ISSN 2278-3091
Volume 9, No.1.3, 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse4391.32020.pdf
https://doi.org/10.30534/ijatcse/2020/4391.32020

Overview of Network Dataset and Data Mining Technique

Noor SuhanaSulaiman, NurSukinah Aziz, NooraidaSamsudin, Wan AinulAlyani, AzlizaYacob, LukmanulhakimNgah
University College TATI, Malaysia, suhana@tatiuc.edu.my

ABSTRACT

Network dataset has been widely used in network research. Network dataset has big challenges lead to certain matters, as an example huge time consumtion in data processing and transmission and false alarm rate increased in attack detection gaining the harm and unsafe condition to online user. Hence the nature of network dataset need to be analyzed. Many researchers try to come out with better algorithm to handle the network dataset, consequently to come out with better finding result. This paper is about to analyze the nature of network dataset and review the data mining techniques in supervised and unsupervised approach to handle and process the network dataset.

Key words: Data mining, Supervised, Unsupervised.

1. INTRODUCTION

An object and device connected to each other through the Internet (i.e. Internet of Things (IoT)) are increased, have seen a steady increase of cloud systems to store data including credential data and paperless transaction as a result of online transactions. All these online technological dependencies expose users to high risk of public network / Internet breach of trusts such as compromised online credential data, unsafe communication between senders and receivers, and unauthorized access to communication session. In fact, given current rate of online technological growth, IoT based technologies are increasingly prone to vulnerabilities.

The latest data on Internet usage across the world [1] indicates that the world saw a billion of Internet users on June of 2017. Refer to Figure 1, Asia has gained the highest Internet user population with 2023 million users; meanwhile, the lowest Internet user is from Oceania/Australia with 28 million users. The increase in networked machines quantity has led to the expansion in illegitimate activities of internal and external attacks; for example, users gaining unprivileged access for individual gain [2].

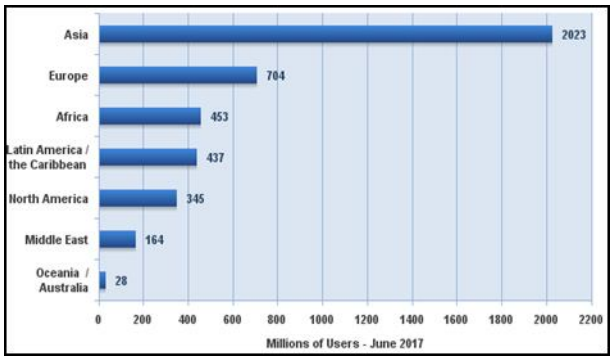

Figure 1: Statistic of Internet User

Mean time, Intrusion Detection System (IDS) dataset consist in large amount records, usually generated from different sources (network traffic, log of network or application, etc), [3] which possibly include the repetitive data. Because of the factor of noise, insignificant and irrelevant feature, the capability to process the less precise and conflict data lead to one of the crucial requirement in feature selection area. Large audit data included in IDS which need to be examined become crucial because of large features become more difficult in detecting abnormal behavior pattern. In network data set, a standout amongst the most imperative lacks is tremendous redundant records number, which makes the learning algorithms to bias to frequent records.

2. NSL KDD CUP 99 DATASET

IDS dataset as KDD Cup 99 dataset is actively used among researcher in Intrusion Detection System area. New dataset, NSL-KDD [4], is proposed, which consist of chosen record of total KDD dataset. NSL-KDD dataset is a second version of KDD 99 dataset. NSL-KDD is better than the previous version as the dataset does not include the redundant records in the train and test set, so that the classifier will not bias towards the frequent record. NSL-KDD dataset consists of 41 condition and one decision class feature attribute. The class attribute has 21 classes that categorize under four attacks types: Probe, User to Root (U2R), Remote to Local (R2L) and Denial of Service (DoS). NSL-KDD dataset also has a binary class attribute.

Table 1: Type of Features in NSL KDD Cup 99

Type	Features
Binary	Protocol_type(2), Service(3), Flag(4)
Nominal	Land(7),logged_in(12),root_shell(14), su_attempted(15),is_host_login(21),is_guest_login(22)
Numeric	Duration(1), src_bytes(5), dst_bytes(6), wrong_fragment(8),urgent(9), ot(10),num_failed_logins(11), num_compromised(13), num_root(16), num_file_creations(17), num_shells(18), num_access_files(19), num_outbound_cmds(20) , count(23) srv_count(24), serror_rate(25), srv_serror_rate(26), rerror_rate(27), srv_rerror_rate(28), same_srv_rate(29)diff_srv_rate(30), srv_diff_host_rate(31), dst_host_count(32), dst_host_srv_count(33), dst_host_same_srv_rate(34), dst_host_diff_srv_rate(35), dst_host_same_src_port_rate(36), dst_host_srv_diff_host_rate(37), dst_host_serror_rate(38),dst_host_srv_serror_rate (39), dst_host_rerror_rate(40), dst_host_srv_rerror_rate(41)

