JOURNAL LA MULTIAPP

VOL. 05, ISSUE 06 (824-832), 2024
DOI: 10.37899/journallamultiapp.v5i6.1538

Using Machine Learning to Detect Unauthorized Access in Database's Log Files

Israa Jihad Abed[1]

[1]*University of Al-Anbar, Iraq*

**Corresponding Author: Israa Jihad Abed*
Email: israajehaad@gmail.com

Article Info

Article history:
Received 17 August 2024
Received in revised form 06 September 2024
Accepted 19 September 2024

Keywords:
Anomaly Detection
Log Files
Machine Learning
Unauthorized Access

Abstract

The paper investigates the use of machine learning techniques to detect unauthorized access in database log files. Results show that most algorithms of supervised machine learning performed well in identifying normal cases but struggled to detect anomalies, with the exception of Naïve Bayes and Random Forest which gave mediocre results by identifying one out of twenty anomalies. In the semi-supervised machine learning methods, Local Outlier Factor showed an accuracy of 0.98 in detecting normal cases and 0.7 in detecting anomalies. One Class Support Vector Machine had an accuracy of 0.89 for normal cases and 0.05 for anomalies, while Isolation Forest had an accuracy of 0.98 for normal cases and 0.0 for anomalies. These findings suggest that semi-supervised techniques may be more effective in detecting unauthorized access in database log files.

Introduction

In today's digital age, databases contain a wealth of sensitive information, making them prime sources of unauthorized access by malicious people. Unauthorized access to databases can have serious consequences, from data breaches to lost revenue and reputational damage. Consequently, it is important for organizations to implement strong security measures to detect and prevent unauthorized access. One way to increase database security is to use machine learning techniques to analyze log files for suspicious activity. Machine learning algorithms can help identify patterns and anomalies in log data that may indicate unauthorized access. By real-time detection and unauthorized action, organizations can reduce the risks associated with data breaches and protect their valuable information.

This paper explores the importance of recognizing access to databases and the role of machine learning to improve security measures. We will also discuss existing methods for detecting unauthorized access to databases, highlighting their strengths and limitations. Unauthorized access to databases can have significant consequences for organizations, including financial loss, legal liabilities, and reputational damage. Determining accessibility is essential to protect sensitive information and ensure compliance with data protection laws. By effectively monitoring database activity and detecting suspicious behavior, organizations can prevent data breaches and reduce the risks associated with cyber threats. Several methods have been developed to detect unauthorized access to databases, such as rule-based policies, anomaly detection, and audit log analysis s and sources.

Literature Review

Unauthorized database access is a serious security risk that can result in sensitive data theft, data breaches, and other malicious activity. Devices with learning have become a common

ISSN: 2716-3865 (Print), 2721-1290 (Online)

tool for spotting illegal access to database systems in recent years. To find access to database systems, machine learning methods including clustering, classification, and anomaly detection have been applied. Anomaly detection techniques, such One-Class SVM and Isolation Forest, are frequently used to find unusual values in database log files. These systems can detect anomalous user activity and notify administrators of possible security threats.

To categorize people as permitted or prohibited, classification methods like random forest and support vector machines have also been employed. Through the use of past log data, analysts may train these classifiers to identify whether specific user actions are suspicious or normal. K-means and DBSCAN are two clustering algorithms that can aggregate related functions and identify outliers that might point to unwanted access.

The application of machine learning to determine access to database log files has been the subject of numerous studies. To find abnormalities in user behavior, Liu et al., for instance, created a method based on K-means clustering and random forest. The authors showed that in terms of identifying unapproved channels, their approach performed better than conventional rule-based approaches.

In a different study, Wang et al. found anomalies in database log files using a deep learning model called long- and short-term memory (LSTM). Researchers have discovered that by identifying intricate patterns in applications, LSTM can effectively identify unwanted access.

In another study, Kotenko et al. and the author. Researchers found that the LOF model for unsupervised machine learning, and the LOF and OCSVM models for semi-supervised learning were the most preferred and their modifications in (Saenko et al., 2022; Saenko et al., 2022; Kotenko et al., 2018).

A machine learning algorithm-based approach to identify illegal access to database log files was presented in a paper by Jiang et al. To find suspicious activity in log files, the authors combined anomaly detection with user behavior analysis. Their approach found unauthorized access attempts with a high degree of accuracy.

Li et al. investigated the application of deep learning methods for identifying unauthorized access in database log files in a different study. The authors created a deep neural network model that was able to examine data from log files and spot trends that pointed to illegal access. Their methodology performed better at identifying security breaches than conventional techniques.

