**Online Detection of User’s Anomalous Activities using Logs**

Kusumanjali Somisetty

*Abstract*—

Securing the Organization’s Confidential Information is always a concern for any Organization. This Paper implements an Machine Learning approach to monitor the Users activities and determine the anomalous Data.The term anomalous data refers to data that are different from what are expected or normally occur. Detecting anomalies is important in most industries. For example, in network security, anomalous packets or requests can be flagged as errors or potential attacks. In customer security, anomalous online behavior can be used to identify fraud. And in manufacturing and the Internet of Things, anomaly detection is useful for identifying machine failures.

Keywords— Anomalous Detection, Machine Learning, PCA, Random Forest

# Introduction

Information system has become an icon of this era. From enterprises even to restaurants, they use information system to support their business. The internet we use every day is also a huge collection of information systems. As information system starts to manage the more important aspects of human lives such as banking and social needs, the security of information system also becomes increasingly important. Many information systems record their users’ activities on log files. One way to ensure an information system’s security is by analyzing its log files for suspicious or anomalous user activity. Yet the textual nature of log files makes it difficult for administrators to comprehend the meaning of its records and its big size makes log analysis a tedious and time-consuming task . Fortunately, machine learning based approach to detect user activities could serve as a solution for this problem.

# Analysis

## Paper I

*Information System Log Visualization to Monitor Anomalous User Activity Based on Time*

In this paper they tried to find out the key characteristics of data visualization methods which can help detect anomalous user activities recorded in log files. They depicted user activities through dot plot and heatmap visualization methods while leaving the decision whether there are anomalies or not to the administrator’s discretion entirely. The data visualization can be a lot more helpful if they add machine learning feature to the dot plot and heatmap visualization methods. That way, the administrators do not have to find out for themselves whether there are anomalies in a user activity but are instead notified when anomalous pattern is detected. The most appropriate data visualization method combined with machine learning could ultimately lead to the most helpful tool on detecting anomalous user activities recorded on log files. Therefore, adding the machine learning feature to the data visualization is the most suitable step to extend this study further.

## Paper II

*Anomalous User Activity Detection in Enterprise Multi-Source Logs*

They have proposed a practical PCA-based user activity

behavior anomaly detection method. They showed the method works for practical purposes. While that method is for event based anomaly detection, it can be extended to other access based anomaly detection. Example types of access include machines, documents on the network, or applications on the cloud.

## Paper III

Loglizer

Loglizer is a machine learning-based log analysis toolkit for automated anomaly detection.

Logs are imperative in the development and maintenance process of many software systems. They record detailed runtime information during system operation that allows developers and support engineers to monitor their systems and track abnormal behaviors and errors. Loglizer provides a toolkit that implements a number of machine-learning based log analysis techniques for automated anomaly detection.

However, traditional anomaly detection that relies heavily on manual log inspection be- comes impossible due to the sharp increase of log size. To reduce manual effort, automated log analysis and anomaly detection methods have been widely studied in recent years. However, developers are still not aware of the state-of-the- art anomaly detection methods, and often have to re-design a new anomaly detection method by themselves, due to the lack of a comprehensive review and comparison among current methods. In this paper, we fill this gap by providing a detailed review and evaluation of six state-of-the-art anomaly detection methods. We also compare their accuracy and efficiency on two representative production log datasets. Furthermore, we release an open-source toolkit of these anomaly detection methods for easy reuse and further study.

## Paper IV

*Cyber Anomaly Detection Using Graph-node Role-dynamics*

Intrusion detection systems (IDSs) generate valuable knowledge about network security, but an abundance of false alarms and a lack of methods to capture the interdependence among alerts hampers their utility for network defense.

Graph-based anomaly detection is a promising new approach for detecting cyber-attacks from cyber artifacts. In both test scenarios, our method identified anomalies corresponding to the start of the attack. They fused data from several intrusion detection systems into artifact graphs using a novel graph construction based on fields in the triggered alerts, not the network topology. Analyzing the role dynamics in these graphs for simulated Hurricane-Panda- like and Energetic-Bear-like APT datasets, we identified a handful of anomalies, including anomalies that coincide with the start of each attack. Their approach successfully identified simulated attacks through anomalies in IDS alert patterns, and reduced the number of false-positive alerts from thousands to just three false-positive anomalies (in one simulation) and two (in the other simulation). Our results illustrate how graph-node role-dynamics analyses can identify anomalies in IDS alerts, however causal analysis will re- quire further investigation by human analysts.

# dataset

Dataset that we have taken is <http://kdd.ics.uci.edu/databases/kddcup99/task.html>

# IV Approach

After carefully Analyzing all the Papers, we have decided to implement the Random Forest algorithm to detect the anomalies in the Data.

Random forests are a powerful prediction method that uses collections of trees with a random parameter holdout to build models that often outperform individual decision trees. However, the random forest is normally a supervised approach, requiring labeled data.

we introduce a method for turning a supervised model into an unsupervised model for anomaly detection. Unsupervised random forests have a number of advantages over k-means for simple detection. First, they are less sensitive to variable scale. Second, they can fit to "normal" behavior and thus can provide a probability of a data point being anomalous.

##### Acknowledgment

My heartful thanks to professor for Guiding me to do this Project.

##### References

1. Data mining log file streams for the detection of anomalies

<https://pdfs.semanticscholar.org/6f31/428d7b0e9e80680370c4b844cfa2bfb31ba1.pdf>

1. Experience Report: System Log Analysis for Anomaly Detection

http://jmzhu.logpai.com/pub/slhe\_issre2016. K. Elissa, “Title of paper if known,” unpublished.

1. Anomaly Detection Using Persistent Homology

<https://www.computer.org/csdl/proceedings-article/cybersecsym/2016/07942418/12OmNx4yvyt>.

1. Hypergraph-Based Anomaly Detection in Very Large Networks

https://www.computer.org/csdl/proceedingsarticle/cybersecsym/2016/07942418/12OmNx4yvyt

[5] Cyber Anomaly Detection Using Graph-node Role-dynamics

https://arxiv.org/pdf/1812.02848.pdf

**.**