



Journal of Soft Computing and Artificial Intelligence

Journal homepage: <https://dergipark.org.tr/en/pub/jscai>

International
Open Access 

Volume 04
Issue 01

June, 2023

Letter to Editor

A User and Entity Behavior Analysis for SIEM Systems: Preprocessing of The Computer Emergency and Response Team Dataset

Yasin Görmez¹ , Halil Arslan² , Yunus Emre Işık¹ , İbrahim Ethem Dadaş³ 

¹ Management Information Systems, Faculty of Economics and Administrative Sciences. Sivas Cumhuriyet University, 58050, Sivas, Türkiye

² Computer Engineering, Engineering Faculty. Sivas Cumhuriyet University, 58050, Sivas, Türkiye

³ Principal Consultant. Detay Danışmanlık, 34674, İstanbul, Türkiye

ARTICLE INFO

Article history:

Received December 2, 2022

Revised February 22, 2023

Accepted March 8, 2023

Keywords:

User and Entity Behavior
Analysis

Preprocessing

Classification

CERT

Security Information and
Event Management

ABSTRACT

A lot of work has been done to prevent attacks from external sources and a great deal of success has been achieved. However, studies to detect internal attacks aren't sufficient today. One of the most important studies for the detection of insider attacks is User and Entity Behavior Analysis (UEBA). In this letter, UEBA studies in the literature were reviewed and The Computer Emergency and Response Team Dataset was analyzed (CERT). For this purpose, preprocessing and feature extraction steps were applied on CERT datasets. Several log files combined with respect to user and for each user the number of activities in the specified time interval were obtained. The python code of these preprocessing and feature extraction steps were shared as open source in GitHub platform. In the final phase, future analysis was described and UEBA system planned to be designed was explained.

1. Introduction

Many of the companies have started to carry out their vital business processes with digital systems thanks to developments in technology. They store all important information such as employee, marketing strategy, application info and project documents in the computer systems. These systems have been made very resistant to external attacks by using applications such as intrusion detection system (IDS). However, IDS like applications do not take precautions against insider attacks.

Insider attacks represent the malicious action that performed by the employee of the organizations. These employees not only have authorization to access companies' own systems but also, they have authorization to access customers' systems. Therefore, insider attacks are very dangerous for the companies. Especially large companies suffer from

insider attackers, and they try to control whole of their system using Security Information and Event Management (SIEM) solutions. SIEM solutions are the platform where all activities in a network are recorded and reported in real time. It is possible to make User and Entity Behavior Analysis (UEBA) with real-time data reported on SIEM platforms. UEBA is the best way to capture malicious action from insider employees. A UEBA integrated with SIEM systems is very helpful in early detection of internal attacks that are very dangerous for companies [1].

To date, many studies have been conducted about attack prediction of SIEM and UEBA. Anumol proposed an intrusion prediction system (IPS), which is called open-source security information management (OSSIM), for SIEM framework to perform event analysis using data mining techniques

