



Vojnotehnicki glasnik/Military Technical
Courier

ISSN: 0042-8469

vojnotehnicki.glasnik@mod.gov.rs

University of Defence
Serbia

Proti, Danijela D.

REVIEW OF KDD CUP '99, NSL-KDD AND KYOTO 2006+ DATASETS

Vojnotehnicki glasnik/Military Technical Courier, vol. 66, núm. 3, 2018, pp. 580-596

University of Defence

Available in: <https://www.redalyc.org/articulo.oa?id=661770389006>

- How to cite
- Complete issue
- More information about this article
- Journal's homepage in [redalyc.org](https://www.redalyc.org)

[redalyc.org](https://www.redalyc.org)


Scientific Information System

Network of Scientific Journals from Latin America, the Caribbean, Spain and Portugal

Non-profit academic project, developed under the open access initiative

REVIEW OF KDD CUP '99, NSL-KDD AND KYOTO 2006+ DATASETS

Danijela D. Protić

Serbian Armed Forces, General Staff,
Department for Telecommunication and Informatics (J-6),
Center for Applied Mathematics and Electronics,
Belgrade, Republic of Serbia,
e-mail: adanijela@ptt.rs,
ORCID iD:  <http://orcid.org/0000-0003-0827-2863>

DOI: 10.5937/vojtehg66-16670; <https://doi.org/10.5937/vojtehg66-16670>

FIELD: Computer Sciences, IT
ARTICLE TYPE: Review Paper
ARTICLE LANGUAGE: English

Abstract:

This paper presents a review of three datasets, namely KDD Cup '99, NSL-KDD and Kyoto 2006+ datasets, which are widely used in researching intrusion detection in computer networks. The KDD Cup '99 dataset consists of five million records, each containing 41 features which can classify malicious attacks into four classes: Probe, DoS, U2R and R2L. The KDD Cup '99 dataset cannot reflect real traffic data since it was generated by simulation over a virtual computer network. In the NSL-KDD dataset, redundant and duplicate records from the KDD Cup '99 dataset are removed from training and test sets, respectively. The Kyoto 2006+ dataset is built on real three year-network traffic data which are labeled as normal (no attack), attack (known attack) and unknown attack. The Kyoto 2006+ dataset contains 14 statistical features derived from the KDD Cup '99 dataset and 10 additional features.

Key words: KDD Cup '99, NSL-KDD, Kyoto 2006+, computer network, intrusion detection.

Introduction

Intrusion can be understood as an attempt to violate information protection, data integrity and resource accessibility (Protić, 2016, pp.483-495). 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. A network-based IDS monitors network traffic and checks for irregular behavior by inspecting the content and header information of all packets to protect the system from network-based threats.

There are two well-known systems for monitoring, analyzing and detecting network security violation. Misuse-based systems rely on pattern recognition and maintain the base of indicators (signatures) extracted from previous attacks. Anomaly-based systems build statistical models of normal network traffic and observe abnormalities in order to detect what is anomalous.

For several decades, a lot of researchers have suggested to use three most known datasets, namely KDD Cup '99, NSL-KDD and Kyoto 2006+ datasets, to design anomaly-based IDSs and develop various tools for computer network security protection. The KDD Cup '99 dataset is a collection of data transferred from virtual environment to be used for the Third Data Mining and Knowledge Discovery Competition on computer network intrusion detection. The task for the learning contest was to learn a predictive model (i.e. classifier) capable of distinguishing between legitimate and illegitimate connections in a computer network (SIGKDD - KDD Cup, 2018). The KDD Cup '99 dataset is the subset of 1998 DARPA dataset that was collected by simulation of the operation of a typical US Air Force Local Area Network (LAN) with multiple attacks classified into four categories: probe, denial of service, user to root and remote to local. KDD Cup '99 dataset records contain 41 features which fall into four categories: basic, traffic, content and host related ones (Aggarwal & Sharma, 2015, pp.842-851).

Since the KDD Cup '99 dataset is a simulation of network traffic, there is a huge number of redundant records in the training set and duplicate records in the test set which prevent classifying the other records which are not redundant. To solve these issues, a new NSL-KDD dataset was proposed (Tavallaee et al, 2009). The NSL-KDD dataset consists of selected features from the KDD Cup '99 dataset but does not include redundant records in the training set and there are no duplicates in the test set. Also, the number of records in the training and test sets is reasonable.