Conversely, Park et al. employed graph-based techniques to ascertain who had access to database log files. The authors employed graph algorithms to find anomalous patterns in the log file data and produced graphical representations of the data. Their approach demonstrated a high degree of accuracy in identifying unwanted access attempts.

Using created data that closely resembles real-world network data, researchers conclude that Ant Colony Optimization (ACO) has the maximum accuracy of 98.15%.

Notwithstanding the encouraging outcomes that machine learning can yield from various learning techniques, researchers should be mindful of a number of constraints. First, supervised learning programs may perform worse in the absence of labeled data. It is difficult to collect and classify enough training data, particularly in real-world settings where harmful action is uncommon.

Second, in security-critical applications, the interpretability of machine learning models may be a problem. Deep-learning algorithms are examples of black-box models that may not give administrators with an interpretation of their predictions, making it challenging to determine why a user's behavior has been labeled as forbidden.

ISSN: 2716-3865 (Print), 2721-1290 (Online)

Furthermore, shifting assault tactics and capable adversaries can impact machine learning performance. By imitating typical user behavior or focusing on attacks that are challenging to identify using conventional anomaly detection techniques, attackers may attempt to avoid detection.

Machine learning and anomaly detection in log files

Overview of Database Log Files: Database log files are an integral part of a database management system. Log files record all activity performed on the database, including user logins, queries, updates, and changes. By analyzing database log files, administrators can monitor and monitor user activity, track changes to the database, and detect anything suspicious or unauthorized.

Challenges in Detecting Unauthorized Access: The volume of data created by database systems makes it difficult to determine the accessibility of database log files. Manual log file analysis requires a lot of work and is prone to human error. Furthermore, it's possible that complex and dynamic security threats are difficult to detect using conventional rule-based techniques. Unauthorized access to database log files can be detected automatically and effectively with the use of machine learning.

Benefits of Using Machine Learning: Large volumes of log data can be analyzed by machine learning algorithms, which can then spot trends and abnormalities that point to unauthorized activity. By training machine learning models on historical log data, administrators can identify deviations from normal user behavior and flag suspicious activity in real time Machine learning can also adapt to new security threats and has continued to improve its recognition capabilities over time.

Types of Machine Learning Algorithms: To find unapproved access to database log files, supervised machine learning methods like logistic regression and support vector approaches can be applied. The methods employing these techniques detect normal and aberrant user action patterns and trends based on labeled training data. Through the use of instances of permitted and prohibited access, the model is trained to reliably identify new log entries as either legitimate or questionable. Additionally, unapproved access to database log files can be detected using semi-supervised machine learning approaches like co-training and self-training. Methods that are semi-supervised Enhanced search accuracy can be achieved by using labeled data, as big quantities of unlabeled data may have limited usefulness. This can be especially helpful when data labeling is costly and time-consuming. Organizations can improve their capacity to identify and thwart threats to database security intensity by merging supervised and semi-supervised machine learning techniques (Gowtham & Promod, 2021).

Challenges and Limitations: The use of machine learning to identify unauthorized access to database log files has many advantages, but there are also drawbacks that machine learning models need to overcome in order to achieve pattern recognition. Rosenberg et al. (2021) These include the need for well-constructed, comprehensive models and the requirement for high-quality, labeled training data. They are also susceptible to deceitful and can be attacked by adversaries. Furthermore, practitioners who need to comprehend the logic behind the model's judgments may find it problematic when it comes to how machine learning models are understood.

Machine learning methods used and their metrics

Logistic Regression: One of the most used classification algorithms in supervised machine learning is logistic regression. It is employed to forecast a binary outcome's probability given one or more independent variables. The model creates an S-shaped curve that connects the likelihood of the result to the input features. Due to its ease of use, speed, and interpretability,

ISSN: 2716-3865 (Print), 2721-1290 (Online)

logistic regression is a widely used technique for classification problems in a variety of industries, including marketing, finance, and healthcare (Hastie et al., 2005).

K-Nearest Neighbor (KNN): KNN is a straightforward yet effective technique for regression and classification applications. A data point needs to be allocated to the average value (for regression) or majority class (for classification) of its 'k' nearest neighbors, as defined by a distance metric like the Euclidean distance, for it to function. Since KNN is a non-parametric technique, it is robust and appropriate for a variety of datasets because it makes no assumptions about the underlying data distribution (Endres & Schindelin, 2003).

