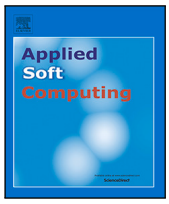
Applied Soft Computing Journal 175 (2025) 113045

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc



Fusing multi-granular-ball fuzzy information to detect outliers

Xinyu Su [a], Shitong Cheng [a], Dezhong Peng [a,c,d], Hongmei Chen [b], Zhong Yuan [a,*]

[a] College of Computer Science, Sichuan University, Chengdu 610065, China
[b] School of Computing and Artificial Intelligence, Southwest Jiaotong University, Chengdu 611756, China
[c] Sichuan National Innovation New Vision UHD Video Technology Co., Ltd, Chengdu 610041, China
[d] Tianfu Jincheng Laboratory, Chengdu 610093, China

ARTICLE INFO

Dataset link: https://github.com/BELLoney/Outlier-detection

Keywords:
Granular computing
Information fusion
Granular-ball computing
Multi-granularity
Outlier detection

ABSTRACT

Outlier detection plays a critical role in data mining and machine learning, and its application value is widely recognized in several industries. However, despite the growing importance of outlier detection, many current outlier detection methods still rely on a single and fine-granularity processing paradigm. Not only does this paradigm lead to inefficient methods, but it also makes the methods vulnerable to noisy data. Furthermore, this processing paradigm ignores the potential multi-granularity information in the data, which may lead to an incomplete understanding of the intrinsic relations and patterns of the data. To further improve the performance of outlier detection, multi-granular-ball fuzzy information granules-based unsupervised outlier detection method (MGBOD) is proposed in this work. In our method, granular-balls with different granularity are first generated and the fuzzy binary relations between the granular-balls with respect to different attributes are computed. Subsequently, two attribute sequences are constructed based on the importance of the attributes. Then, multi-granular-ball fuzzy binary granular structures are constructed based on these two sequences. Finally, the outlier score of the granular-ball is defined by fusing these granules in the granular structures and mapped to the samples in the granular-ball. Experimental results show that, compared with recently proposed methods, our method demonstrates excellent outlier detection performance under a variety of public datasets. The code is publicly available at https://github.com/Mxeron/MGBOD.

1. Introduction

Most tasks in data mining, such as classification, clustering, etc., focus on finding prime patterns in a given dataset. In contrast to these tasks, outlier detection focuses on finding rare and important patterns that consist of a minority of samples in the dataset [1]. Outliers, also known as anomalies, refer to samples in a dataset that deviate so markedly from the normal pattern that it is suspected that they are produced by a different mechanism [2]. As a basic and important task in data mining, outlier detection has received more and more attention in recent years because of its wide application, such as network intrusion detection [3], industrial system monitoring [4], and financial transaction systems [5].

In the real world, complex data is often multi-dimensional and multi-modal, containing rich hierarchical structure and contextual information. Many novel information systems have been proposed to handle complex data [6,7]. To address the challenges faced with complex information systems, information fusion is proposed [8]. Information fusion is a process that collects information from multi-data sources and integrates information in some way to obtain more accurate, comprehensive, and reliable results than a single data source [9,10]. Multi-granularity information fusion is a specific area of information fusion that emphasizes processing and integrating information at different levels or granularities [11,12]. As an effective tool for solving complex problems, multi-granularity information fusion can utilize information with different granularities. This fusion method can significantly improve data processing and analysis by integrating information at different granularities, bringing many benefits to various application areas. However, its application in outlier detection needs to be further explored [11].

In complex data, ignoring the multi-granularity information may lead to an incomplete understanding of the intrinsic relations and patterns of the data, thus affecting the performance of outlier detection [13,14]. Granular computing (GrC) is a widely used set of theories and methods in knowledge discovery that helps us tackle complicated problems by simulating human thinking patterns [15]. Among them, information granulation is an important concept and one of the fundamental problems in the theory of GrC [16–18], and it is closely

* Corresponding author.
E-mail address: yuanzhong@scu.edu.cn (Z. Yuan).

https://doi.org/10.1016/j.asoc.2025.113045
Received 11 August 2024; Received in revised form 3 March 2025; Accepted 12 March 2025
Available online 27 March 2025

related to multi-granularity learning [19]. Information granulation abstracts data to simplify problems and improve processing efficiency by breaking down data into smaller, more manageable units [17]. Information granulation is a key step in understanding complex systems and datasets and has been applied in areas such as rough set theory, machine learning, and databases. The granule and granular structure are two important concepts in information granulation [18]. A granule is a set of objects that are clustered together due to the indistinguishability and similarity of functionality [20]. A granular structure is the mathematical structure of a collection of granules, and the internal structure of each granule is visible [18].

Rough sets, as an important method in information granulation, emphasize the importance of granularity in dealing with problems and have been widely used for outlier detection [21–23]. In rough sets, objects are granulated with equivalence relations. Different equivalence relations induce multi-granular structures. However, these equivalence relations result in rough sets not being able to effectively handle numerical attribute data. It is necessary to discretize numerical attributes before processing them, but this processing changes the potential internal structure of the data and reduces the performance of outlier detection. For this reason, fuzzy rough sets (FRS) have received extensive attention and research, and have been successfully applied to fields such as three-way decision [24], attribute reduction [25], dimensionality reduction [26], and outlier detection [22]. FRS can effectively deal with potential uncertainty information in data, such as fuzziness. Information is granulated through fuzzy binary relations in FRS and information granulation in FRS is called fuzzy information granulation [18]. The applications of FRS in outlier detection have proven beneficial, leading to notable improvements in detection performance [11,21–23].

Considering that existing machine learning methods basically follow a sample-by-sample processing paradigm, this single and fine-granularity processing paradigm is not efficient enough and is susceptible to noisy data. It ignores the inherent multi-granularity of the data [27]. To further improve the performance of different methods, Xia et al. [27] proposed granular-ball computing (GBC). GBC improves the performance of the method by making changes to the input paradigms in conventional machine learning methods [28]. The centerpiece of GBC is the ball structure, a novel data structure with completely symmetric geometric properties and a simple representation. A granular-ball contains only two parameters, the center, and radius in any dimension [28]. The balls can adaptively change the self-generated sizes according to the distributions of different datasets, which in turn results in a multi-granularity representation of the original dataset and being able to adapt to complex datasets. Unlike conventional machine learning methods, GBC starts by generating granular-balls on the dataset and subsequently uses these balls as inputs to the novel method rather than samples. It is the multi-granularity nature of the balls that allows methods constructed based on GBC to process complex data efficiently and to avoid being affected by noise in the data. Existing research on GBC focuses on two aspects, granular-ball generation [28,29] and the application of GBC [27,30]. In granular-ball generation, the quality of balls directly affects the performance of the subsequent methods, so how to efficiently generate high-quality balls is the focus of research in this field. GBC is a novel and efficient method in GrC and currently has many applications such as classification [31,32], clustering [33,34], and attribute reduction [32].

Recently, many efficient rough sets-based outlier detection methods have been proposed. For example, Yuan et al. [35] proposed a novel mixed-attribute outlier detection method based on multigranulation relative entropy. Wang et al. [36] proposed a novel method based on a weighted network model with the similarity and neighborhood relations between samples. Liu et al. [37] proposed the fuzzy granular outlier detection using Markov random walk. Chen et al. [38] proposed a semi-supervised consistency-guided outlier detection method. Chen et al. [39] proposed a method based on improved k-nearest neighbor rough sets. Although these methods have achieved some success, they still have limitations. These methods heavily rely on pairwise distance measures between samples, which are computationally inefficient and not robust to noise. Furthermore, these methods often overlook the multi-granularity information inherent in the data, which may lead to an incomplete understanding of the potential relationships and patterns in the data, which in turn reduces the performance of outlier detection [13,14]. As a novel multi-granularity computing paradigm, GBC addresses these shortcomings. GBC replaces the original samples with a smaller number of granular-balls as the novel multi-granularity processing units in the methods. GBC converts the pairwise distance metric between samples into efficient metrics between granular-balls, which makes the detection methods more efficient and robust. Moreover, GBC enables detection methods to leverage multi-granularity information, thereby improving outlier detection performance.

Based on the above discussion, we construct the multi-granular-ball fuzzy information granules-based unsupervised outlier detection method (MGBOD) by taking advantage of GBC and fuzzy information granulation. Complex data can be effectively simplified through fuzzy information granulation. The original data is transformed into a series of fuzzy sets through fuzzy information granulation. This allows our method to effectively deal with the uncertainty information in the data. The original fine-granularity-based input paradigms are replaced by multi-granularity granular-balls. The above two multi-granularity learning methods can help MGBOD capture the features of the data more comprehensively and improve the performance and robustness of outlier detection by considering both macro and micro views of the data. The outlier detection performance is then improved by multi-granularity information fusion.