¹ Corresponding author
e-mail: ysngrmz@gmail.com
DOI: 10.55195/jscai.1213782

[2]. Laue et al. developed an open source SIEM framework within the GLACIER projects. This system does not require any licensing fees and it contains advance algorithms such as artificial intelligence, data collection and anomaly detection. It is also possible to monitor all event using powerful user interface [3]. Asanger and Hutchison applied k-nearest neighbor algorithm on datasets that contain 15 million Windows security events from various perspectives to show performance of unsupervised anomaly detection systems [4]. Goldstein et al. applied six different algorithms on the same dataset using sliding window technique and they showed that the best algorithm is global k-nearest neighbor algorithm [5]. Lukashin et al. proposed novel UEBA architecture that integrated to SIEM platforms, and they achieved the 97.49%, 47.71% and 54.40% accuracy, precision and recall respectively [6]. Tian et al. proposed long short-term memories (LSTM) for UEBA, and they applied multimodal based system on the CERT dataset. Their system achieved 97% accuracy, 98.84% true positive rate and 14.81% false positive rate [7]. Lee and Zincir-Heywood compared decision tree and self-organizing maps on CERT dataset, and they showed that self-organizing maps is better than the decision tree. In addition to this, they performed these analyses on both daily and weekly data and it is seen that weekly data is better than the daily data to detect anomalies [8]. Sharma et al. extracted activity-based feature from CERT dataset, and they proposed long short-term memories based autoencoder model. They achieved 90.17% accuracy, 91.03% recall, 9.84% false positive rate and 90.15% true negative rate with proposed model [9]. Al-Shehari and Alsowail compared logistic regression, decision tree, random forest, k-nearest neighbor and kernel support vector machines and they showed that the best method is logistic regression for CERT dataset. In addition to this they addressed the imbalance problem of this dataset, and they proposed solutions for it [10]. Dosh tried to show effect of feature selection algorithms on CERT dataset by using random forest, Naïve bayes, and nearest neighbor algorithm and he shows that feature selection algorithm does not work properly on CERT dataset [11]. Shashanka et al. collected network traffic data using Niara platform from Nov 2015 to Jan 2016 and 1.315.895.522 raw data were recorded. They proposed a singular value decomposition model, and they analyzed the UEBA from different perspective [12]. Carlsson and Nabhani generated dataset using amount of traffic sent, the timing of sending packets, direction of the traffic and ports. They applied six different machine learning models

on this dataset, and they showed that k-nearest neighbor and random forest were achieved highest accuracy [13].

Although UEBA has been studied in the literature by researchers, the application area has not reached the desired levels. Many events such as web site connections, temporary device usage, computer usage time, used applications or programs, email activities and file activities are needed to improve performance of insider attacker malicious activities prediction systems. However, logs of these activities are not shared publicly due to security and information privacy reasons. This lack of data affects the number of studies very much. The most common dataset for UEBA is The Computer Emergency and Response Team Dataset (CERT) that produced thanks to support of Carnegie Mellon University [14]. This dataset contains many raw data from device, email, file, http, and logon activities. In addition to this, Lightweight Directory Access Protocol (LDAP) information and result of OCEAN test are also included in these datasets. In the literature, most of UEBA have been done on this dataset, however many of them used different kind of feature extraction techniques. CERT dataset contains several files which contains raw event data. It takes a lot of effort to get this data ready for the feature extraction stage. None of the study in the literature has shared this process or dataset that is ready to extract feature. Because of this reason, CERT dataset preprocessed in this letter and these steps were explained. The python code of these process in shared in GitHub that is open-source web-based storage service for development projects [15]. In addition to this, future prospects of our team for UEBA and SIEM integrated anomaly detection systems were mentioned and the system we are considering designing and missing points of CERT datasets were argued.

2. The Computer Emergency and Response Team Dataset

CERT dataset is insider threat test datasets, which generated synthetically by The Computer Emergency and Response Team [16]. This dataset consists of daily activities from 1000 users for 500 days. Five different activity types, which are temporary device connection, incoming-outgoing email, file transfer, visited websites and computer usage, were tracked. Different files were generated for each activity types. In addition to this, monthly LDAP files from December 2009 to May 2011 with detailed personal

information and result of OCEAN personality test [17] for each user were shared. CERT dataset has different version and version 4.2 was used in this letter, because it contains the largest number of malicious actions. Many actions in these datasets are standard actions and the Computer Emergency and Response Team determined three scenarios, which were listed below, to define malicious action. Therefore, they shared another three files which consist of the malicious actions.

- User who did not previously use removable drives or work after hours begins logging in after hours, using a removable drive, and uploading data to wikileaks.org. Leaves the organization shortly thereafter.
- User begins surfing job websites and soliciting employment from a competitor. Before leaving the company, they use a thumb drive (at markedly higher rates than their previous activity) to steal data.
- System administrator becomes disgruntled. Downloads a keylogger and uses a thumb drive to transfer it to his supervisor's machine. The next day, he uses the collected keylogs to log in as his supervisor and send out an alarming mass email, causing panic in the organization. He leaves the organization immediately.

