THINK INDIA JOURNAL

ISSN:0971-1260

Vol-22-Issue-39-December-2019

Generating Data for Insider Threat Detection for Cybersecurity

Ujwal Sav[1] and Ganesh Magar[2]

[1]Vidyalankar School of Information Technology, Wadala, Mumbai

[2]P.G. Department of Computer Science, S.N.D.T. Women's University, Mumbai

ABSTRACT

Research work needs data. Data collection, data analysis, data cleaning, data feature extractions, data processing, and results, all these processes based on data. Therefore, data is very important for research work and accuracy of results. Insider threat detection based on the anomalous behavior of users in the workplace. Insiders have privileges to access the data legally. If there are disgruntled, careless, and malicious users in an organization, then the data is on risk. This risk factor can remove by identifying the anomalous behavior of the users. For the identification of malicious users, one has to monitor log data to detect abnormal behavior. The violation of the privacy of individuals in an organization and disclose confidential data are not allowing the researcher to collect the primary data for the study. This research paper proposed how to generate data for insider threat detection based on the anomalous behavior of users.

Keywords: Anomalous behavior, Insider Threat, Cyber Security, Data, Machine learning

1 INTRODUCTION

Data is a piece of raw information that is used for processing to generate meaningful data, i.e., final expected outcome. If data is proper and relevant to study and objective, then research gives you fewer errors with more accuracy in the result. Therefore it is essential to generate good data for research. There are two types of data used in research. The first one is primary data, which directly collected from the research domain, and the second one is secondary data, which is collected by other sources for various purposes.

Insiders are the employees who are working in an organization. They have access to all the resources like data, process, network, and system. There are very few are malicious insiders who steal confidential information. There is a need to prevent the data from these insider threats. It observed that processes are unsuccessful due to a lack of data.

Insider threats can be detected by collecting monitoring insider's behavior. It is difficult to collect real data from an organization due to their privacy violation reason. The researcher has other options to use synthetic data. There are the data sources made available by the research centers, government, and other organizations for research. This research paper proposed the study of data sources and data for insider threat detection for cybersecurity.

Paper Outline

Proposed research divided in to mainly five sections. The first section consists of an introduction, second includes related work, third presents the methodology, fourth discussed the result and analysis of data, and the fifth section concludes the research and future work.

2 RELATED WORK

Data collection is the starting task for research. Research needs a proper, reliable dataset. In this section, a literature survey specifically data used by other researchers for insider threat detection research.

The synthetic datasets are useful for a test, which is cost-effective, readily available, and gives quality results. Use of engineered synthetic data improving system quality and reducing program risk in developing record linkage systems [1].

Synthetic data is flexible, economical, and controlled quality data for testing. This data is not original, and the data sets are fully intact, free of privacy restrictions [2].

ISSN:0971-1260

Vol-22-Issue-39-December-2019

UNSW-NB15_3 provided four CSV files of the data records. The names of the CSV files are UNSW_NB15_1.csv, UNSW_NB15_2.csv, UNSW_NB15_3.csv UNSW_NB15_4.csv contains attack and records. It contains 700000 records. It also has an event and email data file for testing [3]. Existing novel methods utilized to generate the features of the UNSWNB15 data set. This data set is available for research purposes, can be accessed from the weblinks [3].

The CERT dataset is available for research in cybersecurity to detect insider threats based on abnormal behavior of the user. The CERT-IT dataset comes from the insider threat center of the CMU (Carnegie Mellon University). For threat detection, TNR (True Negative Rate) is generally used to evaluate the model. [6]. The CERT-IT dataset is from the Insider Threat Center of Carnegie Mellon University. The dataset simulated threat data and large amounts of standard data for five scenarios implemented by malicious insiders, which involve user behavior data from multiple dimensions [20].

3 METHODOLOGY

The survey methodology used for this research paper. There are 30 research papers surveys performed to study different types of data and analyze the dataset used for insider threat detection. It found that the dataset used by the researcher is available in various formats on the website. Therefore, data collected from respective sites that have open access to research.

4 RESULTS AND ANALYSIS

The dataset review result and analysis presented in this section. There are primary and secondary datasets are used for the experiment. Insider threat data is the confidential, therefore existing dataset used for the study.

4.1 Existing Dataset

There are several datasets available for the research and study for insider threat detection. There are six primary datasets selected for the review and research. The datasets are downloaded and extracted and find out the features, records, and data.

4.0.1 UNSW-NB15 DATA SET

Dataset statistics of network traffic: The total number of flows is 987,627 in 16 hours. Attacks identified is 22,215. There are overall 49 features of the UNSW-NB15 datasets elaborated.