However, both KDD Cup '99 dataset and NSL-KDD dataset do not reflect real data flow in computer network since they are generated by simulation over the virtual network. The Kyoto 2006+ dataset is built on real three year-traffic data from November 2006 to August 2009. This dataset is captured using honeypots, darknet sensors, e-mail server and web crawler (Singh et al, 2015, pp.8609-8624). Each record consists of

14 statistical features derived from KDD Cup '99 data set as well as 10 additional features which can be used for the analysis and evaluation of the IDS network. This paper presents a review and a comparative analysis of KDD Cup '99, NSL-KDD and Kyoto 2006+ datasets.

Datasets

KDD Cup '99 dataset

The most known and widely used dataset for experiments on anomaly detection in computer networks is the KDD Cup '99 dataset. The KDD Cup '99 dataset is a collection of data transfer from virtual environment to be used for the Competition of the Third Knowledge Discovery and Data Mining Tools (KDD CUP '99 dataset, 1999). It is the subset of 1998 DARPA dataset that was collected by simulation of the operation of a typical US Air Force LAN with multiple attacks and acquired nine weeks of TCP dump data. The dataset was collected and distributed at the Massachusetts Institute of Technology (MIT) Lincoln Laboratory.

The KDD Cup '99 intrusion detection benchmark consists of three components: the whole KDD Cup '99 dataset contains examples of attacks and normal connections, 10% KDD dataset the purpose of which is to train classifiers, and KDD test dataset designed for testing (Gifty Jeya et al, 2012, pp.28-32.). The whole KDD Cup '99 dataset contains 4,898,431 single connection records, each of which consists of 41 features labeled as normal or attacks (See Table 1).

Table 1 – Features in the KDD Cup '99 dataset
Таблица 1 – Атрибуты в KDD Cup '99 базе данных
Табела 1 – Атрибути у KDD Cup '99 бази података

Index	Feature 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/if not, 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

Index	Feature name	Description
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 root shell=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, then 0
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 current connection at a given interval
25	error rate	% of connections with SYN errors
26	srv error rate	% of connections with SYN errors
27	error rate	% of connections with REJ errors
28	srv error 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 the same service
34	dst host same src rate	% of connections to the same destination that use the same service
35	dst host srv rate	% of connections to different hosts on the same system
36	dst host same srv port rate	% of connections to a system with the same source port
37	dst host srv diff host rate	% of connections to the same service coming from different hosts
38	dst host error rate	% of connections to a host with an S0 error
39	dst host srv error rate	% of connections to a host and specified service with an S0 error
40	dst host error rate	% of connections to a host with an RST error
41	dst host srv error rate	% of connections to a host and specified service with an RST error

The features describing the connections can be classified into four categories:

Basic features are obtained from the packet header, without examining the contents of the packet (duration, protocol type, service, flag and the number of bytes sent from the source to the destination and vice versa).

Content features are determined by analyzing the content of the TCP packet (number of unsuccessful attempts to login to the system).

Time features determine duration of the connection from a source IP address to target IP addresses. The connection is a sequence of data packets starting and ending at some predefined times.

Traffic features are based on a window that has an interval of a given number of connections (not time intervals). This is suitable for describing attacks that last longer than the interval of the stipulated time features.

All attacks in the KDD Cup '99 dataset are classified as one of the four categories given in Table 2 (Al-Dhafian et al, 2015, pp.82-88).

Table 2 – Categories of attacks
Таблица 2 – Категорија атак
Табела 2 – Категорије напада

Category of Attack	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

Probe: the attacker collects information about the system or computer network to find (known) vulnerabilities, by scanning a machine or a networking device in order to determine weaknesses or vulnerabilities that may later be exploited in order to compromise the system.

DoS: the attacker does not allow legitimate users access to computing resources or overloads them so that requests cannot be processed in real time. The result of this attack is the unavailability of resources, i.e. resources are too busy or too full to serve legitimate networking requests and hence denying users access to a machine.

U2R: the attacker explores vulnerabilities in order to acquire administrator privileges (root access to the system). Attacker starts off on

the system with the normal user account and looks for vulnerabilities in order to gain super user privileges (Paliwal & Gupta, 2012, pp.57-62).