The framework diagram of MGBOD is shown in Fig. 1. First, we process the input raw data. Specifically, to achieve multi-granularity representations of the original data, we employ an efficient granular-ball generation method to convert single-granularity samples into a multi-granularity granular-ball set. The multi-granularity property of granular-balls enhances the efficiency of subsequent calculations. Next, we calculate fuzzy binary relations between granular-ball pairs under different attribute subsets to measure their fuzzy similarity, and multiple fuzzy relation matrices are constructed. Based on these relation matrices, we calculate the fuzzy information entropy of the fuzzy relations induced by different attributes, reflecting the importance of each attribute. Subsequently, we sort individual attributes in ascending order of importance to obtain an ordered attribute sequence. Using this ordered attribute sequence, we further construct two distinct sequences, the attribute simplification sequence and attribute complication sequence, by iteratively removing attributes from the original attribute set. With these sequences, we build two groups of multi-granularity granular-ball fuzzy information granules. Finally, we calculate the outlier scores of each granular-ball by weighted fusion of abnormal information from its associated fuzzy information granules. The outlier scores are used to measure the outlier degrees of each granular-ball. The granular-ball outlier scores are then mapped to each internal sample, ensuring every sample receives an outlier score. Based on the outlier scores, we can detect outliers by setting a threshold. Overall, the contribution of our work consists of the following aspects.

(1) A novel fuzzy binary granular structure is proposed by integrating fuzzy information granulation with granular-ball computing.
(2) Based on the attribute importance defined by fuzzy information entropy, we define two attribute sequence sets to construct reasonable granular-ball fuzzy information granules.
(3) A novel unsupervised outlier detection method is proposed by fusing multi-granular-ball fuzzy information granules.
(4) Experiments on datasets of varying scales demonstrate that our method outperforms or is comparable to the state-of-the-art methods.

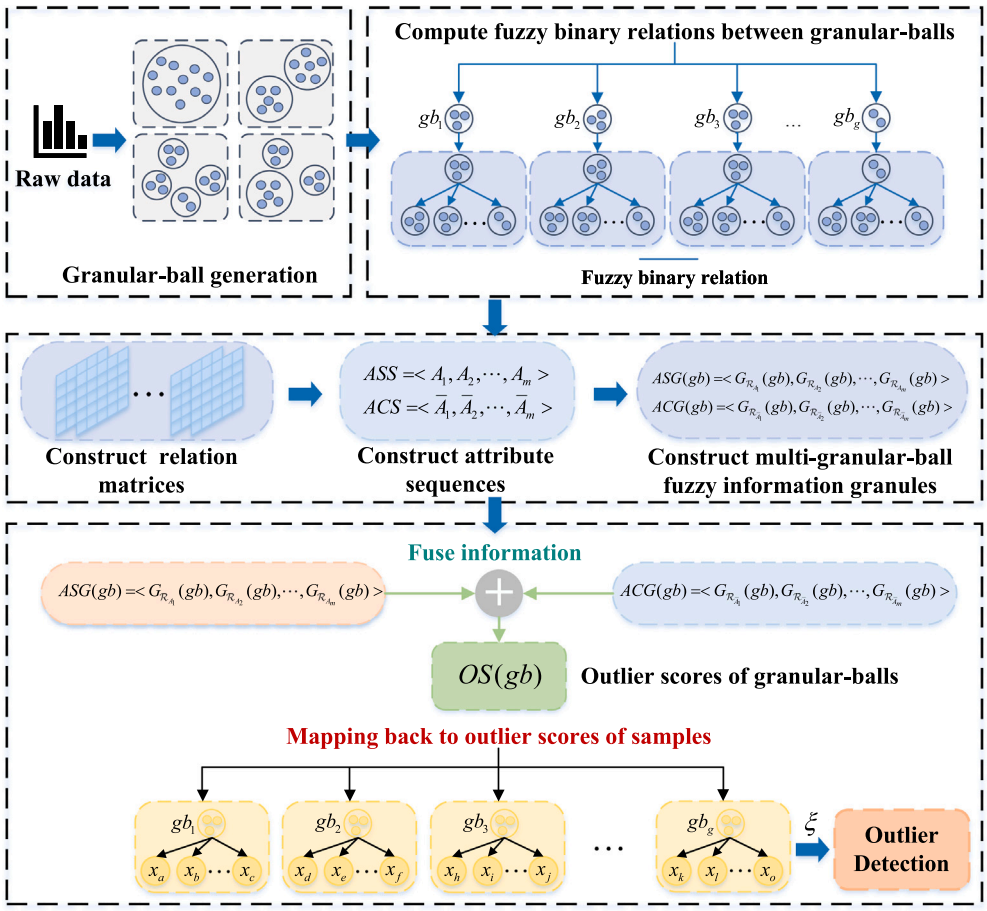


Fig. 1. The framework diagram of MGBOD.

The rest of the paper is organized as follows. Section 2 reviews related works on GBC and rough sets-based outlier detection. Section 3 reviews some preliminary knowledge about fuzzy information granulation and granular-ball generation. Section 4 proposes a novel unsupervised outlier detection method and gives the corresponding algorithm. Section 5 describes the experimental setups and experimental results. Section 6 gives a summary of this work.

2. Related work

In this section, we review related work on GBC theory and rough sets-based methods for outlier detection.

2.1. Granular-ball computing

In the existing work on GBC, the granular-ball is uniformly defined in the following standard form $gb = \{x_i | i = 1, 2, \ldots, k\}$. Each granular-ball has two key parameters, its center and radius. The center of a granular-ball is defined as the mean of the samples contained in the granular-ball, i.e., $c = \frac{1}{k}\sum_{i=1}^{k} x_i$, where x_i and k denote the samples in the granular-ball and the number of these samples, respectively. There are two ways to define the radius of a granular-ball, average distance, and maximum distance. As the name suggests, the radius computed based on the average distance is defined as $r = \frac{1}{k}\sum_{i=1}^{k} \|x_i - c\|$, which is the average distance from all samples in the granular-ball to its center. This radius is often used in classification tasks. The radius based on the maximum distance is defined as $r = \max \|x_i - c\|$, which is the maximum distance between c and the samples in the granular-ball. This radius is often used in clustering tasks. The center and radius of gb_g are denoted as c_g and r_g. In this work, we use Euclidean distance as the above distance metric.

Although GBC has only been proposed for a while, the concept of GBC has driven the development of numerous methods that aim to address challenges related to efficiency, robustness, and interpretability in various areas of artificial intelligence [27,30,31,40–42]. Existing research on GBC is broadly categorized into two areas, one is the study of granular-ball generation methods and the other is the applications of GBC.

The quality of the generated granular-balls affects the performance of the subsequent methods. In terms of granular-ball generation, the early classical generation method is based on k-means [27]. However, this method is rough, the quality of the granular-balls is not high enough and requires manual setting of parameters, making it difficult to adapt to large-scale and complex datasets. Xia et al. [28] proposed a generation method based on k-division, which improves the generation efficiency while ensuring quality. Although these k-means or k-division-based methods have some success, the generation process is random. For this reason, Xie et al. [29] inspired by the hard attention in the attention mechanism, proposed a stable generation method. Xie et al. [33] adopted the weighted distribution measure as a judgment criterion for the splitting of granular-balls. The granular-balls with large radii are refined to further improve the quality of the generated granular-balls.

The high quality of the granular-balls is the key to ensuring that they can be successfully used in a number of applications. Xia et al. [27] proposed two classification models, the granular-ball support vector machine (GBSVM) and granular-ball k-nearest neighbor (GBKNN), by replacing the point input with granular-ball input. Xue et al. [40] proposed granular-ball fuzzy support vector machine (GBFSVM) to efficiently process potentially fuzzy information in classification tasks. Xue

et al. [30] proposed the dual model of GBSVM to solve the problem of the overlap that exists in the heterogeneous samples between heterogeneous granular-balls. Xia et al. [31] proposed a sampling method called granular-ball sampling (GBS) for classification tasks that not only reduces data size but also improve data quality. Peng et al. [32] proposed a robust variable parameter granular-ball model (VPGB) to achieve both attribute reduction and classification in a label noise environment. Zhang et al. [41] proposed a novel granular-ball rough set model (GBRS) and two incremental learning models were constructed based on GBRS. Xia et al. [42] proposed a novel granular-ball neighborhood rough sets (GBNRS) and attribute reduction was implemented based on GBNRS. Cheng et al. [34] proposed a granular-ball-based density peaks clustering algorithm called GB-DP. GB-DP first clustered the granular-balls and then extended the clustering results to the original data. In addition, there are efficient clustering methods that combine granular-ball with density-based spatial clustering of applications with noise (DBSCAN) [43] and spectral clustering [33].