Although version 4.2 of CERT dataset was used in this study, there were more recent version of it. This version was used, because it is more generalized version, and it is the version with highest number of samples belonging to malicious actions. Some of the other versions of CERT datasets contains five scenarios for malicious actions. However, these preprocessing steps can be applied on the other versions of CERT dataset with a few changes.

3. Preprocessing on Cert Dataset

Several feature extraction techniques such as daily, monthly and log-on-log-off based, were used on CERT dataset. It is seen that one of the best performances was achieved with log-on-log-off based. Thus, in this letter, session-based approach was applied during the feature engineering part. Generated action files in CERT dataset are action based and these actions are not separated with respect to user. For example, device file consists of temporary device connection for all users. However malicious action should be detected based on a user. Because of this reason, firstly an empty dictionary

was created where each key of this dictionary represents employee of CERT dataset. Values of each key were consisting of employee's action where each action has four attributes: action id, action date, action personal computer and action type. Device, email, file, http and logon files were processed separately, and actions of user were added to dictionary. Action types are connected or disconnect for device, log-on or log-off for logon, file transfer for file, incoming or outgoing email for email and web browser activities for http files. After this phase, all user activities are sorted by date in themselves. A comma separated values (csv) file was generated for each user where a file consists of all action of user that are ordered by date (the first activity is shown first). As a result of these process a total of 32770222 activities were extracted and 1000 csv file was generated. Table 1 shows the number of events per activity types.

Table 1 Number of Events Per Activity Types

File Name	Activity Type	Number of Event
logon	Logon	470591
	Logoff	384268
device	Connect	203339
	Disconnect	202041
email	Incoming our Outgoing Email	2629979
file	File Transfer to Temporary Device	445581
http	Browser Activity	28434423

As can be seen in the table 1, the most frequent activity is browser activity, and the least frequent activity is device connect or disconnect. In CERT dataset, also there is answer files which contains the list of all malicious events. These files are separated according to the scenarios. 30, 30 and 12 employees committed malicious activity for scenario 1, 2 and 3 respectively. Total number of malicious events are 345 for scenario 1, 6765 for scenario 2 and 213 for scenario 3.

In this letter samples are generated with respect to sessions; thus, it is assumed that an activity sample is all events between a log on and a log off. Thus, for each user all events between a log on and log off were combined. As mentioned in earlier stage, malicious activities were based on event. In a sample, some events may be malicious while the others not. Thus, it is assumed that, if any of events during session was malicious, that session labeled as anomaly. As a result of this process, 384269 samples were generated where 69 of them were anomaly according to

scenario 1, 693 of them were anomaly according to scenario 2, 33 of them were anomaly according to scenario 3 and 383474 of them were normal behavior. As can be seen from the number of samples, the CERT dataset has imbalance problem. This problem can be mitigated by combing all scenarios in one class which is anomaly. However, in anomaly detection systems scenarios were also important, because some scenarios may be more dangerous than the others. As can be seen from the literature review conducted throughout the letter, imbalance problem is not only problem for CERT dataset, but also it is one of the main problems in anomaly detection systems. Thus, the malicious actions were not combined in one class. However, in the future they can be combined easily. In this way, the effect of imbalance problem on the performance of anomaly detection systems also be analyzed using our preprocessed dataset.

Two types of values, which were numerical and action sequence based, were generated for each sample. In action sequence-based values all action during the session were collected in chronological order. In some samples, session duration was too low (for example signed out immediately after signing in) and some sessions had only logon or logoff (Action may not be caught due to some processes such as SIEM connection error). These types of data were not eliminated because it is thought that they can be also used in further process while feature extraction. For example, the combination of sessions also be a malicious event. Some time series model using regression or long short-term memories algorithm can be proposed to make anomaly detection system more complicated.

