ISSN 2347 - 3983
Volume 8. No. 10, October 2020
International Journal of Emerging Trends in Engineering Research
Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter958102020.pdf
https://doi.org/10.30534/ijeter/2020/958102020

A Detailed Study on A Benchmark Intrusion Dataset - Kyoto 2006+

Gaddam Venu Gopal[1], Dr. Gatram Ramamohan Babu[2]
[1]Research Scholar, Dr. Y.S. Rajasekhar Reddy University College of Engineering & Technology, Acharya Nagarjuna University, Nagarjuna Nagar, Guntur, Andhra Pradesh, India.
Email: venugopal.gaddam@gmail.com
[2]Professor, Department of Information Technology, RVR & JC College of Engineering, Chowdavaram, Guntur, Andhra Pradesh, India.
Email: rmbgatram@gmail.com

ABSTRACT

Intrusion detection datasets are playing a vital role in the field of network security. Kyoto 2006+, NSL-KDD, KDD Cup '99 etc., datasets are a few benchmark datasets available for assessing the efficiency of Intrusion Detection Systems for research. These three datasets are widely used for attack prediction approaches in Network Intru sion Detection System (NIDS). The Kyoto 2006+ is developed in a real time environment and data is collected from the three year-network traffic using Honeypots. This paper presents an overview of Kyoto 2006+ dataset and compared with other datasets.

Key words: DDoS, Introduction Detection System, Kyoto 2006+, NSL-KDD.

1. INTRODUCTION

An Intrusion is an act of violating the information protection and Intrusion Detection System (IDS) monitors the network activities and detects intruders and act accordingly. The IDS comprises of software as well as hardware tools that continuously observes computer network traffic, identifies suspicious activities among them, and triggers an alarm whenever a malicious activity founds. There are 14 features that were extracted from KDD Cup'99 and 10 more supplementary features i.e., a total of 24 features are exist in Kyoto 2006+ dataset.

NIDS based on Host (Host-based NIDS) and NIDS based on Network (Network based NIDS) are two different types of systems exist. A Host based NIDS detects the intrusions by observing the logs of Server by the commonality among patterns of the known and existing attacks. Where as in the Network based NIDS identifies the intruder by monitoring the network traffic and by detecting irregular behavior among header and content of each packet in the network traffic.

Network Intrusion Detection System (NIDS), detects the intruders by observing and analyzing the network traffic's security violations. There are two different types of NIDS namely Anomaly-based and Misuse-based NIDS. In Anomaly-based systems, statistical models were built to observe the normal network traffic and monitor abnormalities and detect anomalous. Whereas in misuse based systems, patterns were designed based on the signatures extracted from known attacks, these patterns were used to detect attacks.

Several machine learning techniques can be applicable on these intrusion detection dataset such as Classification, clustering, regression etc. Both supervised and unsupervised approaches can be developed to find the patterns of both attack and normal kind of network traffic. Due to the availability of class labels among most of the intrusion datasets, classification techniques are most preferable by the many researchers to build their intrusion detection systems.

These Intrusion dataset are widely used to predict attacks depending on the class label using machine learning techniques like Support Vector Machines [10], K-nearest Neighbor classifiers [7], Decision trees etc. Few of the authors extracted best features from available 24 features using some feature selection approaches [6][9].

2. DATASET DESCRIPTION

The Kyoto 2006+ dataset [4] was collected from a real-time network traffic during the period of 2006 to 2009 about three years from various types of sources as listed in Table 1. Another collection of the dataset for Kyoto 2006+ also consists of instances from 2006 to 2015. It contains 14 independent variables that were extracted from a well-known benchmark dataset KDD Cup '99 [3], that contains 42 features. Apart from these 14 features, 10 more features that can be useful to analyze and evaluate the IDS network were added in the Kyoto 2006+ dataset. Table 2 provides the list of features that are considered in Kyoto 2006+ from KDD Cup '99[3].

This Kyoto dataset is developed on the combination of various sources like honeypots, darknet sensors, email server and web crawler [1][2]. The Kyoto University built a mechanism that detects the unauthorized attempts to the information. The darknets collected data from many real and virtual machines. The frame work is the process of the deployment of numerous kinds of darknet, honeypots, and a few more systems on the different types of networks and collected network traffic data. These honeypots are used to extract the network Kyoto University. In this period of time, about 50,033,015 instances are collected normal instances and 43,043,225 instances were observed as attacks and as an unknown type of attacks a total of 425,719 instances were generated.

