Unsupervised Anomaly Detection in a Large Corporate Network

Mohammad Ariful Islam
dept. Computer Science
Dalhousie University
Halifax, Canada

Supervisor : Shahrear Iqbal, Ph.D.
Research Officer, Cyber Security
Digital Technologies, National Research Council (NRC) of Canada
Fredericton, NB

I. Abstract

Cyber-security is today a major concern in complex large enterprise networks. A large amount of data is collected from different machines which are usually prone to have anomalous behavior in the dataset. In this project, we have worked with such a large dataset where we are able to detect anomaly using unsupervised anomaly detection techniques. Four different algorithms are evaluated and Isolation forest is selected for final deployment. The algorithm is able to find an anomaly in the dataset specifically for our requirement.

II. Introduction

Nowadays, security solutions is typically deployed in order to protect the network from advanced security threats of both external and internal origin. The solution includes exclusive firewalls, intrusion detection, anti-virus and so on. This solution performs specific tasks on a large number of a dataset which is collected from the events and logs of the system. In order to detect the security breaches, the anomaly in the dataset should be detected first. After that, the detected anomaly cases can be analyzed later for further operations. This task is hard due to the complexity of the dataset in a large corporate network.

For our experiment, we have collected the dataset from a large organization, Los Almos National Laboratory which is one of the prominent network laboratories from the US. The dataset contains logs of events that are generated from the host computers of the enterprise.

III. Dataset preview

In order to help stimulate a larger research effort focused on operational cyber-data, Los Alamos National Laboratory (LANL) has released data set for public use (Kent, 2014, 2016). We have worked with one of the datasets released my LANL entitled, "Unified Host and Network Dataset".

The Dataset is a subset of network flow and computer events from the LANL enterprise network over the course 90 days period. The host (computer) event logs originated from the majority of LANL's computers that run the Microsoft Windows operating system. The network flow data originated from many of the internal core routers within the LANL enterprise network and are derived from router flow records. In our experiment, We have only worked with the host log data for different events. The events are the daily activity represents a specific task for each host.

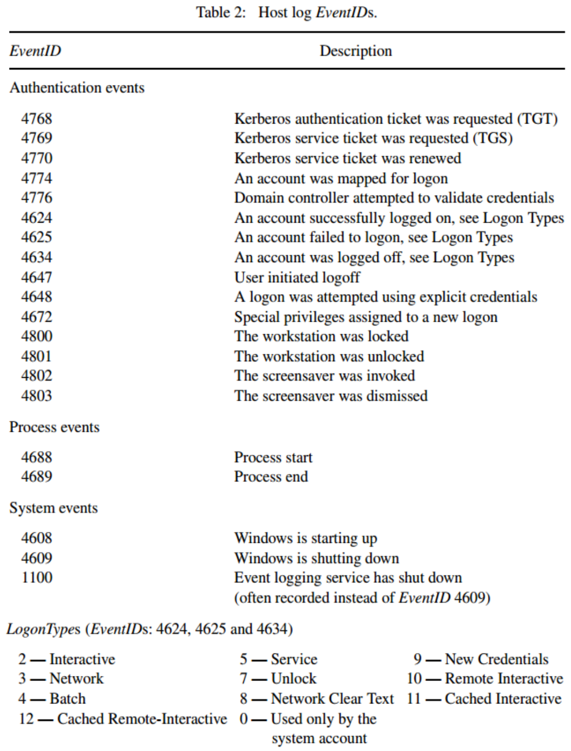


Table 2: Host log *EventIDs*.

EventID	Description
Authentication events	
4768	Kerberos authentication ticket was requested (TGT)
4769	Kerberos service ticket was requested (TGS)
4770	Kerberos service ticket was renewed
4774	An account was mapped for logon
4776	Domain controller attempted to validate credentials
4624	An account successfully logged on, see Logon Types
4625	An account failed to logon, see Logon Types
4634	An account was logged off, see Logon Types
4647	User initiated logoff
4648	A logon was attempted using explicit credentials
4672	Special privileges assigned to a new logon
4800	The workstation was locked
4801	The workstation was unlocked
4802	The screensaver was invoked
4803	The screensaver was dismissed
Process events	
4688	Process start
4689	Process end
System events	
4608	Windows is starting up
4609	Windows is shutting down
1100	Event logging service has shut down (often recorded instead of *EventID* 4609)