Numerical values were computed according to the information package of dataset and literature reviews [3], [9], [14], [16]. Firstly, actions are divided into two categories that are during the working hours and during the work of hours. For this purpose, it is assumed that working hours were between the 8:00 a.m. and 7:00 p.m. and also it is assumed that only weekdays are working hours. Using these information, session duration in work on, session duration in work off, number of email during the work on, number of email during the work off, number of file activity during the work on, number of file activity during the work off, number of browser activity during the work on, number of browser activity during the work off, number of temporary device activity during the work on and number of temporary device activity during the work off were computed for each sample. At the end of the this, process action sequence based, and numerical values

are concatenated and team information, role information, openness value, conscientiousness value, extraversion value, agreeableness value and neuroticism value for employee were added to each sample. The python code of this preprocess steps was shared in GitHub platform [15]. Figure 1 summarizes the step of preprocessing. The python codes of the systems and the functions were developed with respect to these steps.

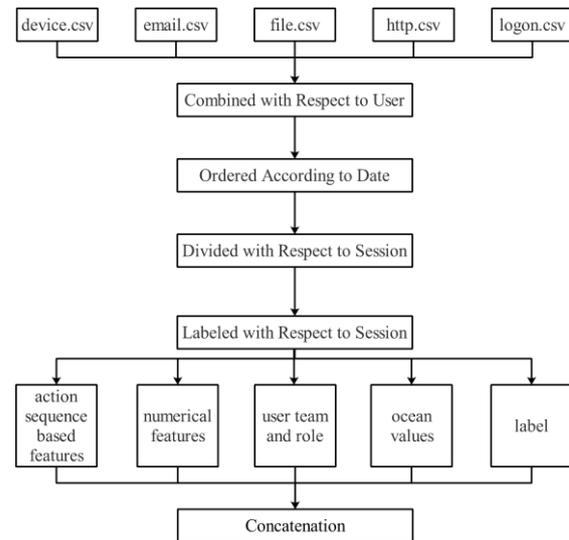


Figure 1 Summary of Preprocessing Steps.

4. Conclusions and Future Prospects

In this letter, importance of SIEM systems and UEBA were mentioned and some studies about anomaly detection on SIEM systems and effect of UEBA to detect insider malicious action were reviewed. Because of the security reason, data collected from SIEM systems are not shared publicly, thus designing an insider attack detection system is hard to research. In the literature, the most common used dataset is CERT [14] for UEBA. There are many studies that developed a machine learning algorithm on CERT dataset, however they did not share their preprocessing steps publicly. In this letter, CERT log files were preprocessed and a file, which is ready to extract feature, generated. It is also possible to use this file without feature extraction to design UEBA system.

In the future work, firstly several machine learning algorithms will be proposed using preprocessed dataset. For this context, multi class and binary class prediction systems will be developed separately. In multi class prediction systems each malicious actions that are generated using three scenarios will be predicted separately. In binary class prediction systems, actions will be labeled as malicious or not malicious. For this purpose, malicious actions

generated using three scenarios will be assigned to insider attack. In addition to this, several feature extraction techniques will be applied to action sequence-based values to extract different kind of features. The effect of features will be analyzed thanks to this stage.

Deep learning techniques are more effective than the traditional machine learning algorithms in many problems including UEBA, especially if the number samples are large [9], [19]– [23]. Thus, instead of using only traditional methods, deep learning approaches will also be proposed. Different kind of layers such as convolutional neural networks, long short-term memories and graph convolutional networks will be used to generate deep models. In this model user, role or department-based approaches will be used. Samples will be grouped with respect to approaches and separate input layers will be generated for each of them. At the final phase all layers will be concatenated for classification layer.

