Anusandhanvallari
Vol 2025, No.1
January 2025
ISSN 2229-3388

Exploring Active Outlier Detection in Big Data and Its Applications Across Diverse Domains

[1]P Ramana Vijaya Kumar, [2]Dr. Renu Chauhan

Department Of School Of Engineering And Technology

1,2shri Venkateshwara University, Gajraula (Uttar Pradesh)

Abstract: Outalier detection is an important component of data mining, especially important in the age of Big data, which is characterized by broad, complex and rapidly changing datasets in various fields including finance, healthcare, monitoring and cyber security. This research examines the active outer identity in large data settings and its practical appropriateness in many areas. The main goal is to investigate the current functioning, understand their shortcomings, and to assess the prevention and efficacy of the outer identity algorithm sophisticated in addressing the complexities of real -world data. A qualitative research technique is used, the papers of scholars published from 2018 to 2024 are enriched by secondary data collected from technical reports and empirical studies. The study conducts a comprehensive examination of several algorithms, including adaptive and hybrid models through material analysis. Conclusions suggest that sophisticated, adapted models' cross traditional techniques in scalability, accuracy and relevance. Conclusions highlight the increasing importance of active outlair detections as a strategic means in making data-powered decision making, providing effective solutions to identify discrepancy in many areas.

Keywords: Active Outlier Detection; Big Data Analytics; Anomaly Recognition; Adaptive Clustering; Statistical Boundary Refinement

INTRODUCTION:

At the age of Big Data, being able to find outlairs - data points that are very different from the rest of the data set - have become more and more important in many areas. Outlier detection is a very important technique to make better decisions and make systems more reliable. It can be used to find fake credit card transactions, suspicious behavior in video surveillance, strange patterns in network security, and even unique medical diseases (Cellia et al., 2021; gin et al., 2024). Traditional methods to find outlair work well with small or structured datasets, but they do not always work well with large data environment because they are very complex, sharp and large. This means that there is a need to come up with smarter and more flexible methods that can work in real-time, high-dimensional situations (faaique, 2024; Mazarei et al., 2025).

To fulfill this difficulty, this study sees in the use of a better method called ADC-SBR (adapted data clustering with statistical boundary refinement). This method combines the dynamic clustering mechanisms with more accurate statistical boundaries to make it easier to find anomalies. The goal of the method is to reduce the amount of work done on the computer while finding rare and microscopic outlair. The purpose of this research is to add the professionals and oppositions of the current methods and to add large data to the growing field of mining by looking at and by making a model which is more scalable and accurate. Its purpose is to provide useful information for different world conditions, which require abilities to detect strong discrepancy. The following section expands the previous literature related to this study.

LITERATURE REVIEW:

The use of large data analytics in many different regions has changed the way to process information, understand and make decisions and find problems to find problems. Yang and GE (2022) take a detailed look at the huge changes in industrial big data analytics, focusing on how it has gone from using a stable model to use stable, real

-time analytics that can bring huge changes in production, automation and system optimization. Their work shows how data-centered strategies have changed over time and emphasize the need for flexible methods to handle the atmosphere with very fast-moving data. Keyogeg et al. (2024) Add this story by focusing on cyber security, especially how to employ machine learning to find ransomware in Windows Active Directory System. Their method, which uses the system log and behavioral indicator, shows that intelligent systems are good in finding discrepancies. Mukherjee et al. (2025) See also the possible uses of large data in health science by analyzing multi-ommics data. They show how high-dimensional biological data can be used to make analog drugs. This suggests that active external identity can be used in important areas of life. On the other hand, Kaulwar (2025), the Enterprise Resource Planning (ERP) system talks about creating a better way by adding AI and Big Data Analytics to find new ways to protect from cyber attack and improve productivity.

These studies show that people are rapidly relying on data mining technologies that are smart, scalable and responsible. But there is still a major hole in integrated framework that uses both adaptive clustering and statistical refinement to find outlars in various large data areas. Most of the studies conducted so far are not a universal, cross-domain solution that works well in changing settings and is both accurate and scalable.

METHODOLOGY:

This qualitative study examines large data in various fields and active outlair detections in its applications using secondary data. The qualitative method reveals complex analytical structures, algorithms, and real -world issues that arise when searching for discrepancies in large datasets. Secondary data comes from 2018-2025 peer-review journal articles, conference proceedings, technical reports and official white paper. These sources suggest how outlair detection systems have developed, used, and are evaluated. Techniques include adaptive clustering, statistical boundary refinement and machine learning-based models. This study employs a systematic review to gather, assess, and combine high-quality, relevant literature. Then, thematic content analysis is used to uncover repeating patterns, key performance measures, and field-specific applications. Real-time, dynamic detection models like ADC-SBR are given specific attention since they differ from static, traditional approaches. The analysis includes experiments that compare techniques based on development, search, accuracy, and processing efficiency. This method provides a complete picture and shows how adaptive algorithms can detect outliers in big data environments in scalable and efficient ways.

