Data Decomposition for Outlier Detection coupled with Information Theoretic Validation

Gouranga Duari [a,*] and Rajeev Kumar [b]
[a] *Data to Knowledge Lab(D2K), School of Computer and System Sciences Jawaharlal Nehru University, Delhi, India*
E-mail: gourangaduari5@gmail.com; ORCID: https://orcid.org/0000-0002-5872-842X
[b] *Data to Knowledge Lab(D2K), School of Computer and System Sciences, Jawaharlal Nehru University, Delhi, India*
E-mail: rajeevkumar.cse@gmail.com; ORCID: https://orcid.org/0000-0003-0233-6563

Abstract. Decomposition for complexity minimization has long been a challenging approach. yet decomposition for outliers has rarely been experimented with. This paper presents a data decomposition approach as a pre-processor for outlier detection. The decomposition of the data using space partitioning makes homogeneous sub-groups. Consequently, it reduces the complexity of data patterns by isolating possible outliers into the sub-groups from monolithic characters. This approach creates sub-groups of homogeneous data points based on the fitness of purpose. They optimize the outlier patterns in the sub-groups for subsequent mapping of outlier detectors onto the sub-groups. This decomposition strategy is found to be effective in reducing the complexity of learning for the detectors without deterioration in the overall detection rate. We experimented with this approach using different benchmark detectors on eight benchmark datasets. Our data decomposition approach is superior for identifying localized patterns in the partitions and offers a better generalization.

Keywords: Decomposition, Space partitioning, Outlier, Entropy

1. Introduction

Outlier detection is one of the most important pattern recognition tasks in data science for data-driven-decision making. Detection of outliers seeks to identify the unique or rare instances that deviate significantly from most data. The objective is to identify and flag the exceptional or uncommon objects compared to most of the data. Misclassification of any such single event can be catastrophic in critical applications, e.g., in social network [11], in Streaming data [13], in medical diagnosis [12], in geoscience [30]. Real-life applications demand that a single outlier instance should not remain undetected even though it is weak.

Various outlier detection techniques have been proposed in the past tailored to different applications' specific characteristics and requirements. Outlier detection tasks are commonly classified into three categories: supervised, semi-supervised, and unsupervised, depending on the availability of outlier labels. Unsupervised methods are extensively used in outlier detection, primarily due to the shortcomings of

*Corresponding author. E-mail: gourangaduari5@gmail.com.

obtaining accurate and representative labels, which are often expensive and scarce. Unsupervised outlier detection methods can be categorized into six major types based on their underlying approaches: Linear models [41], clustering-based methods [29], probabilistic-based method [31] information-theoretic methods [51], neural network-based methods [36], isolation-based methods [28], and nearest neighbour-based methods [3]. These methods have been developed and refined to suit the unique features of each application, taking into account factors such as data distribution, data type, domain knowledge, and desired level of sensitivity.

Although significant development is witnessed in unsupervised outlier detection, it remains a challenging and exciting problem for the pattern recognition community. Outlier detection becomes challenging due to a lack of prior knowledge and a highly imbalanced class. Complexity further inflates because of local irregularities and the boundary effect of defining outliers. In handling such problems, conventional outlier detection methods do not perform effectively. To overcome such limitations, we propose a data decomposition [25] of pattern space aimed at getting a more robust outcome by partitioning the data into sub-groups of homogeneous elements. In Fig. 1, we present the data decomposition effect in $2-d$ projection of d-dimensional pattern space where boldfaced points may be undetected by the conventional method because of local irregularities. By other pattern projections, such as partitioning the monolithic space into sub-groups, those points become potential outliers in that partitioned space because of isolation.

A standard decomposition method isolates anomalous data points into sub-groups based on the inherent characteristics of data points. Such decomposition characteristics are expected to give group-wise detection efficiency to the outlier detectors. To achieve the data decomposition goal, we use the standard clustering [19] method as proof of concept. Here, the clustering method works as a pre-processor for outlier detectors to reduce the data pattern's complexity. It creates a smooth ground for the outlier detectors to learn the patterns in the homogeneous groups of data points in sub-groups and improve the outlier scores. Our approach investigates a pre-processing framework for outlier detection inspired by the *Learning-follows-decomposition* (LFD) [25] strategy through clustering based on the fitness of purpose as it considers homogeneity condition while making sub-groups than other clustering methods. The principle characteristic of our data decomposition modulation is that an outlier detector can take advantage of the decomposition [32]. Experimentally, our approach alleviates the drawbacks mentioned above. We experimented on eight benchmark datasets and six standard outlier detection methods to establish whether the above decomposition strategy produced a conducive environment for the detector to perform better. The proposed approach performs better on almost all the datasets and detectors. Further, to strengthen our proposed approach, we analyze the pattern complexity in the subgroups using an information-theoretic method (i.e., entropy). We do some empirical experimentation to look into the intensity of complexity reduction in the decomposed space to validate our decomposition strategy for outlier detection. This step enhances the acceptability of our proposed approach and gives better logical support for our step.

The research contributions of the proposed work are: (i) data decomposition is more pronounced to create patterns of outliers in sub-groups so that the detection process becomes more accessible for the detectors, and (ii) Outlier-clusters make outlier detection trivial and thus, outliers could be detected effectively from the decomposed clusters.

The rest of the paper is organized as follows. Section 2 tells us about the significance of our approach. Section 3 mentions a few existing methods related to decomposition. Section 4 describes the proposed methodology. Section 5 reports the experimental setup and empirical results. Finally, Section 6 concludes the paper.

Table 1
Abbreviation and Description

Abbreviation	**For**
O-cluster	Outlier cluster (proposed)
AD	After Data Decomposition (proposed)
WD	Without Data Decomposition (proposed)
LOF	Local outlier factor [4]
COF	Connective-based outlier factor [43]
IForest	Isolation Forest [28]
COPOD	Copula-Based Outlier Detection [26]
PCA	PCA-based outlier detector [41]
*k*NN	*k*-Nearest Neighbor (*k*NN) based outlier detector [3]

2. Motivation

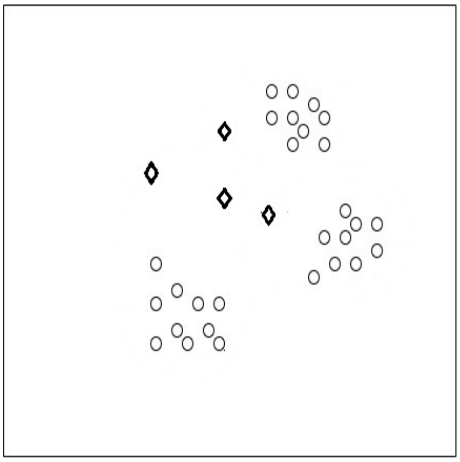
Is there any pre-processing approach concerning the outlier detection that makes the outlier detectors more robust? Our motivation is based on the following factors:

- Understanding the inherent pattern of data is crucial before detection and how data-centric information can help the outlier detection process through partitioning. We want to examine the inherent specific outlier pattern (Fig. 1), which can cause systematic measurement failure for the detectors.
- Outliers are rare events to identify in the different data types. The border effect and dense data locality adversely affect the identification of outliers. How the data modulation in homogeneous sub-groups (Fig. 1) before detection can help smooth and effective the detection process.
- As a part of the complexity reduction and quality enhancement process, how data decomposition [25] can reduce learning complexity by increasing decision surface and as a consequence, it can reduce local irregularities for the outlier detector and increase classification accuracy in different data distribution.

