Decomposition for Outlier Detection Using Space Partitioning

GOURANGA DUARI[1]
RAJEEV KUMAR[2]

Data to Knowledge (D2K) Lab
School of Computer and System Science
Jawaharlal Nehru University, Delhi, India 110 067
[1]gourangaduari5@gmail.com
[2]rajeevkumar.cse@gmail.com

Abstract. Decomposition for complexity minimization has long been a challenging approach. 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 of monolithic character. 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 data sets. Our data decomposition approach is superior for identifying localized patterns in the partitions and offers a better generalization.

Keywords: Decomposition, Space partitioning, Outliers, Complexity

(Received December 12th, 2023 / Accepted June 28th, 2024)

1 Introduction

Outlier is one of the most significant patterns in data analysis. 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 [18], in intrusion detection systems [14], in medical diagnosis [12], in geoscience [25]. 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 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 [37], clustering-based methods [24, 11, 35], information-theoretic methods [43], neural network-based methods [38, 29, 2, 1, 10, 33, 9, 34], isolation-based methods [23], and nearest neighbour-based methods [4]. 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 interesting problem for the pattern recognition community. Outlier detection becomes tough due to no 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 [20] of pattern space aimed at getting a more robust outcome by partitioning the data into sub-groups of homogeneous elements. In Figure 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 projection through partitioning the space into sub-groups, those points become potential outliers in that space.

Decomposition is one of the powerful tools for managing hard data pattern complexity and improving the recognition process, and it requires careful planning, design, and management to reap its full benefits. Though it has some genuine benefits in pattern recognition, it has few challenges in determining the boundaries between different components or subsystems. Overlapping functionalities or dependencies between modules can complicate the decomposition process. Addressing the challenges associated with decomposition is crucial to ensure that the divided parts work together harmoniously to achieve the desired outcomes. 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 [16] 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) [20] 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, 26]. It is presented experimentally that 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.

The research contributions of the proposed work in this paper 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 used.

Abbreviation	For
O-cluster	Outlier cluster (proposed)
AD	After Data Decomposition (proposed)
WD	Without Data Decomposition (proposed)
LOF	Local outlier factor [5]
COF	Connective-based outlier factor [39]
IForest	Isolation Forest [23]
COPOD	Copula-Based Outlier Detection [21]
PCA-OD	PCA-based outlier detector [37]
kNN-OD	k-Nearest Neighbor (kNN) based outlier detector [4]

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 monolithic 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 [20] 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

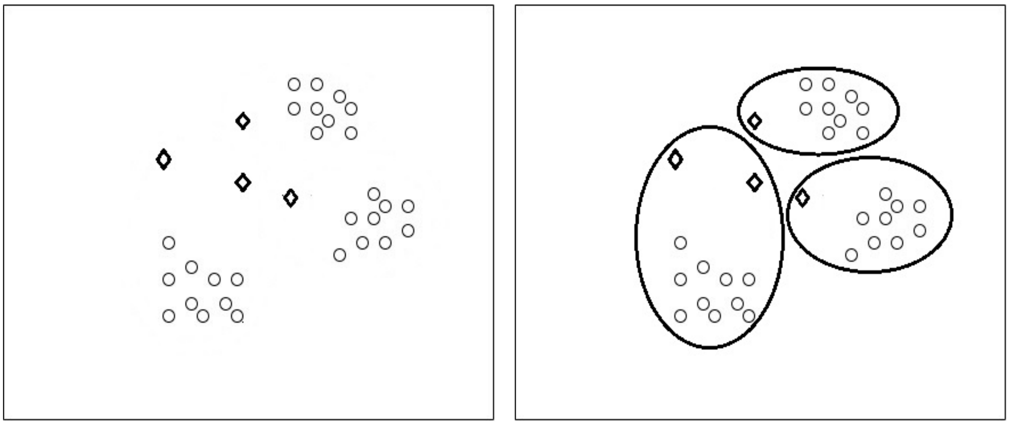
In the past decade, various methods have been developed for outlier detection under the unsupervised category. Among the recent developments, Cheng et al. [7] proposed an ensemble-based detector for global and local outliers. Recently, Li et al. [22] studied ECOD (Empirical-Cumulative-distribution-based Outlier Detection). Wang et al. [41] 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. [42]. 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 (kNN) based approach [4] 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. [30]. They partition the data and remove parts of the data that cannot contain outliers, thus reducing the computation and improving efficiency. Breunig et al. [5] developed a Local Outlier Factor (LOF) to identify outliers based on the density approach. The principle behind the density approach 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. [39] proposed COF, an improved version of LOF [5] based on chaining distance. Liu et al. [23] 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. [37] proposed a PCA-based outlier approach. He et al. [15] designed a cluster-based local outlier factor (CBLOF) based on the concept of a cluster-based local outlier.

