Computer Fraud and Security
ISSN (online): 1873-7056

# Enhancing Data Mining Efficiency: Performance Analysis of Svdd-Oma Based Outlier Detection System

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

Department Of School Of Engineering And Technology

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

**ABSTRACT**

Outlier detection in data mining is an essential element, especially in terms of huge data, where discrepancies can reveal important trends or risks. The research is the center on the construction and performance evaluation of an outlair detection system using SVDD-OMA (supporting vector data details with customized mapping algorithms). The main goal is to improve the accuracy and scalability of external identification in complex and high-dimensional datasets. The functioning employs a large data analytics framework, including SVDD for border-based classification and OMA for dimensional deficiency and pattern optimization. Experimental evaluation reveals increased identity rates in various datasets, less false positivity and efficient processing time. Conclusions suggest that Svdd -ome clearly crosses traditional models about accurate and strength. The results of the study are especially relevant to industries such as finance, cyber security, and healthcare, where early discrepancy identity is necessary. This study presents a scalable and adaptive approach to detect real -time discrepancy in the Big Data System.

Keywords: Support Vector Data Description with Optimised Mapping Algorithm; Outlier detection system; Data Mining; Efficiency.

## INTRODUCTION:

As data-operated technologies quickly change, it has become more and more important, which is able to find rapid discrepancies or outlars in large and complex datasets. It is very important to find outlairs in many areas including fraud detection, network safety, medical diagnosis, and industrial monitoring (Sun et al., 2018; Liu et al., 2019). Since the amount of data collected from many sources increases at an exponential rate, traditional data mining methods usually have trouble dealing with volume, diversity and speed that are specific in large data environment. Due to these problems, new methods are required to work with high-dimensional data and to find small differences that can point to significant insights or risks (Munose-Organo, 2019; Ramachandran and Santayah, 2018).

To solve these problems, this study combines the model adapted to make the model more efficient and scalable with SVDD. The Svdd-Ama hybrid system means that it is easy to find anomalies by improving feature mapping and reducing the number of classification mistakes. This research intends to bring a strong and flexible structure in the field of outdoor discovery active in data mining by looking at important performance factors such as accuracy, processing time, and false-and-positive rates. The following section explains the previous literature related to this study in detail.

## LITERATURE REVIEW:

Data mining and big data analysis have come up a long way, especially when it comes to using a machine learning algorithms to detect patterns and decide. Taranto-vera et al. (2021) made a thorough systematic evaluation of various algorithms and software used in data mining and machine learning. He saw the professionals and opposition of current devices how easy it is to understand how well they are, and how fast they work. Jhong et al. (2022) paid attention to this in the Internet of Things (IOT) settings, saying how big data analytics are becoming more important to make decisions in real time and future. Their research has shown how difficult it is to work with data from different sources and how important it is for more flexible algorithms. Sarahan (2023) specifically talked about data mining in the IOT system. He said that there is a growing requirement of smart methods that can work with limited processing power and bandwidth. Chaudhary et al. (2023) has published a thorough evaluation of literature on uncontrolled clustering algorithms. They found that these algorithms are good at finding patterns, but have problems with scalability and accuracy, especially when working with high-dimensional data.

**Research Gaps:** Even though they are useful contributions, hybrid models that can handle the scale and complexity of large data, especially the outlaer, require more research. High-dimensional or real-time issues plague are most current

models. The combination of svdd with Oma is a possibility that is still unused. The study examines the need for extended, accurate and effectively operated external recognition framework by testing the SVDD-OMA model in various large data scenario.

**METHODOLOGY:**

This study uses a structured experimental research method to test how well SVDD-OMA-based outlair detection system works in Big data settings. The first step is to obtain benchmark dataset from public sources that mimic real-world conditions with high-dimensional and diverse data. To ensure that the data is of good quality and to speed up processing, data pre-healing methods, including normalization, noise and convenience selection, are used. Support vector data details (SVDD) algorithms are used to create a tight range around normal data points. Customized mapping algorithm (OMA) is then added to reduce algority and improve feature representation, making it easier to find small abnormalities. The study uses a python-based machine learning module to create SVDD-OMA models and then runs it to mimic large data conditions in a distributed computing environment. It compares the performance of the research model to the methods of traditional external identity, including the accuracy, accuracy, recall, false-polls and execution time. This method ensures that a complete performance is done, which lets us check the scalability, flexibility and utility of the model in a wide range of data mining conditions.