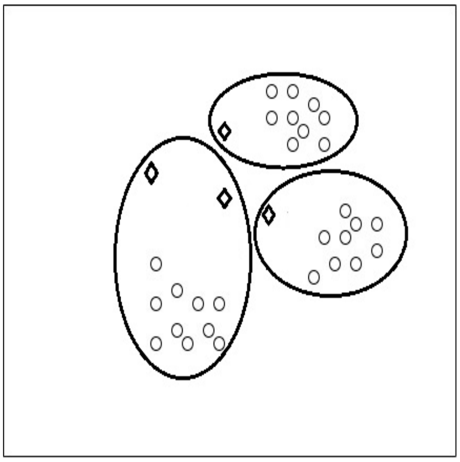
3. Related Work

In the past decade, various methods have been developed for outlier detection under the unsupervised category. Among the recent developments, Cheng et al. [8] proposed an ensemble-based detector for global and local outliers. Recently, Li et al. [27] studied ECOD (Empirical-Cumulative-distribution-based Outlier Detection). Wang et al. [46] used a virtual graph-based outlier detection method. An exclusive survey of model-based outlier detection techniques has been presented recently by Wang et al. [48]. As our work concentrates on data decomposition and subsequent mapping of outlier detectors, we restrict the rest of the related work to the same category.

A well-known *k*-nearest neighbor (*k*NN) based approach [3] computes the distances between data points, and a data point with a significantly higher distance value from its nearest neighbors based on a threshold is regarded as an outlier. An efficient version of the distance-based method is proposed by Ramaswamy et al. [39]. They partition the data and remove parts that cannot contain outliers, thus reducing the computation and improving efficiency. Breunig et al. [4] developed a Local Outlier Factor (LOF) to identify outliers based on the density approach. The principle behind the density approach



(a) Synthetic data



(b) Data decomposition

Figure 1. Data decomposition effect on outliers. Outliers and inliers are presented by boldfaced diamond and circle symbols, respectively.

is that outlier data points are likely to occur in the low-density region while the normal data points are found in dense spaces. Tang et al. [43] proposed COF, an improved version of LOF [4] based on chaining distance. Liu et al. [28] proposed a unique isolation-based model, and they observed that outliers are present in the vicinity of the trees' roots due to their isolation. Inliers are found closer to the terminal nodes of the trees. Shyu et al. [41] proposed a PCA-based outlier approach. He et al. [17] designed a cluster-based local outlier factor (CBLOF) based on the concept of a cluster-based local outlier.

A Learning-follows-Decomposition (LFD) strategy [25] for hierarchical learning of pattern spaces uses a multi-objective genetic algorithm followed by (near-) optimal learning of pattern sub-spaces. Their technique is a generic solution to complex high-dimensional problems where clusters are generated based on the fitness of purpose. This strategy splits a problem into a series of sub-problems; it then assigns a set of function approximators to each sub-problem such that each module specializes in a subdomain to learn the pattern. Maimon et al. [32] outlined a brief overview of the decomposition methods by presenting the essential properties that characterize various decomposition frameworks and their respective benefits. In a different vein, Paulheim and Meusel introduced an alternative method for outlier identification called ALSO (attribute-wise learning for scoring outliers) [37]. Rather than relying on density-based measures, ALSO examines patterns within the data. The authors decompose the outlier detection problem into supervised learning tasks, enabling the identification and evaluation of patterns' strengths within each attribute. Weight assignments are made to attributes based on these strength estimations, with weaker or nonexistent patterns receiving lower weights. Outliers are identified by comparing each data point against the established patterns, considering the attribute weights. Any data point deviating significantly from the patterns is classified as an outlier. Jiang et al.

[21] proposed a K-means a clustering-based two-phase method to detect outliers. The first phase involves partitioning data points. The second phase consists of constructing a minimum spanning tree (MST) based on the cluster centers obtained in the first phase. Outliers are identified as clusters located in small sub-trees. Gan et al. [15] proposed an approach that combines data clustering and outlier detection by augmenting the k-means algorithm with an extra "cluster" to accommodate outliers. To optimize the objective function of this enhanced algorithm, They have developed an iterative procedure and

demonstrated its convergence. Chawla et al. [7] proposed a *k*-means-based solution to cluster the data and identify outliers simultaneously. They apply the *k*-means algorithm; the data points in the clusters far from their nearest cluster centers are considered outliers. This method converges to the local optimum of outliers by an iterative method. However, *K*-means the algorithm is vulnerable to outliers, and such outliers may have a disproportionate impact on the final cluster configuration. This can result in many classification errors. Jiang et al. [21] proposed a *K*-means a clustering-based two-phase method to detect outliers. The first phase involves partitioning data points. The second phase consists of constructing a minimum spanning tree (MST) based on the cluster centers obtained in the first phase. Outliers are identified as clusters located in small sub-trees.

Gan et al. [15] proposed an approach that combines data clustering and outlier detection by augmenting the *k*-means algorithm with an extra "cluster" to accommodate outliers. To optimize the objective function of this enhanced algorithm, They have developed an iterative procedure and demonstrated its convergence. Chawla et al. [7] proposed a *k*-means-based solution to cluster the data and identify outliers simultaneously. They apply the *K*-means algorithm; the data points in the clusters far from their nearest cluster centers are considered outliers. This method converges to the local optimum of outliers by an iterative method. However, *k*-means the algorithm is vulnerable to outliers, and such outliers may have a disproportionate impact on the final cluster configuration. This can result in many classification errors. Jiang et al. [21] proposed a *k*-means a clustering-based two-phase method to detect outliers. The first phase involves partitioning data points. The second phase consists of constructing a minimum spanning tree (MST) based on the cluster centers obtained in the first phase. Outliers are identified as clusters located in small sub-trees.

A K Jain [20] presented a comprehensive examination of pattern clustering methods, focusing on statistical pattern recognition principles. The objective was to offer practical guidance and references to foundational concepts that were accessible to a wide range of clustering practitioners. A taxonomy of clustering techniques was presented alongside an exploration of overarching themes and recent developments. Kashef and Warraich [23] introduced a novel taxonomy of distributed data clustering algorithms, shedding light on their distinct distributed modeling strategies. Their taxonomy categorized these clustering processes into homogeneous and heterogeneous approaches. Additionally, they investigated diverse distributed performance and quality measures, offering a comprehensive understanding of their effectiveness. Buhmann and Kühnel [5] presented a complexity-optimized approach using a clustering strategy that explicitly considered the balance between the simplicity and accuracy of data representation. This approach involved jointly optimizing distortion errors and complexity costs within the clustering algorithm. By employing a maximum entropy estimation of the clustering cost function, they determined the optimal number of clusters, their positions, and their corresponding probabilities. This method ensured an effective trade-off between precision and simplicity in data representation. Wu and Wang [49] introduced a novel approach to outlier detection, which involved formulating a rigorous definition of outliers and devising an optimization model using the concept of holoentropy. This approach incorporated both entropy and total correlation, offering a comprehensive perspective on outlier identification. Leveraging this model, they presented a streamlined outlier factor computation that was uniquely identifiable and amenable to efficient updates.

4. The Proposed Methodology

The existing detection method works on the assumption that outliers can only be identified using deviation characters, but it ignores entirely the local structure of data. Moreover, it becomes tough to identify outliers in a monolithic space because of its pattern complexity and above mentioned assumption. This paper proposed a decomposition-based unsupervised method to detect outliers very effectively. First, we partition the data into sub-groups or clusters using standard clustering. As clustering technique creates clusters of homogeneous elements, and the homogeneous grouping of data points fits the outliers in the different clusters based on the clustering criteria. Second, we check the subgroup's structure and characteristics for possible fitting of detectors onto the subgroups. Third, we employ a standard outlier detector in each cluster based on defined criteria. Finally, we outline the evaluation framework for our approach based on the standard metrics. We present extensive empirical results over eight benchmark data sets. We establish the competency of our approach in the detection of outliers. We also show that the generalization of our approach using six heterogeneous standard outlier methods is quite effective.

