Evaluating unsupervised data mining methods to assess the utility of such approaches for SIEM event analysis.

Autorzy: inż. Marcin Dydo, dr inż. Waldemar Graniszewski, mgr inż. Krzysztof Sosnowski

Afiliacje
[1] marcin.dydo.stud@pw.edu.pl
[2] waldemar.graniszewski@pw.edu.pl
[3] krzysztof.sosnowski@pw.edu.pl

Modern organizations generate vast amounts of data, a significant portion of which consists of system, network, and application logs. In our paper, we are focusing on unsupervised data mining methods for anomaly detection in time-series JSON data. This study compares and explores the utility of several anomaly detection algorithms applied to preprocessed data. Among the various methods, Isolation Forest and probabilistic algorithms like ECOD, have been proven to perform well on multidimensional semi-structured text datasets. Furthermore, we assessed the performance of selected methods using relevant Key Performance Indicators and compared them in diffrent scenarios. Our findings suggest that some of these methods could be adapted to effectively support security analysts working with SIEM systems.

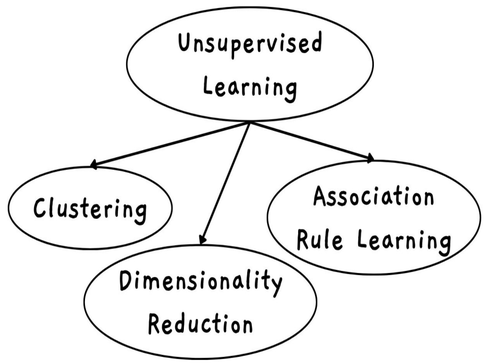
Introduction

SIEM systems are key tools for security teams, integrating data from various sources for monitoring and analysis. These systems often process terabytes of data per day, making manual analysis virtually impossible. In the face of such a huge amount of information, it becomes crucial to extract knowledge efficiently. This paper investigates adoption and practical usage of anomaly detection algorithms within the context of log data analysis. The study is motivated by the need for a tool capable of autonomously identifying statistically significant deviations from normal behavior in datasets derived from security-relevant log messages.

"Existing unsupervised approaches often suffer from high computational cost, complex hyperparameter tuning, and limited interpretability, especially when working with large, high-dimensional datasets." [1]

Unsupervised learning

Due to the dynamic nature of cyber threats and the lack of fixed attack patterns, the use of unsupervised algorithms seems optimal. This type of approach allows detection of anomalies in a changing environment without the need for training data. Association rules are a popular data analysis method for identifying relationships between different events. They are used in data mining to detect patterns of co-occurrence of certain events or event characteristics. Goal of dimenisonality reduction is to eliminate redundant information and focus on key patterns, thus allowing more efficient detection of outliers. Clustering is yet anoter branch closely related to anomaly detection. Many of the basic approaches to it are proximity-based methods. Their key assumption is that normal events cluster in areas of high density. Traditional method involve algorithms such as k-Nearest Neighbors (kNN)[2].



Algorithms used and KPIs

Among the many anomaly detection techniques available, the literature points to the particular effectiveness of the Isolation Forest[4] (IF) algorithm when working with various type of data. Its extension, adapted to the analysis of high-dimensional data, is the Random Projection Isolation Forest (RP-IF), which, by using random projections of the feature space, allows for better scaling and more efficient detection of outliers in complex data sets. Also, ECOD[1] has ben selected, since this algorithm leverages the fact fact that outliers are often the "rare events" that appear in the tails of a distribution.

For each detected outlier, the program calculates several key performance indicators to quantify its statistical significance. These KPIs include: percentage of occurrences; ratio between the observed frequency of the outlier and its expected frequency under assumed uniform distribution; and z-score, which measures how many standard deviations the outlier's occurrence deviates from the mean frequency.

Scope of research

Solution in this paper is to utilize unsupervised algorithms to generate filters for logs, and to evaluate their suitability in powering analysis. The work focuses on using algorithms that operate on unclassified, historical datasets. A lightweight anomaly detection implementation was developed using PyOD[3] and scikit-learn - to identify irregularities in semi-structured SIEM log data.

It successfully identified anomalies in both categorical and numerical fields on benchmark datasets, and demonstrated the ability to detect rare network behaviors in suricata events. Due to modularity of the approach, it supports integrating additional algorithms from pyOD[3] and it is well suited for explorative analysis.

Results and summary

For each of the 2 labeled datasets, two algorithms were compared. As displayed in Table I. - "Acc" column shows accuracy on whole dataset and "TP" is a ratio of true positives. Both algorithms performed well on categorical data which suggests high practical utility in isolating suspicious protocol usage patterns and other similar fields.

Future work includes expanding detection to feature sets for context-rich anomalies or integrating large language models to semantically interpret the results. Overall, these findings demonstrate the feasibility of lightweight, interpretable anomaly detection using open-source tools, offering a robust foundation for production deployment.

Wisdom extraction pyramid (DIKW)

	Iforest Acc	Iforest TP	ECOD Acc	ECOD TP
ADFA – protocol	27,00%	74,00%	28,00%	78,00%
ADFA – service	11,00%	45,00%	6,50%	54,00%
ADFA - sbytes	21,00%	60,00%	21,50%	99,00%
ADFA – dbytes	0,10%	0,10%	9,60%	79,00%
ADFA – attack_cat	3,60%	100,00%	11,50%	100,00%
CSIC – Method	2,00%	100,00%	2,00%	100,00%
CSIC – URL	0,10%	30,00%	0,30%	56,00%
CSIC – host	2,00%	100,00%	2,00%	100,00%

TABLE I
RESULTS OF AUTOMATED TESTS ON LABELED DATASET

Sources

[1] "ECOD: Unsupervised Outlier Detection Using Empirical Cumulative Distribution Functions" - Zheng Li, Yue Zhao, Xiyang Hu, Nicola Botta, Cezar Ionescu, George H. Chen
[2] "A Comparative Study for Outlier Detection Methods in High Dimensional Text Data" - Cheong Hee Park
[3] "Pyod 2: A python library for outlier detection with llm-powered model selection," - Sihan Chen, [..].
[4] "Web log anomaly detection based on isolated forest algorithm. " - Wei Zhang; Lijun Chen