[1]Girish Reddy Ginni

[2]Srinivasa Chakravarthi Lade

Unsupervised Machine Learning Framework for Efficient Detection of Outliers from High Dimensional Datasets

Journal of Electrical Systems

Abstract:- A basic idea in data mining and machine learning applications is outlier detection. Outlier identification and clustering frequently go hand in hand since the former can find outliers. Outlier identification was the primary focus of the majority of current research projects and clustering as two different aspects and their intimate relationship is less explored. However, considering such relationship could leverage cluster quality besides detection of outliers leading to dual benefits. Towards this end, we proposed an unsupervised machine learning (ML) framework for efficient detection of outliers from high dimensional datasets. An objective function is defined to improve cluster compactness leading to efficiency in outlier detection process. Further improvement of clustering process with problem transformation and usage of enhanced K-Means could result in an integrated approach that jointly archives quality clustering and outlier identification. We proposed an algorithm known as Learning based Outlier Detection (LbOD). Novelty of our algorithm lies in simultaneous approach in partition space, objective function and cluster optimization. A prototype is built to evaluate the proposed framework and algorithm for its ability to discover outliers considering multiple benchmark high dimensional datasets. Our empirical study has revealed that the LbOD algorithm outperforms many existing outlier detection methods.

Keywords – Outlier Detection, Clustering, Unsupervised Learning, Machine Learning, High Dimensional Data

1. INTRODUCTION

In many real world applications large volumes of data to be processed. The dataset contains datapoints that are represented and used appropriately for deriving business intelligence. However, there may be some data points that are abnormal when compared with other data points. Such data points are known as outliers and detecting them has many useful applications. Outlet detection research has been around with various methods such as heuristic methods and learning based methods. With the emergence of artificial intelligence, the usage of machine learning is increasing to solve problems in applications of different domains. In this context outlier detection not only helps in solving problems but also improve the quality of data for machine learning and other data driven applications. Literature has rich information about various heuristic and other outlier detection methods.

There is a connection between clustering and outlier identification. Enhancing cluster validity and outlier identification, the COR method effectively combines both objectives [6]. Hilal *et al.* [8] focused on anomaly detection in recent times are unsupervised models. Current mechanisms are being challenged by the rise of financial fraud. Because financial crime presents serious risk, fraud detection technologies are always being improved. Machine learning approaches such as regression for detection, grouping, and classification are highlighted in recent publications [10]. Meng *et al.* [12] provided ideas for future study by reviewing trajectory outlier identification systems based on multi-attribute representation, distance measurements, and algorithm improvements. Erhan *et al.* [15] examined anomaly detection in sensor systems, classifying approaches into data-driven and traditional categories while taking into account topologies such as Cloud, Fog, and Edge. It draws attention to obstacles and effective solutions. Credit scoring, a hybrid ensemble model that combines balanced sampling with voting-based outlier detection performs better. Outperforming benchmark models, the model tackles unbalanced data difficulties and outlier adaptation [18]. Dhiman *et al.* [19] purposed of detecting anomalies in wind turbine gearboxes using SCADA data, an adaptive threshold and TWSVM approach is suggested. Outcomes demonstrate better performance compared to baseline classifiers. Avci *et al.* [20] examined and contrasts ML and DL techniques for structural damage detection (SDD) based on vibration. ML techniques that concentrate on feature extraction and classification perform better than conventional ones. Many existing methods dealing with outlier detection considering machine learning techniques showed deteriorated performance for many reasons. Moreover, there is an issue with scalability of the model besides its accuracy. Our contributions in this paper are listed below.

1. We proposed an unsupervised machine learning (ML) framework for efficient detection of outliers from high dimensional datasets.
2. We proposed an algorithm known as Learning based Outlier Detection (LbOD) whose novelty lies in simultaneous approach in partition space, objective function and cluster optimization.

[1] Research Scholar, Computer Science Engineering, GITAM University, Gandhi Nagar, Rushikonda, Visakhapatnam, Andhra Pradesh-530 045.girishloshankar@gmail.com

[2] Assistant Professor, Department of Computer Science and Engineering,College GITAM University, Gandhi Nagar, Rushikonda, Visakhapatnam, Andhra Pradesh-530 045.chakri.ls@gmail.com

3. A prototype is built to assess the suggested algorithm and framework for its ability to discover outliers considering multiple benchmark high dimensional datasets.

This is the format for the rest of the paper. Section 2 examines the most current research on several techniques for detecting outliers. In Section 3, an automated machine learning-based approach for identifying outliers in high-dimensional data is presented. Conversely, Section 4 showcases the findings from our empirical investigation using many high-dimensional datasets. In addition to outlining potential directions for further study, Section 5 presents the results drawn.

2. RELATED WORK

This section examines current research on a range of techniques of outlier detection. Yang *et al.* [1] observed that stranger in groups are missed by conventional outlier detectors, which concentrate on individual items. The NR framework improves performance of current detectors by utilizing representative objects. Mensi *et al.* [2] found that outlier to other data points are used to identify them. To find outliers based on pairwise distance, Proximity Isolation Forest expands upon Isolation Forest. Enes *et al.* [3] explored and said that for many applications, time series are essential. In order to facilitate anomaly identification, a pipeline for grouping multivariate time series is presented in this study. Qais *et al.* [4] used a clustering approach using K-means and fuzzy c-means for outlier identification, induction heating safety was improved and 96% accuracy was attained. Osman *et al.* [5] regulated Biofeedback is included into serious games to help with good player behavior regulation and mental stress reduction.

Liu *et al.* [6] found that there is a connection between clustering and outlier identification. Enhancing cluster validity and outlier identification, the COR method effectively combines both objectives. Carreno *et al.* [7] observed that unclosed impedes research in areas such as rare event, anomaly, novelty, and outlier identification. In this study, standardization is suggested. Hilal *et al.* [8] focused on anomaly detection in recent times are unsupervised models. Current mechanisms are being challenged by the rise of financial fraud. Brito *et al.* [9] utilized unsupervised techniques and SHAP for explainability, a novel approach to defect identification and diagnosis for rotating equipment is proposed. Sadgail *et al.* [10] investigated and found that because financial crime presents such a serious risk, fraud detection technologies are always being improved. Machine learning approaches such as regression for detection, grouping, and classification are highlighted in recent publications.

Stetco *et al.* [11] examined artificial intelligence models for monitoring wind turbine status, validation, and data source categorization and regression. Model optimization and dataset problems are areas of future development. Meng *et al.* [12] provided ideas for future study by reviewing trajectory outlier identification systems based on multi-attribute representation, distance measurements, and algorithm improvements. Subbian *et al.* [13] showed how to overcome obstacles and successfully implement a robotic instrument for mTBI patient evaluation in an urban ED. Bashar and Nayak [14] observed that standard and neural network models are outperformed by TAnoGan, a GAN-based technique for anomaly identification in time series with sparse data. Erhan *et al.* [15] examined anomaly detection in sensor systems, classifying approaches into data-driven and traditional categories while taking into account topologies such as Cloud, Fog, and Edge. It draws attention to obstacles and effective solutions.

Baur *et al.* [16] compared the effectiveness of deep spatial auto encoders to patch-based approaches in the unsupervised brain MR image anomaly identification process. Latent space constraints and absence of adversarial training requirements. Accurate and swift segmentations between slices point to potential uses as previous knowledge and in unsupervised lesion segmentation. Ruff *et al.* [17] developed in deep learning for anomaly detection enhance the detection of complicated datasets, bringing methods together and examining relationships between traditional and deep techniques. Zhang *et al.* [18] for credit scoring, a hybrid ensemble model that combines balanced sampling with voting-based outlier detection performs better. Outperforming benchmark models, the model tackles unbalanced data difficulties and outlier adaptation. Dhiman *et al.* [19] purposed of detecting anomalies in wind turbine gearboxes using SCADA data, an adaptive threshold and TWSVM approach is suggested. Outcomes demonstrate better performance compared to baseline classifiers. Avci *et al.* [20] examined and contrasts ML and DL techniques for structural damage detection (SDD) determined by vibration. ML techniques that concentrate on feature extraction and classification perform better than conventional ones.

Yang *et al.* [21] addressed the issue of outlier pollution in conventional approaches, a mean-shift outlier detector is proposed. By eliminating the bias associated with outliers, the mean-shift approach enhances performance in outlier identification tasks. Zubaroglu *et al.* [22] processed is gaining popularity as more and more devices are connected and produce constant data streams. Accuracy, complexity, and basic method of recent algorithms are examined; popular tools and open problems are also covered. Chakraborty *et al.* [23] suggested to use ensemble probabilistic neural networks and stacked auto encoders to solve situations involving numerous outliers and class imbalance. The goal of future research is to expand to unsupervised techniques for various kinds of outliers. Thangaramya *et al.* [24] presented a novel secure routing method, FRCSROD, for WSNs that use outlier detection and fuzzy criteria. By identifying hostile nodes, FRDOA enhances energy efficiency, dependability, and security.