A Learning-follows-Decomposition (LFD) strategy [20] 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. [26] 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) [28]. 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. [17] proposed a K-means a clustering-based two-phase method to detect outliers. The first phase involves partitioning data points. The second phase consists in 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.

4 The Proposed Methodology

Our proposed method is composed of a two-stage process. First, we partition the data into sub-groups or clusters using the decomposition mechanism by creating homogeneous sub-groups. We recursively take the best optimal sub-group centers, add the closest possible points of the respective centers in sub-groups, and optimize the entire decomposed systems with an optimizing function to make homogeneous sub-groups. These homogeneous sub-groups of similar characters have reduced pattern complexity because of pattern modulation based on some objective. We design our algorithm and objective function in such a way that it can help overcome the boundary effect of making sub-groups. The decomposition technique creates clusters of homogeneous elements, and the homogeneous grouping of data points fits the outliers in the different sub-groups based on the criteria. Second, we employ a standard outlier detector in each sub-group. We outline the evaluation framework for our approach based on the standard metrics. We present extensive empirical results



(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.

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.

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 pre-processing 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 [19], local [5] type, and unnatural data points distributed in the clusters according to their characteristics. Since *K*-means [40] 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 [5]. 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 Satimage 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 deviate 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 Mathematical 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 $d : X \times X \rightarrow \mathbf{R^d}$ 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 $\mathbf{R^d}$.

4.2 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 $f(X)$.

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

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. Figure 1 depicts the benefits of decomposition into sub-groups.

Here, our objective $f(X)$ is to increase the homogeneity (H) of sub-groups and decrease the learning complexity, i.e., data points should have the same characteristics. 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 $D_j(A)\ \forall j \in \{2, 3, 4, 5\}$.

4.3 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 $K$-means considering Eqn. 1 and Eqn. 2
 5:     calculate $n_K\ \forall K$
 6:     if ($n_K \leq 0.02N$) then
 7:         $z_K = a_K \subseteq A\ \forall K$
 8:     else
 9:         assign any $M_m$
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.

4.4 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: $O(n_1^2) + O(n_2^2) + ... + O(n_k^2) \leq O(N^2)$. 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 [13] 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.

5 Experimental Setup and 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

This work uses 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%)
Optdigit	5216	64	150 (3.00%)
Waveform	3443	21	100 (2.90%)
Thyroid	7200	6	536 (7.42%)
Letter	1599	32	100 (6.25%)
Satimage	5803	36	71 (1.20%)
Ecoli	336	7	9 (2.60%)
ALOI	49534	27	1508(3.04%)

The datasets are briefly described below in Table 2. The pendigits (Pen-Based Recognition of Handwritten Digits) dataset is a multi-class classification dataset that has ten classes (0,1,...,9) of different handwritings, and class 4 is taken as an outlier. 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. [31]. SatImage dataset is also a multi-class dataset, where class 2 of 71 instances is labeled as an outlier. The original Ecoli dataset [36] from UCI machine learning repository is a multiclass classification dataset having eight attributes. Among the eight classes omL, imL, and imS are the minority classes and are used as outliers. The Amsterdam Library of Object Images (ALOI) dataset is a collection of images we use as given by Campos et al. [6].

5.2 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 [23] as an ensemble-based method, PCA-based outlier detector [37] as a linear model, kNN [8] as a distance-based method, LOF [5] as a density-based method, COF [39], and COPOD [21] 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.