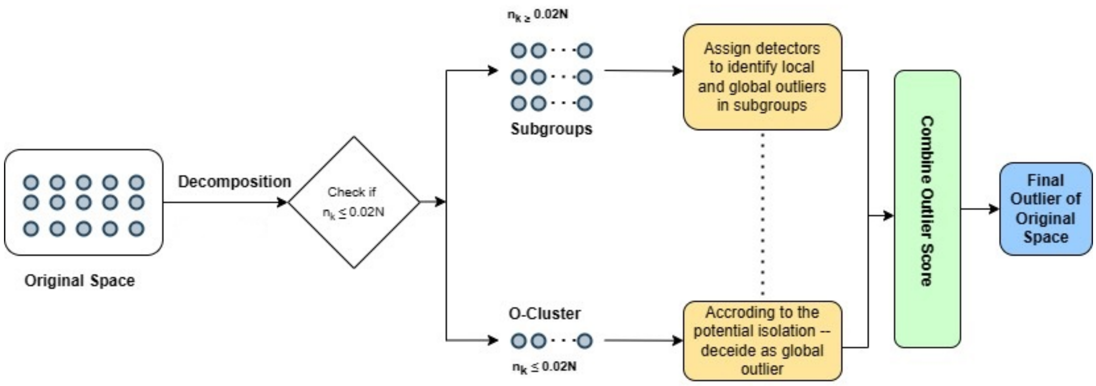


Figure 2. Workflow of the algorithm.

We aim to pre-process the input data by decomposing it into sub-groups of homogeneous data points and detect outliers in the subsequent sub-groups using a standard outlier detection method. Such preprocessing of input data is expected to create a conducive environment for the detector to yield effective and efficient output. In this work, we assume the heterogeneity of unnatural input data with global [24], local [4] type, and unnatural data points distributed in the clusters according to their characteristics. Since k-means [44] clustering tries to separate data points in k groups of equal variance by minimizing inertia or within-cluster sum-of-squares criterion. As inertia assumes that clusters are convex and isotropic, it is expected that separation is appropriately done by the k-means clustering method. But it is not always the case. Sometimes, k-means clustering responds poorly to irregular shapes. Considering this fact, We consider the following two cases:

- **Case 1**: If a few isolated data points create a separate cluster of tiny size (less than equal to 2% of total data points), we treat these data points as an outlier. As outliers are usually located in low-density regions than the normal data points, these highly isolated data points are strong candidates to be outliers compared to other data points. So, they can be treated as outliers by definition [4]. Here, we refer to these clusters as outlier-cluster (O-cluster), i.e., O-cluster is one of the sub-groups containing only outliers. So, there is no need to assign any detector as it categorically classifies the potential outliers. The reasoning behind using a 2% limit for the O-cluster is based on experiments on benchmark datasets, such as the Pageblock dataset, which shows that using a 2% limit results in all data points in the O-cluster being outliers according to the dataset's ground truth. Using a higher limit can result in some normal data points becoming part of the O-cluster, which does not satisfy the criteria for an O-cluster and can lead to unexpected results in the overall decomposition process.

- **Case 2**: Clusters with a significant number of data points that have complex data patterns with the local and global patterns of outliers, which signifies Case 2. This type of cluster requires treatment for depth dive to identify more localized patterns in homogeneous data. This homogeneous data modulation makes few localized outlier points potentially deviated in decomposed space, and their isolation from normal points is quite significant. That is why we consider those clusters suitable for assigning standard outlier detectors to identify more unnatural events or outliers.

4.1. Principle of the Proposed Approach

The proposed approach is based on the assumption that local outliers are the potential global outliers in the subgroup after decomposition. Local outliers can have more precise, stronger isolation in the subgroups after decomposition than its monolithic space. As showcased in Fig. 3, O_1, O_2, and O_3 are the localized outlier patterns that are very tough to identify in monolithic space, but they can behave strong outliers in the partitioned space considering its isolation from inliers. The isolation radius (IR), which is the radius of the isolation region, has great importance in the context of outliers here. Smaller IR may indicate strong outliers in their individual partitioned subgroups, but higher IR values may indicate strong global outliers for all the partitioned subgroups, and here we mention it as an O-cluster or Outlier cluster. This is presented here as O_4 in Fig. 3 (a) and (c). This way, the original space can be transformed into a better, smoother, less complex space for the outlier detectors to identify, like cake and butter.

Based on the above-mentioned concept, the original monolithic space can be presented as a set of a few well-designed subgroups of similar characters. As shown in the Fig. 3 (d), $S_i(i \in [1,k])$ is the isolation region comprised of X and $x_j(j \in [1,N])$ as center of region. $B_j(j \in [1,N])$ are the modified isolation measure in the partitioned subgroups.

4.2. Analytical Formulation

Here, we define our proposed method using mathematical notions. To outline our algorithm, let $X = \{x_1, x_2, ...x_N\}$ be a dataset containing of N data points with d dimension and we also consider a distance function $d : X \times X \to \mathbf{R^d}$ in the d-dimensional Euclidean space $\mathbf{R^d}$. The Euclidean distance is represented between the pairs of data points in X as: $d(x_i, x_j) = (\sum_{t=1}^{d}(x_{it} - x_{jt})^2)^{\frac{1}{2}}$, where $x_i = (x_{i1}, x_{i2}....., x_{id})$ is the representation of x_i in $\mathbf{R^d}$.

4.3. Decomposition Approximation

In our approach towards reducing complexity for the outlier detectors, we intend to decompose input data into sub-groups and subsequently map detectors in the sub-groups for learning. If we consider input data decomposition as mapping F from an (N, d)-shaped input data (X) to k sub-groups (n_k, d), then the following formulation is a (N, d)-shaped function decomposition into many (n_k, d)-shaped sub-groups subject to meeting the criteria Objective $f(X)$.

$$F : \mathbf{R^d_N} \to \bigcup_k \mathbf{R^d_{n_k}}, n_k \leqslant N \tag{1}$$

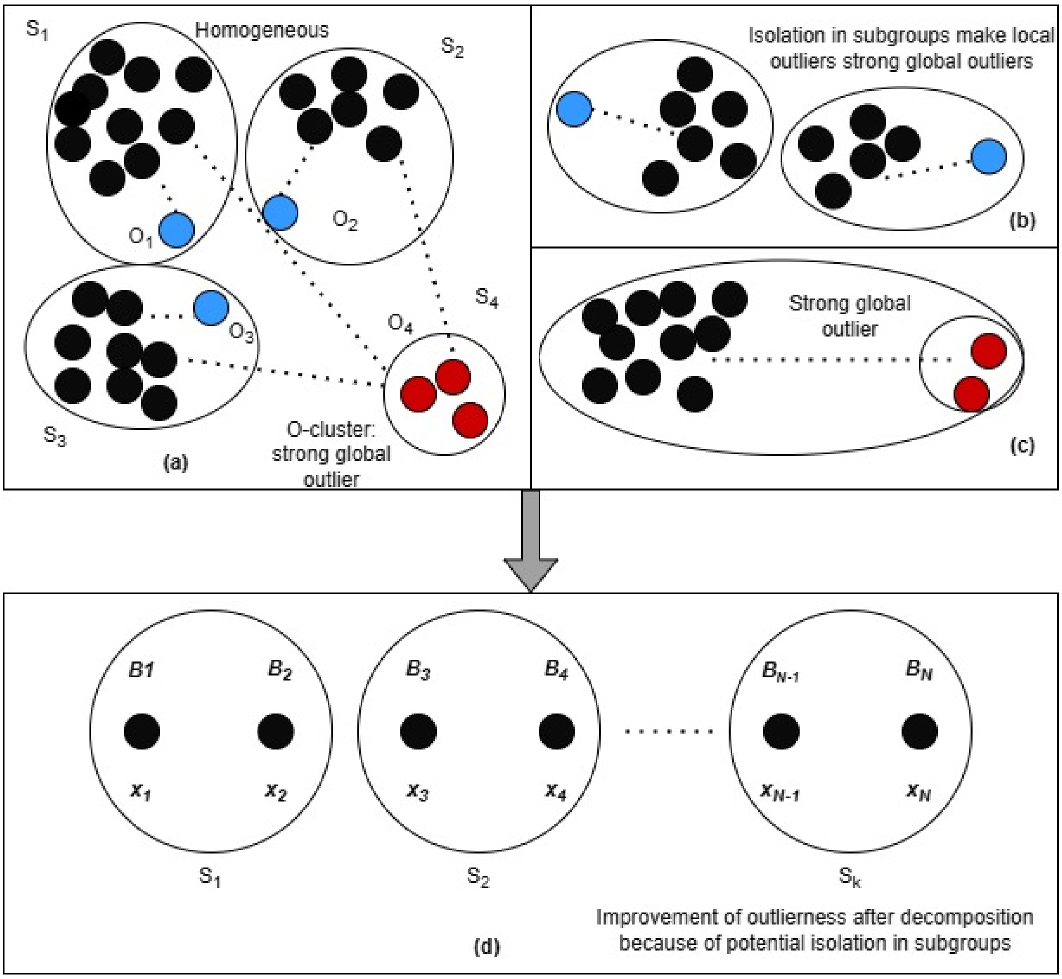


Figure 3. Principle of the proposed approach.