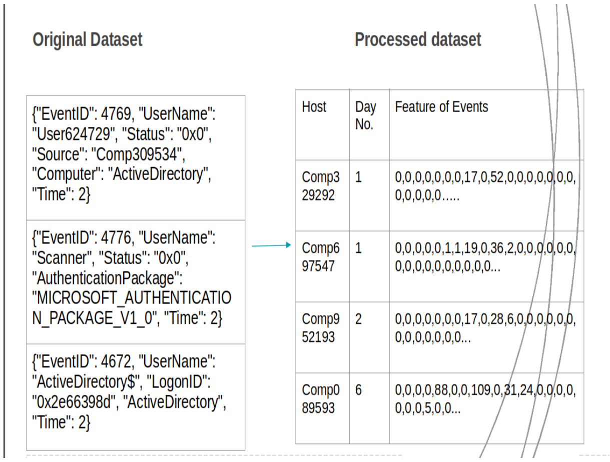
LogonTypes (*EventIDs*: 4624, 4625 and 4634)

2 — Interactive	5 — Service	9 — New Credentials
3 — Network	7 — Unlock	10 — Remote Interactive
4 — Batch	8 — Network Clear Text	11 — Cached Interactive
12 — Cached Remote-Interactive	0 — Used only by the system account	

Fig. 1. Events of host

IV. Dataset preprocessing

An unsupervised anomaly detection algorithm requires an appropriate dataset preprocessing. It's one of the crucial steps in the whole process which does the preparation and transformation of the original dataset. There are various steps in dataset preprocessing that helps to prepare the dataset as per our requirement. Feature extraction, feature construction, feature summation, dimensionality reduction are the part of our process in dataset preprocessing.



Original Dataset

{"EventID": 4769, "UserName": "User624729", "Status": "0x0", "Source": "Comp309534", "Computer": "ActiveDirectory", "Time": 2}

{"EventID": 4776, "UserName": "Scanner", "Status": "0x0", "AuthenticationPackage": "MICROSOFT_AUTHENTICATION_PACKAGE_V1_0", "Time": 2}

{"EventID": 4672, "UserName": "ActiveDirectory$", "LogonID": "0x2e66398d", "ActiveDirectory", "Time": 2}

Processed dataset

Host	Day No.	Feature of Events
Comp329292	1	0,0,0,0,0,0,0,17,0,52,0,0,0,0,0,0,0,0,0,0,0,0,0,.....
Comp697547	1	0,0,0,0,0,1,1,19,0,36,2,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0...
Comp952193	2	0,0,0,0,0,0,0,17,0,28,6,0,0,0,0,0,0,0,0,0,0,0,0...
Comp089593	6	0,0,0,0,88,0,0,109,0,31,24,0,0,0,0,0,0,0,5,0,0...

Fig. 2. Dataset preprocessing

Summation of features : As previously mentioned, the dataset has logs of thousands of hosts over approximately 90 days period of time. Each host stored with 50-dimensional feature events for each day. This produces a considerable amount of data to apply and test the algorithm efficiency. Moreover, the lack of time and resources also makes it harder to work with.

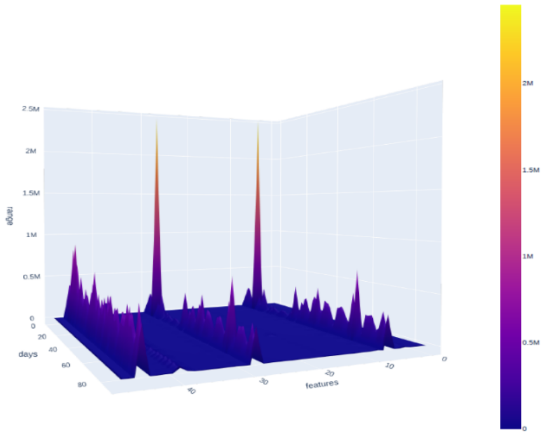
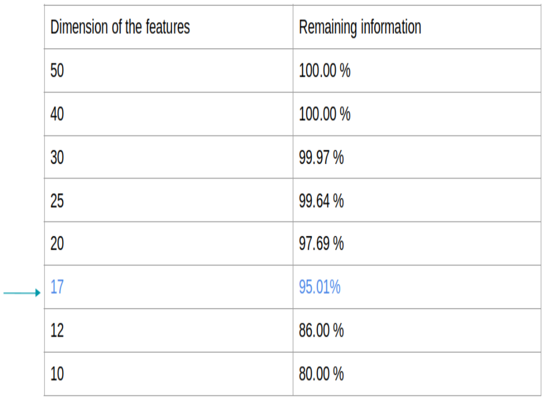


Fig. 3. Feature summation

So as a part of our experiment, we modified our raw dataset to distribute the features according to the day count. And then we have calculated the summation for each event over the number of hosts. It generated a single row dataset with the feature summation for each day over all the hosts in the dataset. The output dataset also helps to visualize the property of the dataset without losing any information.

