2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 979-8-3503-8622-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/BIBM62325.2024.10821743

MGOD: Multi-Granular Outlier Detection with Clustlier Analysis

Qingsheng Chen[1], Mingjie Zhao[1], Yuzhu Ji[1*], Xiaopeng Luo[2], Yiqun Zhang[1*], Yue Zhang[3]
[1]*School of Computer Science and Technology, Guangdong University of Technology*
[2]*School of Computer Engineering, Guangzhou Huali College*
[3]*School of Computer Science, Guangdong Polytechnic Normal University*
{2112205080, 2112205249}@mail2.gdut.edu.cn, yuzhu.ji@gdut.edu.cn
gordonlok@foxmail.com, yqzhang@gdut.edu.cn, zhangyue@gpnu.edu.cn

***Abstract*—Unsupervised Outlier Detection (UOD) is crucial for the analysis of biomedical and health data with undesirable outliers. However, the complex distribution of real data often brings difficulties to UOD where the "masking effect", i.e., only a small number of densely distributed outliers (also called clustliers) can collectively mask themselves from being detected, is particularly challenging. Another difficulty derived from this is how to distinguish clustliers from small clusters. Therefore, we propose a novel Multi-Granular Outlier Detector (MGOD). It first partitions the dataset into subsets with natural neighbor topological relationships to circumvent the non-trivial neighbor range setting. Then it effectively detects both clustliers and isolated samples (also called scatliers) based on a newly designed anomaly score. The score comprehensively takes into account the density and connectivity of samples to reflect different extents and types of abnormality. It turns out that MGOD is accurate and highly interpretable. The performance of MGOD is also robust to the involved hyper-parameters, which are easy to set. Comprehensive evaluations have been conducted to compare seven counterparts on 15 datasets, most of which are biomedical datasets. The results of significance tests confirm the effectiveness and superiority of MGOD. The source code is opened at https://anonymous.4open.science/r/MGOD-C531.**

***Index Terms*—Outlier detection, anomaly detection, clustered outliers, scattered outliers, unsupervised learning**

I. Introduction

Unsupervised Outlier Detection (UOD) aims to mine samples that deviate noticeably from the majority [1], which has a wide range of applications, including medical diagnosis [2], health monitoring [3], urban management, and so on. UOD is also widely used in many data-driven AI tasks to eliminate data outliers for data quality enhancement [4]. In real complex data distributions, outliers manifest in diverse types, broadly categorized as scattered and clustered outliers (as shown in Figure 1), which are also called *scatliers* and *clustliers*, respectively, to better distinguish them. A scatlier is an isolated data point that is easy to detect while clustlier can be viewed as a micro-cluster whose size is much smaller than the normal clusters. Since the data points within a clustlier are with relatively high distribution density, a detector can easily confuse them with normal cluster samples, known as the "masking effect" [5], which is more challenging in UOD.

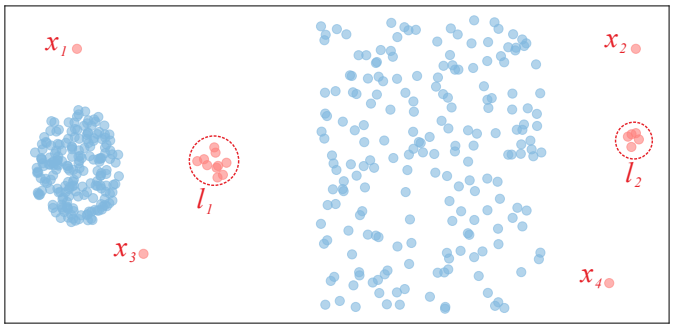

Fig. 1. Comparison of *scatlier*, *clustlier*, and normal data points (blue dots). From the perspective of the prominent clusters, scatlier is an isolated data point (e.g., x_1, x_2, x_3, and x_4) and clustlier is a collection of locally clustered outliers with relatively small scale (e.g., l_1 and l_2).

Both scatlier and clustlier are concepts relative to the granularity [6] [7] of "reference set" [8], i.e., to which samples or sample sets are treated as outliers. Existing UOD approaches can be roughly categorized into global and local according to their adopted reference sets. From the challenging clustlier detection perspective, the performance of global approaches including IFOREST [9] and COPOD [10] are somewhat limited as they adopt the entire dataset as their reference set [11]. That is, when the clustliers are located among several prominent clusters, samples within the clustliers will have relatively high global density [12], and therefore can easily be excluded as outlier candidates.

