TEST Engineering & Management

March - April 2020
ISSN: 0193-4120 Page No. 2188 - 2195

Data Preprocessing for Anomaly Detection

Ujwala Sav
P. G. Dept. of Computer Science, S.N.D.T. Women's University, Mumbai, India. ujwalasav@gmail.com

Ganesh Magar
P. G. Dept. of Computer Science S.N.D.T. Women's University, Mumbai, India.
gmmagar@gmail.com

Article Info
Volume 83
Page Number: 2188 - 2195
Publication Issue:
March - April 2020

Article History
Article Received*: 24 July 2019*
Revised*: 12 September 2019*
Accepted*: 15 February 2020*
Publication*: 18 March 2020*

Abstract
The security of data is challenging in the business sector due to its availability in cyberspace. Our data is most valuable, and it is the asset of an organization. Insider threats can be detected based on the anomalous behavior of inside users. There is a need to divide the data into two parts normal data and abnormal data. Therefore, it is required to find out the specific features based on which the researcher can train the dataset, perform analysis, and conclude that this converted into potential cyberattacks. A cyberattack may leak or damage the data, data theft, data sharing with the externals. These incidences may cause a considerable loss, spoil the image, or creditability; it may close the organization forever. This research paper proposed the data preprocessing process used for insider threat detection, which based on user behavior. It includes a survey of existing data sources, data quality, selection of the datasets for insider threats detection, data cleaning, feature extraction, and check data relevancy for further implementation. Data preprocessing is useful for the research to get accuracy and consistency in the result during implementation.

Keywords – *Anomalous behavior, Insider Threat, Cyber Security, Data preparation, Machine learning, Data preprocessing*

I. INTRODUCTION

Data selection is the most crucial pre-processing step of the research. In this research, due to the confidentiality and privacy rights of employees, organizations are unable to provide primary data for the research study. Therefore, it is required to work on existing data, i.e., secondary dataset. There are so many data sources are available for review and use. Universities, research centres, private organizations have made available data for the research study. This data has free open access. User can download the data from the respective site and use to process research work. In this research, the researcher has selected ten different types of data sources and studied its data relevancy for insider threat detection. Every kind of dataset is unique with various types of characteristics. Data preprocessing for insider threat detection. Datasets review is necessary for finalizing the dataset selection for actual algorithm implementation. After data review, select the appropriate and relevant dataset for the implementation of an algorithm in the research domain. This dataset helps you to get the expected outcome. This processed dataset outcome will further process in machine learning and deep learning algorithm.

Research paper content arranged in 5 sections. Section 1 is an introduction, which presents a brief description of research work and its need. A literature review included in Section 2. Section 3 is of research methods and data processing. Section 4 consists of system architecture for data preprocessing, and section 5 provides an analysis of results, and section 6 includes the conclusion and future work.

II. RELATED WORK

Data preprocessing is the beginning stage for research. The research result is dependent on the data used for processing. The raw data is necessary to preprocess before actual

March - April 2020
ISSN: 0193-4120 Page No. 2188 - 2195

implementation. This section presents a similar study was done in insider threat detection and dataset used. While the literature review, it is observed that data is available in various types and formats. Secure Shell (SSH) and Skype used for encrypted traffic, which classified traffic. In this, classifiers are tutored on data from one network but tested on data from an entirely different system. Here five learning algorithms - RIPPER, SVM, AdaBoost, Naïve Bayesian, and C4.5 - are assessed applying flow-based features, where IP addresses, source/destination ports, and payload data are not used. Findings indicate the C4.5 created approach works well as compare to other algorithms on the detection of both SSH and Skype traffic on totally separate networks. [1],[2].

The researcher aims to find out anomalies in the MAWI (Measurement and Analysis on the WIDE Internet) archive implementing a different procedure that merges separate and individual detectors. It worked to evaluate the alarms created by sensors, although it works at separate traffic granularities. Anomaly detectors to increase over time the condition and array of labels [3]. It is necessary to advancing the state in detecting threats, but it is not easy to obtain suitable data for research. Therefore, the data generator is applied to allow research development [4]. Some of the researchers use synthetic data generation, and it helps to control flexibility. It is also economical as compared to other data collection methods. This method is useful to set the data with the required size, quality, and relative traits characteristics [5]. Not all data preprocessing techniques are valid for any one data set; it proposes a fundamental approach to determine the effectiveness of a data preprocessing method. In the paper, MatPCA, MatFLDA, SVM, NNC, K-means, AHC, INMKMHKS, NNC, KNN, Fisher, and Pseudo Inverse used to decide whether the structure is changed or not [6]. There are two steps to know about the dataset, and step one normalizes the input and output data. The second step computed the difference between input/output values and standardized data. These two values merge employing a weighting act to approximate the learning conflict in the given dataset. This proposed algorithm used to optimize the performance of the system in real-world applications [7].