In the final phase, we are planning to develop anomaly detection systems that is integrated to SIEM platforms. One of the most important modules of this system will be detection of insider attackers. As mentioned before, CERT dataset is the most common used dataset to detect insider attackers. It is very comprehensive dataset. Because it generated synthetically, it has some shortcomings. The first shortcoming is number of scenarios are not enough to cover some malicious event for large companies. Anomaly actions in version 4.2 of CERT dataset were produced using three different scenarios, however in real world applications there are many several scenarios. Thus, we are planning to collect new data that consist of more real-world scenarios. For this context, we will collect data from our employees, which is more than 700, using our SIEM platform. The second shortcoming is CERT dataset does not consist of some information from real world applications. The most important one is it does not collect connection information such as remote desktop or secure shell (SSH). In our example, these deficiencies will be identified, and data collected through extensive research. The other shortcoming is because CERT dataset generated synthetically some information in features are not real. For example, many connect of websites in the http request files are generated randomly (They are not real website). However, this information is crucial for anomaly detection, and it is vital to have real data. In addition to this, visited web site depends on location. For these reasons, frequently used websites will be identified and categorized for Türkiye.

It is thought that it can be done in different studies

from the malicious action detection problem using these data. The first important analysis is identifying employees who will be fired using the visited web site information. The other important analysis is revealing work performance using visited website and log-on and log-off information.

References

- [1] P. Slipenchuk and A. Epishkina, "Practical User and Entity Behavior Analytics Methods for Fraud Detection Systems in Online Banking: A Survey," in *Biologically Inspired Cognitive Architectures 2019*, Cham, 2020, pp. 83–93. doi: 10.1007/978-3-030-25719-4_11.
- [2] E. T. Anumol, "Use of Machine Learning Algorithms with SIEM for Attack Prediction," in *Intelligent Computing, Communication and Devices*, New Delhi, 2015, pp. 231–235. doi: 10.1007/978-81-322-2012-1_24.
- [3] T. Laue, T. Klecker, C. Kleiner, and K.-O. Detken, "A SIEM Architecture for Advanced Anomaly Detection," vol. 6, no. 1, p. 17, 2022.
- [4] S. Asanger and A. Hutchison, "Experiences and Challenges in Enhancing Security Information and Event Management Capability Using Unsupervised Anomaly Detection," in *2013 International Conference on Availability, Reliability and Security*, Sep. 2013, pp. 654–661. doi: 10.1109/ARES.2013.86.
- [5] M. Goldstein, S. Asanger, M. Reif, and A. Hutchison, "Enhancing Security Event Management Systems with Unsupervised Anomaly Detection:," in *Proceedings of the 2nd International Conference on Pattern Recognition Applications and Methods*, Barcelona, Spain, 2013, pp. 530–538. doi: 10.5220/0004230105300538.
- [6] A. Lukashin, M. Popov, A. Bolshakov, and Y. Nikolashin, "Scalable Data Processing Approach and Anomaly Detection Method for User and Entity Behavior Analytics Platform," in *Intelligent Distributed Computing XIII*, Cham, 2020, pp. 344–349. doi: 10.1007/978-3-030-32258-8_40.
- [7] Z. Tian, C. Luo, H. Lu, S. Su, Y. Sun, and M. Zhang, "User and Entity Behavior Analysis under Urban Big Data," *ACMIMS Trans. Data Sci.*, vol. 1, no. 3, p. 16:1–16:19, Sep. 2020, doi: 10.1145/3374749.
- [8] D. C. Le and A. N. Zincir-Heywood, "Evaluating Insider Threat Detection Workflow Using Supervised and Unsupervised Learning," in *2018 IEEE Security and Privacy Workshops (SPW)*, May 2018, pp. 270–275. doi: 10.1109/SPW.2018.00043.
- [9] B. Sharma, P. Pokharel, and B. Joshi, "User Behavior Analytics for Anomaly Detection Using LSTM Autoencoder - Insider Threat Detection," in *Proceedings of the 11th International Conference on Advances in Information Technology*, New York, NY, USA, Jul. 2020, pp. 1–9. doi: 10.1145/3406601.3406610.
- [10] T. Al-Shehari and R. A. Alsowail, "An Insider Data Leakage Detection Using One-Hot Encoding, Synthetic