Local approaches that adopt subsets of data samples as their reference sets are proven to be more competent in clustlier detection. KNN (K Nearest Neighbor)-based approaches [13], [14] are a representative stream of local methods. They adopt certain measures such as distance, density, etc. to compute the anomaly score of a sample w.r.t. its k nearest neighbors, and treat the samples with higher anomaly scores as outliers. Intuitively, when k is set to an appropriate larger value, KNN-based methods will have the ability to detect clustliers, even

This work was supported in part by the NSFC under Grants 62476063, 62302104, 62102097, and 62172112, the Natural Science Foundation of Guangdong Province under grants 2023A1515012855 and 2023A1515012884, the Science and Technology Program of Guangzhou under grant SL2023A04J01625, and the General University Youth Innovation Talent Program of Guangdong under grant 2024KQNCX094. Yuzhu Ji and Yiqun Zhang are the corresponding authors.

if they appear among several prominent clusters. To tackle the k selection issue, several recent works [15], [16] have been proposed to evaluate the optimal k. Some approaches [17], [18] further propose to adaptively select k based on the natural neighbor relationship among samples [19]. However, all the above-mentioned attempts in k selection are not focusing on the detection of clustliers. Most recently, a KNN-based approach DGOF [20] proposes to employ both the density and distance as measures to more comprehensively detect both scatliers and clustliers. Nevertheless, as DGOF considers normal samples to be with similar density to its neighbors and shorter distance to the density peak, it will unavoidably miss the clustliers with higher density than the reference sets. In summary, global UOD methods struggle to handle clustliers in complex distributions, while local ones subjectively introduce bias to certain clustliers due to their corresponding reference sets. We claim that the root cause of these problems is that the existing approaches detect outliers through the reference sets, overlooking the possible abnormality of the reference sets themselves. To well-address these issues, a UOD method that can simultaneously detect scatliers and clustliers, and is robust to various clustliers, is in urgent need.

This paper, therefore, proposes to hierarchically evaluate the abnormality of local compact subsets and samples within each subsets. More specifically, our method partitions dataset into natural subsets with natural neighbor relationships and considers natural subsets as reference sets. By treating clustliers as outliers in the level of reference sets, anomaly scores of subsets are measured according to their global connection to all the subsets. Simultaneously, scatliers are finely considered within their corresponding subsets to measure their local anomaly scores. To ensure a comprehensive abnormality evaluation, the local anomaly score of a sample and the anomaly score of its corresponding subset are combined for the final abnormality ranking. It turns out that separately evaluating the multi-granular subset-level and sample-level abnormalities can effectively facilitate the detection of scatliers and clustliers. The three main contributions of this paper are as follows:

- This paper presents an easy-to-use UOD method called MGOD, which is capable of accurately detecting scatliers and clustliers in complex data distributions, robust to hyper-parameters, and highly interpretable.
- Natural subsets are defined as reference sets to reveal the distributions of clusters and clustliers as naturally connected subsets and isolated subsets, respectively, laying the statistical foundation for measuring the abnormality.
- A comprehensive measure is proposed based on the connection of natural subsets and density of data points, which assigns data points within isolated subsets with higher anomaly scores to mitigate the "masking effect".

II. Preliminaries

Natural Neighbor (NB) defines a natural mutual relationship among neighbors inspired by human friendship [19]. Only when two samples are neighboring each other, do they establish an NB relationship.

Definition 1 (Natural Neighbor - NB)**.** *If sample x_j is a natural neighbor of x_i, their relationship is defined as follows:*

$$x_j \in \mathrm{NB}(x_i) \Leftrightarrow (x_i \in \mathrm{KNN}_\lambda(x_j)) \wedge (x_j \in \mathrm{KNN}_\lambda(x_i)). \quad (1)$$

where λ represents the scope of the KNN neighborhood, an intrinsic value of a dataset discovered by the natural neighbor search algorithm [19].

NB is parameter-free and suitable for representing inter-object relations for various types of data. Since samples and their NBs are more likely to belong to the same cluster [19], NB is utilized for natural subset partition in our work.

Definition 2 (Natural Neighborhood Graph - NBG)**.** *NBG is a structure for representing the NB relationship of the dataset. Each vertex v_i in this graph represents a sample x_i. Two samples x_i and x_j are connected by an edge if x_i is a natural neighbor of x_j, where the edge is defined as:*

$$\mathrm{NBG}_{ij} = \begin{cases} 1 & if\ x_i \in \mathrm{NB}(x_j)\ or\ x_j \in \mathrm{NB}(x_i) \\ 0 & otherwise \end{cases}. \quad (2)$$

III. Proposed Method

For clustliers detection, local data regions with fewer aggregated samples are more anomalous. However, in scenarios with complex outliers, it is difficult to accurately obtain ideal regions using existing clustering techniques. An intuitive way is to divide the dataset into fine subsets and then estimate the abnormality of each subset by analyzing the distribution relationship among subsets. That is, a subset that is adjacent to fewer other subsets will be considered more anomalous.