There are data mining uses clean, excellent, and reliable data. If data is wrong, then it will give you a false result and misguide in public as well as private scale. The sources and formats of the data are different for different processes. Sometimes it is not complete. In this study, an overview of data cleaning approaches, issues, comparison tools, and data quality is provided [8]. Data cleaning is useful for missing values, remove the wrong data, detect anomalies, and remove data inconsistencies. It is necessary to converts the data into an appropriate format for processing. Data cleaning help in reducing the loss of data and present the data in standard form. One can use analysis of histograms, analysis of clustering, and data segmentation [9].

Insider threats are responsible for Data leakage in an organization. Data leakage can be detected by identifying the transformation of a considerable amount of data. The researcher presents this challenge of data leakage in three steps. It tests the sensitivity by using adaptive weighted graphs. This method finds out the data leakage of data transformation [10]. In the area of the Intrusion Detection System, the research NSL-KDD dataset is used for the adaptive ensemble learning model. It works on the NSL-KDD dataset to verify the model, and the MultiTree algorithm accuracy is 84.2% [11]. Existing novel methods utilized to generate the features of the UNSWNB15 data set. These datasets are available on websites, and researchers are allowed to use data to find out the solution and explore the new methods [13].

III. Research Methodology

This section consists of details of the methodology used for data preprocessing. In the

March - April 2020
ISSN: 0193-4120 Page No. 2188 - 2195

above section of the literature survey, various data source references collected. Data is available on respective websites, and they are freely available for research purposes.

A. *Data Preprocessing System*

In the preprocessing, a literature survey is used to find out available data sources. The researcher has collected the data source references from existing research. Data is obtained from various sources in their raw format. Then data is extracted and transformed in readable form csv format. This data is reviewed to check the suitability for the study. A suitable dataset is selected for further processing.

System architecture for data preprocessing for Insider Threat Detection is as given below.

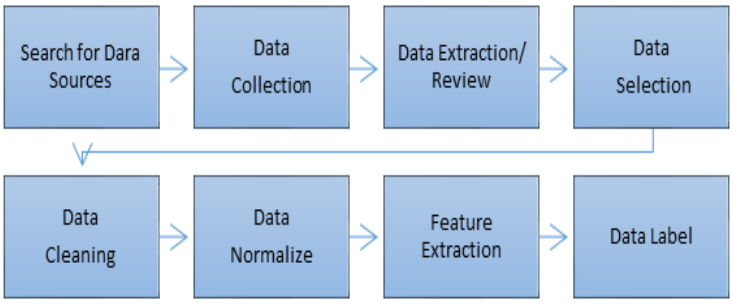

Fig. 1. Data Preprocessing System

B. *Data Preprocessing for Insider Threat Detection*

In the data preprocessing process, data is prepared for the actual experiment, evaluation, or implementation of the algorithm. This preprocessing helps to achieve the objectives of study, accuracy, and consistency in the result. In this research paper, data has preprocessed for Insider threat identification basis on the anomalous behavior of the inside user.

C. *Data Source, Data Collection, and Data Review*

There are many online data sources available on the website. Relevant data is required for the processing of Insider threat identification, which is based on the anomalous behavior of the inside user model. Research is based on primary and secondary data. In this insider threat prediction research, the secondary data collection method is adopted. Because the necessary data for this research is confidential. It is not possible to collect the primary data as a privacy violation of the employee (insider). During the literature review, it observed that previous researchers had used the sample data available on the research website. This data is made possible by university and research organizations, especially for research purposes.

Websites of private and public sectors have checked to find out the existing data. The data collected only from the insider threat detection based on anomalous behavior domain. Data Source and its review is follows in Table 1.