Table 1. Sources of Network traffic generated

S.No	Generator Types	Systems that were deployed
1.	Honeypots	Intel Solaris 8
		Windows XP
		Nepenthes and other systems
2.	Darknet	this sensors (it detects configuration, software, or authorization of non-standard communication protocols and ports) used network packets
3.	Other Environments	Mail servers in order to connect mails of different types.
		Web crawler
		Windows XP in order to observe malware activities

3. FEATURE DESCRIPTION

This section provides the description of each feature of the Kyoto 2006+ dataset. Among the total number of features from the dataset, 14-features were extracted from KDD Cup'99 dataset. These features and their description are tabulated in Table 2

1. IDSdetection: This feature stores a numeric value related whether IDS set off an alarm for the association; '0' signifies any alerts were not set off, and an Arabic numeral (aside from '0') implies the various types of the alerts. Enclosure shows the quantity of a similar alarm saw during the association.

2. Malware detection: This feature is used to represent the malicious software (malware) is identified in the connection; '0' represents no attack is found, whereas when a non-zero literal is present in this feature it is the specific attack that is detected. A software named as 'clamAV' is used to detect malware. Parenthesis is used to represent the total observation of the same malware at the time of the existence of the connection.

3. Ashuladetection: This feature is a numeric and its value is '0' if the packet didn't contain shell codes otherwise it contains non-zero numeral when shell code is observed and each numeral is meant for different type of shell codes. A special character i.e., Parenthesis is used to represent the total shellcodes that were identified at the time of the existence of the connection.

Table 2: List of features considered in Kyoto 2006+ from KDD Cup '99.

S. No	Feature	Description
1	Duration	Connection length in seconds
2	Service	Type of the connection's service like http, telnet
3	Source bytes	# of data bytes that source IP address sent
4	Destination bytes	# of data bytes that Destination IP address sent
5	Count	# of connections have the same source and destination IP addresses to those of the current connection in the past two seconds
6	Samesrvrate	% of connections to the same service in Count feature
7	Serrorrate	% of connections that have "SYN" errors in Count feature
8	Srvserrorrate	% of connections that have "SYN" errors in Srvcount(the # of connections for which service type is the same in the past two seconds)
9	Dsthostcount	among the past 100 connections whose destination IP address and source IP address are the same to that of the current connection.
10	Dsthostsrvcount	among the past 100 connections whose destination IP address and service type are the same to that of the current connection
11	Dsthostsamesrcportrate	% of connections whose source port is the same to that of the current connection in Dsthostcount feature
12	Dsthostserrorrate	% of connections that "SYN" errors in Dsthostcount feature
13	Dsthostsrvserrorrate	% of connections that "SYN" errors in Dsthostsrvcount
14	Flag	A summary is written when any specific connection is terminated. There may be different type of states for different connections and those states while the connection termination is observed and stored in this feature. Table 3 presents a description of flag entries.

The description of each feature from the additional ten features that were extracted in Kyoto 2006+ dataset apart from KDD Cup'99 [5] are presented here.

4. Label: This is the class label feature of the dataset. It contains three distinct values those are 1, -1 and -2. '1' indicates a normal session, -1 indicates as an observation of a known attack, and the numeral '-2' represents an unknown attack in this session.

5. SourceIPAddress: This feature consists of the source machine's session IP address. The IP address is a sanitized exclusive local IPv6 Address of its corresponding IPV4 due to some security issues.

6. SourcePortNumber: This feature holds the port number of the source utilized by the session.

7. DestinationIPAddress: This feature consists of the session's destination IP address. The IP address is a sanitized exclusive local IPv6 Address of its corresponding IPV4 due to some security issues.

8. DestinationPortNumber: This feature holds the port number of the destination utilized by the session.

9. StartTime: represents the staring time of the session.

10. Duration: The total duration of the session being connected.

Table 3: Description of each flag value

S.No.	Flag Value	Description
1	S0	An attempt to establishment of the connection is observed but no reply found
2	S1	An observation of the establishment of the connection is found, but the connection was not closed with a byte count of zero
3	S2	Establishment of the connection is found but originator attempt is closed but reply from the recipient is not observed
4	S3	Establishment of the connection is observed but the response attempt was closed and reply is not observed from originator
5	SF	Connection establishment as well as a connection termination are observed as a normal activity
6	SH	Originator forwarded sync as well as acknowledge packets but form the recipient no observation of sync and acknowledge packets the connection opened in one-way
7	SHR	Recipient forwarded sync as well as acknowledge packets but form the originator no initiation of sync packet.
8	REJ	The attempt of the connection is rejected
9	RSTO	Establishment of the connection is observed, but originator is aborted
10	RSTOS	Originator forwarded sync and RST packets but from the recipient no sync as well as acknowledge packets are not observed
11	RSTR	Connection is established between originator and the recipient but the recipient aborted
12	RSTRH	Recipient forwarded sync, acknowledge and RST packets but no sync packet is observed from the originator
13	OTH	Observation of any other type of problems like no sync or establishment of partial connection etc.