Belhadi *et al*. [25] presented herein for discerning anomalous human conduct from pedestrian data in smart cities. In less than 50 seconds, deep learning achieves 88% accuracy compared to data mining.

Landauer *et al*. [26] observed that, though their massive, unstructured data makes them difficult to analyze, log files are essential for cyber security. Logging techniques and goals are reviewed in this paper's classification of log clustering approaches. Djenouri *et al*. [27] examined the detection of outliers in urban traffic, classifying techniques into flow and trajectory detection. Different strategies are spoken about, emphasizing patterns. Tang *et al*. [28] approached to multi-kernel SVM with K-means clustering for large-scale data categorization is presented. The process chooses typical examples, decreases the amount of human labelling, and greatly increases accuracy and efficiency. Organero [29] stated that for outlier identification and sub-activity classification, a unique method integrates Human Activity classification (HAR) with DRNN. The approach is tested in many settings and yields encouraging outcomes. Chen *et al*. [30] experimented on actual datasets demonstrate that LRTG outperforms existing techniques. Adaptive neighbors, l2,1-norm, and Tucker decomposition are integrated in a unique multi-view clustering technique called Low-Rank Tensor Graph (LRTG).

Fitriyani *et al*. [31] used the Cleveland and Statlog datasets, the HDPM achieved accuracy rates of 98.40% and 95.90%. The goal of HDCDSS is to enhance early detection of cardiac disease. Rogers *et al*. [32] offered a framework for structural health monitoring (SHM) using non-parametric grouping based on Bayesian principles. This method adjusts live, provides excellent accuracy, and does not require pre-collected training data. Population-level applicability will be included in future development. Thole *et al*. [33] indicated the flows of dust and export production close to the Kerguelen Plateau, with export production being reduced during glaciers. During interglacials, the Antarctic Zone shows increased export production, highlighting Fe fertilization and changes in the Southern Ocean's upwelling. Fitriyani *et al*. [34] suggested use ensemble learning, iForest, and SMOTETomek to create a Disease Prediction Model (DPM) regarding hypertension and type 2. On four datasets, the DPM achieved good accuracy. Deepak *et al*. [35] provided a an autoencoder variant that outperforms current techniques in identifying abnormalities in surveillance footage.

Mishra *et al*. [36] expanded of IoT presents security threats because of power and cost limitations, particularly the potential for DDoS assaults. Future security initiatives and intrusion detection models are examined in this study. Kraus *et al*. [37] examined the detection of clusters in scatterplots on 2D, 3D, and virtual reality displays. Better overview was achieved with restricted VR regions, while scatterplot representations benefited from 3D VR's increased memory and orientation for cluster recognition. Liu *et al*. [38] presented SO-GAAL, a method for detecting outliers that directly generates prospective outliers to overcome high-dimensional data sparsity. Performance is further improved by expanding to MO-GAAL with numerous generators, especially on a variety of datasets. Subsequent research endeavors to incorporate group learning for stability and investigate distinct network configurations for a range of data kinds. Wang *et al*. [39] approached for detecting outliers, highlights their advantages and disadvantages, and suggests areas for further study to be improved. Usama *et al*. [40] examined the growing field of unsupervised machine learning in networking and describes its uses, including anomaly detection and traffic engineering. It highlights difficulties and potential research paths while offering insights into current advances. Many existing methods dealing with outlier detection considering machine learning techniques showed deteriorated performance for many reasons. Moreover, there is an issue with scalability of the model besides its accuracy.

3. PRELIMINARIES

In this section, we provide an overview of K-means and entropy. To address K-means' susceptibility to outliers, a variant known as K-means--[17] was created. Few outliers are known to diverge the centroids from their inherent locations. In order to address this, certain data points that are distant from their centroids are considered outlier candidates. These data points are not given a cluster name and are not updated centroidally. K-means and its assigning data points and updating centroid are two iterative phases—are similar. We determine the separations between every data point and the closest centroid throughout the data point assignment process. We then rank the distances, identifying potential outliers as the data points with the highest or lowest distances. Since these outlier candidates aren't given cluster names, K-means updates the centroid in the same way. It is noteworthy that during the iteration, the outlier candidates are evolving. K-means requires two input parameters, whereas K-means just requires the number of clusters and the K and o outliers. Regarding its clean mathematical formulation, convergence, and model effectiveness, it has many qualities with K-means. Ref [19] makes clear that entropy or total correlation alone is insufficient for outlier spotting. They put forth the following new measure of holoentropy. In holoentropy, the overall relationship between the random vector Y and its entropy are added together to form the holoentropy HL(Y), which may be stated as the sum of the entropies for all characteristic. Based on information theory, holoentropy is an outlier identification measure that handles categorical data and accounts for both total correlation and entropy. For a tidy and effective solution, we derive our suggested goal function, which is based on holoentropy associated with K-Means algorithm.

4. PROPOSED METHODOLOGY

This section presents the proposed methodology used for outlier detection in high dimensional data efficiently. The methodology includes proposed outlier detection framework, it's mechanisms and underlying algorithm.

4.1 Problem Definition

Provided a high dimensional dataset proposing a machine learning based framework to detect outliers efficiently is the challenging problem considered.

4.2 The Proposed Outlier Detection Framework

The unsupervised machine learning model serves as the foundation for the suggested outlier identification approach. The framework's architecture allows it to process high-dimensional data as input and produce output in the form of identified outliers. A particular data city is put through an audition process in order to use clustering to exploit a methodology. Clustering method divides data points into number of categories. Strange values among data points which are dissimilar when compared to other values are known as outliers. There are many applications linked to outlier detection. The outlier detection methods help in solving many real time problems.

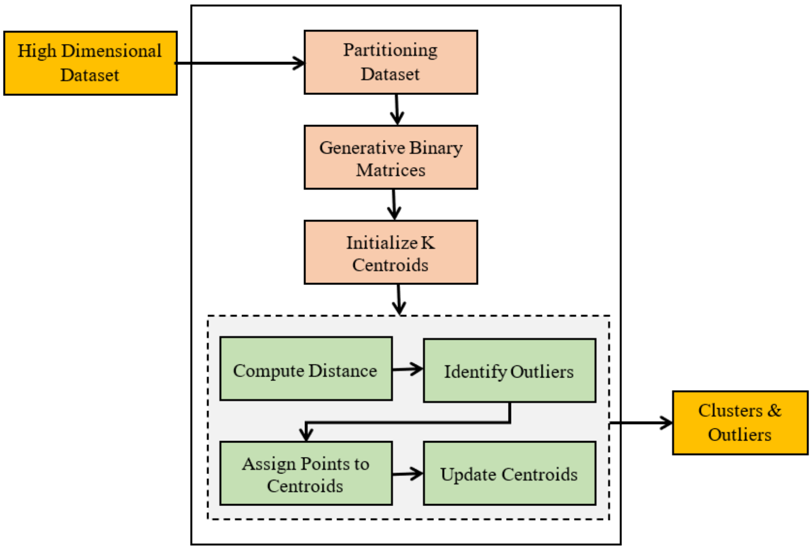

Figure 1: Overview of the proposed outlier detection of framework

Sometimes detection of outliers itself is useful in applications like credit code fraud detection. Figure 1 shows the proposed framework designed for automatic finding outliers in a large-scale data collection.