R2L: the attacker does not have a user account on the victim machine, hence tries to obtain access to the remote system without having the account (Gifty Jeya et al, 2012, pp.28-32.).

Instances in the whole dataset, 10% training set (containing 10% of the total number of instances), and the test set which contains 311,029 instances, according to the categories and datasets, as well as the percentage of the total share of a given category within a particular dataset are shown in Table 3.

Table 3 – Number of instances in the KDD Cup '99 whole dataset, 10% training set and the test set

Таблица 3 – Количество случаев в KDD Cup '99 полной базе данных, 10% в течение обучения и тестирования

Табела 3 – Број инстанци у KDD Cup '99 целој бази података, 10% у тренингу скупу и тест-скупу

Attack category	Whole dataset		10% training set		Test set	
	Number of instances	(%)	Number of instances	(%)	Number of instances	(%)
Normal	492,708	19.86%	97,278	19.69%	60,593	19.48%
Probe	41,102	0.84%	4,107	0.83%	4,166	1.34%
DoS	3,883,370	79.30%	391,458	79.24%	229,853	73.94%
U2R	52	0.00%	52	0.01%	70	0.02%
R2L	1,126	0.02%	1,126	0.23%	16,347	5.26%

There are various criticisms of the KDD Cup '99 dataset. The primary criticism is that the KDD Cup '99 dataset is not an authentic simulation of real network traffic. In addition, authors outline the following issues (Kolez et al, 2003), (Maček & Milosavljević, 2013), (Bukola & Adetunmbi, 2016):

- complexity of the calculations,
- complexity of the training and test sets,
- impact of duplicate to machine learning (ML) algorithms,
- number of instances of attack is too high in relation to the number of instances of normal traffic,
- relationship between individual categories of attack is not realistic,
- R2L instances of individual attacks are similar to normal traffic instances, which is a consequence of transforming data from the DARPA dataset to the KDD Cup '99 dataset,
- low accuracy of detecting the distribution of attacks, etc.

For these reasons, one can create alternative sets for training and testing in the following way:

- make a smaller subset of the training set,
- use only the training set,
- compose a union of parts of the training and test sets for training and for testing,
- filter instances in order to achieve proportionality of attacks, etc.

The way in which alternative sets are composed depends on the evaluation of the IDS model.

NSL-KDD dataset

The KDD Cup '99 dataset contains a number of redundant records (78%) and duplicate records (75%) which prevent classifying the other records (Revathi & Malathi, 2013). To fix these issues, a new NSL-KDD dataset was proposed (Tavallaei et al, 2009). The NSL-KDD dataset consists of a reasonable number of selected features from the KDD Cup '99 dataset which do not include redundant records in the training set nor duplicates in the test set (Kavitha & Usha, 2014, pp.77-84). Considering the design of the dataset, there are three important reasons for using it in the experiments:

- elimination of redundant records in the training set helps classifiers to be unbiased toward more frequent records;
- with duplicate records excluded from the test set, a classifier performance will not be biased by the techniques which have better decision rates on the frequent records;
- training and test sets contain a reasonable number of instances which is affordable for the experiments on the entire set without the need to randomly choose a small portion.

The training dataset is made up of 21 different attacks out of 37 present in the test dataset. The known attacks are those present in the training set, while the additional 16 attacks are available only in the test set (see Table 4). The attack types are grouped into Probe, DoS, U2R and R2L categories (Nkiama et al, 2016).

The normal traffic in the training set contains 67,343 instances which brings a total of 126,620 instances. The normal traffic in the test set contains 9,711 instances which brings total of 22,850 instances in the test set.

Table 4 – Total number of attack instances in the training and test sets
Таблица 4 – Общее количество случаев атак в течение обучения и тестирования
Табела 4 – Укупан број инстанци напада у тренинг и тест-скуповима

Attack Classes	Total number of instances in the training set	Total 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)
		Additional attacks
		apache2 (737), udpstorm (2), processtable (685), worm (2), mailbomb (39)
Probe	11,656	2,421
	satan (3,633), ipsweep (3,599), nmap (1,493), portsweep (2,931)	satan (753), ipsweep (141), nmap (73), portsweep (157)
		Additional attacks
		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)
		Additional attacks
		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)
		Additional attacks
		xterm (13), sqlattack (2), ps (5)
Total	59,277	13,139