4.0.2 KDDCup99 Data Set

The KDDCup99 dataset consists of 5 million connections. The group of Lincoln laboratories at MIT university had generated KDDCup99 Data Set DARPA98 [4]. Upgrading DARAP98 network data features comprehensiveness, utilizing the same environment (U.S. Air Force LAN), the simulation ended with 41 features for each connection along with the class label using Bro-IDS tool. The new version of DARAP98 is known as KDDCUP99. This dataset used primarily for the Intrusion detection system i.e., for outsider threats.

4.0.3 NSLKDD data set

NSLKDD dataset preprocessed. Dataset cleaned by removing duplicate records and selected different documents. It does not show a new low footprint attack. It is with 42 features.

MAWI dataset combines different and independent detectors. Anomaly detectors to improve over time the quality and variety of labels [8]. The data preprocess to normalize the input and the target values. Second, the normalized input difference and normalized target difference computed; these two values combine using a weighting function to estimate the learning conflict between two specific cases in the dataset. Machine learning algorithms used to improve the performance of the model of a refrigeration system in a real-world application.[9].

NSL-KDD data set as the research object analyses the latest progress and existing problems in the field of intrusion detection technology, and proposes an adaptive ensemble learning model. It uses NSL-KDD Test+to to verify our approach, the accuracy of the MultiTree algorithm is 84.2%, while the final accuracy 85.2% [6].

4.0.4 NIMS is Network Information Management Security

ISSN:0971-1260

Vol-22-Issue-39-December-2019

NetMate organization is employed to generate flows and compute feature values on the data sets. Packets collected internally at a research test-bed network. Data simulate six SSH services Shell login, X11, Local tunneling, Remote tunneling, SCP, and SFTP.

4.0.5 NETRESEC

NetMate organization is employed to generate flows and compute feature values on insider threart data. Netresec is an independent software vendor with focus on the network security field. We specialize in software for network forensics and analysis of network traffic [32].

4.0.6 CERT Dataset

CERT dataset is existing dataset. The detail information of dataset is as follows.

Table-1: Dataset and Data Features

Dataset files	Features
Logon	id, date, user, pc, activity (Logon/Logoff)
Device	id, date, user, pc, activity (connect/disconnect)
LDAP	employee_name, user_id, email, domain, role
HTTP	id, date, user, pc, url
Email	id, date, to, from
Psychometric	employee_name, user_id, O, C, E, A, N

Insider Threat dataset is useful to identify the user's anomalous behavior on the network. There are six dataset files. These files contain logon.csv, Device, LDAP, HTTP, Email, and Psychometric dataset. These files shows the users anomalous behavior. Device is connected or not connected. Employee login/logoff the system.

4.3 Data Quality Enhancement

The collection of quality data is a challenge for researchers. There is a number of stages to enhance the collected data like transform data into required compatibility format, data cleaning, and data normalize and analysis. All these stages useful to improve the data quality of the dataset. It needs to design proper interface, database, data integrity to improve the data quality so that errors will be minimized during data entry.

Training data and Test data

A researcher needs data to process and to find out the outcome. In machine learning, data is used for training the model. In machine learning, there are so many available models that are useful to find out the anomaly in user behavior. A researcher has divided the data into two parts; one is training the data and other parts to test the data using machine learning supervised unsupervised and semi-supervised algorithms. The results and accuracy of the result are recorded and find out the optimized model. Therefore, it is necessary to collect the appropriate dataset for the research.

Data analysis outcome is that the CERT (Computer Emergency Response Team-Insider Threat (CERT-IT) is used by most of the researchers to detect insider threats by using various machine learning and deep learning algorithms.

ISSN:0971-1260
Vol-22-Issue-39-December-2019

5 CONCLUSION AND FUTURE WORK

The insider threat detection process is itself a very critical and difficult process. Original primary data collection is not possible due to violation of privacy, and also there is no permission to disclose confidential data of the organization. This proposed research studied the challenges of primary data collection for insider threat detection data and studied the existing available secondary data used by the researcher. It also analyzed the data and found our the relevancy of the dataset. In future research, generated data will go under the various stages of data preprocessing. Data cleaning, data standardized, data normalized, data feature extraction, and then use it for modeling the insider threat detection based on anomalous behavior using machine and deep learning algorithm.