As data decomposition modularity minimizes the hypersphere into smaller sub-groups, they represent a comparatively lesser complex domain for the detectors to learn patterns, and a detector exercises less effort to get better output. Fig. 1 depicts the benefits of decomposition into sub-groups.

Here, our objective $f(X)$ is to increase the homogeneity of sub-groups and decrease the learning complexity, i.e., data points should be of the same characteristic. We wish to maximize the homogeneity of data points by distributing them in sub-groups based on defined decomposition criteria. As a consequence of decomposition, outlier data points of homogeneous character cater according to their sub-groups. For the decomposition step, we choose standard k-means clustering with four different configurations of k. Let $Z = \{z_1, z_2, ...z_k\}$ be a set of potential cluster centers that are used for the decomposition of dataset X. The distance of $x \in X$ to its closest cluster center $z(x \mid Z)$ is represented by:

$$d(x \mid Z) = \min_{z \in Z}\{d(x, z)\} \tag{2}$$

Here, the input data (X) is decomposed into k cluster by minimizing the within-cluster sum-of-squares criteria, which is given by:

$$U(X,Z) = \sum_{i=0}^{n} d(x \mid Z) = \sum_{i=0}^{n} \min_{z_K \in Z}(\|x_i - z_K\|^2) \tag{3}$$

Let the number of data points in the clusters be n_k, a decisive factor for assigning a detector to the clusters. Let $A = \{a_1, a_2, ..., a_k\}$ is a set of detected outliers in the respective clusters. We experiment with our data decomposition strategy in four different configuration values of k (number of clusters), so the detected outlier in the four different configurations of decomposition is represented by $D_j(A)\ \forall j \in \{2,3,4,5\}$.

4.4. Algorithmic Description

Our data decomposition method is described in Algorithm 1. We take six different types of outlier detectors and four different configurations for data decomposition, chosen sequentially in a fixed number of clusters. Here, we attempt to find complex outlier patterns for all four decomposition configurations.

Algorithm 1 Data Decomposition

Input: $M_1, M_2 ...., M_m$:= m-sets of heterogeneous outlier detector, Set of points $X = \{x_1, x_2 ..., x_n\}$
Output: A = identified outliers
1: $A = \phi$
2: $D_j(A) = \phi$
3: **for** each ($j \in \{2,3,4,5\}$) **do**
4: perform decomposition of k subgroups *\/* equation 1 and equation 2*
5: calculate $n_k\ \forall k$
6: **if** ($n_k \leqslant 0.02N$) **then** */* check for O-cluster*
7: $z_k = a_k \subseteq A\ \forall k$
8: **else**
9: assign any M_m */* employ detector*
10: Check a_k in each z_k
11: **end if**
12: $A = a_1 \cup a_2 \cup .... \cup a_k$
13: **end for**
14: **return** $D_j(A)\ \forall j \in \{2,3,4,5\}$

For each clustering configuration of k, we decompose the data points in the sub-groups or clusters using k-means (Lines 3-4). Then we calculate the size of these clusters (Line 5). Then, we check the condition of the O-cluster (Line 6); if it satisfies, we do not assign detectors. Otherwise, we assign detectors in the clusters (Line 9) and detect outliers (Line 10) in the clusters. Finally, we integrate all the clusters (Line 12) for evaluation measures.

4.5. Information Theoretic Validation

To further validate the pattern complexity reduction in the subgroups or decomposed space, we use information-theoretic methods (i.e., entropy).

Mathematically, Let the dataset X consists of subgroups $S_1, S_2, ..., S_k$. H_{com} be the combined entropy of the entire data set or monolithic data. Let H_k be the entropy of the k-th subgroup after decomposition. Here, our objective is to prove:

$$H_{com} \geqslant \sum_{k} H_k \tag{4}$$

Where $H_{com} = -\sum_{i=1}^{N} P_x log_2 P_x$, and $H_k = -\sum_{x_i \in S_k} P_{x_i} log_2 P_{x_i}$. P_x is the probability of the event $x \in X$ and P_{x_i} is the probability of the event in $x_i \in S_k$. So, we get from equation 4 that

$$-\sum_{i=1}^{N} P_x log_2 P_x \geqslant \sum_{k} (-\sum_{x_i \in S_k} P_{x_i} log_2 P_{x_i}) \tag{5}$$

Now, we will calculate the P_x and P_{x_i} from the dataset and subgroups by using Euclidean distance. Our primary goal is to validate the homogeneity concept and further reduce complexity. We will calculate the Euclidean distance of each data point from the respective mean, and that value will be normalized using $Norm_x = \frac{x-\mu}{\sigma}$. We will transform the normalized value into the predictive probability of that data point by using the formula $P_x = \frac{1}{1+\exp^{-x}}$. Here, we try to establish our fitness of purpose by looking into homogeneous characteristics of subgroups by checking their distance values from the respective mean. It is assumed that subgroups may have lower pattern complexity than entire data after decomposition. Let's express Euclidean distance for this purpose, and that is $d_X = \sqrt{(X-\mu)^{\frac{1}{2}}}$. Where μ is the mean of the data points. We calculate distance values for X and subgroups $S_1, S_2, ..., S_k$. Then, these distance values feed in the $Norm_x$ for the normalization, and then these are transformed into probabilities using P_x. Then we calculate entropy H_{com} and H_k using the probability values, and we do our analysis on complexity reduction.

The proposed decomposition method for outlier detection is validated with the help of entropy measures for all the datasets to build a strong foundation for our claim and give a logical sequence of our investigation. We take the decomposition setting of four subgroups for all the datasets. The algorithm 2 of the data decomposition validation is explained below:

4.6. Learning Complexity

The complexity of most of the unsupervised outlier detectors is approximate of the order $O(N^2)$, where N is the number of data points. For any data pattern, learning complexity after using data decomposition is: $O(n_1^2) + O(n_2^2) + ... + O(n_k^2) \leqslant O(N^2)$. Our objective behind data decomposition is to reduce the sum of squares using sub-groups. We assume that each sub-group is an independent event in a statistical sense. Separability measure [14] of data into sub-groups preserve the inherent pattern space intact, increasing the decision surface's regularity. Consequently, this data decomposition step surges classification accuracy.

Algorithm 2 Data Decomposition Validation

Input: $S_1, S_2 ...., S_m$:= k-sets of subgroups, Set of points $X = \{x_1, x_2 ..., x_n\}$, k:= number of subgroups, μ:= mean
Output: H_{com} = calculated combined entropy of X and H_k = calculated entropy of subgroup(S_k)

```
1:  H_com = ϕ
2:  H_k = ϕ
3:  for each S_k ∀k and X do
4:      Calculate μ
5:      Calculate d_X                         /* compute distance from group mean
6:      Standardize d_X using Norm_x
7:      Get P_x from Norm_x                   get probabilities
8:      Calculate H_com and H_k using P_x     /* get entropy
9:      Check H_com ⩾ Σ_k H_k                 /* equation 5
10: end for
11: return H_com and Σ_k H_k
```

5. Experimental Setup and Empirical Results

In this section, We present the experimental setup of our proposed approach and experimental results on meaningful benchmark datasets, which can be easily traceable at the UCI repository. We use six heterogeneous outlier detectors in our approach. We have done our entire experiment using the *Jupyter* notebook [1], and visualization is generated using the *Plotly* library.