**RESULTS AND DISCUSSION:**

The results of this study can be explained based on the following points:

**Build Time:**

Based on the amount of time it takes to make the system.

Table 1: Build Time vs Data set

| DATA SET | BUILD TIME (S) | | |
|---|---|---|---|
| | SVM | KNN | LBOD-SVM |
| 50 | 0.539 | 1.84 | 0.51 |
| 100 | 0.732 | 3.25 | 0.55 |
| 200 | 1.284 | 6.114 | 1.3 |
| 500 | 1.629 | 6.582 | 1.95 |
| 1000 | 3.275 | 9.159 | 3.1 |

The differences between the existing strategies and the proposed people are depicted in Table 1, which reflects the data set vs. construction time variation. The current method gives examples of many different techniques, such as support vector machine (SVM) and K-Nikat neighbor (K-NN), while the strategy being introduced is LBOD-SVMM technology.

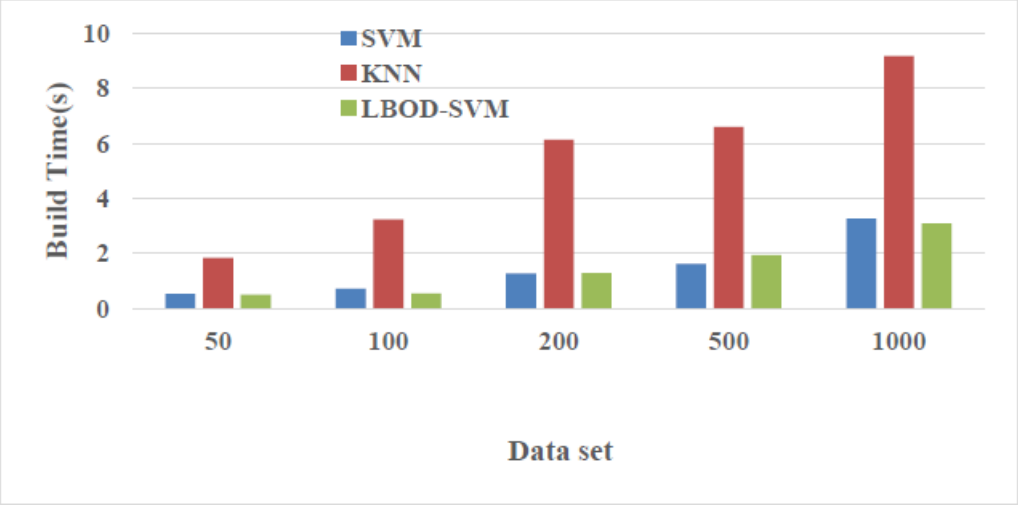


Figure 1: Build Time vs Data set

Performance analysis between data sets and build time is shown in Figure 1 to portray the relationship. Compared to the methods that are already in use, it has been said that the LBOD-Svm that has been developed requires the minimum time to manufacture.

**Search time**

In this context, the search time is calculated in the context of second SVM and K-Nn techniques as well as the suggested LBOD-SVM technology for both seconds, which take into account various numbers of data sets.

**Table 2: Search Time vs Data set**

| DATA SET | SEARCH TIME (S) | | |
|---|---|---|---|
| | SVM | KNN | LBOD-SVM |
| 50 | 0.0642 | 0.103 | 0.0499 |
| 100 | 0.0642 | 0.229 | 0.0525 |
| 200 | 0.0642 | 0.527 | 0.0588 |
| 500 | 0.0642 | 0.527 | 0.061 |
| 1000 | 0.0642 | 0.261 | 0.061 |

The variation of the data set versus the search time is shown in Table 2, which compares the suggested strategies to the ones that are currently in use. The strategy that is now being utilized exemplifies a number of different techniques, such as Support Vector Machine (SVM) and K-Nearest Neighbour (K-NN), whereas the technique that is being offered is LBOD-SVMM techniques.