5.3 Decomposition Results

Here, we demonstrate the effectiveness of our decomposition strategy using three widely used performance measures, precision, recall, and ROC-AUC [27], 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 the Satimage dataset, many outlier data points generate separate subgroups as the O-cluster satisfying the case 1 condition after $K = 3$. In the Satimage data, case 1 applies to a few clusters on $K = 4$ and $K = 5$ configuration, where $n_K \leq 0.02N$ and consequently $a_K = z_K \; \forall K$. 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 also considered a

[1] https://github.com/gourangaduari1995/outlier-decomp

[2] http://archive.ics.uci.edu/ml/datasets.html

[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	0.71	1.82	**2.02**	1.82	1.82
	Recall	35.00	90.00	**100.00**	90.00	90.00
Optdigit	Precision	4.98	12.64	**14.53**	10.11	9.54
	Recall	17.33	44.00	**50.67**	35.33	33.33
Waveform	Precision	3.45	9.86	11.30	11.27	**11.45**
	Recall	12.00	34.00	39.00	39.00	**40.00**
Thyroid	Precision	16.53	16.67	18.75	**22.71**	18.01
	Recall	22.33	22.57	25.33	**30.77**	24.39
Letter	Precision	10.62	22.36	24.22	25.47	**28.40**
	Recall	17.00	39.00	39.00	40.00	**46.00**
Satimage	Precision	11.88	**12.22**	12.05	10.94	10.94
	Recall	97.18	**100.00**	98.59	98.59	98.59
ALOI	Precision	3.31	3.67	3.71	3.69	**3.79**
	Recall	10.08	12.07	12.20	12.14	**12.47**
Ecoli	Precision	20.59	**26.47**	0.00	0.00	2.86
	Recall	77.78	**100.00**	0.00	0.00	11.11

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	85.02	96.59	95.9	96.39	**96.71**
Optdigits	68.22	87.59	89.75	**89.94**	87.06
Waveform	67.96	76.68	66.97	73.07	**85.69**
Thyroid	63.40	67.74	69.57	**73.50**	68.19
Letter	61.75	74.99	81.44	81.39	**83.55**
Satimage	98.14	**99.84**	98.26	99.58	99.58
ALOI	53.41	54.78	**54.90**	54.58	54.03
Ecoli	86.54	**94.02**	55.76	30.75	47.40

smaller dataset, named Ecoli, to evaluate our method, and we found out that the method works well only $K = 2$. It infers that data decomposition is unsuitable for small data, and the model gives better benefits for bigger datasets.

We have taken the ROC-AUC score in Table 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 Annthyroid datasets. We have also tested our algorithm twenty times to confirm performance stability.

5.4 Comparative Analysis

We exhibit our data decomposition approach using three widely used methods, namely, precision, recall, and ROC-AUC in Tables 3 and Table 5 using IForest as a base detector. In the case of highly imbalanced data with a large number of negative samples (inlier), the false-positive rate (FPR) remains relatively small, even for plenty of false positives. However, in dealing with imbalanced data, the accuracy of the positive class (minority class) is more important. So, we take true positive (TP) for comparing our approach globally in Table 5. We can find that data decomposition emerges as a clear winner in all three datasets. After using data decomposition, many true outliers are detected, which significantly benefits the overall efficacy.

Comparing the data decomposition approach using the ROC-AUC curve in Figure 2 with different decomposition configurations with varying values of K, we have found that the data decomposition approach is a strategic winner in detecting outliers in complex data patterns. The performance improvement by data decomposition approach for the Waveform dataset using IForest (Figure 2(a)) and Optical digits dataset using COPOD (Figure 2(b)) and the Letter dataset using Principle Component Analysis (PCA) based outlier detector (Figure 2(d)) is advantageous. 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 reduces due to breaking out into smaller sub-groups. So, the decomposition strategy has the edge over the raw use of the outlier detectors. LOF and COF are both density-based local outlier detectors, and they have not had substantial improvement, which needs more investigation in future research.

As our approach is to give detection efficiency using data decomposition rather than the regular use of detectors, our approach is directly comparable to the concerned detectors. So, we restrict our results to comparing with varying values of K, not with other outlier algorithms.