GBC has great scalability and application potential due to its ability to improve method performance directly from the data input paradigm. However, the application of GBC in outlier detection needs to be further explored and investigated.

2.2. Rough sets-based outlier detection

Outlier detection methods based on rough sets exploit the imprecision, incompleteness, and uncertainty of data to identify anomalies or outliers in datasets.

The first is the classical rough sets-based outlier detection methods. Jiang et al. [44] proposed a novel outlier detection method based on exceptional sets constructed on rough sets. Jiang et al. [45] defined a novel outlier based on a rough membership function and proposed an algorithm to detect such outliers. Jiang et al. [46] proposed an outlier detection method based on GrC and rough sets. Jiang et al. [47] proposed an outlier detection algorithm based on approximation accuracy entropy within the rough sets. Singh et al. [48] proposed a novel outlier detection method for streaming data by introducing concepts such as relative cardinality and entropy outlier factor theory. Since these classical methods are implemented based on equivalence relations, they cannot handle numerical data effectively.

Many extended rough sets methods have been proposed to support outlier detection under numerical and mixed-attribute data. Yuan et al. [35] proposed an outlier detection method based on neighborhood rough sets, extending the traditional distance-based and rough sets-based methods. Then a novel mixed-attribute outlier detection method based on multigranulation relative entropy is proposed using neighborhood rough sets [49]. Wang et al. [36] proposed a mixed-attribute data outlier detection method based on a weighted network model with the similarity and neighborhood relations between samples. Yuan et al. [22] proposed an outlier detection method based on fuzzy-rough density that can effectively handle mixed-attribute data. Yuan et al. [11] realized FRS-based outlier detection by constructing multi-fuzzy granules in order to achieve multi-granularity information fusion. The application of rough sets and their related theories enables the construction of more efficient outlier detection methods. However, these existing methods basically use a sample-by-sample processing method, and the performance of the methods needs to be further improved.

3. Preliminaries

In this section, we review the preliminary knowledge about fuzzy information granulation and granular-ball generation.

3.1. Fuzzy information granulation

A given complex dataset can usually be denoted by an information system. An information system is a 2-tuple $IS = \langle S, A\rangle$. In $IS = \langle S, A\rangle$, $S = \{x_1, x_2, \ldots, x_n\}$ is a set of samples; $A = \{a_1, a_2, \ldots, a_m\}$ is a set of condition attributes; for any $x \in S$ and $a \in A$, $a(x)$ denotes the value of x with respect to attribute a.

A fuzzy binary relation $\mathcal{R}$ on S is defined as $\mathcal{R} : S \times S \to [0, 1]$. Fuzzy information granulation is usually realized through a fuzzy binary relation. Given a dataset S, $\mathcal{R}$ denotes a fuzzy binary relation on S, which is represented by the following relation matrix.

$$M(\mathcal{R}) = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ r_{n1} & r_{n2} & \cdots & r_{nn} \end{pmatrix}, \tag{1}$$

where $r_{ij} = \mathcal{R}(x_i, x_j) \in [0, 1]$ denotes the similarity between samples x_i and x_j.

Given two fuzzy binary relations $\mathcal{R}_1, \mathcal{R}_2$. For any $x, y \in S$, some operations between them are given as follows.

(1) $\mathcal{R}_1 = \mathcal{R}_2 \Leftrightarrow \mathcal{R}_1(x, y) = \mathcal{R}_2(x, y)$;
(2) $\mathcal{R} = \mathcal{R}_1 \cup \mathcal{R}_2 \Leftrightarrow \mathcal{R} = \max\{\mathcal{R}_1(x, y), \mathcal{R}_2(x, y)\}$;
(3) $\mathcal{R} = \mathcal{R}_1 \cap \mathcal{R}_2 \Leftrightarrow \mathcal{R} = \min\{\mathcal{R}_1(x, y), \mathcal{R}_2(x, y)\}$;
(4) $\mathcal{R}_1 \subseteq \mathcal{R}_2 \Leftrightarrow \mathcal{R}_1(x, y) \leq \mathcal{R}_2(x, y)$.

The granulation of a dataset into a set of information granules using a general binary relation is called a binary granular structure [18]. Given a nonempty finite sample set S, for any fuzzy binary relation $\mathcal{R}$, $\mathcal{R}$ can correspondingly induce a set of fuzzy information granules called a fuzzy binary granular structure [18]. A fuzzy binary granular structure is formally denoted as

$$F(\mathcal{R}) = (G_{\mathcal{R}}(x_1), G_{\mathcal{R}}(x_2), \ldots, G_{\mathcal{R}}(x_n)), \tag{2}$$

where $G_{\mathcal{R}}(x_i) = \frac{\mathcal{R}(x_i,x_1)}{x_1} + \frac{\mathcal{R}(x_i,x_2)}{x_2} + \cdots + \frac{\mathcal{R}(x_i,x_n)}{x_n}$. $G_{\mathcal{R}}(x_i)$ denotes the fuzzy information granule determined by x_i with respect to $\mathcal{R}$, and $\mathcal{R}(x_i, x_j)$ denotes the similarity between x_i and x_j. The plus operation here denotes the union of samples. The cardinality of the fuzzy information granule $G_{\mathcal{R}}(x_i)$ can be computed with

$$|G_{\mathcal{R}}(x_i)| = \sum_{j=1}^{n} r_{ij}. \tag{3}$$

Using general binary relations, objects in a dataset can be granulated into a set of information granules, referred to as a binary granular structure [16]. Equivalence relations are used for granulation in rough sets. In GrC methods based on fuzzy set theory, such as FRS, a fuzzy binary relation is usually used for granulation. Under different fuzzy binary relations, objects can be granulated into multi-fuzzy information granules [16]. Through fuzzy information granulation in FRS, complex datasets can be effectively simplified, key information can be extracted, and data analysis methods capable of handling uncertainty can be constructed.

From the above definitions, it can be seen that the fuzzy information granules obtained from fuzzy information granulation in FRS are the fuzzy sets corresponding to each sample. The fuzzy granular structure is the set of these fuzzy sets. Many relation matrices built on different fuzzy binary relations can be viewed as multi-fuzzy binary granular structures, and a row in each relation matrix is a fuzzy information granule. The original dataset is transformed through fuzzy information granulation into a series of fuzzy sets that can better represent the uncertainty information in the data. In this work, we investigate an unsupervised outlier detection method by fusing multi-fuzzy information granules obtained by information granulation of multi-fuzzy binary relations induced by different attribute subsets in FRS.

3.2. Granular-ball generation

Generating granular-balls is one of the key steps in GBC proposed by Xia et al. [27]. The classical granular-ball generation process is as follows. In the beginning, it treats the whole dataset as a coarse ball, then splits the coarse ball to obtain subsequent fine balls by 2-means or 2-division until the quality of all the balls reaches the threshold [27]. Considering the problems in classical granular-ball generation methods, Xie et al. [33] improved the process of generating the granular-balls, and proposed a more robust and reasonable granular-ball generation method. The method first proposes to compute the quality of the granular-ball by distribution measure.

Definition 1. Given a set of granular-balls $GB_S = \{gb_1, gb_2, \ldots, gb_g\}$ generated on the nonempty finite sample set S. For any granular-ball $gb_t \in GB_S$, its distribution measure is defined as

$$DM(gb_t) = \frac{s_t}{|gb_t|}, \tag{4}$$

where $s_t = \sum_{x_i \in gb_t} \|x_i - c_t\|$, and $|gb_t|$ denotes the number of samples in gb_t.

The distribution measure essentially computes the average radius of the granular-balls and these balls are split according to their respective distribution measure. First, the whole dataset is considered as a large ball gb. Subsequently, the two farthest points p_1 and p_2 are chosen to split gb into two sub-balls gb_1 and gb_2. We compute the distribution measures of the above three balls, i.e. gb, gb_2, and gb_2, to characterize the quality of each ball. In turn, we determine whether the balls need to be divided or not based on the above quality. In the previous work [50], if gb is to be split, then both $DM(gb_1)$ and $DM(gb_2)$ should be smaller than $DM(gb)$. However, this method leads to the failure of ball splitting when there is a lot of noise [33]. In Eq. (4), the premise for granular-ball splitting is that the $DM(gb)$ value of the parent ball gb reasonably reflects the tightness. However, if the parent ball gb contains noise points (i.e., points far from the center), these noise points will increase the value of s in $DM(gb)$. The DM values of the sub-balls gb_1 and gb_2 are computed in their respective local regions, which are usually unaffected by the noise points. As a result, $DM(gb_1)$ and $DM(gb_2)$ are typically smaller. Therefore, if $DM(gb)$ is very large, even if the internal structure of the sub-balls is highly sparse (or of poor quality), their $DM(gb_1)$ and $DM(gb_2)$ value can easily satisfy the splitting condition. This is because the large distances of noise points from the center inflate $DM(gb)$, thereby undermining the effectiveness of the splitting rule.