4.3 Objective Function

Outlier identification and cluster analysis are closely related activities. A few outlier points may quickly destroy a cluster's structure; in contrast, outliers are specified by the cluster idea and are identified as points that belong to none of the clusters. We concentrate on the clustering based outlier detection, the proposed approach, in order to address this difficulty. In particular, after identifying o points as outliers, the remaining occurrences are split into K clusters by performing the outlier identification and clustering tasks in parallel. Corresponding symbols used in the next sections are displayed in Table 1. A small number of outliers might undermine the cluster structure, and these outliers want to be recognized by the cluster boundary. It is similar to a chicken-and-egg dilemma due to the coupling relationship between outlier identification and cluster analysis. We draw inspiration from consensus clustering [23], which combines many fundamental divisions created to prevent the circular dependency problem in combined approach considering outlier detection and clustering, the data should be thoroughly fused to minimize the bad effects of outliers. Furthermore, the clusters are necessary for the definition of outliers. The two considerations mentioned above encourage us to create several simple partitions in order to convert the information into partition space from the original feature space. This procedure is comparable to consensus clustering's fundamental partition generation approach [45], [46]. With n points and d features, let X be the data matrix. When X is divided into K distinct clusters, it may be shown as a set of K object subsets including a vector label $\pi = (L_\pi(x_1),\cdots,L_\pi(x_n))$, where x_1 is mapped by $L_\pi(x_1)$ a label in K. The r fundamental partitions $\Pi = \{\pi_i\}, 1 \leq i \leq r$, may be obtained by applying certain basic partition generation strategies, including K-means clustering with varied cluster counts. For π_i, let K_i denotes cluster number and let $R = \sum_{i=1}^{r} K_i$. Then, using Π, the following B = {b_l},1 < l ≤ n, binary matrix, may be obtained:

$$b_l = (b_{l,1}, \dots, b_{l,i}, \dots, b_{l,r}), \quad \text{with}$$
$$b_{l,i} = (b_{l,i1}, \dots, b_{l,ij}, \dots, b_{l,iKi}), \quad \text{and} \qquad (1)$$

$$b_{l,ij} = \begin{cases} 1, \ if \ L_{\pi_i}(x_l) = j \\ 0, otherwise \end{cases}$$

It is important to remember that creating fundamental divisions does not need the use of a particular method. To construct basic partitions, K-means with varying cluster counts is advised for efficiency and simplicity. Based on the fundamental divisions produced by K-means, our proposed approach nonetheless yields encouraging results despite K-means' susceptibility to outliers. The binary value represents the information unique to cluster membership, which is created in accordance with the definition of outliers. Because of these two characteristics, the binary space is preferable to the continuous space in that it facilitates the identification of outliers due to its categorical traits. For instance, a popular measure for detecting outliers in categorical data is holoentropy [19].

The authors of Ref [19] sought to reduce the dataset's Holoentropy after removing the outliers. Here, we're assuming that the entire dataset has a cluster structure. As a result, minimizing each cluster's Holoentropy makes more sense. In this method, instead of the complete dataset becoming compact once the outliers are eliminated, the clusters do. Therefore, we provide our intended purpose for the proposed approach is as follows:, depending on the Holoentropy of each cluster.

$$\min_{\pi} \sum_{k=1}^{K} p_k HL(C_k), \qquad (2)$$

Where π is the cluster indicator and $p_{k+} = |C_k|/(n - o)$. HL(•) referred in a definition earlier covers K clusters C1∪•••∪C_K = X\O, with $C_k \cap C_{k^{'}} = \emptyset$ if $k \neq k^{'}$. In Eq. 2, the objective function is the weighted Holoentropy linked to every cluster and p_k is based on size of cluster. Beyond this work, we think of discovering K and o is associated with an orthogonal issue. Here, the variables of our suggested method—which uses the same setup as K-means—are the K and o denoting number of clusters and outliers respectively [17]. An efficient solution is provided in the next section by addressing the issue expressed in Eq. 2 by providing an objective function based on the binary matrix as expressed below.

$$\sum_{k=1}^{K} p_k HL(C_k) \propto \sum_{k=1}^{K} \sum_{b_l \in C_k} \sum_{i=1}^{r} \sum_{j=1}^{K_i} H(C_{k,ij}), \text{ and}$$

$$H(C_{k,ij}) = -(1 - p_{k,ij}) log(1 - p_{k,ij}) - p_{k,ij} \log p_{k,ij},$$

where the Shannon entropy is denoted by H and the probability that $b_{l,ij} = 1$ in the ij-th column of C_k is represented by $p_{k,ij}$. We offer the following lemma to clarify the meaning of $p_{k,ij}$ in Eq. (3).

Lemma 1. In K-Means clustering applied on binary dataset, k-th centroid is computed to satisfy:

$$m_k = (m_{k,1}, \dots, m_{k,i}, \dots, m_{k,r}), \quad \text{with}$$

$$m_{k,i} = (m_{k,i1}, \dots, m_{k,ij}, \dots, m_{k,iK_i}), \quad \text{and} \qquad (4)$$

$$m_{k,ij} = \sum_{b_{l,ij} \in C_k} \frac{b_{l,ij}}{|C_k|} = p_{k,ij}, \forall \ k, i, j.$$

Lemma 1's proof is clear from the centroid's arithmetic mean in clustering process. Lemma 1 leads us to conclude that the issue linked to Eq. (3) is related to clustering process.

Theorem1. When applying K-means to B's $n - o$ inliers, we obtain

$$\max \sum_{k=1}^{K} p_k \sum_{i=1}^{r} \sum_{j=1}^{K_i} p_{k,ij} \log p_{k,ij} \Leftrightarrow min \sum_{k=1}^{K} \sum_{b_l \in C_k} f(b_l, m_k), \qquad (5)$$

where m_k *is the* $k - th$ *centroid as expressed in Eq. (4) while the distance function* $f(b_l, m_k) = \sum_{i=1}^{r} \sum_{j=1}^{K_i} D_{KL}(b_{l,ij} \| m_{k,ij})$, *here* $D_{KL}(\cdot \, || \, \cdot)$ *is the KL-divergence.*

Proof. The Bregman divergence [47] indicates that we have $D_{KL}(s||t) = H(t) - H(s) + (s - t)^T \nabla H(t)$, where two vectors of the same length are denoted by s and t. Next, we begin with the right side of equation (5).

$$\sum_{k=1}^{K} \sum_{b_l \in C_k} f(b_l, m_k) = \sum_{k=1}^{K} \sum_{b_l \in C_k} \sum_{i=1}^{r} \sum_{j=1}^{K_i} \left(H(m_{k,ij}) - H(b_{l,ij}) + (b_{l,ij} - m_{k,ij})^T \nabla H(m_{k,ij}) \right) =$$

$$\sum_{k=1}^{K} |C_k| \sum_{i=1}^{r} \sum_{j=1}^{K_i} H(m_{k,ij}) - \sum_{k=1}^{K} \sum_{b_l \in C_k} \sum_{i=1}^{r} \sum_{j=1}^{K_i} H(m_{k,ij}) \qquad (6)$$

The above equation holds due to $\sum_{b_l \in C_k} (b_{l,ij} - m_{k,ij})$, When considering the binary matrix B, the second term is a constant. Lemma 1 leads us to the conclusion of the proof.

Remark 1. *Theorem 1 reveals how K-means on B and the second component of Eq. (3) are equal. This suggests that the straightforward K-means with KLdivergence on every dimension may effectively tackle a portion of this complicated problem.*

4.4 Proposed Algorithm

We proposed an algorithm known as Learning based Outlier Detection (LbOD). Novelty of our algorithm lies in simultaneous approach in partition space, objective function and cluster optimization.

Algorithm 1: Learning based Outlier Detection (LbOD)

Input: Data X, partitions r, number of clusters K and outliers o

Output: Clusters *K* and outliers O

```
Begin
r←CreatePartitions(X)
$(B, \tilde{B})$←CreateBinaryMatrices()
```

```
4.      K←InitializeCentroids($B, \tilde{B}$)
5.      While objective value is unchanged
6.        M←Compute distance between data points and nearest centroid
7.        outliers←FindPointsOfHighestDistance()
8.        Assign other points to centroids nearest
9.        Compute arithmetic mean to update clusters
10.     End While
11.     End
```

Algorithm 1: Learning based Outlier Detection (LbOD)

As presented in algorithm it takes the data number of partitions number of clusters and outliers as input and produces clusters and also identification of outliers through clustering process. The algorithm is based on means optimization process. The algorithm also addresses two challenges associated with the optimization problem. Since K-means clustering can resolve the second half of Eq. (3), we should focus on turning the issue towards solution with K-Means. Given that Theorem 1 indicates that $1 - p_{k,ij}$ tough to incorporate into K-means clustering, we intend to represent $1 - p_{k,ij}$ by inserting another binary matrix $\tilde{B} = \{\tilde{b_l}\}, 1 \leq l \leq n$ in the following manner.

$$\tilde{b}_l = (\tilde{b}_{l,1}, \dots, \tilde{b}_{l,i}, \dots, \tilde{b}_{l,r}), \quad \text{with}$$
$$\tilde{b}_{l,i} = (\tilde{b}_{l,i1}, \dots, \tilde{b}_{l,ij}, \dots, \tilde{b}_{l,iKi}), \text{ and} \qquad (7)$$
$$\tilde{b}_{l,ij} = \begin{cases} 0, if\ L_{\pi_i}(x_l) = j \\ 1, otherwise \end{cases}$$

Eq. (7) is also used to construct $\tilde{B}$ from Π. $\tilde{B}$ is equivalent to flipping B when compared to the binary matrix B in Eq. (1). The variables like $(K_i - 1)$-of-K_i and 1-of-K_i associated with the initial data are represented by $\tilde{B}$ and B, respectively. By using Equation (4) to define m˜_kk in terms of $\tilde{B}$, we are able to derive

$$\tilde{m}_{k,iij} = 1 - m_{k,iij} = 1 - p_{k,iij} .$$

The issue in Eq. (3) is transformed into clustering optimization based on $\tilde{B}$ and B.