In Denial of Service (DOS), attackers endeavor to stop the authenticate user to access the provided service. Intruder will send overflow messages asking the network or server to authenticate the requests. Meanwhile, in R2L attack the hacker disguise as a normal user attempt to abuse system vulnerabilities to gain super user privileges. Remote to User (U2R) act with packets send over the internet to machine, until user does not have an access to compromise system anymore. The aim is to uncover flaws and exploit privileges of local user computer. Meanwhile probing scanning a machine or a network device searching the system vulnerabilities to be exploited later on [5]. Refer to Table 2 for specific types of attacks.

Table 2: Class of Attack with Type of Attack Categories

Attack Class	Attack Type
Dos	Back, Land, Neptune, Pod, Smurf, Teardrop, Apache2, Udpstorm, Processtable, Worm.
Probe	Satan, Ipsweep, Nmap, Portsweep, Mscan, Saint.
R2L	Guess_Password, Ftp_write, Imap, Phf, Multihop, Warezmaster, Warezclient, Spy, Xlock, Xsnoop, Snmpguess, nmpgetattack, Httptunnel, Sendmail, Named.
U2R	Buffer_overflow, Loadmodule, Rootkit, Perl, Sqlattack, Xterm, Ps.

3DATA MINING

Meanwhile, data mining has been applied widely in the areas of IDS application, education, medical diagnosis, fraud detection and banking [6]. Data mining is one of techniques that can be used to address IDS performance, specifically in Preprocessing and Analysis phase. Several data mining techniques such as Pattern Recognition, Clustering, Association and Classification that can be employed. Knowledge Discovery in Databases (KDD) is a domain which encompasses theory, method and technique to extract meaningful knowledge from dataset. A set of standard is employed including selection, preprocessing, transformation, data mining and interpretation or evaluation to generate the knowledge, as depicted in Figure 2. In KDD data analyzing, data mining is the most important step in data processing [7].

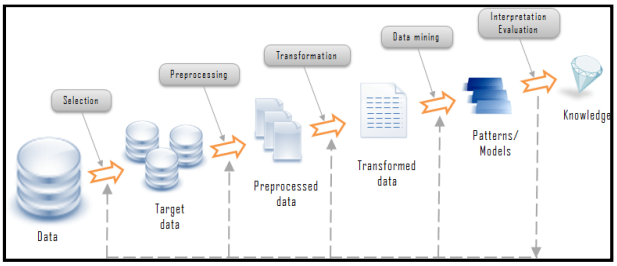

Figure 2: Data Mining Process

To address problems in IDS, classification can be considered as an alternative to increase IDS application performance in terms of accuracy, complexity and precision. The problems in attack to classification are caused by low data quality, incorrect and missing values, attribute types, dominant classes presence. The overfitting (retraining) as well as underfitting (weak model) problems [8]. Performance accuracy is one of data mining concerns as any algorithm can lost the property of accuracy and performance due to several factors. One of the factors is classification algorithm sensitivity to noisy data which causes the processing power of classification to slow down. Consequently, it decreases IDS performance. Hence, to improve classification performance problem, feature selection should be employed to select relevant attributes that improve classification performance [9].

Data mining technique capable to encounter the issues of IDS such as;

- Discard normal activities from alarm dataset to have significant real attacks activities.
- Determine false alarm from “bad” sensor generator signature.
- Search suspicious activity which uncovers a real attacks.

- Identify long, ongoing/continuous patterns.

There are two categories of data mining; supervised and unsupervised method to be employed in IDS dataset, which will elaborate in next subsection.

3.1Supervised Data Mining

In supervised learning, firstly a classifier model needs to be developed. Then, a qualifier makes learning stage to check the expected classification result, which depends on the type of artificial intelligence approach involved. However, if the classification result is inaccurate, additional training of classifier is needed. The stage continues until it reaches the level of quality desired or the algorithm works incorrectly or the data does not have an identified structure [8]. Supervised learning training the algorithm with the labeled sample. Then, the trained algorithm can predict unlabelled sample which is similar to the samples. The process includes knowledge extraction, prediction and compression tasks [10].

One related classifier of supervised learning is K-Means. K-Means algorithm is a simple clustering used in result testing. K-Means algorithm provides an indication to the problem solved or the opposite. However, there is no guarantee that cluster made by K-means algorithm is correct. Another classifier example is One Class Support Vector Machines. Its classifier uses binary classification with fast execution. The classifier pre-processes the data to check for any abnormal behavior condition before passing the condition to other algorithms to be processed into training and testing set [11].

Naive Bayes is another supervised learning classifier, which based on Bayes Theorem. The “naïve” notion refer to the existence of one variable in a problem does not affect on the existence of other variable. The conditional probability is employed to classify the problem by combining previous likelihood calculation and probability to generate next probability by utilize the Bayes rules [12].

