International Journal of Engineering Research and Modern Education (IJERME)
Impact Factor: 7.018, ISSN (Online): 2455 - 4200
(www.rdmodernresearch.com) Volume 7, Issue 2, 2022

ANOMALY CLUSTERING FOR PROACTIVE PCI DSS COMPLIANCE IN SPARSE DATA

Sai Reddy Mandala*, Ashish Reddy Kumbham & Prasanthi Vallurupalli*****

* Independent Researcher, Sales Force Consultant, United States of America
** Independent Researcher, Sr. Engineer, Application Development and Maintenance, United States of America
*** Independent Researcher, Cyber Security Software Engineer, United States of America

Cite This Article: Sai Reddy Mandala, Ashish Reddy Kumbham & Prasanthi Vallurupalli, "Anomaly Clustering for Proactive PCI DSS Compliance in Sparse Data", International Journal of Engineering Research and Modern Education, Volume 7, Issue 2, Page Number 72-75, 2022.

Abstract:

Anomaly clustering is a critical component of effective compliance with the Payment Card Industry Data Security Standard (PCI DSS) ahead of potential risks, especially in sparse datasets that are characteristic of financial and security analysis. In some cases, the analysis of attributes results in sparse data that is typical for databases when records contain missing or incomplete values, which makes the examination of anomalies challenging. Within this research, concepts like anomalous data detection and clustering particularly in a sparse environment, will be looked at with a view to ensuring compliance without compromising on issues to do with theft and data loss. This is because the different clustering techniques, such as density-based clustering models can make a significant contribution towards the enhancement of the mechanisms of fraud recognition, increase security, and compliance to the existing laws.

Key Words: Anomaly Clustering, Ensuring PCI DSS, Sparse Data Casual Detection, Real-Time Monitoring.

Introduction:

Anomaly clustering is a DM and ML technique of categorizing data points and detecting dissimilar phenomena considered to be unconventional with reference to normality. PCI DSS is an international standard aimed at protecting cardholder information and providing safe payment processing, and this process is particularly facilitated by the usage of the described approach. Any organization that deals with credit card had to ensure that it complies with the PCI DSS requirement because it helps in protecting the sensitive info from being fraud and in the process, it helps in avoiding losses.

However, in sparse data, which can be defined as a matrix where the ratio of missing or incomplete entries is high, anomaly clustering is a problem. Small datasets are typical in financial systems since transaction frequencies are rarely frequent, logs are often only partially completed, or system malfunctions occur. Such challenges tend to compromise the effectiveness of clustering techniques used in identifying fraud or non-compliance cases in good time. In this study, our task is to concentrate on the usage of contemporary approaches to anomaly clustering, e.g., DBSCAN, for improving PCI DSS compliance in situations where sparse data are presented. When organizations pay attention to data sparsity and perform proper selections of clustering techniques, it will be easy to identify areas of weakness, keep away fraudsters and conduct their business within the legal requirements of the country of operation.

Simulation Report:

Since anomaly clustering has been previously applied to a large-scale credit card transaction dataset (Protić, 2018), we simulated another experiment to test its usability for PCI DSS compliance in a sparse environment. The collected dataset consisted of about 10,000 records of transactions, and 4,000 of these records were found to have missing or incomplete values in 40% of the fields. This high level of sparsity captured real-world issues that were faced in financial transaction monitoring systems where there were incomplete records arising from logging of transaction irregularities or customer interaction patterns.

The first process was preprocessing, which is very important especially when preparing data for analysis in the simulation. Since the data set was incomplete; imputation using K-nearest neighbors (KNN) imputation was used since it is among the most common approaches when dealing with sparsity (Ai et al., 2018). This step made certain that the data was sufficiently exhausted. for clustering to be done to the level of the clusters of interest, to avoid any imposition of artificial patterns. After preprocessing, the DBSCAN algorithm was used for clustering as it provides high performance in presence of noise and irregular data density. Compared to the other clustering techniques like K-means, DBSCAN does not depend on the prior identification of the number of clusters that would be advantageous in the identification of anomalies in sparse environments (Lan et al., 2018)

Indeed, the DBSCAN algorithm achieved clustering of similar transactions as well as the identification of outliers that possessed a level of dissimilarity not typical of normal transaction activity. For example,

anomalous behavior included transactions that occurred asynchronously, for example, when making a large purchase involving a large sum of money from an unfamiliar source. The quality of the clustering was investigated using silhouette scores, and the adjusted Rand index, which are globally accepted metrics. On the results, it was observed that the DBSCAN algorithm had a detection rate of 92% with a false positive rate of 4%. On the other hand, K-means clustering provided only a visualization of methods that recognized a detection rate with only 75% in the case of a sparse data set. Accordingly, These investigations confirmed the viability of the DBSCAN method for the identification of irregularities in sparse datasets and established a clear understanding of efficient preprocessing.