Theorem2. When $n - o$ inliers associated with $[B\ \tilde{B}]$ are subjected to K-Means, we have

$$\min_{\pi} \sum_{k=1}^{K} p_k HL(C_k) \Leftrightarrow \min \sum_{k=1}^{K} \sum_{b_l \in C_k} \left(f(b_l, m_k) + f(\tilde{b}_l, \tilde{m}_k) \right),$$

where m_k, $\tilde{m}_k$ are the distance function and the k-th centroid determined by Eq. (4).

$f(b_l, m_k) = \sum_{i=1}^{r} \sum_{j=1}^{K_i} D_{KL}(b_{l,ij} \| m_{k,ij})$, $f(\tilde{b}_l, \tilde{m}_k) = \sum_{i=1}^{r} \sum_{j=1}^{K_i} D_{KL}(\tilde{b}_{l,ij} \| \tilde{m}_{k,ij})$, and $D_{KL}(\cdot \,||\, \cdot)$ is theKL-divergence.

Remark 2. K-means cannot be used to solve the issue in Eq. (3) regarding the binary matrix B. To represent $1 - p_{k,ij}$, We nontrivially add an additional binary matrix B˜, which is B flipped. This allows clustering on [B˜] to construct the entire problem, as shown in Theorem 2. Benefits include both inheriting the K-means algorithm's efficiency, and suitability for scalable clustering and detection of outliers, and simplifying the issue with a clean mathematical formulation.

The first difficulty, which is the issue with inliers in Eq. (2) by employing the auxiliary matrix B, is fully resolved by Theorem 2. By doing this, a simple K-means solution is transformed into a whole one. In the subsequent section, we address the second problem, which operates on all data points instead of n-o inliers.

In this work, we investigate the proposed clustering approach, that performs splits to the data in parallel and identifies outliers. As a result, the operations of clustering and outlier identification are carried out using the same framework. As centroids associated with K-means have probability of being outliers, they shouldn't contribute to them. Driven by K-means [17], outliers are defined as places that deviate significantly from the nearest centroid. The problem considered in the solution is linked to partition space' Holoentropy while the K-means focused on feature space in general. Then, utilizing the auxiliary matrix B̃, we formulate the issue as a optimization of K-means. K-means——is a method used to solve the issue in Eq. (2), which yields K clusters and outlier set O— after careful modification and derivation. Algorithm 1 provides a summary of our suggested clustering procedure with outlier elimination. We then examine the time complexity and convergence of Algorithm 1's property. First, we create R basic partitions in Line 1. These are typically completed by K-means clustering with various cluster numbers. The processing time for this phase is $O(rt'\bar{K}nd)$, where t' and $\bar{K}$ stand for the average number of iterations and clusters, respectively. A comparable temporal complexity, O(tKnR), is indicated by lines 5-8 for the typical K-means-- method, where binary matrices' dimension expressed as $R = \sum_{i=1}^{r} K_i$ i. Algorithm fines longest distances in order to find outliers. It is important to remember that parallel computing may be used to create R basic partitions, significantly the duration of execution. Furthermore, when compared to the total number of points (n), $t', t, r,$ and R are quite small. Thus, our technique is easily scalable in clustering to discover outliers, with its time complexity effectively linear in the number of points. Furthermore, Algorithm 1 ensures local optimum with optimized convergence in clustering.

4.5 Dataset Details

Datasets used in this paper are available in [41], [42] and [43]. Each dataset is of specific type with a different number of instances, number of features, number of outliers and number of clusters. These are high dimensional datasets that are widely used in outlier detection research.

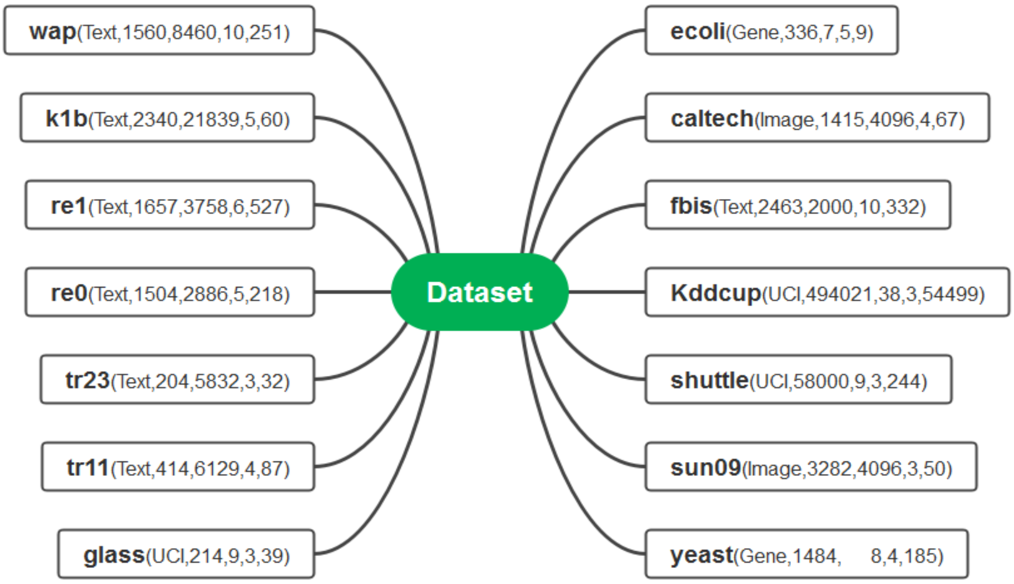
wap(Text,1560,8460,10,251)
k1b(Text,2340,21839,5,60)
re1(Text,1657,3758,6,527)
re0(Text,1504,2886,5,218)
tr23(Text,204,5832,3,32)
tr11(Text,414,6129,4,87)
glass(UCI,214,9,3,39)
Dataset
ecoli(Gene,336,7,5,9)
caltech(Image,1415,4096,4,67)
fbis(Text,2463,2000,10,332)
Kddcup(UCI,494021,38,3,54499)
shuttle(UCI,58000,9,3,244)
sun09(Image,3282,4096,3,50)
yeast(Gene,1484, 8,4,185)

Figure 2: Shows the data distribution dynamics of different benchmark datasets

As presented in Figure 2, each dataset is provided with details such as type of dataset, number of instances, number of features, number of clusters and number of outliers respectively.

4.6 Evaluation Metrics

The performance of proposed outlay detection method is evaluated with different metrics such as Normalized Mutual Information (NMI), Rand Index (Rn), Jaccard index and F-measure as expressed in Eq. 8, Eq. 9, Eq. 10 and Eq. 11 respectively.

$$NMI = \frac{\sum_{i,j} n_{ij} log\frac{n.n_{ij}}{n_{i+} \cdot n_{+j}}}{\sqrt{\left(\sum_i n_{i+} log\frac{n_{i+}}{n}\right)\left(\sum_j n_{j+} log\frac{n_{+j}}{n}\right)}} \tag{8}$$

$$R_n = \frac{\sum_{i,j}\binom{n_{ij}}{2} - \sum_i \binom{n_{i+}}{2} \cdot \sum_j \binom{n_{+j}}{2} \Big/ \binom{n}{2}}{\sum_i \binom{n_{i+}}{2} \Big/ 2 + \sum_j \binom{n_{+j}}{2} \Big/ 2 - \sum_i \binom{n_{i+}}{2} \cdot \sum_j \binom{n_{+j}}{2} \Big/ \binom{n}{2}} \tag{9}$$

$$Jaccard = \frac{|O \cap O*|}{|O \cup O*|} \tag{10}$$

$$F - measure = 2 * \frac{precition \cdot recall}{precition + recall} \tag{11}$$

NMI and Rn are used to validate cluster formation as part of the proposed outlier detection methodology. The Jaccard and F-measure are employed to measure accuracy in detection of outliers.

5. EXPERIMENTAL RESULTS

This section presents the results of our experiments made with number of benchmark datasets described in section 4.5. The performance of the proposed outlier detection method is evaluated in terms of various metrices specified in section 4.6. Performance of the proposed outlier detection method is also compared with number of state of the art methods and found the significance of the proposed approach in outlier detection.