Principal Component Analysis : We have conducted the Principal Component analysis for our dataset as a part of the dataset preprocessing experiment. Principal component analysis (PCA) is a technique for reducing the dimensionality of such datasets, increasing interpretability but at the same time minimizing information loss. It does so by creating new uncorrelated variables that successively maximize variance. We have applied the PCA in our dataset to reduce the dimensionality of the features. Originally the dataset contains 50 dimensions of features for hosts. Our target is to reduce the dimensions to make the calculation easier for the anomaly detection algorithm to apply. In the case of dimensionality reduction, the PCA algorithm retains maximum information from the original dataset as it does not compromise any data rather it calculates the variance and modifies the dataset in a smaller dimension. So that the loss of information is fairly minimum. We have modified our dataset into 17 components applying PCA operation with the percentages of 95 of the information remains intact. This experiment not only reduces the volume of the dataset but also made our understanding clearer about the original dataset.



Dimension of the features	Remaining information
50	100.00 %
40	100.00 %
30	99.97 %
25	99.64 %
20	97.69 %
17	95.01%
12	86.00 %
10	80.00 %

Fig. 4. Principal Component Analysis

V. Anomaly Detection Algorithm

Anomaly detection is the process of identifying unexpected items or events in datasets, which differ from the norm. In contrast to standard classification tasks, anomaly detection is often applied on unlabeled data, taking only the internal structure of the dataset into account. This challenge is known as unsupervised anomaly detection. In contrast to supervised algorithms which are able to learn which features are important, this is not possible in an unsupervised setting. After the preprocessing of the dataset, we have applied several unsupervised anomaly detection algorithms. Our tested anomaly detection algorithms are : Global K-Nearest Neighbour, One class support vector machine, Local outlier factor and Isolation forest. Its difficult to choose and algorithm immediately and working on it from the beginning. The detection is mainly dependent on the property of the dataset. Each of the algorithm we used, has different parameters to fit the dataset into transformation and calculation of the outlier points.

VI. Isolation Forest

After testing our modified dataset with different algorithms, we have selected the ‘Isolation Forest’ as the primary algorithm to deploy for anomaly detection in the dataset.

Isolation forest is a machine learning algorithm for anomaly detection. It’s an unsupervised learning algorithm that identifies anomaly by isolating outliers in the data.

Isolation Forest is based on the decision tree algorithm. It isolates the outliers by randomly selecting a feature from the given set of features and then randomly selecting a split value between the max and min values of that feature. This random partitioning of features produces shorter paths in trees for the anomalous data points, thus distinguishing them from the rest of the data. The Isolation Forest algorithm is based on the principle that anomalies are observations that are few and different, which should make them easier to identify. Isolation Forest uses an ensemble of Isolation Trees for the given data points to isolate anomalies.

At the part of the successfully implementing the Isolation forest, it enables more tools to analyze our dataset. For example, the tweaking of the contamination into the dataset lets us calculate and find varieties of anomalies as well as the rate of anomaly point in the dataset.

VII. Evaluation

A. Anomalous day

At the initial stage of the evaluation, we are able to find out the specific days that encounters the anomaly.We calculated the decision function to determine the total number of outliers in the dataset. Then we calculated the presence of the outliers for each day.

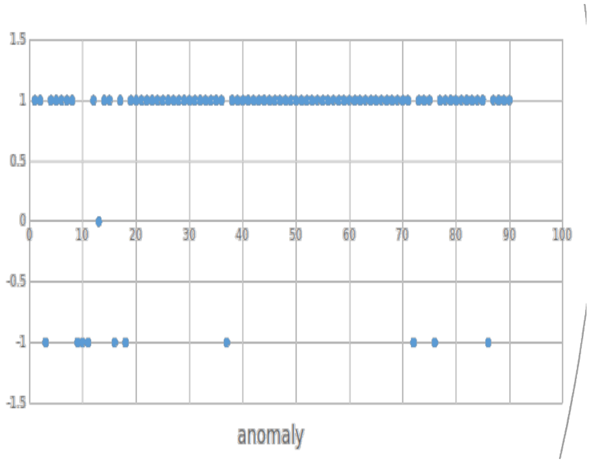


Fig. 5. Anomalous days : 3,9,11,13,16,18,37,72,76,86

B. Successfully Logged on events

Although the algorithm determines the 50 data views specifically, we will describe few interesting results exemplary in this section of work. We find put the anomaly behaviour in the successful logon events by the host computers. There are significant spikes in the dataset view that specifically denotes the anomalous behaviour of the dataset. We have also analyzed the reasoning and found out that two user of the host machine logged on using the local logon promt 47 times within just 71 minutes.It is different behaviour that other hosts on the same day.