Random Forest: Multiple decision trees are joined using an ensemble learning system called Random Forest to produce predictions. To increase accuracy and generalization, each tree in the "forest" is trained using a different subset of the training data and characteristics. Random Forest is a well-liked option for machine learning classification and regression applications because of its strong performance and resilience against overfitting (Breiman, 2001).

Support Vector Machine (SVM): SVM is a powerful supervised learning technique that may be used for both classification and regression issues. It operates by determining the best hyperplane to divide classes in the feature space while increasing the margin of separation between them. By mapping the input characteristics into a higher-dimensional space using kernel methods, SVM is able to handle both linear and non-linear separable data. SVM has been widely used in a wide range of applications, such as bioinformatics, image recognition, and text categorization. Its effectiveness with high-dimensional data is widely recognized (Cortes, 1995).

Naïve Bayes: Based on the Bayes theorem, the Naïve Bayes classifier is a simple probabilistic algorithm that depends on the independence of features. Despite its simplicity, Naïve Bayes is a widely used algorithm for text classification and spam filtering due to its efficiency and scalability. It calculates each class's likelihood based on the supplied features and outputs the class with the highest probability. Naïve Bayes is a common option in machine learning applications because it works well in practice and is appropriate for handling huge datasets with high-dimensional features (McCallum et al., 2000).

Neural Network: For supervised machine learning, a neural network is a computer system that mimics the structure of the human brain and is utilized as a classifier. Layers of networked nodes, or neurons, comprise it. These neurons interpret incoming data and forecast outcomes by identifying patterns in the data. When training on labeled data—that is, input data combined with matching output labels—a neural network is employed in supervised learning. The network modifies the connections amongst neurons during training in order to reduce the discrepancy between the labels it predicts and the actual labels. After it has been trained, fresh, unseen data can be fed into the neural network, and its output can be used to classify it. The network is an effective tool for machine learning classification problems because of its capacity to generalize patterns from the training data to new data (Goodfellow, 2016).

One Class Support Vector Machine (OCSVM): An expansion of SVM intended for applications including novelty or anomaly detection is called OCSVM. It gains knowledge of a hyperplane in high-dimensional feature space that divides typical data points from anomalies. OCSVM recognizes outliers as data points that fall outside of the taught boundaries because it was trained on only one type of data—normal occurrences. OCSVM is extensively utilized in many applications for outlier detection, fraud detection, and network security (Schölkopf et al., 2001).

Local Outlier Factor (LOF): An technique called LOF uses density to find outliers in a dataset. It computes the data point's local density in relation to its neighbors and contrasts it with the density of nearby points. Outliers are those points that have a density that is noticeably lower

ISSN: 2716-3865 (Print), 2721-1290 (Online)

than that of their neighbors. LOF is utilized in many applications, including fraud detection, intrusion detection, and data cleansing, because it is good at finding outliers in noisy, high-dimensional datasets (Breunig et al., 2000).

Isolation Forest: Based on an ensemble method, Isolation Forest is an anomaly detection system that employs isolation trees to distinguish outliers. A feature is chosen at random, then data points are separated at different depths until all instances are isolated, creating an isolation tree. Outliers are defined as those cases that show anomaly by requiring fewer splits to isolate. In high-dimensional datasets with skewed or noisy characteristics, Isolation Forest is a quick, scalable, and efficient method of identifying outliers (Liu et al., 2008).

Several parameters are critical in establishing the efficacy of these algorithms when assessing their performance. Machine learning models are often evaluated using metrics such as area under the curve (AUC), recall, precision, accuracy, F1-score, and confusion matrices. Accuracy is defined as the ratio of correctly predicted events to the total number of instances. The model's performance is fairly evaluated by the F1-score, which accounts for both precision and recall. Accuracy measures the proportion of true positive predictions among all positive predictions, whereas recall shows how well a model can identify all pertinent cases. An AUC graph illustrates a model's performance over a range of thresholds. Confusion matrices summarize the predictions made by the model and allow one to see how well the model performs in terms of true positive, true negative, false positive, and false negative predictions (Hand, 2009; Kohavi, 1995; Provost & Fawcett, 1997).

Methods

This study provides a detailed description of a technique that examines user behavior fingerprints to identify unauthorized access to database log files. The primary idea behind this approach is to identify potentially suspicious activity by analyzing how users typically visit database tables. Initially, the open-source program Orange was used to implement a number of supervised machine learning methods, such as Logistic Regression, Neural Networks, K-Nearest Neighbor, Random Forest, Support Vector Machine, and Naïve Bayes.