Dataset	NMI		
	K-means	K-means--	LbOD
Ecoli	65.05	64.18	64.92
Yeast	20.68	17.33	21.49
Caltech	79.05	77.1	89.73
sun09	20.18	12.17	22.67
Fbis	12.18	33.7	54.98
k1b	52.95	50.17	55.15
re0	20.2	18.06	34.88
re1	19.66	15.49	38.15
tr11	10.29	21.84	62.63

tr23	7.89	12.68	26.03
Wap	43.36	33.17	50.78
Glass	37.25	37.26	39.82
Shuttle	23.55	26.16	36.15
Kddcup	1.46	72.22	86.72

Table 1: Performance comparison in terms of NMI

As presented in Table 1, the performance of outlier detection methods in terms of NMI is provided against number of datasets.

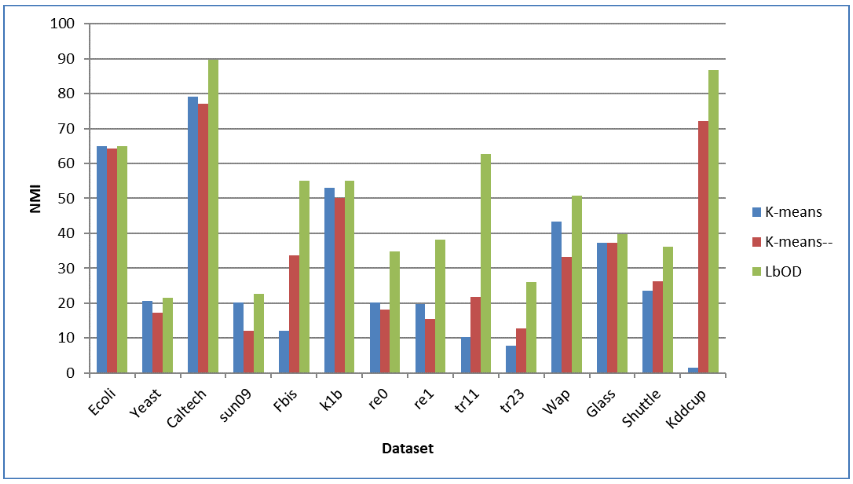

Figure 3: Performance comparison among outlier detection methods in terms of NMI

As president in Figure 3, the performance comparison among outlier detection methods in terms of NMI is provided. It is observed from the results that the proposed outlier detection method showed better performance over existing methods against all the datasets used in experiments.

Dataset	**Rn**		
	K-means	**K-means--**	**LbOD**
Ecoli	67.83	62.95	70.42
Yeast	15.12	13.78	20.11
caltech	63.13	78.2	89.43
sun09	18.81	10.80	22.2
fbis	-0.67	12.65	40.68
k1b	43.99	44.22	42.01
re0	11.66	13.28	25.59
re1	4.15	5.4	23.3
tr11	0.52	8.63	59.5
tr23	-0.3	4.33	22.5
wap	14.34	12.66	36.64
glass	23.53	25.56	26.58
shuttle	40.85	33.44	60.29
Kddcup	0.04	81.21	94.76

Table 2: Performance comparison in terms of Rn

As presented in Table 2, the performance of outlier detection methods in terms of Rn is provided against number of datasets.

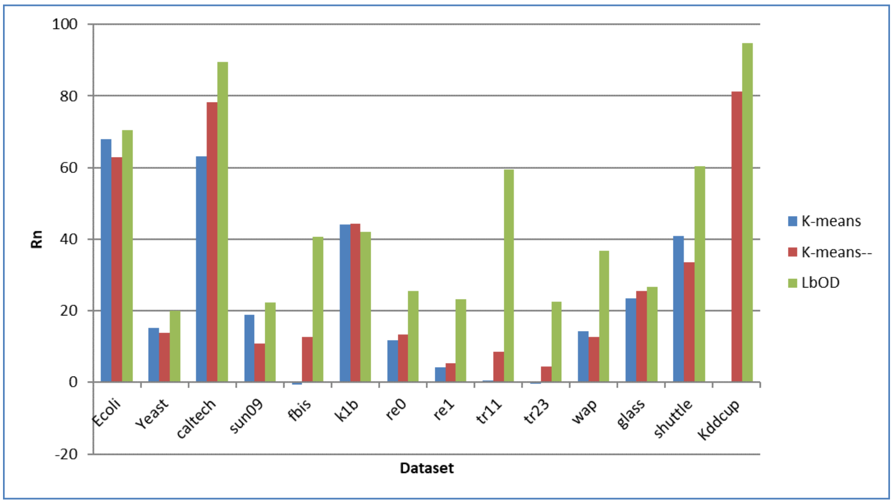

Figure 4: Performance comparison among outlier detection methods in terms of Rn

As president in Figure 4, the performance comparison among outlier detection methods in terms of Rn is provided. It is observed from the results that the proposed outlier detection method showed better performance over existing methods against all the datasets used in experiments.

Dataset	Jaccard		
	K-means	K-means--	LbOD
ecoli	4.36	58.54	51.12
Yeast	6.25	20.52	51.92
caltech	19.68	45.81	98.58
sun09	1.93	3.71	2.49
fbis	0.09	5.36	26.01
k1b	0	0	21.35
re0	5.56	9.5	29.70
re1	0.54	17.09	29.52
tr11	0	10.35	37.09
tr23	0	6.89	15.01
wap	1.11	11.29	23.31
glass	13.64	32.28	35.54
shuttle	0	5.39	6.51
Kddcup	0.01	18.32	16.61

Table 3: Performance comparison in terms of Jaccard

As presented in Table 3, the performance of outlier detection methods in terms of Jaccard is provided against number of datasets.

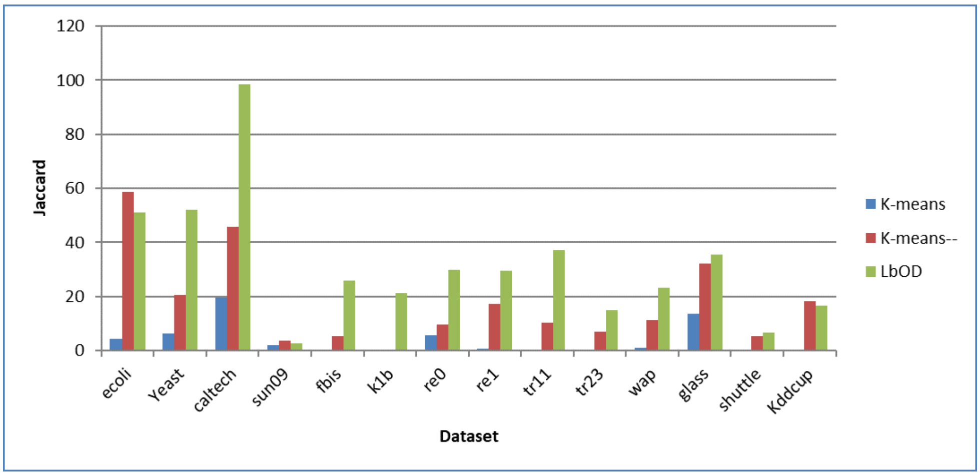

Figure 5: Performance comparison among outlier detection methods in terms of Jaccard

As president in Figure 5, the performance comparison among outlier detection methods in terms of Jaccard is provided. It is observed from the results that the proposed outlier detection method showed better performance over existing methods against all the datasets used in experiments.

Dataset	**F-measure**		
	K-means	**K-means--**	**LbOD**
ecoli	8.21	76.18	67.45
Yeast	11.79	33.61	68.36
caltech	31.47	64.21	99.29
sun09	3.78	7.15	4.86
fbis	0.17	10.18	41.3
k1b	0	0	35.16
re0	10.52	17.35	45.78
re1	1.09	29.21	45.58
tr11	0	18.76	54.18
tr23	0	12.92	26.19
wap	2.17	20.28	37.80
glass	23.64	49.56	52.42
shuttle	0	10.22	12.29
Kddcup	0.02	31.59	28.51

Table 4: Performance comparison in terms of F-measure

As presented in Table 4, the performance of outlier detection methods in terms of F-measure is provided against number of datasets.