Kyoto 2006+ dataset

The Kyoto 2006+ dataset was built on the three years of real traffic data from November 2006 to August 2009. A new version of the dataset contains additional data collected from November 2006 to December

2015. It consists of 14 statistical features derived from the KDD Cup '99 dataset as well as 10 additional features which can be used for the analysis and evaluation of the IDS network. The Kyoto 2006+ dataset is captured using honeypots, darknet sensors, email server and web crawler (Singh et al, 2015, pp.8609-8624). Song et al (2011, pp.29-36) provided a detailed analysis of honeypots (i.e. computer network security mechanisms which detect attempts of unauthorized use of information) and darknets data collected on many real and virtual machines as honeypots. They have deployed various types of honeypots, darknet and other systems on the five networks inside and outside of the Kyoto University, and collected all traffic data to and from honeypots (Table 5). During the observation period, there were 50,033,015 normal sessions, 43,043,225 attack sessions and 425,719 sessions related to unknown attacks.

Table 5 – Deployed honeypots, darknet and other systems
Таблица 5 – Установленные honeypots, darknet и другие системы
Табела 5 – Инсталирани honeypots, darknet и други системи

Deployed systems	
Honeypots	Solaris 8 for Intel
	Windows XP (no patch, SP2, fully patched)
	Nepenthes
	Others
Darknet	Darknet sensors (for detection of software, configuration, or authorization that use non-standard communication protocols and ports)
Other systems	Mail server (to collect various types of mails)
	Web crawler (developed by the NTT Information Sharing Platform Laboratories)
	Windows XP (to evaluate malware activities)

Based on 41 original features of the KDD Cup '99 dataset, the authors extracted the statistical features from the honeypot data, ignoring other features that contain redundant data (see Table 6).

The authors excluded substantially redundant and insignificant features as well as contents features (number of file creation operation, number of operation on access control files), because they are not suitable for network-based IDSs and it is time consuming to extract them without the domain knowledge. In addition to the above 14 statistical features, the authors also extracted additional 10 features (Table 7), which enabled them to investigate what kinds of attacks happened on computer networks.

Table 6 – Statistical features in the Kyoto 2006+ dataset derived from the KDD Cup '99 dataset

Таблица 6 – Статистические характеристики в Kyoto 2006+ базе данных, полученных из KDD Cup '99 базы данных

Табела 6 – Статистички атрибути у Kyoto 2006+ бази података који су преузети из KDD Cup '99 базе података

Index	Feature name	Description
1	Duration	The length of the connection (seconds).
2	Service	The connection's server type (http, telnet).
3	Source bytes	The number of data bytes sent by the source IP address.
4	Destination bytes	The number of data bytes sent by the destination IP address.
5	Count	The number of connections whose source IP address and destination IP address are the same to those of the current connection in the past two seconds.
6	Same_srv_rate	% of connections to the same service in the Count feature.
7	Serror_rate	% of connections that have 'SYN' errors in Count feature.
8	Srv_serror_rate	% of connections that have 'SYN' errors in Srv_count (% of connections whose service type is the same to that of the current connections in the past two seconds) feature.
9	Dst_host_count	Among the past 100 connections whose destination IP address is the same to that of the current connection, the number of connections whose source IP address is also the same to that of the current connection.
10	Dst_host_srv_count	Among the past 100 connections whose destination IP address is the same to that of the current connection, the number of connections whose service type is also the same to that of the current connection.
11	Dst_host_same_src_port_rate	% of connections whose source port is the same to that of the current connection in Dst_host_count feature.
12	Dst_host_serror_rate	% of connections that have 'SYN' errors in Dst_host_count feature.
13	Dst_host_srv_serror_rate	% of connections that have 'SYN' errors in Dst_host_srv_count feature.
14	Flag	The state of the connection at the time of connection was written.

