD Cup ’99 data set and research methodology for machine learning. In: Proceedings of the 57th ETRAN conference, Zlatibor, pp.(VI 2.3.1-4.).
11. Bukola, O. & Adetunmbi, A.O. 2016. Auto-Immunity Dendritic Cell Algorithm. In: International Journal of Computer Applications, 137(2), pp.10-17. Available at:

https://doi.org/10.5120/ijca2016908689.

12. Revathi, S. & Malathi, A. 2013. A Detailed Analysis on NSL-KDD Dataset Using Various Machine Learning Techniques for Intrusion Detection. International Journal of Engineering Research & Technology, 2(12), pp.1848-1853.
13. Kavitha, P. & Usha, M. 2014. Anomaly based intrusion detection in WLAN using discrimination algorithm combined with Naïve Bayesian classifier. Journal of Theoretical and Applied Information Technology, 62(1), pp.77-84. Available at: *http://www.jatit.org/volumes/Vol62No1/11Vol62No1.pdf.*
14. Nkiama, H., Said, S.Z.M. & Saidu, M. 2016. A Subset Feature Elimination Mechanisms for Intrusion Detection System. International Journal of Advanced Computer Science and Application, 7(4), pp.148-157.Available at: *https://doi.org/10.14569/IJACSA.2016.070419.*