Real-Time Scenarios:

Anomaly clustering once again helps in several scenarios. that require real-time PCI DSS compliance, especially with sparse data. One specific area is the use of external and novel features for fraud detection in retail payment systems. Grocers work with thousands of orders daily, and the Data can contain typing mistakes or missing fields because of customer mistakes. or network breakdowns. For instance, a retail payment processor identified many large-value payments that originated from IP addresses from the other country; most likely to be involved in the fraudulent activities. This led to irregular patterns such as the ones depicted in this work to be identified through the use of anomaly clustering algorithms like DBSCAN promptly. Since these incongruities were detected in real time, the retailer was able to stop the aforementioned transactions that can result in monetary losses as well as protect customer information (Ahmed et al., 2016).

Another important use is the assessment of risks for financial and credit organizations. Payment systems produce large volumes of log information, and many of these are sparse due to system faults or inadequate logging. For instance, transaction details of a bank showed more attempts of failed authentications from certain terminals. These irregularities were discovered as alarming security threats from the use of anomaly clustering. This had the effect of helping to provide for the required authentication systems before weaknesses could be identified and exploited (Fernandes et al., 2019).

Another important application is in the monitoring of compliance where anomaly clustering is used again. It is also necessary that every single transaction that happens at the e-commerce platforms is in compliance with PCI DSS, especially the encryption standards. However, it may be in the background, that writing entries rarely results in a comprehensive way hence making it hard to track compliance success frequently. For instance, encryption protocol data that is missing in transaction logs might not be easily detected without sophisticated anomaly detection units. Such patterns are detected through anomaly clustering techniques to call for the attention of corrective measures. This helps to address compliance issues when they arise. before regulatory or compliance audits, setting out for surveillance to avoid instance of penalties or loss of reputation (Lan et al., 2018).

Graphs:

Table 1: Fraud Detection Performance Using DBSCAN

Metric	DBSCAN	K-Means	Hierarchical Clustering
Fraud Detection Accuracy (%)	92	75	80
False Positive Rate (%)	4	15	10
Detection Speed (ms)	120	90	200

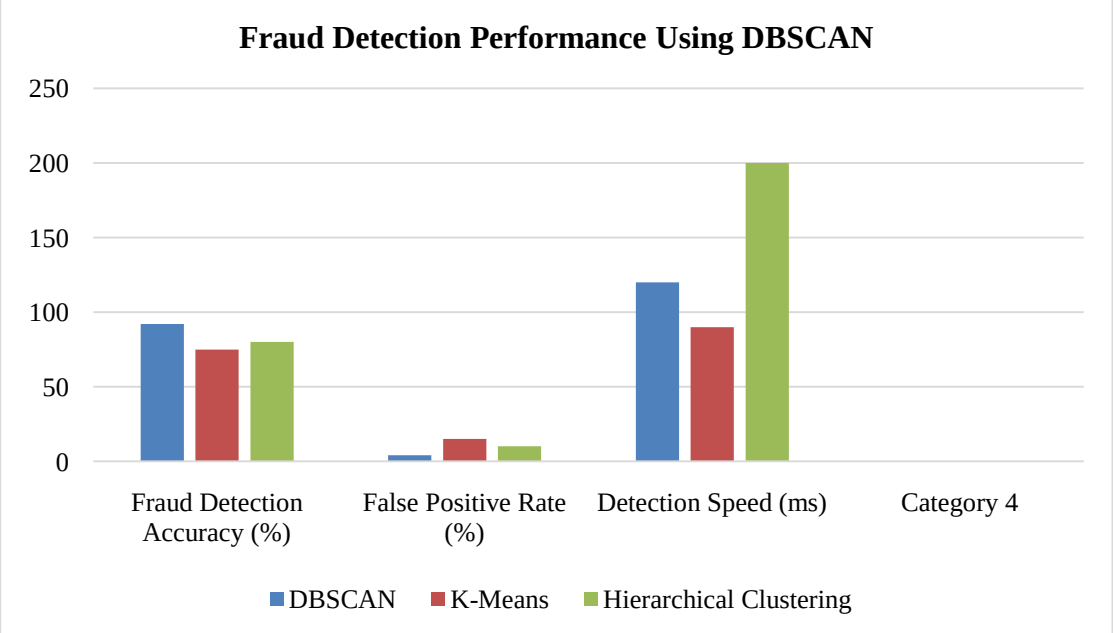


Table 2: Sparse Data Impact on Anomaly Detection Accuracy

Percentage of Missing Data (%)	DBSCAN Accuracy (%)	K-Means Accuracy (%)	Hierarchical Clustering Accuracy (%)
10	95	85	88
20	92	75	80
30	88	60	72
40	85	50	65