Based on features taken out of log files, these algorithms were utilized to categorize the behavior of users. Furthermore, outliers and odd patterns in user behavior were found using three semi-supervised machine learning techniques: One Class Support Vector Machine, Local Outlier Factor, and Isolation Forest. This strategy seeks to improve overall security measures and the detection of unwanted access in databases by integrating supervised and semi-supervised algorithms.Both datasets in supervised machine learning methods were divided into 70% for training and 30% for testing with cross-validation equal to five in Orange. As is known in semi-supervised machine learning, the classifiers were trained only on normal cases by Python.

Dataset Description

The data set is an Excel file with the CSV suffix, and it is the features taken from a data log file for 12 working hours from the PostgreSQL database for one of the universities. The number of features is 153, consisting of the user name, which is a numerical identifier, and 11 features are some basic words in the SQL language, and 141 table names in the database. Each row of the data set indicates the number of these features in the SQL statement. The total number of rows in the dataset is 507,210.

We added a final feature, which is the result. The values are 0 if the behavior is authorized, and 1 if the behavior is abnormal. We generated two data sets, each with 20 anomalous behaviors, by changing the value corresponding to the table that the user usually uses.

ISSN: 2716-3865 (Print), 2721-1290 (Online)

After preprocessing the two datasets and removing duplicate rows, each dataset had 34357 rows with normal behavior, and 20 rows with anomaly. Fig. (1) shows the distribution of the normal and anomaly cases in the data set.

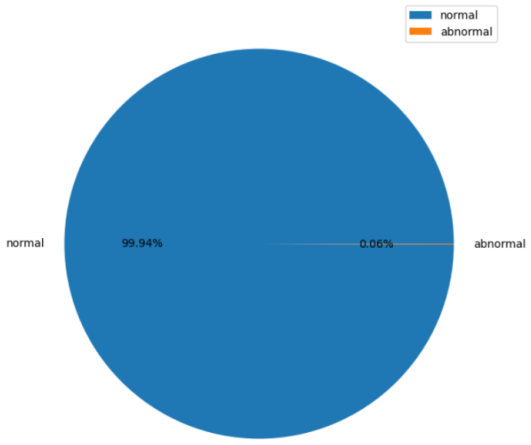

Figure 1. Pie chart distribution of normal and anomaly cases

Results and Discussion

Supervised machine learning methods showed very good results in terms of accuracy and F1 score as shown in Figure 2, but since the number of anomalies is very small compared to normal cases, we checked the results of the confusion matrices of the applied algorithms.

Model	AUC	CA	F1	Prec	Recall
Random Forest	0.574	0.999	0.999	0.999	0.999
Naive Bayes	0.794	0.999	0.999	0.999	0.999
SVM	0.550	0.999	0.999	0.999	0.999
Logistic Regression	0.587	0.999	0.999	0.999	0.999
kNN	0.625	0.999	0.999	0.999	0.999
Neural Network	0.745	0.999	0.999	0.999	0.999