- Minority Oversampling and Machine Learning Techniques,” *Entropy*, vol. 23, no. 10, Art. no. 10, Oct. 2021, doi: 10.3390/e23101258.
- [11] M. Dosh, “Detecting insider threat within institutions using CERT dataset and different ML techniques,” *Period. Eng. Nat. Sci. PEN*, vol. 9, no. 2, Art. no. 2, May 2021, doi: 10.21533/pen.v9i2.1911.
- [12] M. Shashanka, M.-Y. Shen, and J. Wang, “User and entity behavior analytics for enterprise security,” in 2016 IEEE International Conference on Big Data (Big Data), Dec. 2016, pp. 1867–1874. doi: 10.1109/BigData.2016.7840805.
- [13] O. Carlsson and D. Nabhani, “User and Entity Behavior Anomaly Detection using Network Traffic,” p. 52.
- [14] “Insider Threat Test Dataset.” Carnegie Mellon University, Sep. 30, 2020. doi: 10.1184/R1/12841247.v1.
- [15] “Arge-Preprocessing-CERT.” Detaysoft, Oct. 23, 2022. Accessed: Oct. 23, 2022. [Online]. Available: <https://github.com/Detaysoft/Arge-Preprocessing-CERT>
- [16] W. R. Claycomb and A. Nicoll, “Insider Threats to Cloud Computing: Directions for New Research Challenges,” in 2012 IEEE 36th Annual Computer Software and Applications Conference, Jul. 2012, pp. 387–394. doi: 10.1109/COMPSAC.2012.113.
- [17] “Big Five personality traits,” Wikipedia. Oct. 07, 2022. Accessed: Oct. 20, 2022. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Big_Five_personality_traits&oldid=1114671408
- [18] J. Glasser and B. Lindauer, “Bridging the Gap: A Pragmatic Approach to Generating Insider Threat Data,” in 2013 IEEE Security and Privacy Workshops, May 2013, pp. 98–104. doi: 10.1109/SPW.2013.37.
- [19] R. Abdulhammed, M. Faezipour, A. Abuzneid, and A. AbuMallouh, “Deep and Machine Learning Approaches for Anomaly-Based Intrusion Detection of Imbalanced Network Traffic,” *IEEE Sens. Lett.*, vol. 3, no. 1, pp. 1–4, Jan. 2019, doi: 10.1109/LSENS.2018.2879990.
- [20] A. Hassan and A. Mahmood, “Deep Learning approach for sentiment analysis of short texts,” in 2017 3rd International Conference on Control, Automation and Robotics (ICCAR), Apr. 2017, pp. 705–710. doi: 10.1109/ICCAR.2017.7942788.
- [21] Y. Görmez, M. Sabzevar, and Z. Aydın, “IGPRED: Combination of convolutional neural and graph convolutional networks for protein secondary structure prediction,” *Proteins Struct. Funct. Bioinforma.*, vol. 89, no. 10, pp. 1277–1288, 2021, doi: 10.1002/prot.26149.
- [22] X. Hou, T. Arslan, A. Juri, and F. Wang, “Indoor Localization for Bluetooth Low Energy Devices Using Weighted Off-set Triangulation Algorithm,” presented at the Proceedings of the 29th International Technical Meeting of the Satellite Division of The Institute of Navigation (ION GNSS+ 2016), Sep. 2016, pp. 2286–2292. doi: 10.33012/2016.14720.
- [23] Z. Wang, S. Sugaya, and D. P. T. Nguyen, “Salary Prediction using Bidirectional-GRU-CNN Model,” p. 4, 2019.