TABLE 1. DATA SOURCES AND DATA REVIEW DETAILS

Data Source	Data Name	Data Format	Tool	Description	Data Review
NetMate	NIMS is Network Information Management Security	ARFF/ CSV Packets (ARFF-Attribute Rich File Format)	Weka	NetMate organization is hired to create flows and calculate feature values on the datasets [13].	Packets internally gathered at an exploration test-bed network. Data suggest six SSH services like SFTP, X11, Remote, Local tunneling, Shell login.

March - April 2020
ISSN: 0193-4120 Page No. 2188 - 2195

NLANR	NLANR National Laboratory for Applied Network Research	TSH (Time Sequenced Header) PCAP Convertible	Weka Note pad	The goal of NLANR is to give technical, engineering, and traffic support of NSF [14].	It is network packets captured for traffic analysis. NLANR is with a high performance and services connections site. HPNSP
Data Source	Data Name	Data Format	Tool	Description	Data Review
UCI .	kaptail.dat It is a social network repository	HTML	notepad	The repository of the network is available with UCI It scientifically helps network study [16].	The dataset is in a matrix format of 39x14 with row labels.
MAWI	MAWI (Measurement and Analysis on the WIDE Internet)	PCAP	Wire shark	The MAWI acting company has held out the measurement of network traffic analysis, WIDE Project [15].	MAWI [15] calculates whether the network acts as per the designed model and realizes anomalous behavior.
Canadian Institute for Cyber Security	NSL-KDD	.ARFF .TXT	Notepad	NSL-KDD gives solution to the problems of KDD99 dataset [17].	In all 42 fields in the dataset, including character type dataset like label, protocol, and flag fields are character types. In the algorithm, we need a numeric data type.
NETRESEC	Smia2011	PCAP	Wire shark	Netresec is working for network security. Its' main specialization in network forensics and analysis of users network data. [18].	It maintains a comprehensive list of publicly available PCAP files.
Numenta	Anomaly Detection	CSV	Notepad	Numenta organization is developing theory, software, and applications based on neocortex principles. Numenta is focused on machine intelligence [19].	Numenta has timestamped, ordered, and single-valued matrices data. Data files consist of notified anomalies.
UNSW-NB15	UNSWNB15	CSV, PCAP, BRO, Argus, and the reports filed	Notepad	The UNSW-NB15 source files can be downloaded from the website [21].	The traffic analysis clarifies the aggregate activities when the time of recreation while producing the UNSW-NB15 information set.
MAWILab v1.1	Anomaly classification	XML CSV	Notepad IE python	This is consists of four separate detectors, which are based on distinct background: They are Gamma distribution, Kullback Leibler, PCA, and the Hough transform [22].	MAWILab annotates traffic anomalies in the MAWI archive with four different labels: anomalous, suspicious, notice, and benign.

March - April 2020
ISSN: 0193-4120 Page No. 2188 - 2195

Carnegie Mellon University	CERT data	Tar.bz files to be converted in,csv	Em-Editor Notepad, Wordpad	Carnegie Mellon University has a CERT Division. This division has collaborated with ExactData, LLC, produced synthetic data for insider threat detection study [23].	There is large dataset generated which is used for the various research study. r1.1 to r6.2 versions are available.

The above data sources have been studied, reviewed, and compared the dataset for anomalous behavior identification of insider threat detection based on the user's irregular detection. CERT data selected for insider threat detection.

In this data preprocessing process, the author has focused on the anomalous behavior of the users who are in an organization and not on the outside users. Therefore malware or network data, payload not considered. Multiple connections derived features are used to compare normal and abnormal traffic, and characteristics are considered from packet headers. This feature helps to know about unusual traffic, which includes packet size, average flow duration, and algorithms like LOF, KNN, Clustering, SVM, mining are used.

IV. Results And Analysis

In section 3, data sources, data collection, data Extraction-Transformation, and data selection covered in table 1. Data cleaning, data normalization, feature extraction, and data labeling, along with the required tools, techniques are evaluated for the selected dataset.

A. *Data Selection*

The data sources and features analyses and CERT dataset selected for the machine and deep learning algorithms implementation in the future study. These are supervised and unsupervised learning algorithms. The proper result required to extract the right features and to label the data. As per the data preprocessing system architecture, to ensure that algorithms were tuned to find anomalous data and not just artificial data, benign events were injected as part of user histories as well.

This dataset is available on this site[25]. The critical data is generating by the Software Engineering Institute of Carnegie Mellon University in partnership with Exact data LLC. Privacy violation is the main issue while working on insider threat detection. Therefore, a synthetic test dataset that provides a dataset of a malicious user. Insider Threat Test Dataset selected for the Modelling of Insider Threat Detection for cybersecurity.