4. COMPARISON OF KYOTO 2006+ DATASET:

There are numerous intrusion detection datasets that are available; the following is the overall description of some of these datasets.

The DARPA dataset was prepared in 1998, a simple models were used to generate the network traffic data through simulators. The limitation of this dataset is the network traffic generated by the simulator far from the real traffic.

The KDD Cup '99 dataset is developed in 1999. This dataset is a benchmark dataset widely using by many researchers to evaluate their intrusion detection systems[7][8] through machine learning algorithms. This dataset comprises of 41 features and 25 distinctive attack types. One of the limitations of this dataset is, this dataset contains duplicate instances.

NSL-KDD dataset is a dataset extracted from KDD Cup'99 dataset by eliminating duplicate instances. It is available in CSV formats and also divided into test and training sets with different number instances for the flexibility of researchers. The limitations of NSL-KDD and KDD Cup'99 datasets are that, these datasets doesn't reflects modern environment and not suitable for the current real network.

KYOTO 2006+ dataset is developed in 2006 and extracted upto 3 years and contains 24 distinct features and one of the limitations is that it doesn't provide the attack type information. Intrusion Detection Systems

5. CONCLUSION

In this paper a detailed study of Kyoto 2006+ dataset is presented. It is also compared with the KDD Cup'99, DARPA 1998, NSL KDD benchmark datasets. Kyoto 2006+ dataset selected specific set of useful features derived by the KDD Cup'99 dataset and ten more features were also extracted to describe the characteristics of the network traffic. It is a dataset collected from real-time environment through honeypots. The main advantage of this dataset is, it is developed in the real time environment and addresses current attacks.

REFERENCES

[1] Aburomman, Abdulla Amin, and Mamun Bin Ibne Reaz. "A survey of intrusion detection systems based on ensemble and hybrid classifiers." *Computers & Security*, *Vol.* 65, pp. 135-152, 2017.

[2] KDD Cup'99: http://kdd.ics.uci.edu/ databases/kddcup99/kddcup99.html

[3] Dhanabal, L., and S. P. Shantharajah. "A study on NSL-KDD dataset for intrusion detection system based on classification algorithms." *International Journal of Advanced Research in Computer and Communication Engineering*, Vol. 4, no. 6, pp. 446-452, 2015.

[4] Kyoto 2006+ dataset: http://www.takakura.com/Kyoto_data/

[5] Divekar, A., Parekh, M., Savla, V., Mishra, R., & Shirole, M. " Benchmarking datasets for anomaly-based network intrusion detection: KDD CUP 99 alternatives". In *2018 IEEE 3rd International Conference on Computing, Communication and Security (ICCCS)*. IEEE, pp. 1-8, 2018.

[6] Singh, Amrit Pal, and Arvinder Kaur. "Flower Pollination Algorithm for feature analysis of Kyoto 2006+ data set." *Journal of Information and Optimization Sciences* Vol. 40, no.2, pp. 467-478, 2019.

[7] K., & Rao, B. B. (2019). "Impact of PDS Based kNN Classifiers on Kyoto Dataset." *International Journal of Rough Sets and Data Analysis (IJRSDA)*, Vol.*6, no.* 2, PP. 61-72, 2019.

[8] Canbay, Yavuz, and Seref Sagiroglu. "A hybrid method for intrusion detection.", *IEEE 14th International Conference on Machine Learning and Applications (ICMLA)…2015*

[9] Najafabadi, Maryam M., et al., " Evaluating feature selection methods for network intrusion detection with kyoto data." *International Journal of Reliability, Quality and Safety Engineering, Vol.* 23, no. 01, sp. 1650001, 2016.

[10] Agarap, A. F. M. (2018, February). "A neural network architecture combining gated recurrent unit (GRU) and support vector machine (SVM) for intrusion detection in network traffic data." In *Proceedings of the 2018, 10th International Conference on Machine Learning and Computing*, pp. 26-30, 2018.