Figure 2. Comparing parameters of supervised machine learning methods

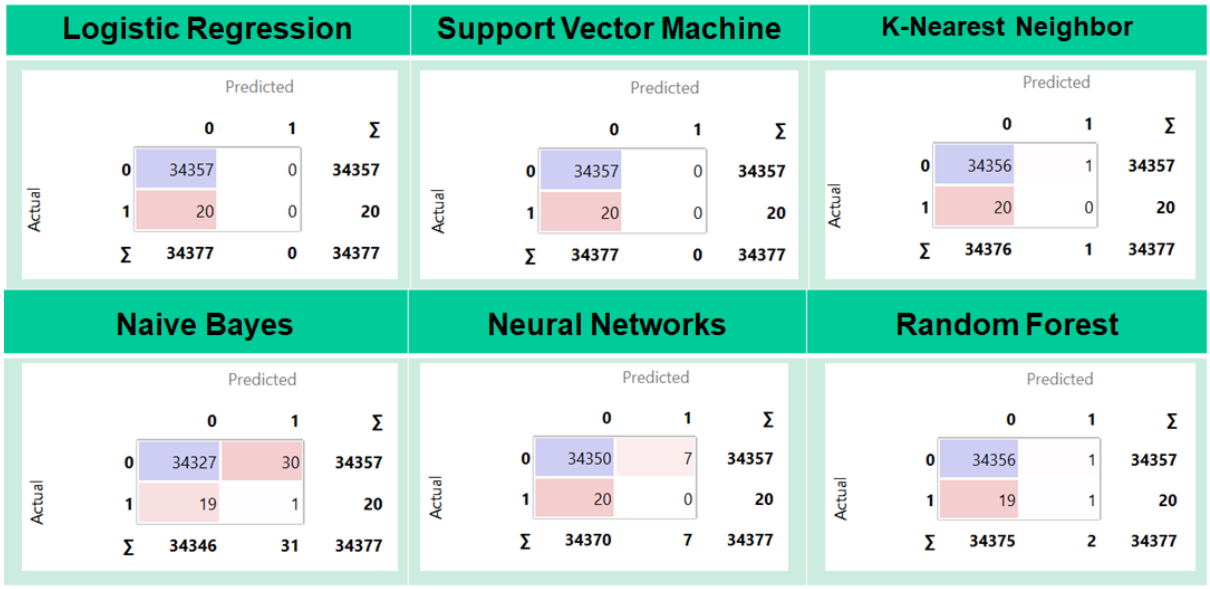

Figure 3. The confusion matrices for the supervised machine learning methods used

ISSN: 2716-3865 (Print), 2721-1290 (Online)

The confusion matrices in turn showed that most algorithms gave very good results for normal cases, but bad results for anomalies, with the exception of Naïve Bayes and Random Forest which gave very mediocre results (they identified one out of twenty anomalies). Figure (3) shows the confusion matrices for the supervised machine learning methods used.

For the first data set, as illustrated in figure (4), the semi-supervised machine learning methods yielded the following results: Local Outlier Factor provided an accuracy of 0.9838 for normal cases and an accuracy of 0.7 for anomaly cases; One Class Support Vector Machine provided an accuracy of 0.8889 for normal cases and an accuracy of 0.05 for anomaly cases; and finally, Isolation Forest provided an accuracy of 0.9775 for normal cases and an accuracy of 0.0 for anomaly cases.

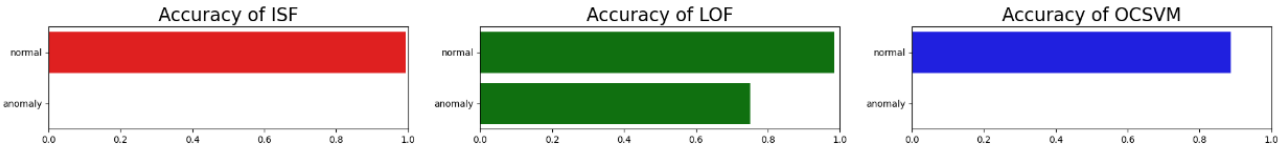

Figure 4. The results of semi-supervised machine learning methods for the first data set

In contrast, in the second data set, as seen in figure (5), the Local Outlier Factor provided an accuracy of 0.9839 for normal cases and an accuracy of 0.75 for anomaly cases. One Class Support Vector Machine provided an accuracy of 0.887 for normal cases and an accuracy of 0.0 for anomaly cases, and Isolation Forest provided an accuracy of 0.9923 for normal cases and an accuracy of 0.0 for anomaly cases.

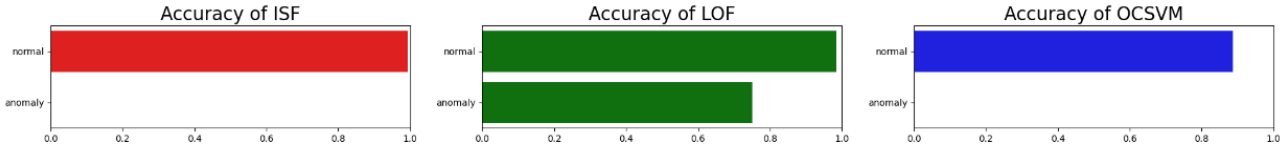

Figure 5. The results of semi-supervised machine learning methods for the second data set

The results of the study on using machine learning to detect unauthorized access in database log files show that supervised machine learning methods performed well in identifying normal cases but struggled to detect anomalies, with Naïve Bayes and Random Forest performing particularly poorly. Conversely, different degrees of accuracy were demonstrated by semi-supervised machine learning techniques as Isolation Forest, One Class Support Vector Machine, and Local Outlier Factor when it came to identifying typical and anomalous cases. Local Outlier Factor had the highest accuracy in detecting normal cases but struggled with anomalies, while One Class Support Vector Machine had a lower accuracy overall. Isolation Forest had high accuracy in detecting normal cases but failed to identify anomalies. These results have important implications for detecting unauthorized access in databases, as they highlight the limitations of certain machine learning algorithms in detecting anomalies.