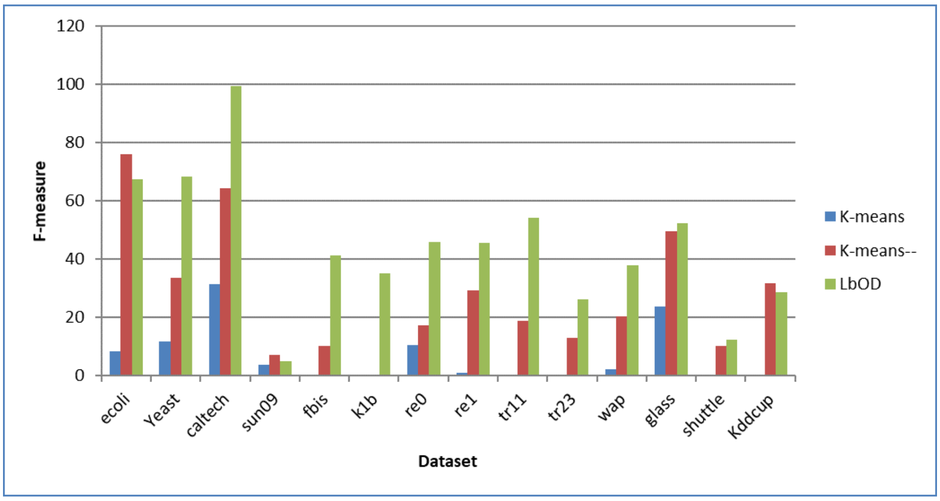

Figure 6: Performance comparison among outlier detection methods in terms of F-measure

As president in Figure 6, the performance comparison among outlier detection methods in terms of F-measure is provided. It is observed from the results that the proposed outlier detection method showed better performance over existing methods against all the datasets used in experiments.

Dataset	**Jaccard**								
	LOF	**COF**	**LDOF**	**FABOD**	**iForest**	**OPCA**	**TONMF**	**K-means--**	**LbOD**
ecoli	20.00	38.46	5.88	20.00	38.28	5.88	0.00	45.76	47.37
yeast	11.45	11.45	5.11	13.85	23.75	26.71	8.66	14.38	50.47
caltech	2.29	0.75	1.52	8.06	27.62	0.00	0.00	30.36	97.19
sun09	1.01	2.04	0.00	2.04	2.04	0.00	0.00	3.27	2.27
fbis	8.32	5.56	4.90	6.41	5.40	4.40	8.32	5.21	23.77
k1b	0.00	0.00	0.00	0.84	0.00	0.00	1.69	0.00	20.53
re0	2.59	5.31	3.07	6.34	2.83	11.79	7.13	8.82	28.50
re1	21.85	15.44	15.19	18.83	16.85	17.77	16.98	16.75	27.64
tr11	10.13	8.75	19.18	10.83	8.75	8.75	12.99	9.93	34.06

tr23	4.92	4.92	6.67	10.34	6.67	1.59	10.34	5.87	12.35
wap	10.82	12.30	6.36	12.81	11.31	6.58	7.49	10.98	22.01
glass	16.42	36.84	4.00	25.81	13.04	14.71	0.00	24.00	32.67
shuttle	12.44	12.96	0.21	7.25	1.46	3.61	0.00	5.39	5.58
Kddcup	11.54	15.26	3.40	8.50	21.22	15.66	8.66	15.06	15.98

Table 5: Performance comparison among many outlier detection methods in terms of Jaccard index

As presented in Table 5, the performance of more outlier detection methods in terms of Jaccard is provided against number of datasets.

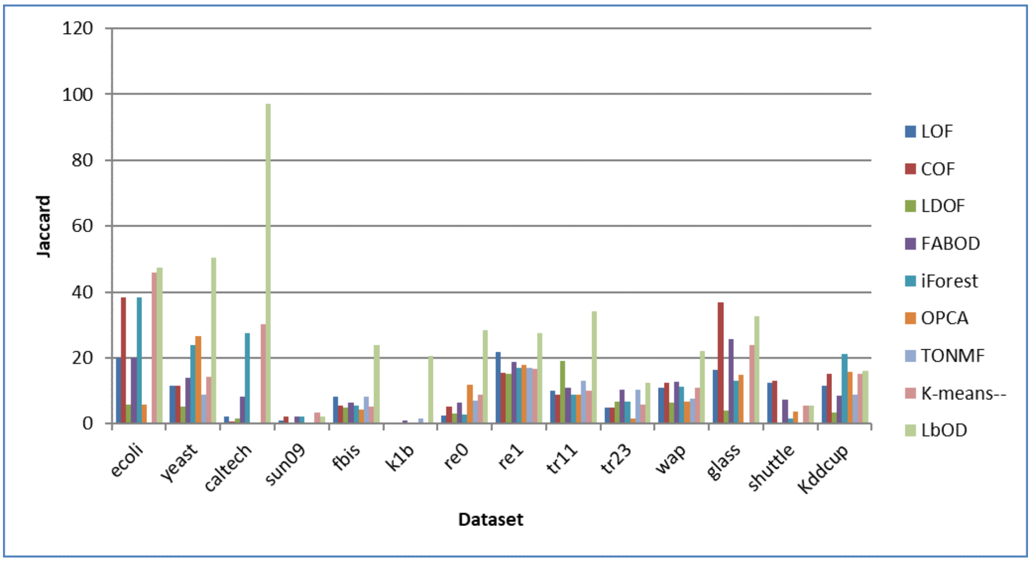

Figure 7: Performance comparison among more outlier detection methods in terms of Jaccard

As president in Figure 7, the performance comparison among more outlier detection methods in terms of Jaccard is provided. It is observed from the results that the proposed outlier detection method showed better performance over existing methods against all the datasets used in experiments.

Dataset	**F-measure**								
	LOF	**COF**	**LDOF**	**FABOD**	**iForest**	**OPCA**	**TONMF**	**K-means--**	**LbOD**
ecoli	33.33	55.56	11.11	33.33	55.56	11.11	0.00	61.58	64.21
yeast	20.54	20.54	9.73	24.32	38.38	11.11	8.11	24.69	67.07
caltech	4.48	1.49	2.99	14.93	43.28	0.00	1.49	44.37	98.57
sun09	2.00	4.00	0.00	4.00	4.00	0.00	6.00	6.34	4.44
fbis	15.36	10.54	9.34	12.05	43.28	8.43	15.36	9.91	38.35
k1b	0.00	0.00	0.00	1.67	0.00	0.00	3.33	0.00	34.06
re0	5.05	10.09	5.96	11.93	5.50	21.10	13.30	16.20	44.34
re1	35.86	26.76	26.38	31.69	28.84	30.17	29.03	28.70	43.28
tr11	18.39	16.09	32.18	19.54	16.09	16.09	22.99	18.06	50.74
tr23	9.37	9.37	12.50	18.75	12.50	3.12	18.75	11.08	21.88
wap	19.52	21.91	11.95	22.71	23.75	12.35	13.94	19.78	36.06
glass	28.21	53.85	76.90	22.71	23.08	25.64	0.00	37.97	49.18
shuttle	22.13	22.95	0.41	13.52	2.87	6.97	0.00	10.22	10.56
Kddcup	20.53	18.65	0.31	11.62	35.01	27.08	15.94	26.03	27.55

Table 6: Performance comparison among many outlier detection methods in terms of F-measure index

As presented in Table 6, the performance of more outlier detection methods in terms of F-measure is provided against number of datasets.

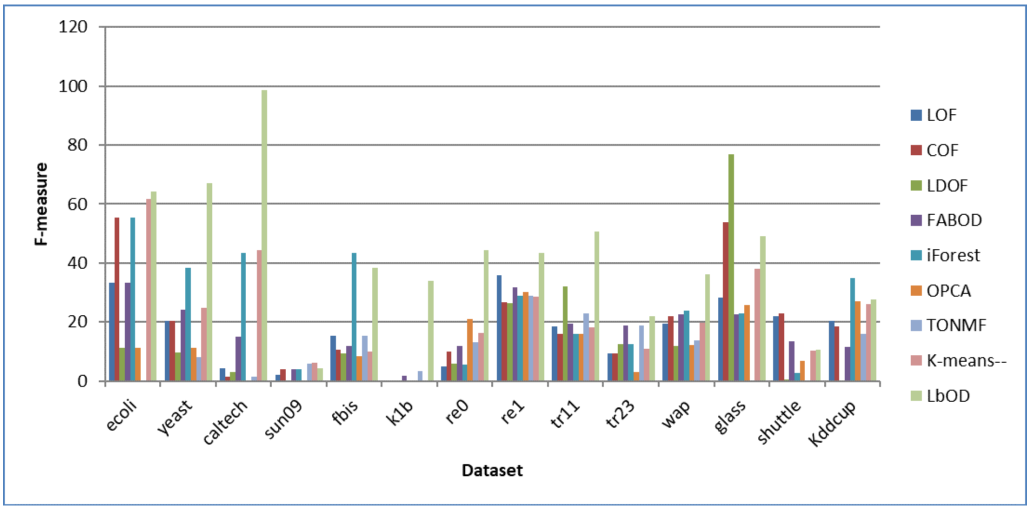

Figure 8: Performance comparison among more outlier detection methods in terms of F-measure

As president in Figure 8, the performance comparison among more outlier detection methods in terms of Jaccard is provided. It is observed from the results that the proposed outlier detection method showed better performance over existing methods against all the datasets used in experiments.