To better adapt to noisy environments, a novel splitting criterion of granular-balls is introduced in [33] by a weighted distribution measure. The definition of weighted distribution measure is as follows.

Definition 2. Given a granular-ball G. The granular-balls G_1 and G_2 are two sub-balls generated by the splitting of G, then the weighted DM of G is defined as

$$DM^*(G) = \frac{|G_1|}{|G|} DM(G_1) + \frac{|G_2|}{|G|} DM(G_2), \tag{5}$$

where $|G|$, $|G_1|$, and $|G_2|$ denote the number of samples in the corresponding granular-balls, respectively. If $DM^*(G) \geq DM(G)$, then splits G.

In Eq. (5), the weighted distribution measure DM^* is introduced, which integrates the tightness of the sub-balls by weighting them according to their size proportions. The weighted distribution measure makes the splitting judgment more flexible and avoids the shortcomings of solely relying on the parent ball. The noise points that are far from the center of the parent ball gb may cause the $DM(gb)$ to increase. However, noise points have less impact on the sub-balls because the DM values of the sub-balls are computed based on their own local regions. Even if noise points affect the parent ball's DM value, the final weighted distribution measure DM^* will make a more reasonable judgment based on the actual conditions of the sub-balls, thereby improving the robustness of the splitting process.

However, some granular-ball with too large radii may still be affected by boundary points and noise points and need to be further split [33]. Therefore, this method also needs to determine the radii of the granular-balls at the end. If the radius of a granular-ball is too large, it needs to be further split: if $r_j > 2 \times \max(\text{mean}(r), \text{median}(r))$, gb_j needs to be split. $\text{mean}(r)$ and $\text{median}(r)$ represent the mean and median of the radii of all granular-balls [33].

Algorithm 1: Granular-ball generation

Input: $IS = \langle S, A \rangle$
Output: Granular-balls GB

$GB \leftarrow \{S\}$;
for *each granular-ball* gb_t *in* GB **do**
  Compute $DM(gb_t)$ and $DM^*(gb_t)$ by Eq. (4) and Eq. (5);
  if $DM^*(gb_t) \geq DM(gb_t)$ **then**
    Split gb_t into gb_{t_1} and gb_{t_2};
    Remove gb_t from GB;
    $GB \leftarrow GB \cup \{gb_{t_1}, gb_{t_2}\}$;
  end
end
Compute mean(r) and median(r) in GB;
for *each granular-ball* gb_t *in* GB **do**
  if $r_t > 2 \times \max(\text{mean}(r), \text{median}(r))$ **then**
    Split gb_t into gb_{t_1} and gb_{t_2};
    Remove gb_t from GB;
    $GB \leftarrow GB \cup \{gb_{t_1}, gb_{t_2}\}$;
  end
end
return GB.

A diagram of the granular-ball generation process based on Algorithm 1 is shown in Fig. 2, where the blue dots are samples in the raw dataset and the red forks are the centers of the granular-balls. As can be seen from the figure, the granular-balls generated at the beginning have a large radius and coarse granularity. As the number of iterations increases, more and more granular-balls are generated and the radii of the granular-balls decreases, which in turn creates a reasonable coverage of the raw dataset.

Based on the above definitions, the specific granular-ball generation algorithm is given in Algorithm 1. The time complexity of this algorithm is $O(n\log n)$ [33], where n is the number of samples in the dataset. With this algorithm, the input raw dataset can be transformed into a novel data representation with multi-granular-balls. In the subsequent processing, we employ these multi-granularity granular-balls as a novel input paradigm for our method.

4. Outlier detection based on multi-granular-ball fuzzy information

In this section, we propose a novel outlier detection method by fusing multi-granular-ball fuzzy information granules in different fuzzy binary granular structures, and the corresponding algorithm pseudo-code of the method is given.

4.1. Outlier detection method

In this work, samples are granulated through multi-fuzzy binary relations with respect to different attribute subsets to construct the corresponding fuzzy binary granular structures. The outlier detection method is designed by fusing multi-granular-ball fuzzy information granules in these granular structures. Multi-fuzzy binary relations induced by different subsets of attributes are given below.

It is important to define a proper fuzzy binary relation before fuzzy information granulation. The definition of the fuzzy binary relation between samples is given below first, followed by the definition between granular-balls.

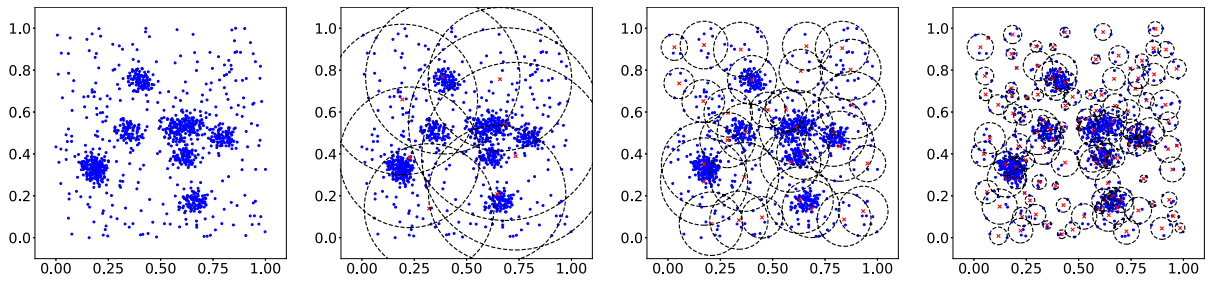


Fig. 2. The granular-ball generation process.

Definition 3. Given a attribute subset B, for any $x_i, x_j \in S$, the fuzzy binary relation $\mathcal{R}_B$ is defined as

$$\mathcal{R}_B(x_i, x_j) = \begin{cases} \frac{1}{1+D_B(x_i,x_j)}, & \text{if } \frac{1}{1+D_B(x_i,x_j)} \geq \sigma; \\ 0, & \text{else;} \end{cases} \tag{6}$$

where $D_B(x_i, x_j) = \sqrt{\sum_{a_k \in B}(a_k(x_i) - a_k(x_j))^2}$; σ is a hyperparameter with a value range of $[0, 1]$. After normalization, the values of $D_B(x_i, x_j)$ are always larger than or equal to 0. After adding 1 to the denominator, the value of the denominator is always larger than or equal to 1. Therefore, it is clear that the values of $\mathcal{R}_B(x_i, x_j)$ always remain in the interval $[0, 1]$. Adding 1 to the denominator is to avoid the denominator being zero when $D_B = 0$. It can also ensure the value of fuzzy relation can be guaranteed to be within the interval of $(0, 1]$ to avoid large fluctuations.

Given a fuzzy binary relation $\mathcal{R}_B$ induced by $B \subseteq A$, we can use the following relation matrix to denote $\mathcal{R}_B$.

$$M(\mathcal{R}_B) = \begin{pmatrix} r_{11}^B & r_{12}^B & \cdots & r_{1n}^B \\ r_{21}^B & r_{22}^B & \cdots & r_{2n}^B \\ \vdots & \vdots & \ddots & \vdots \\ r_{n1}^B & r_{n2}^B & \cdots & r_{nn}^B \end{pmatrix}, \tag{7}$$

where $r_{ij}^B = \mathcal{R}_B(x_i, x_j) \in [0, 1]$ denotes the similarity between samples x_i and x_j with respect to $\mathcal{R}_B$. Fuzzy binary relations induced by different subsets of attributes can induce corresponding fuzzy binary granular structures.

Similarly, through these fuzzy binary relations induced by different subsets of attributes, the samples can be granulated differently, which in turn induces multi-fuzzy binary granular structures. Given a fuzzy binary relation $\mathcal{R}_B$, a fuzzy binary granular structure with respect to $\mathcal{R}_B$ is denoted as

$$F(\mathcal{R}_B) = (G_{\mathcal{R}_B}(x_1), G_{\mathcal{R}_B}(x_2), \ldots, G_{\mathcal{R}_B}(x_n)), \tag{8}$$

where $G_{\mathcal{R}_B}(x_i) = \frac{\mathcal{R}_B(x_i,x_1)}{x_1} + \frac{\mathcal{R}_B(x_i,x_2)}{x_2} + \cdots + \frac{\mathcal{R}_B(x_i,x_n)}{x_n}$; $G_{\mathcal{R}_B}(x_i)$ denotes the fuzzy information granule determined by x_i with respect to $\mathcal{R}_B$. A fuzzy binary relation induced by a subset of attributes can induce a fuzzy binary granular structure that contains multi-fuzzy information granules.