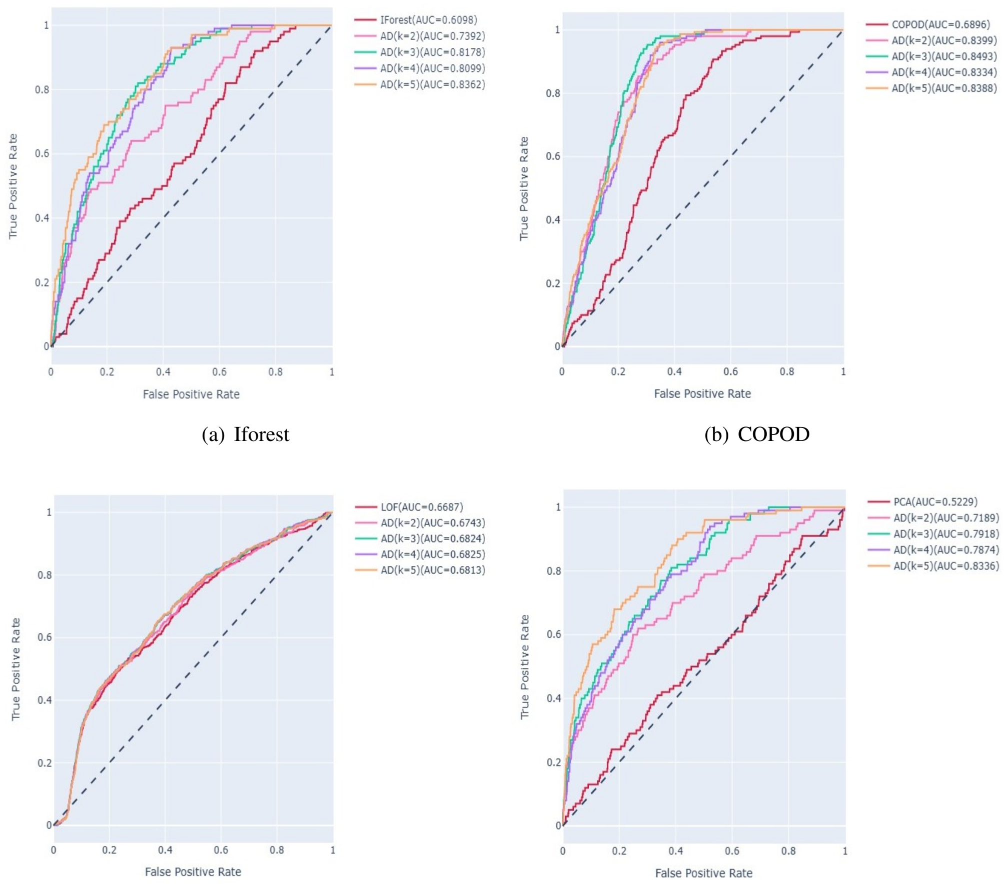
5.5 Visual Analysis

Figure 3 represents t-distributed stochastic neighbor embedding (t-SNE) plots to show the effectiveness of performances in the Waveform dataset. We conduct comparative performance between without data decomposition and after decomposition using the kNN-based outlier detector. Here, we use the default value of perplexity (30) and iteration (1000) for all the plots. Figure 3(a) and Figure 3(b) display exclusively detected true-positive (TP) outliers by red dots. In the OptDigits dataset, our data decomposition approach detects 22 true outliers, and 11 are detected without data decomposition. Evidently, data decomposition has the edge over the usual use of an outlier detector.

5.6 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 (Figure 1), and k-means clustering 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 subgroup, 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.

Further investigation is required to explore how clusters of outliers (Figure 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



(a) Iforest

(b) COPOD

(c) LOF

(d) PCA

Figure 2: ROC-AUC score of IForest, COPOD, LOF, and PCA-OD *after decomposition (AD)* and without decomposition in different configurations of k.