6. DISCUSSION

We introduce many topics on clustering with outlier reduction in this section. Traditionally, clustering separates a large number of points into discrete groups according to the degree of similarity between the points inside the same cluster. A firm or soft label is applied to each spot. Despite the fact that robust clustering is intended to lessen the influence of outliers, the cluster label is applied to every point, even outliers. In contrast, the issue we tackle in this work uses clustering to identify the outlier set and only assign labels to inliers. In technical terms, our method is a part of the non-exhaustive clustering, in which certain data points may belong to more than one cluster and not all data points are given labels. Our method has a different feature space difference from K-means. In addition to easily meeting the concepts of outliers and holoentropy, the partition space also facilitates the K-means optimization process' spherical structure assumption.

Extensive attempts have been done to flourish the hot study field of outlier detection from several angles. Very few of them do outlier identification and cluster analysis at the same time. With the exception of K-meansthe issue of grouping with outliers is expressed as an integer programming challenge using Langrangian Relaxation (LP) [18], where the input parameter is the cluster building costs. In addition to having extremely sophisticated algorithms, LP also has difficulty setting this parameter in real-world situations, which causes LP to produce impractical answers. For this reason, we are unable to disclose LP's performance within the part dedicated to experimentation. Our approach begins with the outlier detection objective function and uses a clustering tool to solve the problem. This shows how closely related the domains of cluster analysis and outlier identification are.

Several fundamental divisions are intended to be combined into one cohesive one via consensus clustering. An adaptive KCC utility function for a K-means system is provided for the difficult consensus clustering issue by our earlier work, Consensus Clustering (KCC) [49], [50]. An analogous collection of fundamental partitions serves as the input for our method, which uses K-means to produce the partition containing outliers. Combining basic partitions to establish consensus clustering and identifying outliers are made possible by the proposed partition space, which is formed from basic partitions. This perspective views holoentropy as the utility function that quantifies the degree of similarity between the final partition and the fundamental partition in B or $\tilde{B}$. The absence of values in fundamental partitions inside the KCC framework does not contribute to the centroid update and is not useful. For the proposed method, we are able to outliers automatically.

6. CONCLUSION AND FUTURE WORK

Our solution for effective detection involved the use of unsupervised machine learning (ML) of outliers from high dimensional datasets. An objective function is defined to improve cluster compactness leading to efficiency in outlier detection process. Further improvement of clustering process with problem transformation and usage of enhanced K-Means could result in an integrated approach that jointly archives quality clustering and outlier identification. We proposed an algorithm known as Learning based Outlier Detection (LbOD). Novelty of our algorithm lies in simultaneous approach in partition space, objective function and cluster optimization. A prototype is built to evaluate the proposed framework and algorithm for its ability to discover outliers considering multiple benchmark high dimensional datasets. Our empirical study has revealed that the LbOD algorithm outperforms many existing outlier detection techniques. We want to get better in the future our framework by exploiting the usage of ensemble of multiple best performing unsupervised learning models with novel selection strategy.

REFERENCES

[1] Jiawei Yang, Yu Chen, Sylwan Rahardja. (2023). Neighborhood representative for improving outlier detectors. Elsevier. 623, pp.192-205. [Online]. Available at: https://doi.org/10.1016/j.ins.2022.12.041.

[2] Antonella Mensi, David M.J. Tax, Manuele Bicegoa. (2023). Detecting outliers from pairwise proximities: Proximity isolation forests. Elsevier. 138, pp.1-12. [Online]. Available at: https://doi.org/10.1016/j.patcog.2023.109334.

[3] Jonatan Enes, Roberto R. Expósito, José Fuentes, Javier López Cacheir. (2023). A pipeline architecture for feature-based unsupervised clustering using multivariate time series from HPC jobs. Elsevier. 93, pp.1-20. [Online]. Available at: https://doi.org/10.1016/j.inffus.2022.12.017 [Accessed 27 February 2024].

[4] Mohammed H. Qais, Seema Kewat, K.H. Loo, Cheung-Ming Lai. (2024). Early outlier detection in three-phase induction heating systems using clustering algorithms. Elsevier. 15, pp.1-14. [Online]. Available at: https://doi.org/10.1016/j.asej.2023.102467 [Accessed 27 February 2024].

[5] Chen, Tingting; Liu, Xueping; Xia, Bizhong; Wang, Wei; Lai, Yongzhi (2020). Unsupervised Anomaly Detection of Industrial Robots Using Sliding-Window Convolutional Variational Autoencoder. IEEE Access, 8, pp.47072–47081. doi:10.1109/access.2020.2977892.

[6] Liu, Hongfu; Li, Jun; Wu, Yue; Fu, Yun (2019). Clustering with Outlier Removal. IEEE Transactions on Knowledge and Data Engineering, pp.1–11. doi:10.1109/TKDE.2019.2954317

[7] Carreño, Ander; Inza, Iñaki; Lozano, Jose A. (2019). Analyzing rare event, anomaly, novelty and outlier detection terms under the supervised classification framework. Artificial Intelligence Review, pp.1 –20. doi:10.1007/s10462-019-09771-y

[8] Waleed Hilal, S. Andrew Gadsden, John Yawney. (2022). Financial Fraud: A Review of Anomaly Detection Techniques and Recent Advances. Elsevier. 193(.), pp.1-34. [Online]. Available at: https://doi.org/10.1016/j.eswa.2021.116429

[9] Brito, L. C., Susto, G. A., Brito, J. N., & Duarte, M. A. V. (2022). An explainable artificial intelligence approach for unsupervised fault detection and diagnosis in rotating machinery. Mechanical Systems and Signal Processing, 163, pp.1-21. doi:10.1016/j.ymssp.2021.108105

[10] SADGALI, I.; SAEL, N.; BENABBOU, F. (2019). Performance of machine learning techniques in the detection of financial frauds. Procedia Computer Science, 148, pp.45–54. doi:10.1016/j.procs.2019.01.007.

[11] Stetco, Adrian; Dinmohammadi, Fateme; Zhao, Xingyu; Robu, Valentin; Flynn, David; Barnes, Mike; Keane, John; Nenadic, Goran (2018). Machine learning methods for wind turbine condition monitoring: A review. Renewable Energy, 133, pp.620-635, S096014811831231X–. doi:10.1016/j.renene.2018.10.047

[12] Meng, Fanrong; Yuan, Guan; Lv, Shaoqian; Wang, Zhixiao; Xia, Shixiong (2018). An overview on trajectory outlier detection. Artificial Intelligence Review, pp.1-20. doi:10.1007/s10462-018-9619-1

[13] Raghavan, Pradheepan; Gayar, Neamat El (2019). [IEEE 2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE) - Dubai, United Arab Emirates (2019.12.11-2019.12.12)] 2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE) - Fraud Detection using Machine Learning and Deep Learning. , pp. 334–339. doi:10.1109/ICCIKE47802.2019.9004231

[14] Md Abul Bashar;Richi Nayak; (2020). TAnoGAN: Time Series Anomaly Detection with Generative Adversarial Networks . 2020 IEEE Symposium Series on Computational Intelligence (SSCI), pp.1778-1785. doi:10.1109/ssci47803.2020.9308512

[15] Erhan, L.; Ndubuaku, M.; Di Mauro, M.; Song, W.; Chen, M.; Fortino, G.; Bagdasar, O.; Liotta, A. (2021). Smart anomaly detection in sensor systems: A multi-perspective review. Information Fusion, 67,pp. 64–79. doi:10.1016/j.inffus.2020.10.001

[16] Crimi, Alessandro; Bakas, Spyridon; Kuijf, Hugo; Keyvan, Farahani; Reyes, Mauricio; van Walsum, Theo (2019). [Lecture Notes in Computer Science] Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries Volume 11384 (4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II) || Deep Autoencoding Models for Unsupervised Anomaly Segmentation in Brain MR Images. , 10.1007/978-3-030-11726-9(Chapter 16), pp.161–169. doi:10.1007/978-3-030-11723-8_16

[17] Lukas Ruff;Jacob R. Kauffmann;Robert A. Vandermeulen;Gregoire Montavon;Wojciech Samek;Marius Kloft;Thomas G. Dietterich;Klaus-Robert Muller; (2021). A Unifying Review of Deep and Shallow Anomaly Detection . Proceedings of the IEEE, 109(5),pp.756–792,. doi:10.1109/jproc.2021.3052449