Table 7 – Additional features in Kyoto 2006+ dataset
Таблица 7 – Дополнительные атрибуты в Kyoto 2006+ базе данных
Табела 7 – Додатни атрибуту у Kyoto 2006+ бази података

Index	Feature name	Description
1	IDS_detection	Reflects if IDS triggered an alert for the connection; '0' means any alerts were not triggered and an arabic numeral means the different kind of alerts. Parenthesis indicates the number of the same alert.
2	Malware_detection	Indicates if malware, also known as malicious software, was observed at the connection; '0' means no malware was observed, and string indicates the corresponding malware observed at the connection. Parenthesis indicates the number of the same malware.
3	Ashula_detection.	Means if shellcodes and exploit codes were used in the connection; '0' means no shellcode nor exploit code were observed, and an arabic numeral means the different kinds of the shellcodes or exploit codes. Parenthesis indicates the number of the same shellcode or exploit code
4	Label	Indicates whether the session was attack or not; '1' means normal. '-1' means known attack was observed in the session, and '-2' means unknown attack was observed in the session.
5	Source_IP_Address	Means source IP address used in the session. The original IP address on IPv4 was sanitized to one of the Unique Local IPv6 Unicast Addresses. Also, the same private IP addresses are only valid in the same month; if two private IP addresses are the same within the same month, it means their IP addresses on IPv4 were also the same, otherwise are different.
6	Source_Port_Number	Indicates the source port number used in the session.
7	Destination_IP_Address	It was also sanitized.
8	Destination_Port_Number	Indicates the destination port number used in the session.
9	Start_Time	Indicates when the session was started.
10	Duration	Indicates how long the session was being established.

Datasets comparison

Al-Dhafian et al (2015, pp.82-88) presented a comparison between five datasets: DARPA, KDD Cup '99, CAIDA, NSL-KDD and Kyoto 2006+ datasets. Table 8 shows the results for all datasets except for CAIDA, which is a collection of several different types of data resulting from both

active and passive measurements of the Internet, and is not analyzed here.

Table 8 – Comparison of the standard datasets in IDSs
Таблица 8 – Сравнение стандартных баз данных в системах обнаружения атак
Табела 8 – Поређење стандардних база података у системима за детекцију упада

Dataset (year)	Features	Pros	Cons
DARPA (1998)	–	<ul style="list-style-type: none"> – First standard for evaluating IDS. – Consists of broad range of attacks. 	<ul style="list-style-type: none"> – Models used to generate traffic were too simple. – Synthesized data does not simulate the background traffic in real networks.
KDD Cup '99 (1999)	41 features (32 numeric and 9 categorical)	<ul style="list-style-type: none"> – Used for evaluating anomaly detection systems. – Attack types in training set are distinctive from the testing set. 	<ul style="list-style-type: none"> – Includes redundant and duplicate records. – Does not reflect the modern environment.
NSL-KDD (2009)	41 features (32 numeric and 9 categorical)	<ul style="list-style-type: none"> – 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. 	<ul style="list-style-type: none"> – Not perfect for representing the existing real networks.
Kyoto 2006+ (2009)	24 features (14 statistical derived from KDD Cup '99 and 10 additional)	<ul style="list-style-type: none"> – Ignored features that contain redundant. – Represents the existing real networks. 	<ul style="list-style-type: none"> – Does not mention information on particular attack types.

The DARPA dataset is considered as a popular dataset used in IDSs to measure detection rate and false alarm rate for network traffic which consists of four types of attacks (Probe, DoS, U2R and R2L). However, it faces a lot of criticism primarily because of using very simple

models to create background network traffic. As a result, synthesized data does not look like to be similar to the records of background traffic in real networks.

The KDD Cup '99 dataset is a preprocessed version of the DARPA dataset, which classified records into 41 features. The dataset consists of a huge number of records in both training and tests sets but includes redundant and duplicate records and does not represent real network traffic. However, in the development of new intrusion detection systems and tools for data protection, the KDD Cup '99 dataset is widely used to conduct the experiments on large amounts of data, or whenever the repeatability is a must.

The NSL-KDD dataset contains selected features from the KDD Cup '99 dataset. It is designed to fix problems related to redundant records in the training set and duplicated records in the test set, as well as to reduce quantity of data to a reasonable size.

The Kyoto 2006+ dataset is a comprehensive representation built on real network traffic data through ignoring features that contain redundant records. The dataset is captured using honeypots, darknet sensors, email server, web crawler and other computer network security mechanisms which detect attempts of unauthorized use of information. Researchers from the Kyoto University have deployed various types of honeypots, darknet sensors and other systems on five networks inside and outside the Kyoto University, and collected all traffic data to and from honeypots.