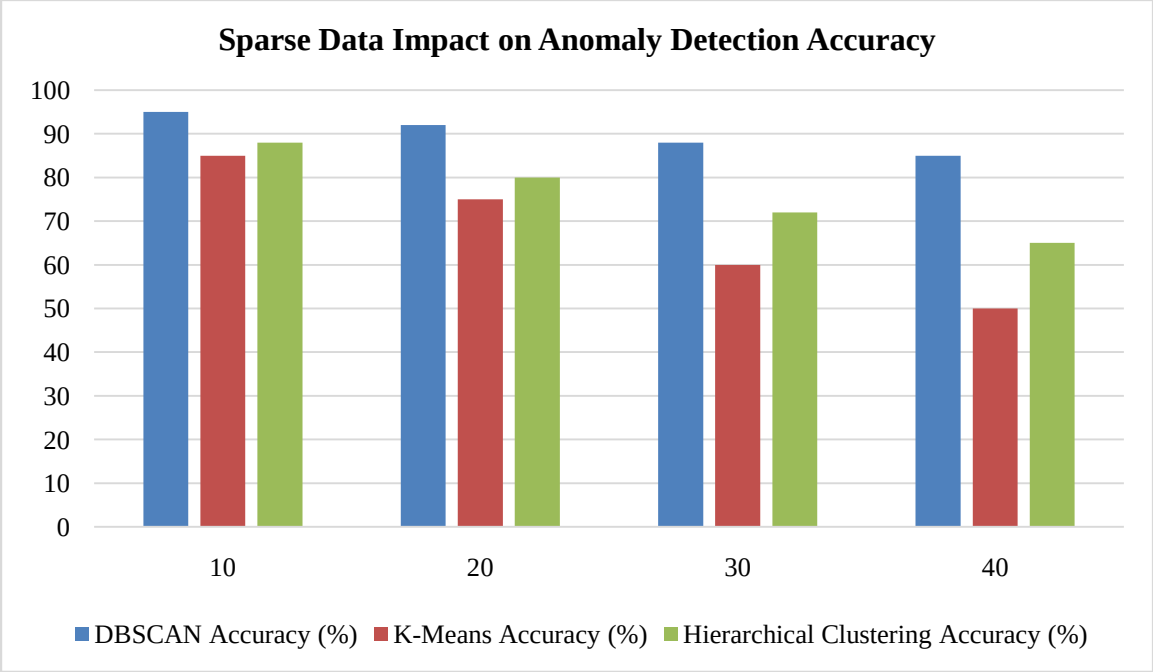
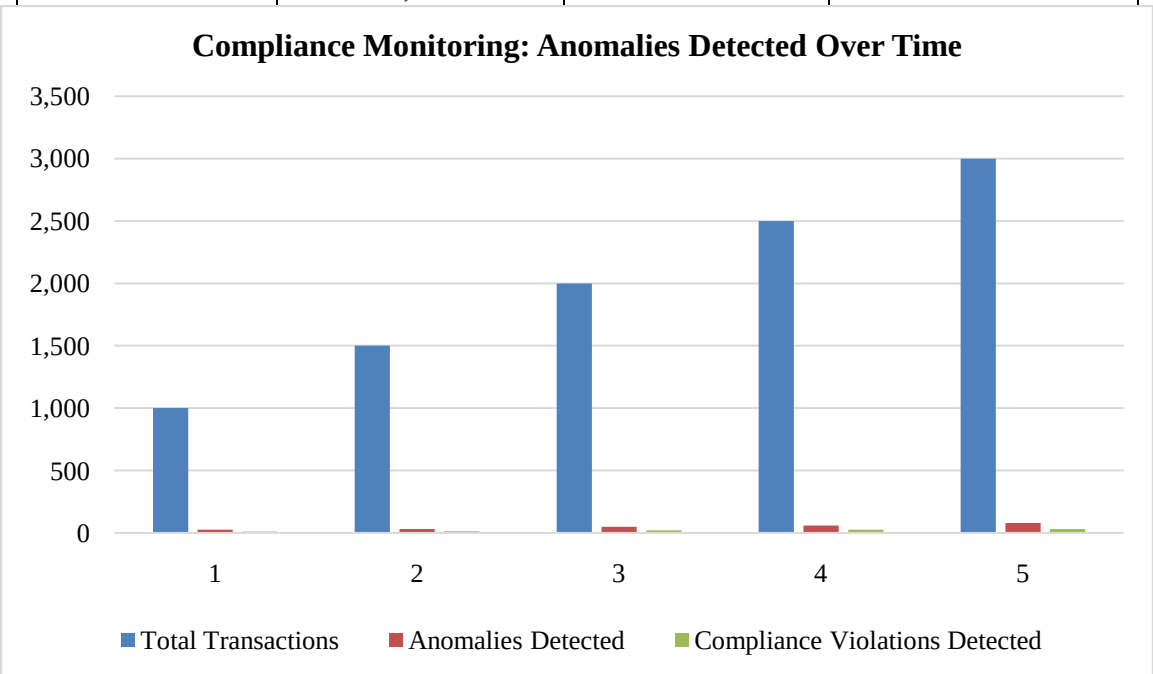