The comprehensive anomaly score of a sample requires not only measuring its abnormality in a local reference set but also considering the anomaly information carried by the reference set it belongs to. The pipeline of the proposed method is illustrated in Figure 2, where the process for scoring the abnormality of samples is outlined in four steps: (1) partition the whole dataset into natural subsets based on natural neighbor relationships (Section III-A), (2) measure the anomaly scores of subsets by analyzing their distribution (Section III-B), (3) treat natural subsets as reference sets and calculate the local anomaly scores of samples in each set (Section III-C), and (4) form the overall anomaly scores by combining the two types of scores obtained in (2) and (3) (Section III-D).

A. Natural Subset Partition

In this subsection, a dataset $X = \{x_1, ..., x_i, ..., x_n\}$ (x_i is a data sample) is divided into a series of natural subsets $S = \{s_1, ..., s_m, ...\}$ (s_m is a natural subset), which are then treated as reference sets. To adapt to the dataset distribution, the expansion of a natural subset starts from dense samples and is based on natural neighbors.

Before subset partition, it is necessary to measure the local density of samples. Traditionally in KNN, the local density of a sample can be simply considered as the inverse of the average distance to its k neighbors. However, for natural neighbors, the number of neighbors for each sample varies,

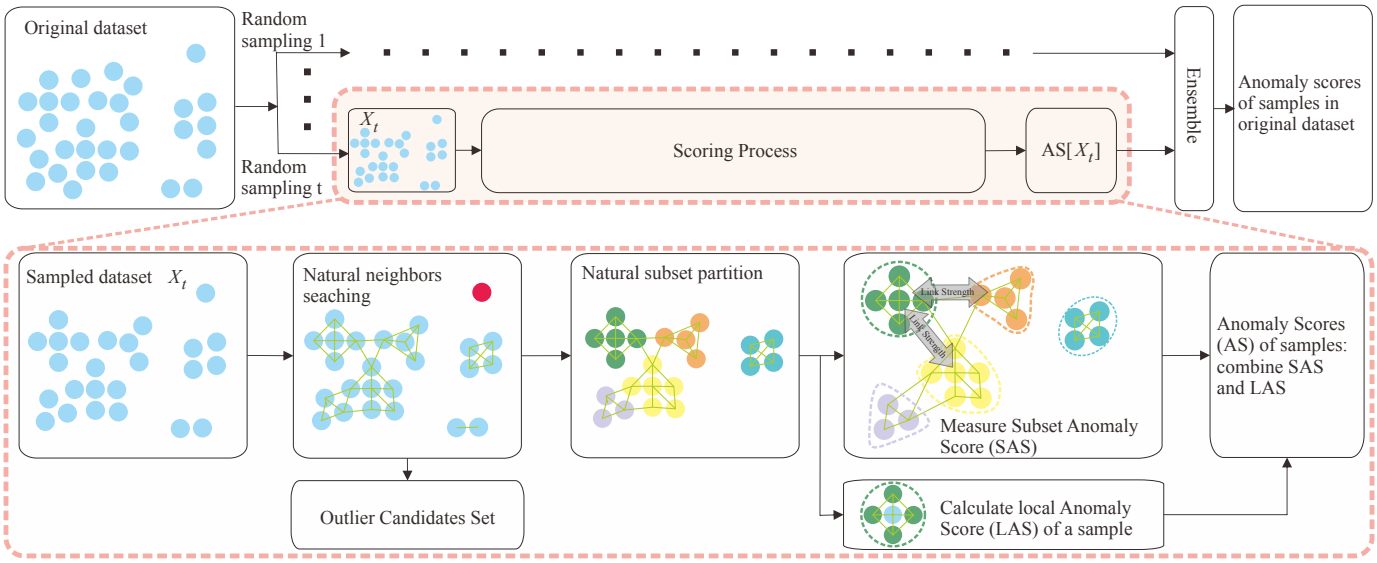

Fig. 2. Overview of the proposed MGOD method. The random sampling phase obtains multiple compact natural subsets from the original dataset to properly reflect data distribution and make potential cluster boundaries clearer. Then in the scoring process, the Subset Anomaly Score (SAS) and Local Anomaly Score (LAS) are computed to comprehensively reflect the sample abnormality under the complex co-occurrence of scatliers and clustliers. SAS is derived from the degree of adjacency (i.e., Link Strength) among subsets, indicating the subset-level abnormality of samples. LAS measures the local abnormality of samples within each subset based on distribution density. These two scores collectively ensure the detection of scatliers without overlooking clustliers.

unlike a relatively fixed k in KNN. Therefore, measuring local density should consider the quantity of natural neighbors, rather than relying solely on average distance. The local density of a sample is defined as follows.