REFERENCES

1. K. B. Paxton and T. Hager, "Use of synthetic data in testing administrative records systems," inProc. of theFederal Committee on Statistical Methodology (FCSM), Washington, D.C., USA, January 2012.
2. https://resources.sei.cmu.edu/library/asset-view.cfm?assetid=508099
3. N. Moustafa,, S. Jill," UNSW-NB15: a comprehensive data set for network intrusion .... https://ieeexplore.ieee.org/document/7348942/" Military Communications and Information Systems Conference (MilCIS), 2015. IEEE, 2015.
4. R. Fontugne, P. Borgnat, P. Abry, K. Fukuda. "MAWILab: Combining diverse anomaly detectors for automated anomaly labeling and performance benchmarking" ACM CoNEXT 2010. Philadelphia, PA. December 2010.
5. S. Ledesma, M. Ibarra-Manzano, E. Cabal-Yepez, D. Almanza-Ojeda, and J. Avina-Cervantes, "Analysis of Data Sets With Learning Conflicts for Machine Learning," in IEEE Access, vol. 6, pp. 45062-45070, 2018.
6. X. Gao, C. Shan, C. Hu, Z. Niu, and Z. Liu, "An Adaptive Ensemble Machine Learning Model for Intrusion Detection," in IEEE Access, vol. 7, pp. 82512-82521, 2019.
7. DARPA98.Available on:http://www.ll.mit.edu/mission/communications/cyber/CSTcorpora/ideval/data/, 1998.
8. KDDCup1999.Available-on:http://kdd.ics.uci.edu/databases/kddcup99/KDDCUP99.html, 2007.
9. NSLKDD. Available-on http://nsl.cs.unb.ca/NSLKDD/, 2009
10. McHugh, John, "Testing intrusion detection systems: a critique of the 1998 and 1999 DARPA intrusion detection system evaluations as performed by Lincoln Laboratory". ACM transactions on Information and system Security, 3, 2000, p 262-294.
11. Alshammari, R.; Zincir-Heywood, A.N., "A flow-based approach for SSH traffic detection," Systems, Man and Cybernetics, 2007. ISIC. IEEE International Conference on, vol., no., pp. 296-301, 7-10 Oct. 2007.
12. Alshammari, Riyad; Zincir-Heywood, A. Nur, "Investigating Two Different Approaches for Encrypted Traffic Classification," PST '08. Sixth Annual Conference on Privacy, Security and Trust, 2008, vol., no., pp.156-166, 1-3 Oct. 2008.
13. Lindauer, B.; Glasser, J.; Rosen, M.; Wallnau, K. C.; and ExactData, L. 2014. Generating test data for insider threat detectors. Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications 5(2):80–94.
14. J. Glasser and B. Lindauer, "Bridging the Gap: A Pragmatic Approach to Generating Insider Threat Data," 2013 IEEE Security and Privacy Workshops, San Francisco, CA, 2013, pp. 98-104.
15. Zhu, C., Gao, D.: Influence of data preprocessing. J. Comput. Sci. Eng. 10(2), 51–57 (2016)
16. Sapna Dev, Dr. Arvind Kalia, Study of Data Cleaning & Comparison of Data Cleaning Tools, IJCMSC, Vol. 4, Issue. 3, pp. 360–370, (2015).

ISSN:0971-1260

Vol-22-Issue-39-December-2019

17. Lakshmi S.: An overview study on data cleaning, its types, and its methods for data mining International Journal of Pure and Applied Mathematics 119(12):16837-16847 (2018).
18. X. Huang, Y. Lu, D. Li, and M. Ma, "A Novel Mechanism for Fast Detection of Transformed Data Leakage," in IEEE Access, vol. 6, pp. 35926-35936, 2018. doi: 10.1109/ACCESS.2018.2851228
19. K. B. Paxton and T. Hager, "Use of synthetic data in testing administrative records systems," inProc. of theFederal Committee on Statistical Methodology (FCSM), Washington, D.C., USA, January 2012.
20. Zhang JG, Chen Y, Ju AK, "Insider threat detection of adaptive optimization DBN for behavior logs," Turkish Journal of Electrical Engineering & Computer Sciences, vol. 26, pp. 792-802, April 2018.
21. https://resources.sei.cmu.edu/library/
22. https://web.cs.dal.ca/~riyad/Site/download.html
23. http://www.nlanr.net/
24. https://mawi.wide.ad.jp/mawi/
25. https://www.unb.ca/cic/datasets/nsl.html
26. https://networkdata.ics.uci.edu/resources.php
27. http://www.netresec.com/?page=PcapFiles
28. ftp://download.iwlab.foi.se/dataset/smia2011/Network_traffic/
29. https://digitalcorpora.org/corpora/scenarios/m57-patents-scenario
30. https://www.unsw.adfa.edu.au/australian-centre-for-cyber-security/cybersecurity/ADFA-NB15-Datasets/
31. http://www.fukuda-lab.org/mawilab/v1.1/index.html
32. http://www.netresec.com/