5.1. Dataset Description

In this work, we use eight benchmark datasets from the UCI machine learning repository [2].

Table 2
Summary of the used datasets.

Dataset	**Instances**	**Dimension**	**Outliers (%)**
Pendigits	6870	16	156(2.27%)
Pageblocks	5471	11	560(10.23%)
Optdigit	5216	64	150(3%)
Waveform	3443	21	100(2.9%)
Thyroid	7200	6	536(7.42%)
Letter	1599	32	100(6.25%)
Satimage	5803	36	71(1.2%)
ALOI	49534	27	1508(3.04%)

The datasets are briefly described below in Table 2. Pendigits (Pen-Based Recognition of Handwritten Digits) dataset is a multi-class classification dataset having ten classes (0,1,...,9) of different handwritings, and class 4 is taken as an outlier. The PageBlocks dataset contains information on various block

[1]https://github.com/gourangaduari1995/outlier-decomp
[2]http://archive.ics.uci.edu/ml/datasets.html

types in document pages, and the task is to separate text from pictures or graphics. Here, other than text blocks are treated as outliers, as in Campos et al. [6]. Optdigits (optical recognition of handwritten digits) is a multi-class dataset of handwritten digits. Here, the digit "0" instances are taken as outliers, and the rest are inliers. As mentioned in the UCI repository, the Waveform dataset represents three classes of waves. Class "0" is labeled an outlier, and the rest of the data instances are inliers. The Thyroid dataset contains information about the hypothyroid patient, where hyperfunction and subnormal functioning are considered outliers, and the remaining instances are inliers. The Letter recognition dataset is a classification dataset, and we use it as in Rayana et al. [40]. SatImage dataset is also a multi-class dataset, where class 2 of 71 instances is labeled as an outlier. The Amsterdam Library of Object Images (ALOI) dataset is a collection of images we use as given by Campos et al. [6].

5.2. Baseline Metrics

Addressing the challenge posed by highly imbalanced data is paramount in the outlier detection task. In this realm, improving detection performance is the most necessary task. The Receiver Operating Characteristic (ROC) curve stands out as the predominant evaluation measure in research literature [35]. A crucial metric derived from this curve is the area under the curve (AUC), which serves as a comprehensive indicator of overall performance. Unlike other metrics, AUC remains unaffected by specific thresholds, directly reflecting the algorithm's classification capability. Other than that, we use precision [22, 34] (a high precision indicates low false positives) and recall [22, 34] (a recall indicates low false negatives) metrics for calculating classification accuracy. For visualization, we use t-SNE plot using true positives.

5.3. Outlier Detectors used for Comparison

In this paper, we consider six heterogeneous standard outlier detectors to check the effectiveness of our decomposition approach as base detectors: IForest [28] as an ensemble-based method, PCA-based outlier detector [41] as a linear model, *k*NN [9] as a distance-based method, LOF [4] as a density-based method, COF [43], and COPOD [26] as a probability-based method for all datasets in Table 2. Here, we use global and local outlier detectors to diversify our analysis. However, more outlier detectors with different characteristics may be experimented with to check the effectiveness of the decomposition strategy on outlier detection. We use the *sci-kit* library for *k*-means clustering, and pyod[3] library for the detectors.

5.4. Decomposition Results

Here, we demonstrate the effectiveness of our decomposition strategy using three widely used performance measures, precision, recall, and ROC-AUC [35], as shown in Table 3 and Table 4, respectively. As mentioned in the proposed methodology, we present the output of four different configurations of k ($j \in \{2, 3, 4, 5\}$). In Table 3 and Table 4, the best parameter-wise performance corresponding to each dataset is boldfaced in the table.

The proposed decomposition strategy works pretty well in detecting outliers in the datasets. This is evident in the results summarized in Table 3 and Table 4. Our data decomposition step gives the best results in all four decomposition configurations (k) compared to those without data decomposition. In

[3]https://pyod.readthedocs.io/en/latest/index.html

Table 3

Precision and Recall with the proposed decomposition strategy. The performance of detectors is shown in two ways: (a) Without data decomposition and (b) with data decomposition.

Dataset	**Parameter**	**IForest**	**Data Decomposition**			
			k=2	k=3	k=4	k=5
Pendigits	Precision	0.0071	0.0182	**0.0202**	0.0182	0.0182
	Recall	0.3500	0.9000	**1.0000**	0.9000	0.9000
Pageblock	Precision	0.3942	0.4173	0.3821	0.4693	**0.5077**
	Recall	0.3857	0.4143	0.4107	0.4911	**0.5321**
Optdigit	Precision	0.0498	0.1264	**0.1453**	0.1011	0.0954
	Recall	0.1733	0.4400	**0.5067**	0.3533	0.3333
Waveform	Precision	0.0345	0.0986	0.1130	0.1127	**0.1145**
	Recall	0.1200	0.3400	0.3900	0.3900	**0.4000**
Thyroid	Precision	0.1653	0.1667	0.1875	**0.2271**	0.1801
	Recall	0.2233	0.2257	0.2533	**0.3077**	0.2439
Letter	Precision	0.1062	0.2236	0.2422	0.2547	**0.2840**
	Recall	0.1700	0.3900	0.3900	0.4000	**0.4600**
Satimage	Precision	0.1188	**0.1222**	0.1205	0.1094	0.1094
	Recall	0.9718	**1.0000**	0.9859	0.9859	0.9859
ALOI	Precision	0.0331	0.0367	0.0371	0.0369	**0.0379**
	Recall	0.1008	0.1207	0.1220	0.1214	**0.1247**

the Satimage dataset, many outlier data points generate separate sub-groups as the O-cluster satisfying the case 1 condition after $k = 3$. In the Pageblock data, case 1 applies to a few clusters, and we detect a few O-cluster in each value of K, where $n_k \leqslant 0.02N$ and consequently $a_k = z_k \ \forall K$. We can see that detection performance after decomposition in the Pageblock dataset at $k = 5$ is best compared to without data decomposition. For the rest of the datasets, there are no O-clusters. Data decomposition works well, and superior performance is achieved in all the decomposition configurations. Meanwhile, our decomposition strategy helps the detector to achieve a 100% recall rate in the Satimage ($k = 2$) and Pendigit ($k = 3$) data, which shows the efficacy of our approach.

We have taken the ROC-AUC score in Table 4 and Fig. 4 as the parameter for checking overall performance competency concerning four decomposition configurations. We can see that overall detection has significantly improved using the data decomposition approach for the Pendigits, Optical digits, Waveform, Letter, and Ann-thyroid datasets. Fig. 4 exhibits the comparison of detectors' outlier identification capability with the proposed decomposition strategy in four configurations. Fig. 4(a), Fig. 4(b), and Fig. 4(d) outline the superiority decomposition strategy for outlier detection against the alone use of detectors, and it gives a clear visual picture of our method's acceptability.

Table 4

ROC-AUC score with the proposed decomposition strategy with varying *k*. The performance of detectors is shown in two ways: (a) Without decomposition and (b) after decomposition. The variance of ROC-AUC in parentheses shows the stability of the results. Bold fonts are the best values.