Definition 3 (Local Density for Natural Neighbors - ρ)**.** *In NB contexts, the local density of x_i is computed by*

$$\rho(x_i) = 1/\left(\frac{\sum\limits_{x_j \in \mathrm{NB}(x_i)} dist(x_i, x_j)}{|\mathrm{NB}(x_i)|} + \frac{1}{|\mathrm{NB}(x_i)|}\right). \quad (3)$$

where $|\mathrm{NB}(x_i)|$ is the number of samples in $\mathrm{NB}(x_i)$, and

$$dist(x_i, x_j) = ||x_i - x_j||_2 \quad (4)$$

calculates the distance between two sample vectors x_i and x_j using the Euclidean distance.

Generating compact natural subsets involves the following steps: (1) From those unassigned samples, select the sample with the highest ρ, and add it to a new natural subset s_{new}. (2) Search for the unassigned natural neighbors of members in s_{new} and add them to s_{new}. (3) Repeat (2) until the size of s_{new} reaches a preset upper limit U, or there are no more unassigned natural neighbors. (4) Add s_{new} into S, and turn to (1) until all the samples are assigned. The above process can be rigorously presented as Algorithm 1.

B. Subset-wise Anomaly Score Measurement

Anomalous subsets are adjacent to fewer subsets. We define link strength (LS) to indicate the degree of adjacency between two natural subsets.

Definition 4 (Link Strength - LS)**.** *For two natural subsets s_m and s_w, the degree of their adjacency considers both the distance between the subsets and the number of natural neighbor relationships between their respective members. The link strength between them is computed by*

$$\mathrm{LS}(s_m, s_w) = \frac{\mathrm{NBP}(s_m, s_w)}{dist(c_m, c_w)} \quad (5)$$

where $\mathrm{NBP}(s_m, s_w)$ counts the number of sample pairs with the natural neighbor relationship between s_m and s_w. c_m and c_w are the centroids of s_m and s_w, respectively.

It is intuitive that when a natural subset is weakly adjacent to fewer other subsets, it is more likely to be an abnormal subset. Therefore, the anomaly score of a subset can be determined by the sum of its link strength with other subsets. We compute subset anomaly scores (SAS) and squish the scores over [0, 1] using Eq. (6). A higher score indicates the higher anomaly level of a subset.

$$\mathrm{SAS}(s_m) = 1 - norm\left(\sum_{s_w \in S, s_w \neq s_m} \mathrm{LS}(s_m, s_w)\right). \quad (6)$$

where $norm(\cdot)$ represents using the min-max normalization.

For samples in s_m, SAS (s_m) represents their shared base anomaly scores. Additionally, we need to vary the abnormality of every sample in each subset, as detailed in Section III-C.

C. Sample-wise Anomaly Score Measurement

After partitioning, natural subsets serve well as reference sets. In each reference set, we assess the abnormality of each sample based on density, where lower density indicates a higher anomaly. The local anomaly score (LAS) of a sample

Algorithm 1: Natural subset partition

Input: Dataset $X = \{x_1, x_2, ..., x_n\}$.
Output: A series of natural subsets S, a set of outlier candidates O, NBG.

Initialize: $S = \emptyset$, upper limitation $U = \sqrt{n}$;
Obtain NBG, O by the natural neighbor search algorithm [19]; Obtain density $\rho(x_i)$ of each sample x_i by Eq. (3);
Initialize unassigned set $R = \{x_i | x_i \in X \cap \rho(x_i) > 0\}$;
while $R \neq \emptyset$ **do**
    $x_m \leftarrow \arg\max_{x_i \in R} \rho(x_i)$;
    $R \leftarrow R \setminus x_m$;
    initialize a new subset $s_{new} = \emptyset$;
    $s_{new} \leftarrow s_{new} \cup x_m$;
    for *each* x_j *in* s_{new} **do**
        obtain NB(x_j) by NBG;
        add unassigned neighbors in NB(x_j) to s_{new} and remove them from R;
        if $size(s_{new}) > U$ **then break**;
    end
    $S \leftarrow S \cup s_{new}$;
end
Return S, O, NBG.

Algorithm 2: MGOD

Input: Original dataset $X_0 = \{x_1, x_2, ..., x_N\}$, sampling rate η, sampling time t.
Output: Anomaly score AS_{X_0} of samples in X_0.