Conclusion

KDD Cup is an annual conference for Data Mining and Knowledge Discovery, intended for competition in the field of machine learning and data mining. In 1999, competitors had to solve the problem of protection against attacks on computer networks. For the purpose of competition, the KDD Cup '99 dataset had been created. The KDD Cup '99 benchmark consists of the whole dataset, 10% training set and the test set. Each record is made up of 41 features which describe the network traffic of a simulated computer network. The dataset, among other things, contains data on the following attacks: Probe, DoS, U2R and R2L.

The KDD Cup '99 dataset is widely used as a reference for researching IDSs and for the development of new tools for protection against various attacks on computer networks. However, there are shortcomings which can affect the research such as complexity, the effect of duplicates and redundant records, unbalanced number of

attacks relative to each other and disproportion between the number of attacks and normal traffic. One way to avoid these problems is to use the NSL-KDD dataset which does not contain redundant records in the training set and duplicates in the test set. However, researchers have to be aware that both KDD Cup '99 and NSL-KDD datasets are a simulation of a virtual computer network and, consequently, experiments can give contradictory results (especially if the number of features describing the attack is small). The Kyoto 2006+ dataset represents selected features of real network traffic which is captured using honeypots, darknet sensors, email server and web crawler deployed on five networks inside and outside the Kyoto University. It does not contain information on particular attacks and ignore features that contain redundant records.

Since rapid development of computer networks and information systems has led to a large number of sophisticated attacks, researchers from all around the world develop new IDSs to protect computer networks from hackers by using known datasets and their pre- and post-processed versions. KDD Cup '99, NSL-KDD and Kyoto 2006+ datasets are widely used in the experiments to develop various tools for protection against malicious attacks. Which of the bases is used depends on the purpose of a particular IDS and security goals in specific problem solving.

References

- 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>.
- 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, July 27-30.
- Bukola, O. & Adetunmbi, A.O. 2016. Auto-Immunity Dendritic Cell Algorithm. In: *International Journal of Computer Applications*, 137(2), pp.10-17, March 2016. New York: Foundation of Computer Science. Available at: <https://doi.org/10.5120/ijca2016908689>.
- 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. Available at: <http://www.ijcaonline.org/archives/volume45/number21/7076-9751>. Accessed: 10.01.2018.
- 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>. Accessed: 11.01.2018.

- KDD CUP '99 dataset*. [Internet] Available at: <http://kdd.ics.uci.edu/dataset/kddcup'99/kddcup'99.html>. Accessed: 12.02.2018.
- Kolez, A., Chowdhury, A. & Alspector, J. 2003. Data duplication: an imbalance problem? In: *ICML 2003. Workshop on Learning from Imbalanced Data Sets (II)*, Whashington, August 21.
- 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.), June 3-6.
- Nkama, 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>.
- 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>. Accessed: 12.02.2018.
- Protić, D. 2016. Neural Cryptography. *Vojnotehnički glasnik/Military Technical Courier*, 64(2), pp.483-495. Available at: <https://doi.org/10.5937/vojtehg64-8877>.
- 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. Available at: file:///C:/Users/Intel/Downloads/V2I12_IJERTV2IS120804.pdf. Accessed: 12.02.2018.
- SIGKDD - KDD Cup. *KDD Cup 1999: Computer network intrusion detection*. [Internet]. Available at: www.kdd.org. Accessed: 13.02.2018.
- Singh, R., Kumar, H. & Singla, R.K. 2015. An intrusion detection system using network traffic profiling and online sequential extreme learning machine. *Expert Systems With Applications*, 42(22), pp.8609-8624. Available at: <https://doi.org/10.1016/j.eswa.2015.07.015>.
- Song, J., Takakura, H., Okabe, Y., Eto, M., Inoue, D. & Nakao, K. 2011. Statistical Analysis of Honeypot Data and Building of Kyoto 2006+ Dataset for NIDS Evaluation. In: *Proc. 1st Work-shop on Building Anal. Datasets and Gathering Experience Returns for Security*. Salzburg, pp.29-36. April 10-13. Available at: <https://doi.org/10.1145/1978672.1978676>.
- 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, July 8-10. Available at: <https://doi.org/10.1109/CISDA.2009.5356528>.