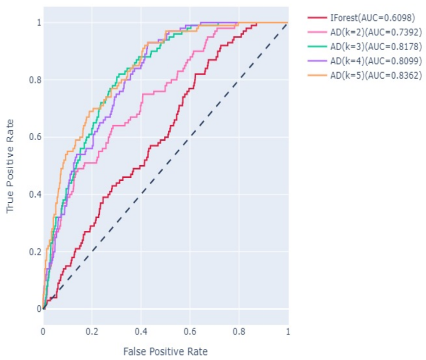
Dataset	**IForest**	**Data Decomposition**			
		k=2	k=3	k=4	k=5
Pendigits	0.8502	0.9659	0.9590	0.9639	**0.9671**
Pageblock	0.8855	0.8931	0.8800	0.8956	**0.9199**
Optdigits	0.6822	0.8759	0.8975	**0.8994**	0.8706
Waveform	0.6796	0.7668	0.6697	0.7307	**0.8569**
Thyroid	0.6340	0.6774	0.6957	**0.7350**	0.6819
Letter	0.6175	0.7499	0.8144	0.8139	**0.8355**
Satimage	0.9814	**0.9984**	0.9826	0.9958	0.9958
ALOI	0.5341	0.5478	**0.5490**	0.5458	0.5403

5.5. *Comparative Analysis*

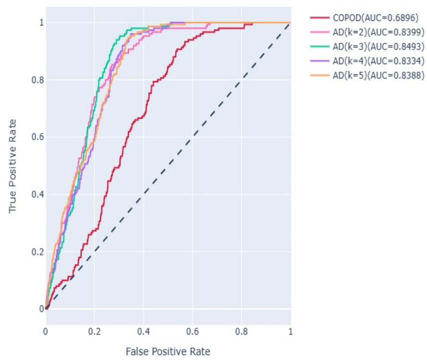
We compare the data decomposition approach in Fig. 5 with state-of-the-art detectors using the ROC-AUC curve. To have a broader compartive analysis with State-of-the-art (SOTA) methods, we take Pageblock, Waveform, OptDigit, and ANN-Thyroid data. We have found that the data decomposition approach is a strategic winner in detecting outliers in complex data patterns. Overall, we can say that the data decomposition approach performs substantially better than the regular use of detectors. The reasoning behind our claim is very subtle. First, detectors are more potent in filtering out outliers from homogeneous sub-groups. Second, the learning complexity of data patterns is reduced due to breaking them out into smaller and lower complex sub-groups. So, the decomposition strategy has the edge over the raw use of the outlier detectors.

5.6. *Visual Analysis*

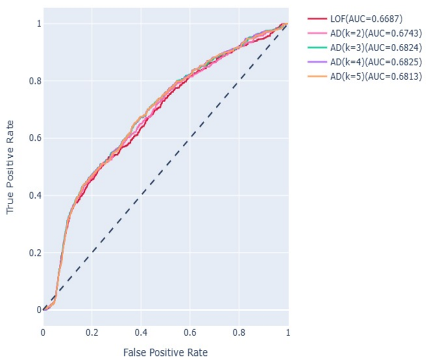
We present better visibility and solidify our method‘s effectiveness by showcasing *t*-SNE plot of true outliers in comparative mode between without decomposition strategy and after decomposition of data. *t*-SNE (*t*-distributed stochastic neighbor embedding) is a dimensionality reduction technique customarily used for visualizing high-dimensional data in a lower-dimensional space, typically 2D or 3D. It preserves the local structure of the data by modeling the similarities between data points in the high-dimensional space and embedding them into a lower-dimensional space while minimizing the divergence between their pairwise similarities. Fig. 6 represents *t*-distributed stochastic neighbor embedding (*t*-SNE) plots to show the effectiveness of performances in the Waveform dataset. We conduct comparative performance between the PCA-based outlier detector (without data decomposition) and after decomposition. Here, we use the default value of perplexity (30) and iteration (1000) for all the plots. Fig. 6(a) and Fig. 6(b) display exclusively detected true-positive (TP) outliers by red dots. In the Waveform dataset, our data decomposition approach detects 49 true outliers, and 12 are detected without data decomposition. Evidently, data decomposition has the edge over the usual use of an outlier detector.



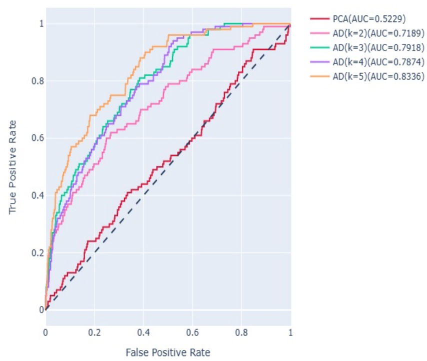
(a) Decomposition vs. Iforest



(b) Decomposition vs. COPOD



(c) Decomposition vs. LOF



(d) Decomposition vs. PCA

Figure 4. ROC-AUC score of IForest, kNN, LOF, and PCA *after decomposition (AD)* and without decomposition in different configurations of k.

5.7. Empirical Results of Complexity Reduction

We also have experimented with the complexity reduction aspects by measuring the entropy of the decomposed data, and We do the comparisons with the monolithic data so that we can have empirical and logical proofs of our claim. We present the comparisons through Table 5. We do the decomposition in four subgroups for all the datasets in this presentation. We observe that total entropy has been reduced by the decomposition strategy significantly, as exhibited by looking at the numbers in Table 5. We can see that five datasets show evidence of our claim, and two datasets (Pageblock and ALOI) show strong support. So, complexity reduction is achieved in most of the data datasets by looking into the Total ($\sum_k H_k$) compared with the Monolithic data (H_{com}).

It is evident from this empirical presentation that decomposition gives a boost to the detection process by grouping the homogeneous points into subgroups, and this pattern modification gives a linear space or

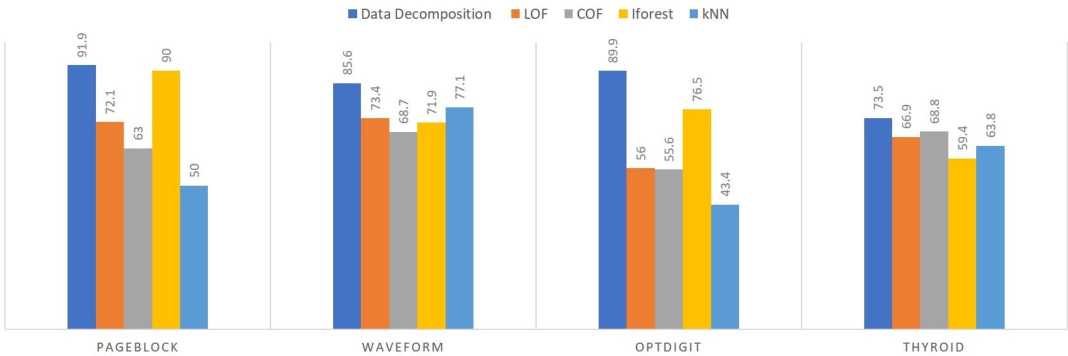
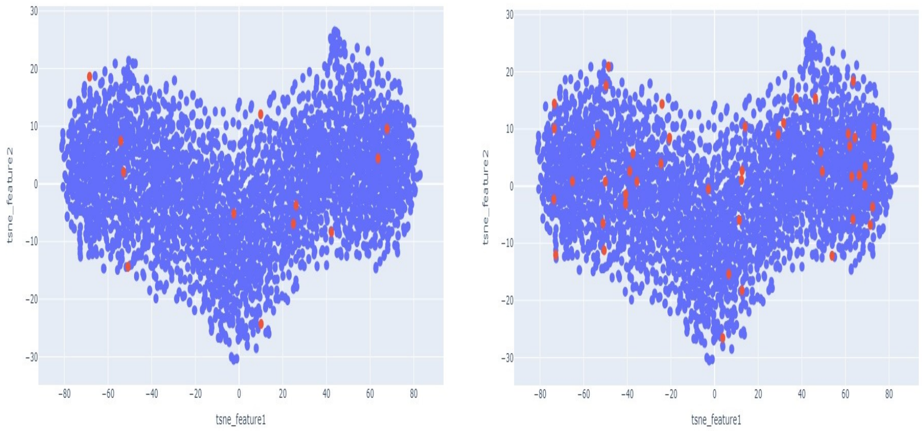


Figure 5. Comparative analysis of data decomposition for outlier detection with state-of-the-art detectors.