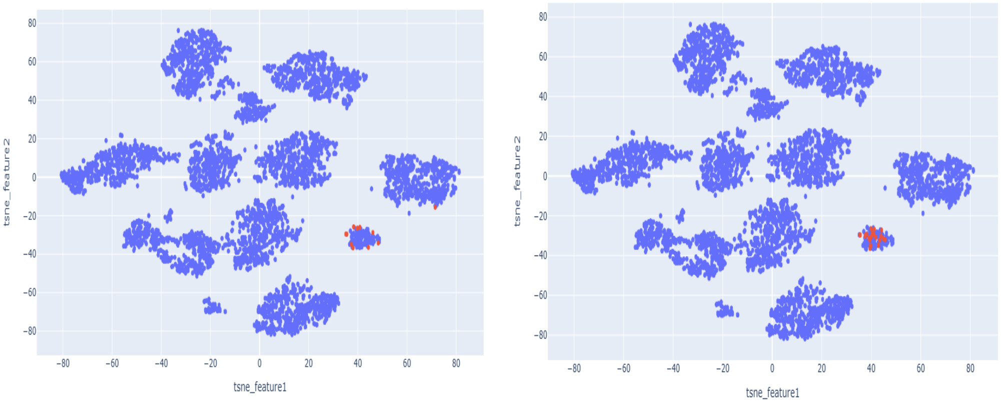
such outlier patterns can be very valuable in the pattern recognition community. The handling and understanding of these outliers remain a topic of ongoing study.

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 [3]. 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 the 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 subclusters. Our pre-processing approach works as a catalyst for the detector to maximize performance, Table 4 and Figure 2. 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

Table 5: A brief presentation of exclusively detected true outliers without decomposition (WD) and after decomposition (AD). Benchmark datasets with respective outliers are also mentioned.

Detector	Optical Digits (150)		Waveform (100)		Letter (100)	
	WD	AD	WD	AD	WD	AD
IForest	25	**52**	18	**41**	16	**44**
kNN-OD	11	**22**	35	**38**	54	**57**
PCA-OD	5	**44**	12	**49**	13	**49**
LOF	31	**38**	27	**40**	53	**58**
COF	59	**61**	25	**31**	59	**61**
COPOD	17	**60**	3	**42**	11	**40**



(a) Without data decomposition, #exclusively detected outliers: 11

(b) After data decomposition #exclusively detected outliers: 22

Figure 3: t-SNE plot for Waveform dataset. True outliers and inliers are presented by red and blue dots, respectively.

positives (When inliers are detected as outliers). The significance of our method lies in there. Table 4, Table 5, and Figure 3(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. The reduced complexity in the data space makes the entire system compact and conducive to the detectors' doing smooth and effective operations to recognize the complex pattern. Our approach can work well for a wide range of complex unsupervised problems where prior knowledge is not readily available to analyze data patterns.

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) as 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. Experiments indicate that our proposed approach outperforms almost every dataset and each detector. We have recorded around 1.7% to 30% improvement in AUC score and significant improvements in the detection of positive class, especially local outliers.

In the future, we would like to investigate a more efficient objective function for decomposition so that we can improve the homogeneity conditions for complex data patterns. We can also focus on optimized multi-layered homogeneity, considering diverse data characteristics that can unveil more knowledge to the pattern recognition community.

Acknowledgement

The authors thank the editor and reviewer(s) for their valuable comments.

References

[1] Affonso, E. T., Nunes, R. D., Rosa, R. L., Pivaro, G. F., and Rodriguez, D. Z. Speech quality assessment in wireless voip communication using deep belief network. *IEEE Access*, 6:77022–77032, 2018.

[2] Affonso, E. T., Rosa, R. L., and Rodriguez, D. Z. Speech quality assessment over lossy transmission channels using deep belief networks. *IEEE Signal Processing Letters*, 25(1):70–74, 2017.

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

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

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

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

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

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

[9] Dantas Nunes, R., Lopes Rosa, R., and Zegarra Rodríguez, D. Performance improvement of a non-intrusive voice quality metric in lossy networks. *IET Communications*, 13(20):3401–3408, 2019.

[10] de Almeida, F. L., Rosa, R. L., and Rodriguez, D. Z. Voice quality assessment in communication services using deep learning. In *2018 15th International Symposium on Wireless Communication Systems (ISWCS)*, pages 1–6. IEEE, 2018.