Future studies in this field might concentrate on enhancing the efficacy of supervised machine learning algorithms for identifying irregularities in database log files, in addition to investigating alternative semi-supervised techniques that might offer superior precision in identifying unapproved entry. The using of deep learning may be a good future work in this area. Additionally, further investigation could be done on different feature selection techniques and data preprocessing methods to improve the overall performance of machine learning models in detecting unauthorized access in databases. Overall, this study opens up new avenues for research in the field of using machine learning for improving database security.

ISSN: 2716-3865 (Print), 2721-1290 (Online)

Conclusion

The domain of machine learning is rapidly developing new algorithms and strategies for detecting unauthorized access to database logs. One potential answer to the expanding problem of database security is the use of machine learning to identify access to database logs. Administrators may more effectively monitor user activity and safeguard sensitive data from security risks by utilizing sophisticated algorithms and data-driven insights. The advantages of using machine learning to database security exceed the hazards, nevertheless certain difficulties and restrictions to take into account. Machine learning has the potential to significantly improve database security and privacy in the digital era with more research and development.

The use of deep learning models, such as neural networks, for more intricate micropattern recognition is one of the future directions. Managers will be able to comprehend and have confidence in their decisions when machine learning models are more transparent and comprehensible. Organizations may exchange threat intelligence and enhance their security protection as a group by utilizing collaborative techniques like integrated learning and international secure computing.

References

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5-32.

Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000, May). LOF: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data* (pp. 93-104). https://doi.org/10.1145/342009.335388

Cortes, C. (1995). Support-Vector Networks. *Machine Learning*.

Endres, D. M., & Schindelin, J. E. (2003). A new metric for probability distributions. *IEEE Transactions on Information theory*, *49*(7), 1858-1860. https://doi.org/10.1109/TIT.2003.813506

Goodfellow, I. (2016). Deep learning.

Gowtham, M., & Pramod, H. B. (2021). Semantic query-featured ensemble learning model for SQL-injection attack detection in IoT-ecosystems. *IEEE Transactions on Reliability*, *71*(2), 1057-1074. https://doi.org/10.1109/TR.2021.3124331

Hand, D. J. (2009). Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Machine learning*, *77*(1), 103-123. https://doi.org/10.1007/s10994-009-5119-5

Hastie, T., Tibshirani, R., Friedman, J., & Franklin, J. (2005). The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, *27*(2), 83-85.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Morgan Kaufman Publishing*.

Kotenko, I., Saenko, I., & Branitskiy, A. (2018). Framework for mobile Internet of Things security monitoring based on big data processing and machine learning. *IEEE Access*, *6*, 72714-72723. https://doi.org/10.1109/ACCESS.2018.2881998

Liu, F. T., Ting, K. M., & Zhou, Z. H. (2008, December). Isolation forest. In *2008 eighth ieee international conference on data mining* (pp. 413-422). IEEE. https://doi.org/10.1109/ICDM.2008.17

ISSN: 2716-3865 (Print), 2721-1290 (Online)

McCallum, A., Nigam, K., & Ungar, L. H. (2000, August). Efficient clustering of high-dimensional data sets with application to reference matching. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 169-178). https://doi.org/10.1145/347090.347123

Provost, F., & Fawcett, T. (1997). Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions In: Proc of the 3rd International Conference on Knowledge Discovery and Data Mining.

Rosenberg, I., Shabtai, A., Elovici, Y., & Rokach, L. (2021). Adversarial machine learning attacks and defense methods in the cyber security domain. *ACM Computing Surveys (CSUR)*, *54*(5), 1-36. https://doi.org/10.1145/3453158

Saenko, I. B., Kotenko, I. V., & Al-Barri, M. H. (2022). The use of artificial neural networks to detect anomalous behavior of users of data processing centers. *Voprosy kiberbezopasnosti*, (2), 48.

Saenko, I. B., Kotenko, I. V., & Al-Barri, M. H. (2022, December). Research on the possibilities of detecting anomalous behavior of data center users using machine learning models. In *Twentieth National Conference on Artificial Intelligence with International Participation, KII-2022 (Moscow* (pp. 232-241).

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., & Williamson, R. C. (2001). Estimating the support of a high-dimensional distribution. *Neural computation*, *13*(7), 1443-1471. https://doi.org/10.1162/089976601750264965

ISSN: 2716-3865 (Print), 2721-1290 (Online)