The fuzzy binary relation between the finest-granularity samples is given above, and the same can be given between multi-granularity granular-balls.

Given a set of granular-balls $GB_S = \{gb_1, gb_2, \ldots, gb_g\}$ generated on S. For any $gb_i, gb_j \in GB_S$, the fuzzy binary relation $\mathcal{R}_B$ between gb_i and gb_j induced by B is computed as $\mathcal{R}_B(gb_i, gb_j) = \mathcal{R}_B(c_i, c_j)$ where c_i and c_j denote the centers of gb_i and gb_j respectively. The number of attributes in the center of a granular-ball is the same as the number of attributes in a sample, and the granular-ball contains one or more samples, with each attribute in its center being the average of the corresponding attributes of all the samples inside the granular-ball. Therefore, the fuzzy binary relations between granular-balls can be computed directly according to Eq. (6) by their centers. Similarly, the fuzzy binary relation $\mathcal{R}_B$ between granular-balls induced by B can be denoted by the following relation matrix.

Table 1
A data table.

S	a_1	a_2	a_3	a_4
x_1	3	120	3	280
x_2	2	20	2	180
x_3	2	180	5	150
x_4	1	160	6	50
x_5	2	130	4	80
x_6	3	30	1	100
x_7	1	40	3	200
x_8	3	90	5	140

$$MG(\mathcal{R}_B) = \begin{pmatrix} \hat{r}_{11}^B & \hat{r}_{12}^B & \cdots & \hat{r}_{1g}^B \\ \hat{r}_{21}^B & \hat{r}_{22}^B & \cdots & \hat{r}_{2g}^B \\ \vdots & \vdots & \ddots & \vdots \\ \hat{r}_{g1}^B & \hat{r}_{g2}^B & \cdots & \hat{r}_{gg}^B \end{pmatrix}, \tag{9}$$

where $\hat{r}_{ij}^B = \mathcal{R}_B(gb_i, gb_j) \in [0, 1]$ denotes the similarity between granular-balls gb_i and gb_j with respect to $\mathcal{R}_B$.

Definition 4. Given a set of granular-balls $GB_S = \{gb_1, gb_2, \ldots, gb_g\}$ generated on S. The granular-ball fuzzy binary granular structure induced by $\mathcal{R}_B$ is defined as

$$GF(\mathcal{R}_B) = \{G_{\mathcal{R}_B}(gb_1), G_{\mathcal{R}_B}(gb_2), \ldots, G_{\mathcal{R}_B}(gb_g)\}, \tag{10}$$

where $G_{\mathcal{R}_B}(gb_i) = \frac{\mathcal{R}_B(gb_i,gb_1)}{gb_1} + \frac{\mathcal{R}_B(gb_i,gb_2)}{gb_2} + \cdots + \frac{\mathcal{R}_B(gb_i,gb_g)}{gb_g}$ is a granular-ball fuzzy information granule determined by gb_i with respect to $\mathcal{R}_B$.

Note that the granular-ball fuzzy binary granular structure here is relative to the classical object (or sample) fuzzy binary granular structure. The subject of discussion here changes from the classical object to the granular-ball. The cardinality of the granular-ball fuzzy information granule $G_{\mathcal{R}_B}(gb_i)$ can be computed with

$$|G_{\mathcal{R}_B}(gb_i)| = \sum_{j=1}^{n} \hat{r}_{ij}^B. \tag{11}$$

Example 1. Let $IS = \langle S, A \rangle$ be a example data table for explanation (as shown in Table 1).

After normalization, we can get 4 granular-balls $GB_S = \{gb_1, gb_2, gb_3, gb_4\}$ according to the granular-ball generation method in Algorithm 1. gb_1 contains samples x_2 and x_6; gb_2 contains samples x_1 and x_8; gb_3 contains sample x_7; gb_4 contains samples x_3, x_4, x_5. We compute the granular-ball centers $c_1 = [0.750, 0.031, 0.100, 0.462], c_2 = [1, 0.531, 0.6, 0.731], c_3 = [0, 0.125, 0.400, 0], c_4 = [0.333, 0.854, 0.800, 0.282]$. Then, we set $\sigma = 0.6$ and compute the fuzzy relation matrix

by Eq. (6) as $MG(\mathcal{R}_{a_1}) = \begin{pmatrix} 1 & 0.800 & 0 & 0.706 \\ 0.800 & 1 & 0 & 0.600 \\ 0 & 0 & 1 & 0.750 \\ 0.706 & 0.600 & 0.750 & 1 \end{pmatrix}$,

$MG(\mathcal{R}_{a_2}) = \begin{pmatrix} 1 & 0.667 & 0.914 & 0 \\ 0.667 & 1 & 0.711 & 0.756 \\ 0.914 & 0.711 & 1 & 0 \\ 0 & 0.756 & 0 & 1 \end{pmatrix}$,

$MG(\mathcal{R}_{a_3}) = \begin{pmatrix} 1 & 0.667 & 0.769 & 0 \\ 0.667 & 1 & 0.833 & 0.833 \\ 0.769 & 0.833 & 1 & 0.714 \\ 0 & 0.833 & 0.714 & 1 \end{pmatrix}$,

$MG(\mathcal{R}_{a_4}) = \begin{pmatrix} 1 & 0.788 & 0.684 & 0.848 \\ 0.788 & 1 & 0 & 0.690 \\ 0.684 & 0 & 1 & 0.780 \\ 0.848 & 0.690 & 0.780 & 1 \end{pmatrix}$.

The information entropy reflects the degree of disorder or confusion in the data distribution, and such disorder and confusion can be regarded as manifestations of outliers. Therefore, we use fuzzy information entropy to measure the potential uncertainty of a fuzzy information granule [51].

Definition 5. Let $GF(\mathcal{R}_B) = \{G_{\mathcal{R}_B}(gb_1), G_{\mathcal{R}_B}(gb_2), \dots, G_{\mathcal{R}_B}(gb_g)\}$. The fuzzy information entropy with respect to $\mathcal{R}_B$ on GB_S is defined as

$$EN(\mathcal{R}_B) = EN(B) = -\frac{1}{g}\sum_{i=1}^{g} \log_2 \frac{|G_{\mathcal{R}_B}(gb_i)|}{g}. \tag{12}$$

The fuzzy information entropy on granular-balls is obtained by computing the mean value of the cardinality of multi-granular-ball fuzzy information granules with respect to a certain fuzzy binary relation. This entropy can quantitatively represent the discriminative ability of a certain fuzzy binary relation, and thus the importance of the corresponding attributes [11]. The larger the fuzzy information entropy, the stronger the discriminative ability and the more important the attribute is. Therefore, We can characterize the importance of attributes by fuzzy information entropy. Based on the above analysis, for any attribute a_p, the importance of a_p is denoted as $imp(a_p) = EN(\mathcal{R}_{a_p}) = EN(\{a_p\})$.

Example 2. From Example 1, we can compute the importance of each attribute in Table 1. $imp(a_1) = EN(\mathcal{R}_{a_1}) = -1/4(\log_2(2.506/4) + \log_2(1.4/4) + \log_2(1.75/4) + \log_2(3.056/4)) \approx 0.748, imp(a_2) \approx 0.695$, $imp(a_3) \approx 0.475, imp(a_4) \approx 0.482$.

After computing the importance of each attribute in A, a sequence of attributes can be obtained by sorting their importance.

Definition 6. Given a set of attributes $A = \{a_1, a_2, \dots, a_m\}$, the ordered attribute sequence is defined as

$$AS = \langle \overline{a}_1, \overline{a}_2, \dots, \overline{a}_m \rangle, \tag{13}$$

where $imp(\overline{a}_k) \le imp(\overline{a}_{k+1})$.

Based on the sequence of attributes, removing one attribute from the set of attributes A gradually in a forward–backward or backward-forward manner until only one attribute remains at the end, two sequences of attribute subsets can be obtained, i.e., the attribute simplification sequence and attribute complication sequence.

Definition 7. The attribute simplification sequence ASS and attribute complication sequence ACS are respectively constructed as

$$ASS = \langle A_1, A_2, \dots, A_m \rangle; \tag{14}$$

$$ACS = \langle \overline{A}_1, \overline{A}_2, \dots, \overline{A}_m \rangle, \tag{15}$$

where $A_k \subseteq A, A_1 = A, A_m = \{\overline{a}_m\}$ and $A_{k+1} = A_k - \{\overline{a}_k\}$; $\overline{A}_k \subseteq A, \overline{A}_1 = A, \overline{A}_m = \{\overline{a}_1\}$ and $\overline{A}_{k+1} = \overline{A}_k - \{\overline{a}_{m-k+1}\}$.