RESULTS AND DISCUSSION:

The suggested ADC-SBR (Adaptive Data Clustering with Statistical Boundary Refinement) approach is compared to SVM and K-Nearest Neighbour outlier identification methods in this part. To demonstrate ADC-SBR's practicality in big data situations, datasets of various sizes were used to analyse construction time, search time, accuracy, and processing time.

Build time is critical to detecting system initialisation performance. Table 1 shows that ADC-SBR-based LBOD-SVM outperforms SVM and KNN at all data scales. SVM took 3.275 seconds and KNN 9.159 seconds to develop a model on a 1000-record dataset, but LBOD-SVM did it in 3.1 seconds. Intelligent clustering decreases computational overhead by efficiently grouping data before classification, resulting in this huge improvement. Reducing duplicate data processes speeds system start up.

Search time reflects system responsiveness while spotting outliers in real-time activities. Table 2 shows that ADC-SBR performs better again. For the largest dataset, LBOD-SVM had a constant and reduced search time of 0.061 seconds, while SVM had 0.0642 seconds and KNN had 0.527 seconds. Because ADC-SBR uses statistical boundary refinement to apply learnt thresholds and quickly locate anomalies, it is efficient.

LBOD-SVM provided 96.8% detection precision, beating MPSO-LS-SVM (84%), and FCM with outlier detection (93%). The dual-layer mechanism of ADC-SBR, which combines global clustering and localised

statistical refining, works. This layered structure helps the system adapt to varied data patterns and find outliers that other methods miss. Critical applications like fraud detection and health monitoring require more accuracy to improve system reliability and eliminate false alarms.

ADC-SBR's integration with Apache Spark shows its scalability and appropriateness for real-time big data applications. Table 4 shows that Spark with LBOD-SVM processed in 0.88 seconds, surpassing Spark + Logistic Regression (0.96s) and Hadoop + Logistic Regression (80s). This substantial processing time decrease shows ADC-SBR's optimization for parallel and distributed processing, making it ideal for large-scale industrial environments.

CONCLUSION:

The results clearly show that the ADC-SBR technique outperforms traditional models in every important area of performance. Its adaptive clustering process and statistical boundary refinement work together to make outlier detection faster, more accurate, and able to handle more data. These benefits make it useful in many fields, including finance, cybersecurity, smart healthcare, and industrial monitoring. This shows that it is flexible and strong enough to find anomalies in massive data.

REFERENCES:

[1] "Seliya, N., Abdollah Zadeh, A., & Khoshgoftaar, T. M. (2021). A literature review on one-class classification and its potential applications in big data. *Journal of Big Data*, *8*, 1-31."

[2] "Jin, L., Zhai, X., Wang, K., Zhang, K., Wu, D., Nazir, A., ... & Liao, W. H. (2024). Big data, machine learning, and digital twin assisted additive manufacturing: A review. *Materials & Design*, 113086."

[3] "Faaique, M. (2024). Overview of big data analytics in modern astronomy. *International Journal of Mathematics, Statistics, and Computer Science*, *2*, 96-113."

[4] "Mazarei, A., Sousa, R., Mendes-Moreira, J., Molchanov, S., & Ferreira, H. M. (2025). Online boxplot derived outlier detection. *International journal of data science and analytics*, *19*(1), 83-97."

[5] "Yang, Z., & Ge, Z. (2022). On paradigm of industrial big data analytics: From evolution to revolution. *IEEE Transactions on Industrial Informatics*, *18*(12), 8373-8388."

[6] "Keyogeg, B., Thompson, M., Dawson, G., Wagner, D., Johnson, G., & Elliott, B. (2024). Automated detection of ransomware in windows active directory domain services using log analysis and machine learning. *Authorea Preprints*."

[7] "Mukherjee, A., Abraham, S., Singh, A., Balaji, S., & Mukunthan, K. S. (2025). From data to cure: A comprehensive exploration of multi-omics data analysis for targeted therapies. *Molecular biotechnology*, *67*(4), 1269-1289."

[8] "kumar Kaulwar, P. (2025). Enhancing ERP Systems with Big Data Analytics and AI-Driven Cybersecurity Mechanisms. *Journal of Artificial Intelligence and Big Data Disciplines*, *2*(1), 27-35."