$\mathrm{AS}_{X_0} \leftarrow 0 \in R^N$; // N:the size of X_0.
while $t > 0$ **do**
    obtain sampled dataset X_t by random sampling with rate η from X_0;
    $\mathrm{AS}_{X_t} \leftarrow 0 \in R^n$; // n:the size of X_t.
    obtain natural subsets S, outlier candidates O, and NBG on X_t by Algorithm 1;
    calculate the anomaly score (SAS) of each subset in S by Eq. (6);
    compute local anomaly score (LAS) of each sample in $X_t \setminus O$ by Eq. (7);
    calculate anomaly score (AS_{X_t}) of each sample in $X_t \setminus O$ by Eq. (8);
    for *each candidate* x_c *in* O **do**
        $\mathrm{AS}_{X_t}[x_c] \leftarrow \max_{x_i \in X_t} \mathrm{AS}_{X_t}[x_i]$;
    end
    for *each* x_i *in* X_t **do**
        $\mathrm{AS}_{X_0}[x_i]$ = $\mathrm{AS}_{X_0}[x_i]$ + $\mathrm{AS}_{X_t}[x_i]$;
    end
    $t = t - 1$;
end
Return AS_{X_0}.

is determined by the difference between its density and the density peak in the subset to which it belongs, calculated by

$$\mathrm{LAS}(x_i) = \rho_{max} - \rho(x_i). \tag{7}$$

where ρ_{max} is the max local density in s_m that x_i belongs to.

Additionally, LAS undergoes the min-max normalization (i.e., $norm(\mathrm{LAS})$) and then combines with SAS to obtain the overall anomaly scores of samples. Detailed discussion is provided in Section III-D below.

D. Outlier Detection with Sampling Enhancement

LAS represents the internal anomaly information of samples in a local subset, while SAS indicates subset abnormality at the global level. The overall anomaly score (AS) of a sample is computed by

$$\mathrm{AS}(x_i) = \mathrm{SAS}(s_m) + \beta(s_m) \cdot \mathrm{LAS}(x_i). \tag{8}$$

where s_m is the subset to which x_i belongs. $\beta(s_m)$ controls the contribution of LAS from the corresponding s_m, as determined by $\beta(s_m) = \mathrm{SAS}(s_m)$.

We use the subset anomaly scores (SAS) to control the contribution of LAS from corresponding subsets, with SAS values normalized to [0, 1]. The aim is to reduce the likelihood of anomaly scores (AS) from members in low-anomalous subsets exceeding those from members in high-anomalous subsets due to LAS involvement. Besides, low-anomalous subsets are typically surrounded by other subsets, where their internal anomaly information (i.e., LAS) is relatively less significant compared to high-anomalous subsets.

We propose MGOD to detect scatliers and clustliers by scoring the abnormality of each sample using Eq. (8). Additionally, MGOD conducts multiple random samplings to enhance detection, detailed in Algorithm 2. MGOD calculates anomaly scores in each sampled dataset, and integrates the scores obtained from every sampling epoch. The necessity of sampling is to enhance the boundary among sample cluster distributions by removing more potential noises that may degenerate the partition performance. The time complexity of MGOD is $O(t \cdot N \cdot log\ N)$, which is relatively efficient compared to the advanced counterparts.

IV. Experiments

Three experiments are conducted: 1) outlier detection performance evaluation, 2) significance tests, and 3) hyper-parameter evaluations.

A. Experimental Settings

Datasets: 15 real datasets [21] are used in the experiments. Their detailed statistics are provided in the left part of Table I.

Metrics: AUC, the area under the Receiver Operating Characteristic curve, is the most popular measurement in outlier detection studies as it indicates the performance of a detector in ranking true outliers ahead of normal data without setting thresholds [8]. AUC ranges from 0 to 1, with larger values indicating better performance. Besides, the Friedman test and the Bonferroni Dunn (BD) post-hoc test [22] are used for significance analysis.

TABLE I

INFORMATION FOR 15 PUBLIC REAL DATASETS AND AUC OF 8 METHODS ON THESE DATASETS. WE SHOW THE PERFORMANCE RANK IN PARENTHESIS (THE LOWER, THE BETTER), AND MARK THE BEST AND THE SECOND-BEST PERFORMING METHOD(S) IN **BOLD** AND UNDERLINE, RESPECTIVELY.