Insider Threat Test dataset consists of employee data. There is logon, device, LDAP, HTTP, email, psychometric data.

B. *Data Cleaning, Data Normalized, Feature Extraction*

Insider Threat data is the data that is required to make ready for the process. The insider threat test data presented in Table 2.

Insider Threat dataset checked and cleaned by removing corrupt, redundant, and inconsistent data from each dataset. In logon.csv, the domain name column is dropped. LDAP username field is dropped as if it is not used at the time of processing. 'Id' field can apply for the identification of the user. As data is enormous, the researcher has re-moved the unwanted, incomplete records from the dataset. Data normalized by managing missing values using statistical calculations. In this research, sklearn used to label and handle string data like logon/logoff to integers value for processing.

March - April 2020
ISSN: 0193-4120 Page No. 2188 - 2195

Table 2 describes the file used to extract the features from the collected data and data preprocessing process.

TABLE 2. DATASET AND DATA FEATURES

Dataset	Description	Features
Logon	1000 Users logon & logoff records. 13 IT admin with global access.	id, activity, date, user, pc (Logon and Logoff)
Device	Thumb drive usage by assigned users	id, activity, date, user, pc, (connect and disconnect)
LDAP	Monthly list of Active Users with details	employee_name, user_id, email, domain, role
HTTP	List of domain names to identify malicious websites.	id, date, user, pc, url
Email	Users to and from list	id, date, to, from
Psychometric	Psychometric list	employee_name, user_id, O, C, E, A, N

C. Data Preprocessing result for logon dataset

Preprocessed data of the logon dataset is consists of the following results. no.of users: 1000, no.of device: 407908, no.of pc: 947, psychometric records size (1000, 7), Check for missing values in each dataset is null, Separated logon and logoff activities and counted no. of activity per user and pc.

D. Graph Analysis: Inter-relationship between users and pcs

The undirected bipartite graph is constructed using 'users' and 'pc' as nodes, where edges representing the relationship between the users and pc(s) and edge weights representing the total number of Logoff events.

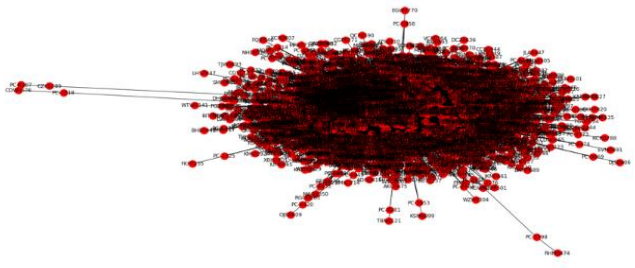

Fig. 2. Undirected bipartite graph

This graph shows the relation and behavior of the user, and it will be helpful to detect insider threats in future research.

V. CONCLUSION

This research paper proposed the data preprocessing for insider threat detection based on the anomalous behavior of the user. It is essential to check the relevancy of data which is used for our experiment and goal of study. Relevant data will go under the preprocessing. The data transformation step is essential if data is not in the required format. Data cleaning is useful for removing data noise, duplicate data and not relevant data from the dataset. There is a need for data normalization to fill the missing values in the dataset by using techniques like removing the record, calculate missing values using statistical techniques like mean, mode, and median. As data is enormous, we use sampling methods for the selection of data from the dataset. Features extracted for the specification of the algorithm. For Data labeling, sklearn used for supervised learning algorithms. The undirected bipartite graph is constructed using 'users' and 'pc' as nodes, where edges representing the relationship between the users and pc(s) and edge weights representing the

March - April 2020
ISSN: 0193-4120 Page No. 2188 - 2195

total number of Logoff events similarly another dataset is processed for anomaly detection. As research is focused on insider threat detection based on the behavior of insider, therefore, network payload is not used.

In future research, preprocessed data will use to detect insider threats by identifying anomalous behavior. Malicious significance will identify and prevent the data cyberattack caused by insiders.

ACKNOWLEDGMENT

The authors are thankful to the Software Engg. Institute of the University of Carnegie Mellon in partnership with Exact data LLC. For providing us users' behavior data for research study purposes.