Based on the above-obtained sequences of attribute subsets, where each set can determine a fuzzy binary relation. Then, multi-granular-ball fuzzy binary granular structures can be obtained. These structures are the key to carrying out subsequent outlier detection.

Definition 8. The attribute simplification and attribute complication multi-granular-ball fuzzy information granules determined by gb with respect to the attribute subsets in ASS and ACS are respectively constructed as

$$ASG(gb) = \langle G_{\mathcal{R}_{A_1}}(gb), G_{\mathcal{R}_{A_2}}(gb), \dots, G_{\mathcal{R}_{A_m}}(gb) \rangle; \tag{16}$$

$$ACG(gb) = \langle G_{\mathcal{R}_{\overline{A}_1}}(gb), G_{\mathcal{R}_{\overline{A}_2}}(gb), \dots, G_{\mathcal{R}_{\overline{A}_m}}(gb) \rangle. \tag{17}$$

Similarly, for any granular-ball $gb_t \in GB_S$, the corresponding $ASG(gb_t)$ and $ACG(gb_t)$ can be obtained.

Example 3. From Examples 1 and 2, we can get the ordered attribute sequence $AS = \langle a_3, a_4, a_2, a_1 \rangle$ by Eq. (13). Then, we can construct the attribute simplification sequence ASS and attribute complication sequence ACS by Definition 7. We have

$ASS = \langle \underline{A}_1, \underline{A}_2, \underline{A}_3, \underline{A}_4 \rangle = \langle \{a_1, a_2, a_3, a_4\}, \{a_1, a_2, a_4\}, \{a_1, a_2\}, \{a_1\} \rangle$,
$ACS = \langle \overline{A}_1, \overline{A}_2, \overline{A}_3, \overline{A}_4 \rangle = \langle \{a_1, a_2, a_3, a_4\}, \{a_2, a_3, a_4\}, \{a_3, a_4\}, \{a_3\} \rangle$.
Next, for $gb_1 \in GB_S$, we have
$ASG(gb_1) = \langle G_{\mathcal{R}_{A_1}}(gb_1), G_{\mathcal{R}_{A_2}}(gb_1), G_{\mathcal{R}_{A_3}}(gb_1), G_{\mathcal{R}_{A_4}}(gb_1) \rangle \approx \langle (1,0,0,0), (1,0.617,0,0), (1,0.641,0,0), (1,0.8,0,0.706) \rangle$,
$ASG(gb_1) = \langle G_{\mathcal{R}_{\overline{A}_1}}(gb_1), G_{\mathcal{R}_{\overline{A}_2}}(gb_1), G_{\mathcal{R}_{\overline{A}_3}}(gb_1), G_{\mathcal{R}_{\overline{A}_4}}(gb_1) \rangle$
$\approx \langle (1,0,0,0), (1,0.642,0,0), (1,0.638,0.645,0), (1,0.667,0.769,0) \rangle$.

It is intuitive to construct a novel subset of attributes through the above attribute sequence and further construct granular-ball fuzzy information granules. Constructing the granules on the whole attribute subsets would incur a huge time and space overhead, which is impractical for existing devices. This is because an exponential number of subsets of attributes are generated for the set of attributes. Therefore, in this work, we construct granules on the above-constructed attribute simplification sequence ASS and attribute complication sequence ACS. This construction not only obtains as much information as possible but also reduces the time complexity.

In the following, we define the outlier scores of the samples by fusing multi-granular-ball fuzzy information granules in different fuzzy binary granular structures to characterize the outlier degrees of samples. These fuzzy binary granular structures are constructed on the attribute simplification and attribute complication sequences.

First, the outlier scores of granular-balls are defined and then mapped to the samples inside the granular-balls.

Definition 9. The multi-granular-ball fuzzy information-based outlier score of sample x is defined as

$$OS(x) = OS(\hat{gb}) = 1 - W(\hat{gb}) \frac{\sum_{k=1}^{m} \left(\frac{|G_{\mathcal{R}_{A_k}}(\hat{gb})| + |G_{\mathcal{R}_{\overline{A}_k}}(\hat{gb})|}{2g} \right)}{m}, \tag{18}$$

where the weight $W(\hat{gb}) = \sqrt[3]{\sum_{k=1}^{m} \frac{|G_{\mathcal{R}_{A_k}}(\hat{gb})|}{g}/m}$; $\hat{gb}$ denotes the granular-ball to which sample x belongs.

In the above definition, the outlier score of the sample x is equal to the outlier score of the granular-ball $\hat{gb}$ to which it belongs. The outlier score of a granular-ball is computed by fusing the mean value of previously constructed granular-ball fuzzy information granules and the weight corresponding to that granular-ball. The smaller mean value of multi-granular-ball fuzzy information granules of $\hat{gb}$, suggests that $\hat{gb}$ has a smaller similarity with other granular-balls and tends to belong to a minority class. Therefore, $\hat{gb}$ is more likely to be an outlier granular-ball, and in turn, for any samples $x \in \hat{gb}$ are more likely to be outliers. Therefore, a relatively larger outlier score should be assigned to x.

Definition 10. Given a threshold ξ. For any $x \in S$, if $OS(x) > \xi$, then x is called a multi-granular-ball fuzzy information-based outlier in S.

Example 4. From Examples 1, 2, and 3, for $gb_1 \in GB_S$, the weight is computed as $W(gb_1) = \sqrt[3]{\frac{2.506/4+2.581/4+2.436/4+3.32/4}{4}} \approx 0.878$. The multi-granular-ball fuzzy information-based outlier score of gb_1 is computed as $OS(gb_1) = 1 - 0.878 \times \frac{(1+1)/8+(1.617+1.642)/8+(1.641+2.283)/8+(2.506+2.436)/8}{4} \approx 0.612$. Similarly, we can get $OS(gb_2) \approx 0.584, OS(gb_3) \approx 0.650$, $OS(gb_4) \approx 0.629$. Finally, we map the outlier score of the granular-ball to each sample in the granular-ball, and we can get $OS(x_1) \approx 0.584, OS(x_2) \approx 0.612, OS(x_3) \approx 0.629, OS(x_4) \approx 0.629, OS(x_5) \approx 0.629, OS(x_6) \approx 0.612, OS(x_7) \approx 0.650, OS(x_8) \approx 0.584$. If we set threshold $\xi = 0.63$, then x_7 is detected as the outlier by MGBOD.

4.2. Outlier detection algorithm

We give the corresponding outlier detection algorithm based on the method introduced above. As shown in Algorithm 2, the whole process is intuitive. Firstly, granular-balls are generated by Algorithm 1. Subsequently, we compute the fuzzy binary relations between the granular-balls and construct the corresponding relation matrices. Then, we compute the importance of the attributes and construct the attribute sequences. Multi-granular-ball fuzzy binary granular structures are constructed based on attribute simplification and attribute complication sequences. Based on multi-granular-ball fuzzy information granules in these granular structures, we compute the outlier scores of each granular-ball and map them to the samples inside the granular-balls.

Algorithm 2: MGBOD

Input: $IS = \langle S, A \rangle$, σ
Output: Outlier scores *score*

```
score ← ∅;
Generate GB_S on S by Algorithm 1;
for i ← 1 to m do
    // Iterate over all attributes a_i, where m
       denotes the number of attributes
    Compute MG(R_{a_i}) by Eq. (6);
    Compute EN(a_i) by Eq. (12);
end
Construct AS by Definition 6;
Construct ASS and ACS by Definition 7;
for i ← 1 to m do
    Compute MG(R_{A_i}) and MG(R_{A_i^-}) by Eq. (6);
end
for i ← 1 to |GB_S| do
    Construct ASG and ACG by Definition 8;
    Compute OS(gb_i) by Definition 9;
    for x ∈ gb_i do
        score(x) = OS(gb_i);
    end
end
return score.
```

The time complexity of the algorithm is analyzed as follows. First is the granular-ball generation, which has a time complexity of $O(|S|\log|S|)$ and the number of generated granular-balls is $|GB_S|$. Subsequently, the relation matrices between granular-balls are constructed and the importance of attributes is computed with a time complexity of $O(|A||GB_S||GB_S|)$. Attribute sequences and multi-granular-ball fuzzy binary granular structures are constructed with a time complexity $O(|A|)$. Finally, the outlier scores are computed with a time complexity of $O(|S|)$. Thus the total time complexity of the algorithm is $O(|S|\log|S| + |A||GB_S||GB_S|)$.