Datasets	Samples	Category	Outliers	KNN+KFC	LOF+KFC	DGOF+KFC	IFOREST	CBLOF	OCSVM	COPOD	MGOD
Pageblocks	5393	Document	510	0.5598(8)	0.7263(6)	0.6721(7)	0.8977(3)	0.9083(2)	0.8903(4)	0.8754(5)	**0.9154(1)**
WPBC	198	Biomedical	47	0.5323(2)	0.4932(6)	**0.5600(1)**	0.4938(5)	0.4686(8)	0.4743(7)	0.5233(4)	0.5249(3)
Waveform	3443	Physics	100	0.8009(2)	0.7826(4)	0.7969(3)	0.7071(7)	0.7212(6)	0.5393(8)	0.7339(5)	**0.8755(1)**
cardio	1831	Biomedical	176	0.7333(6)	0.5887(8)	0.6041(7)	0.9299(2)	0.8605(5)	0.9286(3)	0.9219(4)	**0.9368(1)**
Cardioto.	2114	Biomedical	466	0.6003(7)	0.5915(8)	0.6189(6)	0.6940(3)	0.6278(5)	**0.7872(1)**	0.6629(4)	0.7431(2)
landsat	6435	Astronautics	1333	0.5937(2)	0.5512(4)	0.5701(3)	0.4872(6)	0.5170(5)	0.3660(8)	0.4215(7)	**0.6290(1)**
optdigits	5216	Image	150	0.5355(5)	0.3930(7)	0.2831(8)	0.7140(3)	0.7547(2)	0.5336(6)	0.6824(4)	**0.9802(1)**
pendigits	6870	Image	156	0.8723(6)	0.4763(8)	0.8305(7)	0.9481(2)	0.8912(5)	0.9354(3)	0.9048(4)	**0.9537(1)**
Pima	768	Biomedical	268	0.6072(5)	0.5396(8)	0.5584(7)	0.6752(2)	0.6700(3)	0.6022(6)	0.6540(4)	**0.7215(1)**
satellite	6435	Astronautics	2036	0.6931(4)	0.5179(8)	0.5701(7)	0.7076(3)	0.7373(2)	0.5972(6)	0.6335(5)	**0.7873(1)**
satimage-2	5803	Astronautics	71	0.9474(6)	0.5989(8)	0.6355(7)	0.9928(3)	**0.9989(1)**	0.9747(4)	0.9745(5)	0.9940(2)
speech	3686	Linguistics	61	0.4787(5)	**0.6917(1)**	N/A(N/A)	0.4905(4)	0.4723(6)	0.4639(7)	0.4911(3)	0.5758(2)
vowels	1456	Linguistics	50	0.9682(2)	0.8684(5)	0.9556(3)	0.7600(6)	0.9214(4)	0.5507(7)	0.4958(8)	**0.9748(1)**
Ionosphere	351	Oryctognosy	126	0.8300(6)	0.8889(2)	0.8379(5)	0.8503(4)	**0.9013(1)**	0.7395(8)	0.7895(7)	0.8598(3)
HeartDis.	270	Biomedical	120	0.6821(3)	0.6511(5)	**0.7151(1)**	0.6233(6)	0.5895(7)	0.5491(8)	0.6946(2)	0.6677(4)
			Avg.Rank	4.60	5.87	5.33	3.93	4.13	5.73	4.73	**1.67**

Compared methods and parameter settings: Some well-known methods are selected for comparison, including IFOREST [9], KNN [23], LOF [24], CBLOF [25], OCSVM [26], COPOD [10], and DGOF [20]. Implementation code for some comparative methods can be found in PyOD [27], a publicly available Python library. The proposed method MGOD defaults to utilizing sampling rate $\eta = 0.8$ and sampling time $t = 60$. Default parameters in PyOD are used for IFOREST, CBLOF, OCSVM, and COPOD. Regarding KNN-based methods, due to the difficulty in manually selecting a k value, we adopt the latest work, KFC [16], for selecting a k. In experimental results, "N/A" denotes out-of-memory.

B. Outlier Detection Performance

Table I presents the AUC of 8 methods on 15 real datasets. It can be observed that the proposed MGOD demonstrates superiority by outperforming most counterparts on all 15 real datasets with an average performance rank of 1.67. More specifically, MGOD achieved an AUC of 0.9802 on the "optdigits" dataset, significantly surpassing the second-best score of 0.7547. It is because the dataset originates from a scenario with clustliers [28]. This exceptional performance indicates that MGOD is more suitable than other methods in scenarios with complex types of outliers. Although MGOD achieves the second on 3 datasets, it obtains AUC values very close to the top. Overall, MGOD demonstrates more stable performance than other counterparts on complex real datasets, without achieving extremely poor results.

The significance test for the experimental results is shown in Figure 3. Firstly, the Friedman test applied to average ranks in Table I yields a p-value = 0.00005, leading to a clear rejection of the null hypothesis. We then proceed with the Bonferroni Dunn (BD) post-hoc test. The Critical Difference (CD) intervals for the two-tailed BD tests at 95% ($\alpha = 0.05$) and 90% ($\alpha = 0.1$) confidence intervals are 2.1413 and 1.9033, respectively, for comparing 8 methods on 15 datasets. Clearly, in Figure 3, all compared methods fall outside the right boundary of the CD intervals, which indicates that MGOD statistically outperforms the other methods.

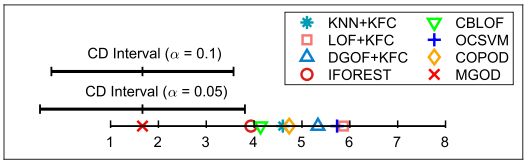

Fig. 3. The two-tailed BD test on the AUC results in Table I.