Random Forest is also a supervised learning classifier that is based on an analysis of group of tree. Each tree produces a random selection. The training set is made from the example of tree classification. Every tree provides a classification and call “votes” for each class. Forest choose from all classification trees in the forest, by the elementary value that a group of “weak learner”' can compose by a “strong learner” [12].

3.2 Unsupervised Data Mining

Unsupervised learning consists of tasks combination and descriptive models. Unsupervised classification can be called as clustering which is also known as data analysis. Clustering process is to separate unlabeled dataset into finite data, hidden data structure and natural discrete set. This technique does not provide characterization accuracy of unobserved samples generated by the same probability distribution [13]. Meanwhile, the advantage of unsupervised learning is that it enables to solve problem occurs without any prior knowledge on the analyzed data [8]. Unsupervised learning does not need any sample of training set. It uses statistical approach of density estimation to find clusters of hidden data or to group similar data. The process includes pattern recognition and outlier detection tasks [10].

Support Vector Machines (SVM) is among popular algorithms in data mining implementation. SVM is able to classify linear and non-linear processes and promises accurate result as classification output. K-Nearest Neighbors (k-NN) is also a promising algorithms in the classification. The classification happens on the basis of different neighbors instead of trying to make classifier seems to fit better the featured data. Nonetheless, there are other related machine learning algorithm such as Decision Tree and Bayesian algorithms. Both classifiers seem less promising for detecting problems. The differences between normal and abnormal data are minimal and it seems that these algorithms produce more errors due to bias variance value and overfit model. Decision Tree and Bayesian are still employed in the implementation of IDS to analyze and classify the correct or incorrect data [11].

4. CONCLUSION

This paper is give the overview of NSL KDD Cup 99 dataset, which is Intrusion Detection System (IDS) dataset. An IDS dataset source is generated from a bundle of network traffic data, system and application log, etc. Normally the IDS dataset consist a large amount of data including repetitive, irrelevant, misleading features and noise that need to be discarded. Data mining technique is employed to encounter the IDS classification issue, also have been detailed out, comprises of supervised and unsupervised data mining technique, which actively employed in IDS dataset.

ACKNOWLEDGEMENT

This research is fully supported by TATIUC Short Term Grant, 9001-1810, which help in completing the research in viable and efficient.

REFERENCES

1. Internet Users, "Internet Live Stats", 2018.
2. S. Choudhury and A.Bhowal, “Comparative Analysis of Machine Learning Algorithms along with Classifiers for Network Intrusion Detection,” Smart Technol. Manag. Comput. Commun. Control. Energy Mater. IEEE, 2015. https://doi.org/10.1109/ICSTM.2015.7225395
3. S.Vijayarani and M. Sylviaa.S, “Intrusion Detection System – A Study,” Int. J. Secur. Priv. Trus t Manag., vol. 4, no. 1, 2014.

4. M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A Detailed Analysis of the KDD CUP 99 Data Set," pp. 1–6, 2009.
5. V. Rampure and A. Tiwari, "A Rough Set Based Feature Selection on KDD CUP 99 Data Set," vol. 8, no. 1, pp. 149–156, 2015. https://doi.org/10.14257/ijdta.2015.8.1.16
6. K. Maheshwar and D. Singh, "A Review of Data Mining based Intrusion Detection Techniques," Int. J. Appl. or Innov. Eng. Manag. (IJ AI E M), vol. 2, no. 2,2013.
7. I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine Learning and Data Mining Methods in Diabetes Research," Comput. Struct. Biotechnol. J. 15, pp. 104–116, 2017.
8. K. Vadim, "Overview of different approaches to solving problems of data mining," Procedia Comput. Sci., vol. 123, pp. 234–239, 2018.
9. P. Ahmad, S. Qamar, and S. Q. A. Rizvi, "Techniques of Data Mining In Healthcare: A Review," Int. J. Comput. Appl., vol. 120, no. 15, 2015. https://doi.org/10.5120/21307-4126
10. Y. Hamid, M. Sugumaran, and L. Journaux, "Machine Learning Techniques for Intrusion Detection: A Comparative Analysis," ACM, 2016.
11. S. A. Repalle, V. R. Kolluru, and 2, "Intrusion Detection System using AI and Machine Learning Algorithm," International Res. J. Eng. Technol., vol. 4, no. 12, 2017.
12. M. and Stampar, "Artificial Intelligence in Network Intrusion Detection," in 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2015. https://doi.org/10.1109/MIPRO.2015.7160479
13. M. Gera and S. Goel, "Data Mining - Techniques, Methods and Algorithms : A Review on Tools and their Validity," International J. Comput. Appl., vol. 113, no. 18, 2015. https://doi.org/10.5120/19926-2042