(a) Without data decomposition, #exclusively detected outliers: 12

(b) After data decomposition #exclusively detected outliers: 49

Figure 6. t-SNE plot for Waveform dataset. True outliers and inliers are presented by red and blue dots, respectively.

smooth ground to the detection process. Consequently, we detect more true positive outliers than without the decomposition process. On the other hand, this process significantly reduces false classifications. This observed reduction in false classification signifies the efficacy of the decomposition approach in outlier detection by mitigating the erroneous classification and bolstering the reliability and accuracy of of outlier detection process.

5.8. *Discussion*

From the results outlined above, we observe two primary categories of clusters emerge as the outcome of decomposition: (I) O-cluster: a cluster of a few significant isolated data points that are potential candidates for outliers; and (ii) cluster with outliers and natural data points. In the first category, we do not assign any detector to classify the data points as outliers further, as the clustering method does not forcefully include isolated data points in the cluster. It is sufficient to consider them as outliers with 100% probability. These data points are significantly dissimilar from the rest (Fig. 1), and data decomposition

Table 5
A brief comparison of complexity reduction in the decomposed space with its monolithic space.

Dataset	Monolithic Data(H_{com})	**Decomposed Data**				
		S1	S2	S3	S4	Total ($\sum_k H_k$)
Pendigit	4330.01	1011.96	1481.12	629.98	1247.50	4370.56
Pageblock	2731.32	2257.79	3.82	276.65	15.26	**2553.52**
Optdigit	2260.39	456.88	458.26	634.95	716.97	2267.06
Waveform	1510.59	312.05	437.40	440.89	308.88	**1499.22**
Thyroid	3115.18	1731.47	388.57	872.52	214.74	3207.3
Letter	712.30	119.81	228.41	124.44	227.43	**700.09**
Satimage	2583.37	773.95	963.19	813.78	28.53	**2579.27**
ALOI	22982.70	11216.30	1259.95	6938.37	2516.46	**21931.08**

alone is sufficient to separate outlier points into clusters. As our approach involves excluding isolated outliers named after themselves, our primary goal is to maximize the utilization of patterns, not to obligate the use of all patterns. Additionally, groups of outliers produced due to systematic measurement failure will form their own separate sub-group, which may be considerably distinct from other patterns with the same classification. So, we can consider them as potential global outliers of distinct nature. The handling of outliers is a topic that requires additional research. The second main category of clusters requires assigning a detector to classify the class of the data points, as these categories of clusters may consist of both standard and outlier data points. These clusters mainly contain local with few global outliers, and detectors can only learn the patterns to classify the outlier class.

In this experiment, we analyze the pattern complexity using an entropy-based method for a better understanding of our proposed approach in Table 5. Outcomes show massive support for our approach that complexity in terms of entropy has reduced in the subgroups than in monolithic space, even for the total entropy ($\sum_k H_k$) of all subgroups. This signifies that the homogeneous pattern modification after decomposition has made crucial improvements in pattern complexity reduction. This step boosts the data decomposition method validation and enhances the acceptance in actual applications.

The sole purpose of the decomposition of input data is to improve the detection efficiency of the detectors compared to regular methods. We also explain the decomposition strategy using the bias-variance tradeoff [2]. As the decomposition strategy creates sub-groups, it helps the detectors perform better by choosing appropriate sub-groups for learning, i.e., finding the best bias-variance tradeoff. The entire input data model is a single decision tree with high variance and low bias. On the other side, decomposition brings a set of potential decision trees. Each tree is less complicated than the entire input data and has a lower bias and low variance. Bias might increase in the case of the O-cluster (Case 1), but the criterion defined for O-cluster is not enough to have a high bias.

Our method is a generic approach to complex outlier detection problems where we use the decomposition strategy of input space as a pre-processor to the outlier detectors. This method is applicable even when we do not have enough prior knowledge of the input space, and clustering is a guiding principle for decomposition. Though the k-means clustering algorithm is based on similarity measures, and it depends on the judgment of the number of clusters (k), which is ill-defined. Here, our approach avoids such biased factors. Instead, we decompose the input space based on the purpose of identifying unnatural data points in sub-clusters. Our pre-processing approach works as a catalyst for the detector to maximize performance, Table 4 and Fig. 4. We also explain theoretically and empirically how data decomposition can reduce learning complexity and enhance classification accuracy. It is crucial to identify true outliers rather than having false positives (When inliers are detected as outliers). The significance of our method lies in there. Table 4, Table **??**, and Fig. 6(b) suggest that our method increases the detection accuracy of true outliers more than any other method. So, the above-outlined results and theoretical foundation make our approach more relevant in the unsupervised outlier detection process. Our approach can work well for a wide range of complex unsupervised problems where prior knowledge is not readily available to analyze data patterns.

Further investigation is required to explore how clusters of outliers (Fig. 1) resulting from systematic measurement failure can create distinct sub-groups that are notably isolated from other patterns of the same class. This outcome gives a significant hint about possible outlier patterns in the data. Categorization of such outlier patterns can be very valuable in the pattern recognition community. The handling and understanding of these outliers remain a topic of ongoing study.

Limitations of Implementation: As stated above, the proposed approach is generic. However, the implementation presented in this work is limited by (i) the decomposition, which is influenced by the k-means clustering. Results may not be as good as above if the clustering is inaccurate. (ii) Since the data decomposition approach limits the number of clusters (k), efficiency deteriorates after $k = 5$ for most of the datasets. So, performance entirely depends on decomposition configuration or the number of clusters (k).

6. Conclusion

In this paper, we proposed a new approach to outlier detection by data decomposition with k-means clustering. Our approach is pragmatic in practical applications of complex data patterns. Our entire framework of research is designed in two phases. In the first stage, we perform data decomposition to partition input data into sub-groups. The second stage assigns the outlier detectors in the sub-groups based on the criteria mentioned in the algorithm. Experiments indicate that our proposed approach outperforms each dataset and each detector.

In the future, we would like to investigate other ways of decomposition with better partitioning conditions for the sub-groups, which can be more suitable for outlier detection. A more robust decomposition strategy is needed for this work's future direction. We would also check the possible ways to use hierarchical decomposition for outlier detection.

References

[1] S.S. Abraham et al., FairLOF: Fairness in Outlier Detection, *Data Science and Engineering* **6**(4) (2021), 485–499.

[2] C.C. Aggarwal and S. Sathe, Theoretical foundations and algorithms for outlier ensembles, *ACM SIGKDD Explorations Newsletter* **17**(1) (2015), 24–47.

[3] F. Angiulli and C. Pizzuti, Fast outlier detection in high dimensional spaces, in: *Proc. European Conf. Principles of Data Mining & Knowledge Discovery*, Springer, 2002, pp. 15–27.

[4] M.M. Breunig, H.-P. Kriegel, R.T. Ng and J. Sander, LOF: identifying density-based local outliers, in: *Proc. ACM SIGMOD Int. Conf. Management of Data*, 2000, pp. 93–104.

[5] J. Buhmann and H. Kühnel, Complexity optimized data clustering by competitive neural networks, *Neural Computation* **5**(1) (1993), 75–88.

[6] G.O. Campos, A. Zimek, J. Sander, R.J. Campello, B. Micenková, E. Schubert, I. Assent and M.E. Houle, On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study, *Data Mining & Knowledge Discovery* **30**(4) (2016), 891–927.

[7] S. Chawla and A. Gionis, K-means–: A unified approach to clustering and outlier detection, in: *Proc. SIAM Int. Conf. Data Mining*, 2013, pp. 189–197.

[8] Z. Cheng, C. Zou and J. Dong, Outlier detection using isolation forest and local outlier factor, in: *Proc. Conf. Research in Adaptive and Convergent Systems*, 2019, pp. 161–168.

[9] T.T. Dang, H.Y. Ngan and W. Liu, Distance-based kNN outlier detection method in large-scale traffic data, in: *Proc. IEEE Int. Conf. Digital Signal Processing*, 2015, pp. 507–510.

[10] T. De Vries, S. Chawla and M.E. Houle, Finding local anomalies in very high dimensional space, in: *Proc. IEEE Int. Conf. Data Mining*, IEEE, 2010, pp. 128–137.