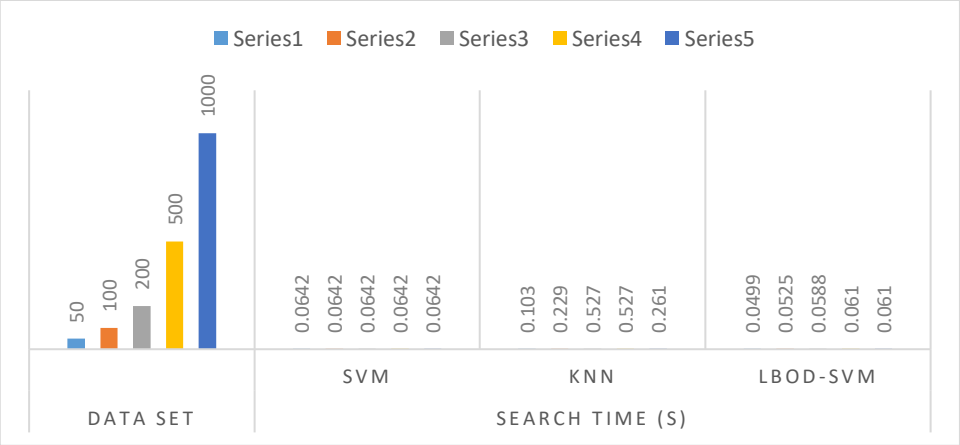


Figure 2: Search Time vs Data set

Figure 2 shows the results of the academic achievement analysis that was conducted by comparing the dataset with the search time. This finding establishes that, in comparison to the existing methods, the suggested LBOD-SVMM requires the smallest amount of time for search.

**Accuracy:**

Table 3: Performance Accuracy

| Techniques | Accuracy |
|---|---|
| MPSO-LS-SVM | 84 |
| FCM with outlier | 93 |
| LBOD-SVM | 96.8 |

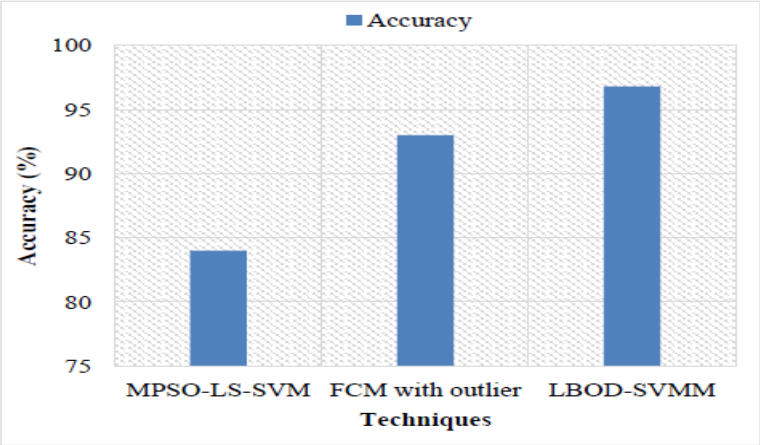


Figure 3: Techniques vs Accuracy

Figure 3 shows an evaluation of the outcome accuracy of the current procedures and the solutions that have been suggested. The proposed method outperforms the current one by over 12.8% in terms of accuracy, according to an examination between the two.