[18] Wenyu Zhang;Dongqi Yang;Shuai Zhang; (2021). A new hybrid ensemble model with voting-based outlier detection and balanced sampling for credit scoring . Expert Systems with Applications, 174,pp.1–13. doi:10.1016/j.eswa.2021.114744

[19] Harsh Dhiman;Dipankar Deb;S. M. Muyeen;Innocent Kamwa; (2021). Wind Turbine Gearbox Anomaly Detection Based on Adaptive Threshold and Twin Support Vector Machines . IEEE Transactions on Energy Conversion, pp.1-8– . doi:10.1109/tec.2021.3075897

[20] Avci, Onur; Abdeljaber, Osama; Kiranyaz, Serkan; Hussein, Mohammed; Gabbouj, Moncef; Inman, Daniel J. (2021). A review of vibration-based damage detection in civil structures: From traditional methods to Machine Learning and Deep Learning applications. Mechanical Systems and Signal Processing, 147, pp1-45,. doi:10.1016/j.ymssp.2020.107077

[21] Jiawei Yang;Susanto Rahardja;Pasi Fränti; (2021). Mean-shift outlier detection and filtering . Pattern Recognition, 115,pp1-11,. doi:10.1016/j.patcog.2021.107874

[22] ZubaroÄŸlu, Alaettin; Atalay, Volkan (2020). Data stream clustering: a review. Artificial Intelligence Review, pp.1-36 –. doi:10.1007/s10462-020-09874-x

[23] Chakraborty, Debasrita; Narayanan, Vaasudev; Ghosh, Ashish (2019). Integration of Deep Feature Extraction and Ensemble Learning for Outlier Detection. Pattern Recognition, 89,pp.161-171, S0031320319300020–. doi:10.1016/j.patcog.2019.01.002

[24] Thangaramya, K.; Kulothungan, K.; Indira Gandhi, S.; Selvi, M.; Santhosh Kumar, S. V. N.; Arputharaj, Kannan (2020). Intelligent fuzzy rule-based approach with outlier detection for secured routing in WSN. Soft Computing, pp.1-15 –. doi:10.1007/s00500-020-04955-z

[25] Belhadi, Asma; Djenouri, Youcef; Srivastava, Gautam; Djenouri, Djamel; Lin, Jerry Chun-Wei; Fortino, Giancarlo (2020). Deep learning for pedestrian collective behavior analysis in smart cities: A model of group trajectory outlier detection. Information Fusion, pp.1-32, S1566253520303316–. doi:10.1016/j.inffus.2020.08.003

[26] Landauer, Max; Skopik, Florian; Wurzenberger, Markus; Rauber, Andreas (2020). System log clustering approaches for cyber security applications: A survey. Computers & Security, 92, pp.1-23–. doi:10.1016/j.cose.2020.101739

[27] Djenouri, Youcef; Belhadi, Asma; Lin, Jerry Chun-Wei; Djenouri, Djamel; Cano, Alberto (2019). A Survey on Urban Traffic Anomalies Detection Algorithms. IEEE Access, 4, pp.1–13. doi:10.1109/ACCESS.2019.2893124

[28] Tang, Tinglong; Chen, Shengyong; Zhao, Meng; Huang, Wei; Luo, Jake (2018). Very large-scale data classification based on K-means clustering and multi-kernel SVM. Soft Computing, pp.1-9 –. doi:10.1007/s00500-018-3041-0

[29] Munoz-Organero, Mario (2019). Outlier Detection in Wearable Sensor Data for Human Activity Recognition (HAR) Based on DRNNs. IEEE Access, pp.1–17. doi:10.1109/ACCESS.2019.2921096

[30] Yongyong Chen;Xiaolin Xiao;Chong Peng;Guangming Lu;Yicong Zhou; (2022). Low-Rank Tensor Graph Learning for Multi-View Subspace Clustering . IEEE Transactions on Circuits and Systems for Video Technology, pp.1-13, doi:10.1109/tcsvt.2021.3055625

[31] Fitriyani, Norma Latif; Syafrudin, Muhammad; Alfian, Ganjar; Rhee, Jongtae (2020). HDPM: An Effective Heart Disease Prediction Model for a Clinical Decision Support System. IEEE Access, 8, pp.133034–133050. doi:10.1109/access.2020.3010511

[32] Rogers, T.J.; Worden, K.; Fuentes, R.; Dervilis, N.; Tygesen, U.T.; Cross, E.J. (2019). A Bayesian non-parametric clustering approach for semi-supervised Structural Health Monitoring. Mechanical Systems and Signal Processing, 119, pp.100–119. doi:10.1016/j.ymssp.2018.09.013

[33] Thöle, Lena M.; Amsler, H. Eri; Moretti, Simone; Auderset, Alexandra; Gilgannon, James; Lippold, Jörg; Vogel, Hendrik; Crosta, Xavier; Mazaud, Alain; Michel, Elisabeth; Martínez-García, Alfredo; Jaccard, Samuel L. (2019). Glacial-interglacial dust and export production records from the Southern Indian Ocean. Earth and Planetary Science Letters, 525(), pp.1-14–. doi:10.1016/j.epsl.2019.115716

[34] Fitriyani, Norma Latif; Syafrudin, Muhammad; Alfian, Ganjar; Rhee, Jongtae (2019). Development of Disease Prediction Model Based on Ensemble Learning Approach for Diabetes and Hypertension. IEEE Access, 7, pp.144777–144789. doi:10.1109/access.2019.2945129

[35] Deepak, K.; Chandrakala, S.; Mohan, C. Krishna (2020). Residual spatiotemporal autoencoder for unsupervised video anomaly detection. Signal, Image and Video Processing, pp.1-8,. doi:10.1007/s11760-020-01740-1

[36] Nivedita Mishra;Sharnil Pandya; (2021). Internet of Things Applications, Security Challenges, Attacks, Intrusion Detection, and Future Visions: A Systematic Review . IEEE Access, 9, pp.59353-59377,. doi:10.1109/access.2021.3073408

[37] Kraus, M.; Weiler, N.; Oelke, D.; Kehrer, J.; Keim, D. A.; Fuchs, J. (2019). The Impact of Immersion on Cluster Identification Tasks. IEEE Transactions on Visualization and Computer Graphics, pp. 1–11. doi:10.1109/TVCG.2019.2934395

[38] Liu, Yezheng; Li, Zhe; Zhou, Chong; Jiang, Yuanchun; Sun, Jianshan; Wang, Meng; He, Xiangnan (2019). Generative Adversarial Active Learning for Unsupervised Outlier Detection. IEEE Transactions on Knowledge and Data Engineering, pp.1–12. doi:10.1109/TKDE.2019.2905606

[39] Wang, Hongzhi; Bah, Mohamed Jaward; Hammad, Mohamed (2019). Progress in Outlier Detection Techniques: A Survey. IEEE Access, 7, pp.107964–108000. doi:10.1109/ACCESS.2019.2932769.

[40] Usama, Muhammad; Qadir, Junaid; Raza, Aunn; Arif, Hunain; Yau, Kok-lim Alvin; Elkhatib, Yehia; Hussain, Amir; Al-Fuqaha, Ala (2019). Unsupervised Machine Learning for Networking: Techniques, Applications and Research Challenges. IEEE Access, pp.1–37. doi:10.1109/ACCESS.2019.2916648.

[41] H. Liu, J. Wu, D. Tao, Y. Zhang, and Y. Fu, “Dias: A disassemble assemble framework for highly sparse text clustering,” in Proceedings of SIAM International Conference on Data Mining, 2015.

[42] H. Liu, M. Shao, S. Li, and Y. Fu, “Infinite ensemble clustering,” Data Mining and Knowledge Discovery, no. 1-32, 2017.

[43] UCI Datasets. Retrieved from https://archive.ics.uci.edu/datasets

[44] S. Chawla and A. Gionis, “k-means-: A unified approach to clustering and outlier detection,” in Proceedings of SIAM International Conference on Data Mining, 2013.

[45] S. Wu and S. Wang, “Information-theoretic outlier detection for largescale categorical data,” IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 3, pp. 589–602, 2013.

[46] A. Strehl and J. Ghosh, “Cluster ensembles — a knowledge reuse framework for combining partitions,” Journal of Machine Learning Research, vol. 3, pp. 583–617, 2003

[47] H. Liu, T. Liu, J. Wu, D. Tao, and Y. Fu, “Spectral ensemble clustering,” in Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2015.

[48] H. Liu, M. Shao, S. Li, and Y. Fu, “Infinite ensemble for image clustering,” in Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.

[49] A. Banerjee, S. Merugu, I. Dhillon, and J. Ghosh, “Clustering with bregman divergences,” Journal of Machine Learning Research, vol. 6, pp. 1705–1749, 2005