5. Experiments

This section presents the experiments performed in this work, detailing the datasets used and the methods compared. The experimental results are presented visually, followed by an explanation and discussion.

5.1. Experimental setups

Before outlier detection, to avoid being affected by a large gap between the magnitudes of different data, all data are first min–max normalized so that all data take values between 0 and 1.

A selection of publicly accessible and comprehensive datasets[1,2] has been chosen for our analysis, which is detailed in Table 2. This table shows key characteristics of each dataset, including their original names, abbreviations, number of attributes, number of outliers, outlier ratios, and number of samples. These datasets simulate a diverse array of practical applications allowing for a realistic assessment of the performance of the methods in this work.

We have undertaken a comparative analysis of MGBOD against several outstanding outlier detection methods, as enumerated in Table 3. This table encompasses diverse methods that have garnered significant recognition within the field due to their proven performance and innovative methodologies. This comparative analysis aims to demonstrate the performance of MGBOD across various scenarios, thereby showcasing its potential advantages and disadvantages in practical applications.

As shown in Table 3, we have selected a range of tuning intervals for the hyperparameters associated with each method. This approach is designed to optimize the performance demonstration of the respective methods and to ensure that the experiment is as fair as possible. We have implemented a min–max normalization for the attributes. Subsequently, during the predictive phase, the outlier score for each sample is computed, thereby quantifying the outlier degree associated with that particular sample.

5.2. Experimental results

The receiver operating characteristic curve (ROC) and area under the curve (AUC) are two popular and commonly used metrics for method performance evaluation [11,21,22]. In our work, we utilize these metrics to measure the effectiveness of methods. A method is considered to perform better when its ROC curve is closer to the upper left corner; higher AUC values (ranging from 0 to 1) indicate better performance.

The ROC curves in the experimental results are shown in Fig. 3. In Fig. 3, the ROC curves of MGBOD are closer to the upper left corner on datasets Cardio, Ecoli, Iono, Musk, Sate, Yeast, and Thyroid_d, demonstrating that MGBOD has relatively desirable detection performance. However, the overlapping ROC curves of the different methods under some datasets resulted in not easy comparisons, so the following further compares the performance of the different methods directly through the AUC metric.

Table 4 shows the AUC results for all methods with different datasets, where the best results in each row are highlighted in bold. The **Rank 1** row shows the number of times each method achieved the first rank. The **Difference** row shows the percentage rise or fall of the other methods compared to MGBOD. From the table, we can see that MGBOD achieves the best AUC results on datasets such as Cardio, Ecoli, Iono, Iris, Musk, Sate, Yeast, and Thyroid_d. MGBOD also achieved the best results in terms of average AUC results, which demonstrates

[1] https://github.com/BELLoney/Outlier-detection.
[2] https://odds.cs.stonybrook.edu/.

Table 2
Basic information of the experimental datasets.

No.	Datasets	Abbr.	# Samples	# Attributes	# Outliers	% Outlier ratios
1	Annthyroid	Ann	7200	6	534	7.4%
2	Cardio	Cardio	1831	21	176	9.6%
3	Cardiotocography_2and3_33_variant1	Cardiot	1688	21	33	2.0%
4	Ecoli	Ecoli	336	7	9	2.7%
5	Ionosphere_b_24_variant1	Iono	249	34	24	9.6%
6	Iris_Irisvirginica_11_variant1	Iris	111	4	11	9.9%
7	Mammography	Mamm	11 183	6	260	2.3%
8	Musk	Musk	3062	166	97	3.2%
9	Pageblocks_1_258_variant1	Page	5171	11	258	5.0%
10	Pendigits	Pend	6870	17	156	2.3%
11	Satellite	Sate	6435	37	2036	31.6%
12	Satimage2	Satimage	5803	36	71	1.2%
13	Spambase_spam_56_variant1	Spam	2844	58	56	2.0%
14	Thyroid	Thyroid	3772	6	93	2.5%
15	Vowels	Vowels	1456	12	50	3.4%
16	Waveform_0_100_variant1	Waveform	3443	21	100	2.9%
17	Yeast_ERL_5_variant1	Yeast	1141	8	5	0.4%
18	Abalone_variant1	Aba	4177	9	79	1.9%
19	Sick_sick_35_variant1	Sick	3576	29	35	1.0%
20	Thyroid_disease_variant1	Thyroid_d	9172	28	74	0.8%

Table 3
Descriptions of compared methods, where ✘ indicates that this method has no hyperparameters or step size.

No.	Methods (Years)	Descriptions	Hyperparameter tuning ranges	Step sizes
1	MGBOD (Ours)	Multi-granular-ball fuzzy information granules-based outlier detection	$[0, 1]$	0.05
2	ADkNRS (2025) [39]	Anomaly detection based on improved k-nearest neighbor rough sets	$[1, 60]$	1
3	DIF (2023) [52]	Deep isolation forest for outlier detection	$r = 50; t = 6; \psi = 256$	✘
4	ILGNI (2023) [53]	Outlier detection based on local–global neighborhood information	✘	✘
5	MFGAD (2023) [11]	Multi-fuzzy granules-based anomaly detection	$[0.1, 2]$	0.1
6	WFRDA (2023) [22]	Weighted fuzzy rough density-based anomaly detection	$[0.1, 2]$	0.1
7	ROD (2022) [54]	Outlier detection using rotations	✘	✘
8	ECOD (2022) [55]	Outlier detection using empirical cumulative distribution functions	✘	✘
9	DCROD (2022) [2]	Outlier detection based on directed density ratio changing rate	$[1, 60]$	1
10	VarE (2020) [56]	Outlier detection using structural scores in a high-dimensional space	$\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3\}$	✘
11	DeepSVDD (2018) [57]	Outlier detection by deep support vector data description	✘	✘
12	NC (2018) [58]	Efficient representation-based outlier detection	$[1, 60]$	1
13	ODGrCR (2015) [46]	Outlier detection based on granular computing and rough set theory	✘	✘
14	WDOD (2014) [59]	Weighted density-based outlier detection	✘	✘

the superior performance of MGBOD and further demonstrates the effectiveness of MGBOD.

We further analyze the running times of different methods. Table 5 shows the running times and average running times of 14 methods on 20 datasets, where **MGBOD w/o gen.** denotes the running time after removing the time of granular-ball generation. The running time of MGBOD mainly includes two parts: granular-ball generation and outlier detection. The granular-ball generation in MGBOD can be regarded as a data preprocessing method. MGBOD uses multi-granular-balls as basic processing units to achieve efficient outlier detection. Multi-granular-balls reduce the time cost of MGBOD in computing the similarity measures and subsequent processing time, allowing MGBOD to detect outliers efficiently. A comprehensive analysis of the results in Tables 4 and 5 shows that MGBOD can detect outliers quickly and effectively. Even in large-scale datasets, MGBOD has relatively high detection accuracy and short running time. Moreover, the running times of MGBOD do not increase rapidly with the expansion of data scale, which shows that MGBOD is suitable for outlier detection in large-scale datasets.

Considering that the granular-ball generation also takes a certain amount of time, we also analyze the running time of MGBOD after excluding the granular-ball generation time. From the last column of Table 5, it can be observed that granular-ball generation has a certain impact on the overall runtime, especially on large-scale datasets. Therefore, selecting an efficient granular-ball generation method can help improve the efficiency of MGBOD. In general, MGBOD has satisfactory detection accuracy and short running time. It can be applied to different scales of data and has a wider application scope.

Finally, we compare the running times of MGBOD with and without granular-ball computing. As shown in Table 6, the average running time of MGBOD across 20 datasets is 3.8346. When excluding the granular-ball generation time, the average running time decreases to 2.6351. For comparison, replacing the processing units in MGBOD from granular-balls to the original samples results in an average running time of 16.2607. After introducing granular-ball computing, the running time of the original sample-based MGBOD is significantly reduced from 16.2607 to 3.8346, demonstrating that granular-ball computing, as an efficient multi-granularity computing framework, can enhance both the efficiency and performance of our method.

5.3. Hyperparameter sensitivity analysis

During the application of MGBOD, as shown in Eq. (6), the hyperparameter σ needs to be determined to compute the fuzzy relations, and the choice of the size of σ affects the result of the fuzzy relations, which in turn affects the detection performance of MGBOD. To reveal the impact of different σ on the performance under different datasets, we plot the corresponding variation figures as shown in Fig. 4.