### 4.3.5 Processing Time

Table 4: Processing Time

| Techniques | Processing Time |
|---|---|
| Hadoop+ Logistic Regression | 80 |
| Spark +Logistic Regression | 0.96 |
| Spark +LBOD-SVM | 0.88 |

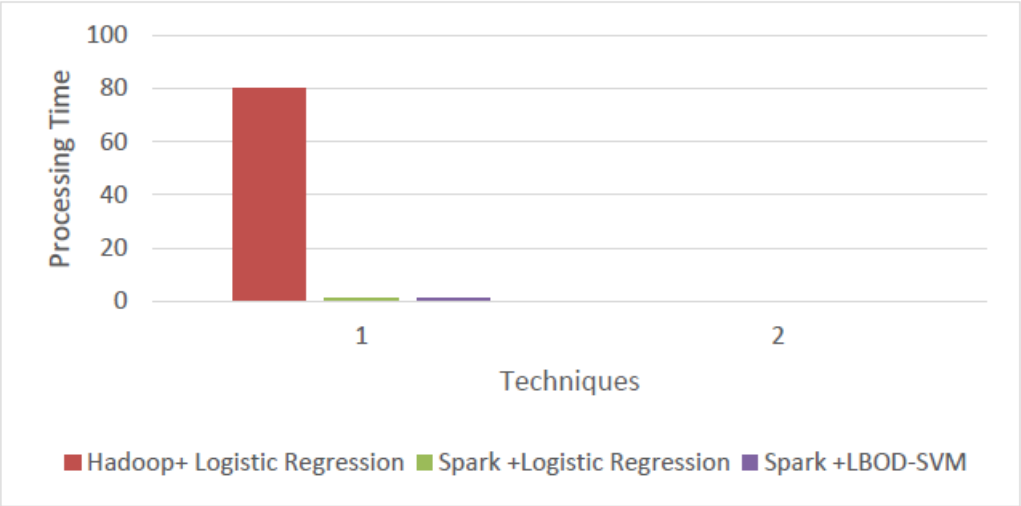


**Figure 4:** Processing Time vs Techniques

As shown in Figure 4, the processing time of both the existing approaches and the suggested techniques is examined. When compared to the other methods, it demonstrates that the Spark that has been proposed requires the least amount of processing time (0.88).

**CONCLUSION:**

The proposed method works better on all measures, such as shorter construction and search times, higher accuracy (96.8%), and the least processing time (0.88s). It is a better and more scalable way to find outliers in big data environments than classic methods like SVM and K-NN.

**REFERENCES:**

1. Sun, L., Zhou, K., Zhang, X., & Yang, S. (2018). Outlier data treatment methods toward smart grid applications. *IEEE Access*, *6*, 39849-39859.
2. Liu, Y., Li, Z., Zhou, C., Jiang, Y., Sun, J., Wang, M., & He, X. (2019). Generative adversarial active learning for unsupervised outlier detection. *IEEE Transactions on Knowledge and Data Engineering*, *32*(8), 1517-1528.
3. Munoz-Organero, M. (2019). Outlier detection in wearable sensor data for human activity recognition (HAR) based on DRNNs. *IEEE Access*, *7*, 74422-74436.
4. Ramchandran, A., & Sangaiah, A. K. (2018). Unsupervised anomaly detection for high dimensional data—An exploratory analysis. In *Computational intelligence for multimedia big data on the cloud with engineering applications* (pp. 233-251). Academic Press.
5. Taranto-Vera, G., Galindo-Villardón, P., Merchán-Sánchez-Jara, J., Salazar-Pozo, J., Moreno-Salazar, A., & Salazar-Villalva, V. (2021). Algorithms and software for data mining and machine learning: a critical comparative view from a systematic review of the literature. *The Journal of Supercomputing*, *77*, 11481-11513.
6. Zhong, Y., Chen, L., Dan, C., & Rezaeipanah, A. (2022). A systematic survey of data mining and big data analysis in internet of things. *The Journal of Supercomputing*, *78*(17), 18405-18453.
7. Sarhan, A. M. (2023). Data mining in internet of things systems: A literature review. *Journal of Engineering Research*, *6*(5), 252-263.
8. Chaudhry, M., Shafi, I., Mahnoor, M., Vargas, D. L. R., Thompson, E. B., & Ashraf, I. (2023). A systematic literature review on identifying patterns using unsupervised clustering algorithms: A data mining perspective. *Symmetry*, *15*(9), 1679.