[11] Duari, G. and Kumar, R. Clustering for global and local outliers. In *Proc. 4th Int. Conf. Machine Intelligence Techniques for Data Analysis & Signal Processing (MISP 2022), Volume 1*, pages 601–610. Springer, 2023.

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

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

[14] Hassaan, M., Maher, H., and Gouda, K. A fast and efficient algorithm for outlier detection over data streams. *Int. Journal Advanced Computer Science & Applications*, 12(11), 2021.

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

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

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

[18] Khan, W. and Haroon, M. An efficient framework for anomaly detection in attributed social networks. *Int. Journal Information Technology*, 14(6):3069–3076, 2022.

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

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

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

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

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

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

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

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

[27] Mukhriya, A. and Kumar, R. Building outlier detection ensembles by selective parameterization of heterogeneous methods. *Pattern Recognition Letters*, 146:126–133, 2021.

[28] Paulheim, H. and Meusel, R. A decomposition of the outlier detection problem into a set of supervised learning problems. *Machine Learning*, 100:509–531, 2015.

[29] Rajalakshmi, S. and Madhubala, P. Centroid stabilized fuzzy tukey quartile and z curve neural network based outlier detection. *INFOCOMP Journal Computer Science*, 21(2), 2022.

[30] Ramaswamy, S., Rastogi, R., and Shim, K. Efficient algorithms for mining outliers from large data sets. In *Proc. ACM SIGMOD Int. Conf. Management of Data*, pages 427–438, 2000.

[31] Rayana, S. and Akoglu, L. Less is more: Building selective anomaly ensembles. *ACM Trans. Knowledge Discovery from Data*, 10(4):1–33, 2016.

[32] Rios, R. A. and de Mello, R. F. A systematic literature review on decomposition approaches to estimate time series components. *INFOCOMP Journal Computer Science*, 11(3-4):31–46, 2012.

[33] Rodriguez, D. Z. and Bressan, G. Video quality assessments on digital tv and video streaming services using objective metrics. *IEEE Latin America Transactions*, 10(1):1184–1189, 2012.

[34] Rodriguez, D. Z. and Junior, L. C. B. Determining a non-intrusive voice quality model using machine learning and signal analysis in time. *INFOCOMP Journal of Computer Science*, 18(2), 2019.

[35] Rodríguez, D. Z., Rosa, R. L., Almeida, F. L., Mittag, G., and Möller, S. Speech quality assessment in wireless communications with mimo systems using a parametric model. *IEEE Access*, 7:35719–35730, 2019.

[36] Sathe, S. and Aggarwal, C. LODES: Local density meets spectral outlier detection. In *Proc. SIM Int. Conf. Data Mining*, pages 171–179. SIAM, 2016.

[37] Shyu, M.-L., Chen, S.-C., Sarinnapakorn, K., and Chang, L. A novel anomaly detection scheme based on principal component classifier. Technical report, Miami Univ Coral Gables Fl Dept of Electrical and Computer Engineering, 2003.

[38] Sun, L., He, M., Wang, N., and Wang, H. Improving autoencoder by mutual information maximization and shuffle attention for novelty detection. *Applied Intelligence*, pages 1–15, 2023.

[39] Tang, J., Chen, Z., Fu, A. W.-C., and Cheung, D. W. Enhancing effectiveness of outlier detections for low-density patterns. In *Proc. Pacific-Asia Conf. Knowledge Discovery & Data Mining*, pages 535–548. Springer, 2002.

[40] Vassilvitskii, S. and Arthur, D. k-means++: The advantages of careful seeding. In *Proc. 18th Annual ACM-SIAM Symp. Discrete Algorithms*, pages 1027–1035, 2006.

[41] Wang, C., Liu, Z., Gao, H., and Fu, Y. VOS: A new outlier detection model using virtual graph. *Knowledge-Based Systems*, 185:104907, 2019.

[42] Wang, H., Bah, M. J., and Hammad, M. Progress in outlier detection tech: A survey. *IEEE Access*, 7:107964–108000, 2019.

[43] Yuan, Z., Chen, H., Li, T., Liu, J., and Wang, S. Fuzzy information entropy-based adaptive approach for hybrid feature outlier detection. *Fuzzy Sets and Systems*, 421:1–28, 2021.