C. Hyper-parameter Evaluation

We set the random sampling rate at $\eta = 0.8$ for all the above experiments. This recommendation is based on the understanding that the sampled data can effectively capture multiple structures of the original data while reducing the adverse effects of potential outliers on the natural subset partition process, thereby making our approach more robust. To evaluate the reasonableness of the parameter setting, we conduct experiments on all 15 real datasets to assess the influence of different sampling rates η from 0.1 to 1 with a step size of 0.1, under a sufficiently large sampling time, such as $t = 100$. The results are depicted in Figure 4 (a).

It is observed that with the increasing of the sampling rate to 0.9, the corresponding AUC for most datasets shows improvement. When the rate η is less than 0.5, the performance of the proposed MGOD is relatively lower and unstable. This is attributed to the inadequacy of excessively low sampling rates in capturing the underlying structure of the data effectively. When the rate η increases from 0.9 to 1, the AUC decreases on 13 datasets. The hint is that compared to the case without sampling (i.e., $\eta = 1$), sampling can effectively enhance the effectiveness of MGOD. When the rate η is between 0.6 and 0.9, performance of MGOD remains stable in general, which verifies that the recommended sampling rate $\eta = 0.8$ in the baseline scenario would be a reasonable choice.

We set the default value of the sampling time to $t = 60$ in the experiments. To also observe the impact of t, we analyzed AUC changes on 15 real datasets by varying sampling times from 10 to 100 with a sampling rate fixed at the proper

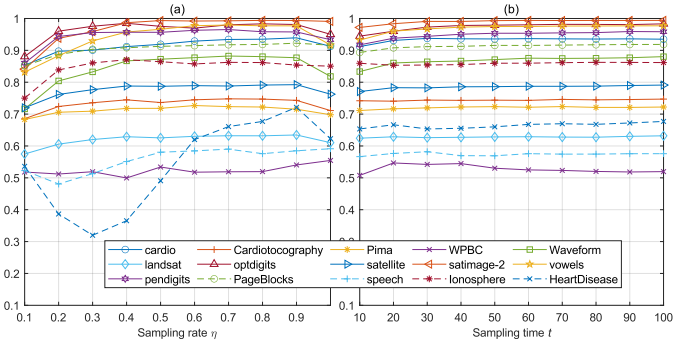

Fig. 4. Hyper-parameter evaluation on 15 real datasets. (a) AUC of MGOD with sampling rates η from 0.1 to 1 with stepsize 0.1 under fixed sampling time $t = 100$. (b) AUC of MGOD on 15 real datasets with a fixed sampling rate $\eta = 0.8$ in different sampling times.

$\eta = 0.8$. The results shown in Figure 4 (b) indicate that for most datasets, AUC gradually increases with additional sampling times. However, when the number of times exceeds 50, extra sampling times do not obviously enhance AUC anymore. Therefore, setting $t = 60$ is a suitable choice.

V. Concluding Remarks

This paper proposes a novel UOD solution called MGOD to detect multi-granular outliers. It first divides the dataset into natural subsets, where the closely distributed subsets are usually connected and the others appear to be isolated. These subsets form a UOD-friendly representation of data, as the clustliers can be effectively partitioned into isolated subsets. Accordingly, a measure that considers both natural subset connection and data point density is proposed to well address the "masking effect" of clustliers while accurately reflecting scatliers. MGOD achieves superior UOD performance, and is robust to hyper-parameter settings. Moreover, it is highly interpretable, and acts as a universal data pre-processing tool for enhancing downstream analysis tasks. Extensive experiments on real datasets verify its effectiveness.

The limitations and future scenarios of MGOD are also worthy of discussion. In this work, we assume that the data is static and complete. Since such a problem setting has no requirements on efficiency, the current time complexity of MGOD is of the same magnitude as most SOTA counterparts. Taking into account the above issues and current research hotspots, we believe that the next avenue would be to extend MGOD to handle streaming or time-series data and improve its time complexity. The corresponding promising application scenarios include the enhancement for downstream tasks like clustering, the detection of medical emergencies, public health events, and mutations of bacteria and virus.

References

[1] A. Boukerche *et al.*, "Outlier detection: Methods, models, and classification," *ACM Computing Surveys*, vol. 53, no. 3, pp. 1–37, 2020.

[2] A. Zeng, H. Rong *et al.*, "Discovery of genetic biomarkers for alzheimer's disease using adaptive convolutional neural networks ensemble and genome-wide association studies," *Interdisciplinary Sciences: Computational Life Sciences*, vol. 13, no. 4, pp. 787–800, 2021.

[3] L. Zhao, Y. Zhang, X. Luo, Y. Zhang *et al.*, "Selecting heterogeneous features based on unified density-guided neighborhood relation for complex biomedical data analysis," in *IEEE International Conference on Bioinformatics and Biomedicine*. IEEE, 2023, pp. 771–778.