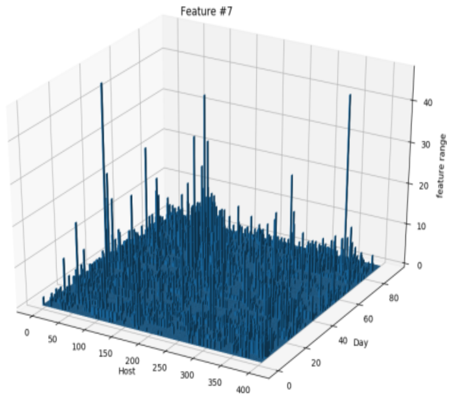


Fig. 6. Successful Logon Events

C. Failed Logged on events

Similarly as previous finding, we have analyzed our algorithm throughput for failed logon events. We have found that two users attempted logging and failed 28 times within only 95 minutes. This behaviour also limited to those days where we found anomaly in our initial evaluation process.

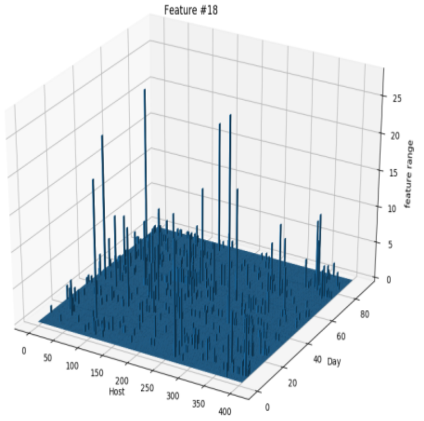
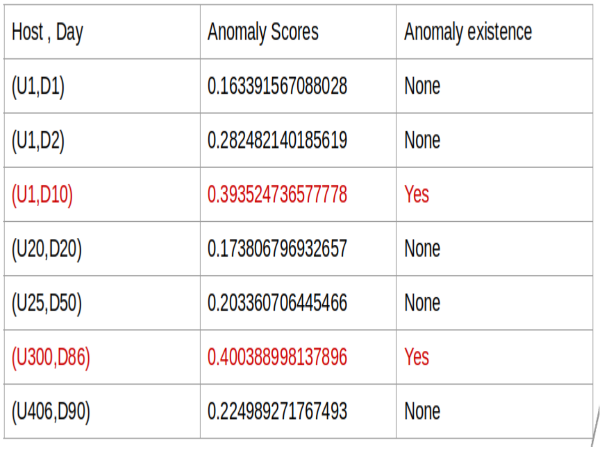


Fig. 7. Failed Logon Events

D. Host and Day specific anomaly detection

After detecting anomaly for specific events in the dataset, we need to find out anomalies for a specific host by day events. We expanded our algorithm to fit the preprocessed dataset to find out the required result. We have calculated the anomaly score for each host and determined the behavior by ranking the anomaly scores. fig-8 shows the chunk of the dataset where (U1, D1) pair denotes the anomaly score and anomaly behavior of the particular for User 1/host 1 and D1 is the number of days.



Host , Day	Anomaly Scores	Anomaly existence
(U1,D1)	0.163391567088028	None
(U1,D2)	0.282482140185619	None
(U1,D10)	0.393524736577778	Yes
(U20,D20)	0.173806796932657	None
(U25,D50)	0.203360706445466	None
(U300,D86)	0.400388998137896	Yes
(U406,D90)	0.224989271767493	None

Fig. 8. Host and day specific anomaly

VIII. Conclusion

We have found a way to implement and determine the anomaly in the large network corporation dataset. We have applied the unsupervised anomaly detection technique to find out the specific anomaly point from the dataset. The data preprocessing is a major part of the experiment which allows us to dig deeper and gain more understanding about the dataset properties. The principal component analysis operation lets us work with the reduced dimensional feature without losing the vital information. We are able to find the reason behind the anomaly behavior of a specific user on a specific day. Although there is plenty of scope to extend our work in the future. The sequence clustering techniques will help us to find the sequence of anomaly in the dataset. The more data viewing will also enable the feature to pinpoint the anomaly source and reasons. The experiment we have done in this large corporation network dataset to detect an anomaly may lead to various future experiments with a large dataset.

References

[1] Goldstein, Markus, et al. ”Enhancing Security Event Management Systems with Unsupervised Anomaly Detection.” ICPRAM. 2013.

[2] Turcotte, Melissa JM, Alexander D. Kent, and Curtis Hash. ”Unified host and network data set.” ArXiv e-prints 1708 (2017).

[3] https://towardsdatascience.com/pca-using-python-scikit-learn-e653f8989e60

[4] https://towardsdatascience.com/outlier-detection-with-isolation-forest-3d190448d45e