ОБЗОР KDD CUP '99, NSL-KDD И KYOTO 2006+ БАЗ ДАННЫХ

Даниела Д. Протич

Вооруженные силы Республики Сербия, Генеральный штаб,
Управление информатики и телекоммуникаций (J-6),
Центр прикладной математики и электроники,
г. Белград, Республика Сербия

ОБЛАСТЬ: компьютерные науки, информационные технологии

ВИД СТАТЬИ: обзорная статья

ЯЗЫК СТАТЬИ: английский

Резюме:

В данной работе представлен обзор трех баз данных: KDD Cup '99, NSL-KDD и Kyoto 2006+ база данных, которые широко используются в исследованиях обнаружения взлома компьютерных сетей. KDD Cup '99 база данных состоит из пяти миллионов записей, каждая из них содержит 41 атрибут, который может классифицировать атаки по следующим четырем видам: Probe, DoS, U2R и R2L. KDD Cup '99 база данных не в состоянии отражать реальные данные, так как она генерирована моделированием на виртуальной компьютерной сети. Из NSL-KDD базы удалены избыточные записи, а дублированные записи удалены из баз обучения и тестирования KDD Cup '99. Kyoto 2006+ база образована на основании данных трехлетнего реального сетевого трафика, которые обозначены, как: нормальный (не атака), атака (известная атака) и неизвестная атака. Kyoto 2006+ база содержит 14 статистических атрибутов, выбранных из KDD Cup '99 базы и дополнительных 10 атрибутов.

Ключевые слова: обнаружение атак, компьютерная сеть, KDD Cup '99, NSL-KDD, Kyoto 2006+.

ПРЕГЛЕД KDD CUP '99, NSL-KDD И KYOTO 2006+ БАЗА ПОДАТАКА

Данијела Д. Протић

Војска Србије, Генералштаб, Управа за телекомуникације и информатику
(J-6), Центар за примењену математику и електронику,
Београд, Република Србија

ОБЛАСТ: рачунарске науке, информационе технологије

ВРСТА ЧЛАНКА: прегледни чланак

ЈЕЗИК ЧЛАНКА: енглески

Сажетак:

У раду је приказан преглед три базе података: KDD Сир '99, NSL-KDD и Kyoto 2006+, које се често користе у истраживању детекције упада у рачунарске мреже. KDD Сир '99 база података састоји се од пет милиона записа, од којих сваки садржи 41 атрибут, који могу да класификују нападе у четири класе: Probe, DoS, U2R и R2L. KDD Сир '99 база података не може да рефлектује реалне податке, јер је генерисана симулацијом на виртуелној рачунарској мрежи. Из NSL-KDD базе уклоњени су редувантни записи и дупликати из KDD Сир '99 тренинг и тест-базе, респективно. Kyoto 2006+ база формирана је на основу података трогодишњег реалног мрежног саобраћаја, који су означени као: нормалан (није напад), напад (познат напад) и непознат напад. Kyoto 2006+ база садржи 14 статистичких атрибута издвојених из KDD Сир '99 базе и додатних 10 атрибута.

Кључне речи: детекција упада, рачунарска мрежа, KDD Сир '99, NSL-KDD, Kyoto 2006+.

Paper received on / Дата получения работы / Датум пријема чланка: 25.02.2018.
Manuscript corrections submitted on / Дата получения исправленной версии работы /
Датум достављања исправки рукописа: 09.04.2018.
Paper accepted for publishing on / Дата окончательного согласования работы / Датум
коначног прихватања чланка за објављивање: 11.04.2018.

© 2018 The Author. Published by Vojnotehnički glasnik / Military Technical Courier (www.vtg.mod.gov.rs, втг.мо.упр.срб). This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/3.0/rs/>).

© 2018 Автор. Опубликовано в «Военно-технический вестник / Vojnotehnički glasnik / Military Technical Courier» (www.vtg.mod.gov.rs, втг.мо.упр.срб). Данная статья в открытом доступе и распространяется в соответствии с лицензией «Creative Commons» (<http://creativecommons.org/licenses/by/3.0/rs/>).

© 2018 Аутор. Објавио Војнотехнички гласник / Vojnotehnički glasnik / Military Technical Courier (www.vtg.mod.gov.rs, втг.мо.упр.срб). Ово је чланак отвореног приступа и дистрибуира се у складу са Creative Commons лиценцом (<http://creativecommons.org/licenses/by/3.0/rs/>).