[11] A. Dey, B.R. Kumar, B. Das and A.K. Ghoshal, Outlier detection in social networks leveraging community structure, *Information Sciences* **634** (2023), 578–586.

[12] T. Fernando, H. Gammulle, S. Denman, S. Sridharan and C. Fookes, Deep learning for medical anomaly detection–a survey, *ACM Computing Surveys (CSUR)* **54**(7) (2021), 1–37.

[13] J. Fondaj, Z. Hasani and S. Krrabaj, Performance measurement with high performance computer of HW-GA anomaly detection algorithms for streaming data, *Computer Science* **23**(3) (2022).

[14] K. Fukunaga, Introduction to statistical pattern recognition, chapter 10, *Academic Press* **2** (1990), 446–451.

[15] G. Gan and M.K.-P. Ng, K-means clustering with outlier removal, *Pattern Recognition Letters* **90** (2017), 8–14.

[16] X. Gao, J. Yu, S. Zha, S. Fu, B. Xue, P. Ye, Z. Huang and G. Zhang, An ensemble-based outlier detection method for clustered and local outliers with differential potential spread loss, *Knowledge-Based Systems* **258** (2022), 110003.

[17] Z. He, X. Xu and S. Deng, Discovering cluster-based local outliers, *Pattern Recognition Letters* **24**(9–10) (2003), 1641–1650.

[18] D. Huang, D. Mu, L. Yang and X. Cai, CoDetect: Financial fraud detection with anomaly feature detection, *IEEE Access* **6** (2018), 19161–19174.

[19] A.K. Jain, Data clustering: 50 years beyond K-means, *Pattern Recognition Letters* **31**(8) (2010), 651–666.

[20] A.K. Jain, M.N. Murty and P.J. Flynn, Data clustering: a review, *ACM computing surveys (CSUR)* **31**(3) (1999), 264–323.

[21] M.-F. Jiang, S.-S. Tseng and C.-M. Su, Two-phase clustering process for outliers detection, *Pattern Recognition Letters* **22**(6–7) (2001), 691–700.

[22] B. Juba and H.S. Le, Precision-recall versus accuracy and the role of large data sets, in: *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33, 2019, pp. 4039–4048.

[23] R. Kashef and M. Warraich, Homogeneous Vs. Heterogeneous Distributed Data Clustering: A Taxonomy, *Data Management and Analysis: Case Studies in Education, Healthcare and Beyond* (2020), 51–66.

[24] E.M. Knorr and R.T. Ng, A unified approach for mining outliers, in: *Proc. Conf. Centre for Advanced Studies on Collaborative Research*, 1997, p. 11.

[25] R. Kumar and P. Rockett, Multiobjective genetic algorithm partitioning for hierarchical learning of high-dimensional pattern spaces: a learning-follows-decomposition strategy, *IEEE Trans. Neural Networks* **9**(5) (1998), 822–830.

[26] Z. Li, Y. Zhao, N. Botta, C. Ionescu and X. Hu, COPOD: copula-based outlier detection, in: *Proc. IEEE Int. Conf. Data Mining (ICDM)*, IEEE, 2020, pp. 1118–1123.

[27] Z. Li, Y. Zhao, X. Hu, N. Botta, C. Ionescu and G. Chen, Ecod: Unsupervised outlier detection using empirical cumulative distribution functions, *IEEE Trans. Knowledge and Data Engineering* (2022).

[28] F.T. Liu, K.M. Ting and Z.-H. Zhou, Isolation forest, in: *Proc. 8th IEEE Int. Conf. Data Mining*, 2008, pp. 413–422.

[29] H. Liu, J. Li, Y. Wu and Y. Fu, Clustering with outlier removal, *IEEE Trans. Knowledge & Data Engineering (TKDE)* **33**(6) (2019), 2369–2379.

[30] W. Liu and M.J. Pyrcz, A spatial correlation-based anomaly detection method for subsurface modeling, *Mathematical Geosciences* **53** (2021), 809–822.

[31] P. Maciołek, P. Król and J. Koźlak, Probabilistic anomaly detection based on system calls analysis, *Computer Science* **8** (2007), 93–108.

[32] O. Maimon and L. Rokach, Decomposition methodology for knowledge discovery and data mining, in: *Data Mining & Knowledge Discovery Handbook*, Springer, 2005, pp. 981–1003.

[33] V. Mednitskii, Y.V. Mednitskii and V.Y. Leonov, Application of decomposition methods in optimization, *Int. Journal Computer & Systems Sciences* **48** (2009), 827–838.

[34] J. Miao and W. Zhu, Precision–recall curve (PRC) classification trees, *Evolutionary intelligence* **15**(3) (2022), 1545–1569.

[35] A. Mukhriya and R. Kumar, Building outlier detection ensembles by selective parameterization of heterogeneous methods, *Pattern Recognition Letters* **146** (2021), 126–133.
[36] S. Pande and A. Khamparia, Explainable deep neural network based analysis on intrusion detection systems, *Computer Science* **24**(1) (2023).
[37] H. Paulheim and R. Meusel, A decomposition of the outlier detection problem into a set of supervised learning problems, *Machine Learning* **100** (2015), 509–531.
[38] S. Pidhorskyi, R. Almohsen and G. Doretto, Generative probabilistic novelty detection with adversarial autoencoders, *Advances in Neural Information Processing Systems (NeurIPS)* **31** (2018).
[39] S. Ramaswamy, R. Rastogi and K. Shim, Efficient algorithms for mining outliers from large data sets, in: *Proc. ACM SIGMOD Int. Conf. Management of Data*, 2000, pp. 427–438.
[40] S. Rayana and L. Akoglu, Less is more: Building selective anomaly ensembles, *ACM Trans. Knowledge Discovery from Data* **10**(4) (2016), 1–33.
[41] M.-L. Shyu, S.-C. Chen, K. Sarinnapakorn and L. Chang, A novel anomaly detection scheme based on principal component classifier, Technical Report, Miami Univ Coral Gables Fl Dept of Electrical and Computer Engineering, 2003.
[42] L. Sun, M. He, N. Wang and H. Wang, Improving autoencoder by mutual information maximization and shuffle attention for novelty detection, *Applied Intelligence* (2023), 1–15.
[43] J. Tang, Z. Chen, A.W.-C. Fu and D.W. Cheung, Enhancing effectiveness of outlier detections for low-density patterns, in: *Proc. Pacific-Asia Conf. Knowledge Discovery & Data Mining*, Springer, 2002, pp. 535–548.
[44] S. Vassilvitskii and D. Arthur, K-means++: The advantages of careful seeding, in: *Proc. 18th Annual ACM-SIAM Symp. Discrete Algorithms*, 2006, pp. 1027–1035.
[45] Y.V. Vizilter, Design of data segmentation and data compression operators based on projective morphological decompositions, *Int. Journal Computer & Systems Sciences* **48** (2009), 415–429.
[46] C. Wang, Z. Liu, H. Gao and Y. Fu, VOS: A new outlier detection model using virtual graph, *Knowledge-Based Systems* **185** (2019), 104907.
[47] H. Wang, H. Li, J. Fang and H. Wang, Robust Gaussian Kalman filter with outlier detection, *IEEE Signal Processing Letters* **25**(8) (2018), 1236–1240.
[48] H. Wang, M.J. Bah and M. Hammad, Progress in outlier detection tech: A survey, *IEEE Access* **7** (2019), 107964–108000.
[49] S. Wu and S. Wang, Information-theoretic outlier detection for large-scale categorical data, *IEEE transactions on knowledge and data engineering* **25**(3) (2011), 589–602.
[50] W. Xu, J. Jang-Jaccard, A. Singh, Y. Wei and F. Sabrina, Improving performance of autoencoder-based network anomaly detection on NSL-KDD dataset, *IEEE Access* **9** (2021), 140136–140146.
[51] Z. Yuan, H. Chen, T. Li, J. Liu and S. Wang, Fuzzy information entropy-based adaptive approach for hybrid feature outlier detection, *Fuzzy Sets and Systems* **421** (2021), 1–28.