VI. REFERENCES

1. R. Alshammari and A. N. Zincir-Heywood, "Investigating Two Different Approaches for Encrypted Traffic Classification," PST '08., 2008, vol., no., pp.156-166.
2. R. Fontugne, P. Borgnat, P. Abry, and K. Fukuda, "MAWILab: combining diverse anomaly detectors for automated anomaly labeling and performance benchmarking," Philadelphia, Pennsylvania, 2010, p. 1, doi: 10.1145/1921168.1921179.
3. B. Lindauer, J. Glasser, M. Rosen, and K. Wallnau, "Generating Test Data for Insider Threat Detectors," p. 15.
4. J. Glasser and B. Lindauer, "Bridging the Gap: A Pragmatic Approach to Generating Insider Threat Data," in 2013 IEEE Security and Privacy Workshops, San Francisco, CA, 2013, pp. 98–104, doi: 10.1109/SPW.2013.37.
5. C. Zhu and D. Gao, "Influence of Data Preprocessing," Journal of Computing Science and Engineering, vol. 10, no. 2, p. 7, 2016.
6. S. Ledesma, M.-A. Ibarra-Manzano, E. Cabal-Yepez, D.-L. Almanza-Ojeda, and J.-G. Avina-Cervantes, "Analysis of Data Sets With Learning Conflicts for Machine Learning," IEEE Access, vol. 6, pp. 45062–45070, 2018, doi: 10.1109/ACCESS.2018.2865135.
7. S. Devi and D. A. Kalia, "Study of Data Cleaning & Comparison of Data Cleaning Tools," p. 11, 2015.
8. S Lakshmi, "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).
9. X. Huang, Y. Lu, D. Li, and M. Ma, "A Novel Mechanism for Fast Detection of Transformed Data Leakage," IEEE Access, vol. 6, pp. 35926–35936, 2018, doi: 10.1109/ACCESS.2018.2851228.
10. X. Gao, C. Shan, C. Hu, Z. Niu Z. Liu, "An Adaptive Ensemble Machine Learning Model for Intrusion Detection," IEEE Access, vol. 7, pp. 82512–82521, 2019, doi: 10.1109/ACCESS.2019.2923640.
11. N. Moustafa and J. Slay, "UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)," in 2015 Military Communications and Information Systems Conference (MilCIS), Canberra, Australia, 2015, pp. 1–6, doi: 10.1109/MilCIS.2015.7348942.
12. M. S. Sarma, Y. Srinivas, M. Abhiram, L. Ullala, M. S. Prasanthi, and J. R. Rao, "Insider Threat Detection with Face Recognition and KNN User Classification," in 2017 IEEE International Conference on Cloud Computing in Emerging Markets (CCEM), Bangalore, India, 2017, pp. 39–44, doi: 10.1109/CCEM.2017.16.
13. www.web.cs.dal.ca
14. www.nlanr.net
15. www.mawi.wide.ad.jp
16. www.networkdata.ics.uci.edu
17. www.unb.ca
18. www.netresec.com
19. www.numenta.com
20. ftp://download.iwlab.foi.se
21. www.unsw.adfa.edu.au
22. www.fukuda-lab.org
23. www.resources.sei.cmu.edu
24. www.digitalcorpora.org

Authors Profile

Mrs. Ujwala M. Sav holds a Bachelor's degree in Computer Science, Master's degrees in M.Sc. (CS), M.Ed., MBA(Marketing), M.Phil (IT), along with a Diploma in Business Management, and Diploma in Marketing Management. She is currently working as an Assistant Professor in the Department of Information Technology at

Vidyalankar School of Information Technology, Wadala, Mumbai. She is currently pursuing her Ph. D. from S.N.D.T Women's University, Mumbai. She is a life member of the Computer Society of India. Her research work focused on Cyber Security and Machine Learning. She has more than 15 years of teaching experience in computer science and information technology courses.

Dr. Ganesh M. Magar holds Bachelor's and Master's degrees in Computer Applications and also a Doctorate in Computer Science. He is currently working as Associate Professor Head in P.G. Department of Computer Science and (Ad-hoc) Dean, Faculty of Science and Technology at S.N.D.T. Women's University, Mumbai. He is a member of IEEE, ACM, ISCA, CSI, and many other scientific societies. He has published more than 20 research papers in peer-reviewed reputed international journals and conferences. His thurst research areas include GIS, databases, and image processing. He has more than 16 years of teaching and research experience and three years of Industry Experience.