[4] M. N. K. Sikder and F. A. Batarseh, "Outlier detection using ai: a survey," *AI Assurance*, pp. 231–291, 2023.

[5] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "On detecting clustered anomalies using sciforest," in *Machine Learning and Knowledge Discovery in Databases*. Springer Berlin Heidelberg, 2010, pp. 274–290.

[6] S. Cai, Y. Zhang *et al.*, "Robust categorical data clustering guided by multi-granular competitive learning," in *IEEE International Conference on Distributed Computing Systems*. IEEE, 2024, pp. 288–299.

[7] Y. Zhang, X. Luo, Q. Chen, R. Zou, Y. Zhang, and Y.-m. Cheung, "Towards unbiased minimal cluster analysis of categorical-and-numerical attribute data," in *International Conference on Pattern Recognition*, 2024, pp. 1–16.

[8] G. O. Campos, A. Zimek *et al.*, "On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study," *Data Mining and Knowledge Discovery*, vol. 30, no. 4, pp. 891–927, 2016.

[9] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation-based anomaly detection," *ACM Transactions on Knowledge Discovery from Data*, vol. 6, no. 1, pp. 1–39, 2012.

[10] Z. Li, Y. Zhao *et al.*, "Copod: Copula-based outlier detection," in *IEEE International Conference on Data Mining*, 2020, pp. 1118–1123.

[11] M. Zhao, Y. Zhang *et al.*, "Unsupervised concept drift detection via imbalanced cluster discriminator learning," in *Chinese Conference on Pattern Recognition and Computer Vision*. Springer, 2023, pp. 31–43.

[12] Y.-m. Cheung *et al.*, "Fast and accurate hierarchical clustering based on growing multilayer topology training," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 3, pp. 876–890, 2018.

[13] B. Tang and H. He, "A local density-based approach for outlier detection," *Neurocomputing*, vol. 241, pp. 171–180, 2017.

[14] J. Xie, Z. Xiong, Q. Dai, X. Wang, and Y. Zhang, "A local-gravitation-based method for the detection of outliers and boundary points," *Knowledge-Based Systems*, vol. 192, p. 105331, 2020.

[15] X. Gu, L. Akoglu, and A. Rinaldo, "Statistical analysis of nearest neighbor methods for anomaly detection," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[16] J. Yang, X. Tan, and S. Rahardja, "Outlier detection: How to select k for k-nearest-neighbors-based outlier detectors," *Pattern Recognition Letters*, vol. 174, pp. 112–117, 2023.

[17] J. Huang, Q. Zhu, L. Yang, and J. Feng, "A non-parameter outlier detection algorithm based on natural neighbor," *Knowledge-Based Systems*, vol. 92, pp. 71–77, 2016.

[18] A. Wahid and C. S. R. Annavarapu, "Nanod: A natural neighbour-based outlier detection algorithm," *Neural Computing and Applications*, vol. 33, pp. 2107–2123, 2021.

[19] Q. Zhu, J. Feng, and J. Huang, "Natural neighbor: A self-adaptive neighborhood method without parameter k," *Pattern Recognition Letters*, vol. 80, pp. 30–36, 2016.

[20] K. Li, X. Gao, X. Jia, B. Xue *et al.*, "Detection of local and clustered outliers based on the density–distance decision graph," *Engineering Applications of Artificial Intelligence*, vol. 110, p. 104719, 2022.

[21] S. Han, X. Hu, H. Huang, M. Jiang, and Y. Zhao, "Adbench: Anomaly detection benchmark," *Advances in Neural Information Processing Systems*, vol. 35, pp. 32 142–32 159, 2022.

[22] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *The Journal of Machine Learning Research*, vol. 7, p. 1–30, dec 2006.

[23] S. Ramaswamy, R. Rastogi, and K. Shim, "Efficient algorithms for mining outliers from large data sets," *SIGMOD Record*, vol. 29, no. 2, p. 427–438, may 2000.

[24] M. M. Breunig, H.-P. Kriegel *et al.*, "Lof: identifying density-based local outliers," *SIGMOD Record*, vol. 29, no. 2, p. 93–104, may 2000.

[25] Z. He, X. Xu, and S. Deng, "Discovering cluster-based local outliers," *Pattern Recognition Letters*, vol. 24, no. 9-10, pp. 1641–1650, 2003.

[26] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Computation*, vol. 13, no. 7, pp. 1443–1471, 2001.

[27] Y. Zhao, Z. Nasrullah, and Z. Li, "Pyod: A python toolbox for scalable outlier detection," *Journal of Machine Learning Research*, vol. 20, no. 96, pp. 1–7, 2019.

[28] C. C. Aggarwal and S. Sathe, "Theoretical foundations and algorithms for outlier ensembles," *SIGKDD Explorations Newsletter*, vol. 17, no. 1, pp. 24–47, 2015.