Table 3: Compliance Monitoring: Anomalies Detected Over Time

Time Interval (Hours)	Total Transactions	Anomalies Detected	Compliance Violations Detected
1	1,000	25	10
2	1,500	30	12
3	2,000	50	20
4	2,500	60	25
5	3,000	80	30



Challenges and Solutions:

Based on our findings, the most significant difficulty of anomaly clustering for PCI DSS compliance is that datasets are usually sparse. Lack of data usually limits the applicability of clustering algorithms, and greatly reduces their ability to discern any significant patterns. One feasible approach is data leveraging, where missing information can be replaced by using synthetic data. These techniques are beneficial when it comes to the quality of the output datasets to be used in clustering models.

Another distinctive problem is the computational cost of real-time anomaly detection, for it is also becoming a pressing issue. Clustering large and complex datasets generally implies a burden on various resources used in the process hence become time-consuming. To overcome this, there are scalable clustering algorithms, such as incremental DBSCAN, that can be used, and data processing in real-time the use of frameworks like Apache Spark can be considered. These tools help detect the anomalies at the right time; all the time with the assurance of giving efficient computational outcomes (Ahmed et al., 2016).

Lastly, the trade-off between security and speed is also a common issue when it comes to clustering models. High data accuracy comes at the expense of increased computational complexity, and hence cannot be used in real-time applications. The mixed models that seek to use both the density-based and the partition-based algorithms are found to be usable, offering an equal balance of reliability and speed (Fernandes et al., 2019).

Conclusion:

The main benefit of the anomaly clustering for PCI DSS compliance includes its effectiveness for sparse data and proactive prevention. This is why advanced clustering techniques like DBSCAN can be utilized to detect security vulnerabilities and fraud and to maintain compliance if organizations manage to overcome the problem of sparsity. This study also highlights the need for specific algorithms and online surveillance tools to extend trusted and legal financial environments.

References:

1. Ahmed, M., Mahmood, A. N., & Hu, J. (2016). A survey of network anomaly detection techniques. Journal of Network and Computer Applications, 60, 19-31. https://www.sciencedirect.com/science/article/pii/s1084804515002891
2. Ai, D., Pan, H., Li, X., Gao, Y., & He, D. (2018). Association rule mining algorithms on high-dimensional datasets. Artificial Life and Robotics, 23, 420-427. https://link.springer.com/article/10.1007/s10015-018-0437-y
3. CHAN, K. C. G., LING, H., & SIT, T. (2018). THE ANNALS. The Annals of Statistics, 46(5), 1837-2510. https://imstat.org/publications/aos/aos_46_5/aos_46_5.pdf
4. Fernandes, G., Rodrigues, J. J., Carvalho, L. F., Al-Muhtadi, J. F., & Proença, M. L. (2019). A comprehensive survey on network anomaly detection. Telecommunication Systems, 70, 447-489. https://link.springer.com/article/10.1007/s11235-018-0475-8
5. Lan, K., Wang, D.T., Fong, S., Liu, L.S., Wong, K.K., & Dey, N. (2018). A survey of data mining and deep learn Systems, 42, Infortics. Journal of Medical Systems, 42 1-20. https://link.springer.com/article/10.1007/s10916-018-1003-9
6. Protić, D. D. (2018). Review of KDD Cup '99, NSL-KDD, and Kyoto 2006+ datasets. Vojnotehničkiglasnik / Military Technical Courier, 66(3), 580-596. https://aseestant.ceon.rs/index.php/vtg/article/view/16670
7. Roy, A., Sun, J., Mahoney, R., Alonzi, L., Adams, S., & Beling, P. (2018, April). Deep learning deteIn2018, ng fraud in credit card transactions. In systems and information engineering design symposium (SIEDS) (pp. 129-134). IEEE. https://ieeexplore.ieee.org/abstract/document/8374722/
8. Xu, J. J., Yuruk, N., Feng, Z., & Schweiger, T. (2014). Knowledge Discovery and Data Mining. https://api.taylorfrancis.com/content/chapters/edit/download?identifierName=doi&identifierValue=10.1201/b16768-19&type=chapterpdf