It can be seen from the figure that at first the AUC results on most of the datasets slowly increase as σ increases, and then the AUC results on each dataset decrease to some extent as the value of σ approaches 1.0. From Definition 3, it can be seen that the values of the fuzzy relations obtained from the computation are located in the interval $[0, 1]$. However, in reality, datasets often have complex distributions, resulting in the values of the fuzzy relations being mostly not close to

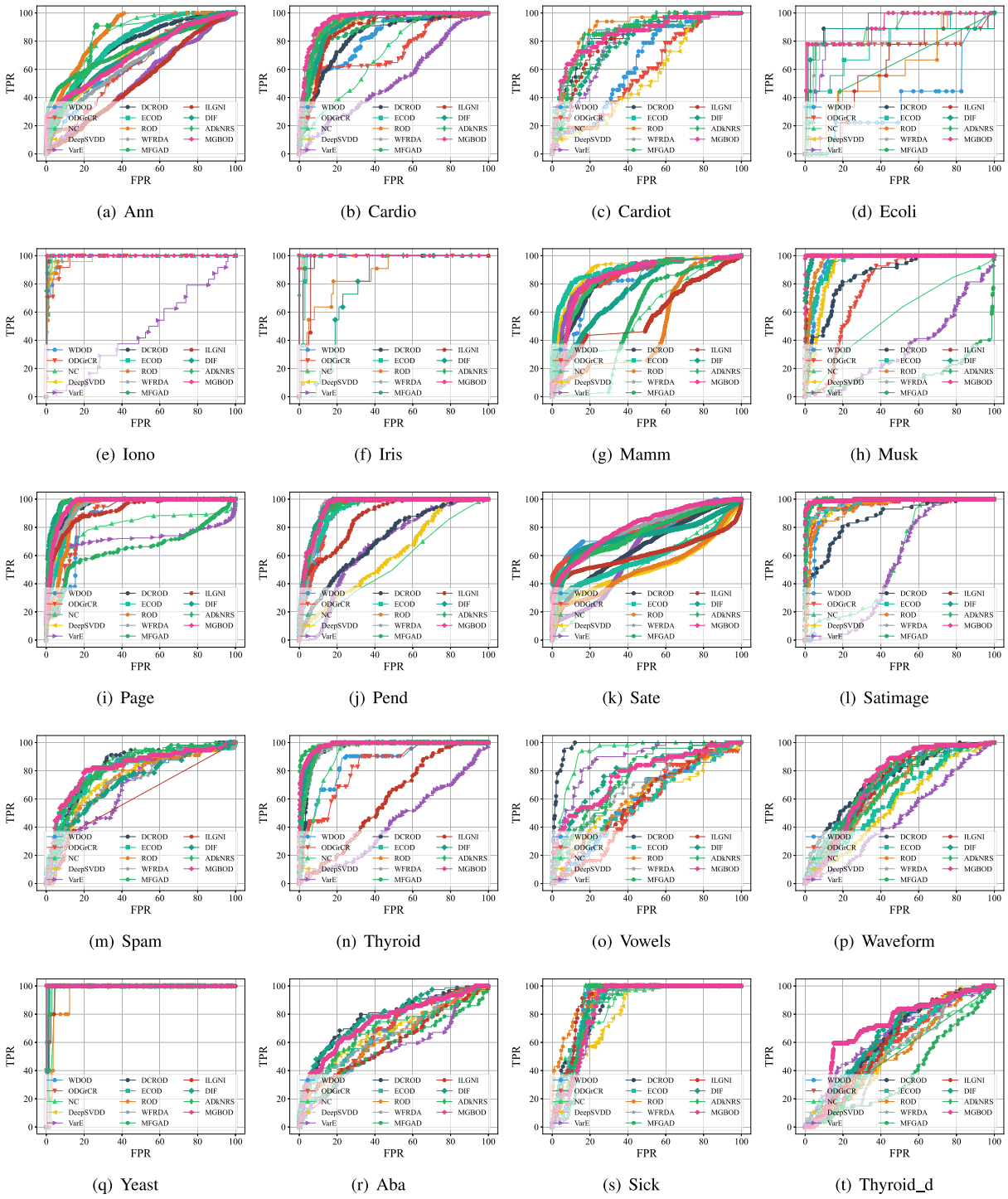


Fig. 3. ROC of different methods on all datasets.

1. Therefore, when the value of σ is close to 1.0, the values of most fuzzy relations are truncated to 0, which makes the differences between granular-balls in the dataset assimilated, and thus leads to a serious degradation of subsequent outlier detection performance. If we exclude the hyperparameter σ close to 1.0, the curve in the remaining figure is relatively smooth. In practical applications, users can directly select hyperparameters in the range of $[0, 0.6]$ to achieve relatively superior results.

5.4. Statistical analysis

Referring to previous studies on outlier detection [11,22,23], in this section, we use the Friedman test and the Nemenyi post-hoc test to evaluate the statistical significance of the experimental results.

As shown in Tables 2 and 3, our experimental analysis encompasses a comparative study of 14 distinct outlier detection methods across 20 datasets. The Friedman test is a non-parametric statistical analysis

Table 4
AUC of different methods on 20 datasets and the best AUC result in each row is highlighted in bold. **Rank 1** denotes the number of times each method achieved the first rank. **Difference** denotes the percentage rise or fall of the other methods compared to MGBOD.

Dataset	WDOD	ODGrCR	NC	DeepSVDD	VarE	DCROD	ECOD	ROD	WFRDA	MFGAD	ILGNI	DIF	ADkNRS	MGBOD
Ann	0.655	0.625	0.826	0.710	0.523	0.757	0.789	**0.860**	0.658	0.746	0.522	0.707	0.809	0.703
Cardio	0.830	0.748	0.703	0.909	0.526	0.834	0.935	0.932	0.922	0.897	0.887	0.918	0.933	**0.956**
Cardiot	0.641	0.581	0.806	0.549	0.780	0.834	0.871	**0.893**	0.865	0.784	0.811	0.796	0.863	0.852
Ecoli	0.380	0.799	0.881	0.879	0.899	0.875	0.781	0.613	0.875	0.849	0.638	0.864	0.602	**0.912**
Iono	0.985	0.976	0.993	0.984	0.456	**1.000**	0.994	0.987	0.993	0.999	0.998	0.999	**1.000**	**1.000**
Iris	**1.000**	**1.000**	0.996	0.999	**1.000**	0.983	0.977	0.866	**1.000**	0.997	0.951	0.779	**1.000**	**1.000**
Mamm	0.794	0.864	0.514	0.905	0.866	0.850	**0.906**	0.497	0.848	0.644	0.587	0.755	0.856	0.851
Musk	0.950	0.767	0.572	0.929	0.346	0.863	0.956	0.973	**1.000**	0.975	0.993	0.994	**1.000**	**1.000**
Page	0.866	0.881	0.803	0.948	0.727	0.953	0.938	0.921	0.944	0.655	0.914	**0.980**	0.976	0.965
Pend	0.921	0.913	0.537	0.603	0.665	0.712	0.927	0.948	0.942	0.941	0.862	**0.952**	0.938	0.949
Sate	0.792	0.787	0.524	0.522	0.643	0.662	0.583	0.513	0.785	0.764	0.614	0.735	0.749	**0.803**
Satimage	0.931	0.961	0.590	0.961	0.537	0.865	0.965	0.933	0.972	0.989	0.994	**0.997**	0.994	0.995
Spam	0.793	0.790	0.767	0.756	0.685	0.825	0.812	0.736	0.725	**0.836**	0.607	0.718	0.819	0.822
Thyroid	0.870	0.829	0.903	0.976	0.432	0.955	0.977	0.977	0.957	**0.989**	0.596	0.975	0.962	0.984
Vowels	0.590	0.552	0.927	0.621	0.878	**0.978**	0.593	0.626	0.686	0.767	0.609	0.786	0.717	0.770
Waveform	0.707	0.726	0.696	0.564	0.473	**0.743**	0.608	0.670	0.704	0.736	0.705	0.714	0.670	0.727
Yeast	0.996	0.999	0.970	0.997	**1.000**	0.990	0.995	0.949	0.998	0.992	0.981	0.987	0.984	**1.000**
Aba	0.649	0.660	0.729	0.677	0.548	0.780	0.653	0.629	0.661	0.565	0.604	**0.787**	0.689	0.748
Sick	0.880	0.870	0.853	0.815	0.842	0.875	0.883	**0.927**	0.868	0.860	0.902	0.879	0.865	0.859
Thyroid_d	0.635	0.604	0.564	0.594	0.650	0.646	0.581	0.533	0.531	0.388	0.610	0.659	0.512	**0.712**
Average	0.793	0.797	0.758	0.795	0.674	0.849	0.836	0.799	0.847	0.819	0.769	0.849	0.847	**0.880**
Rank 1	1	1	0	0	2	3	1	3	2	2	0	4	3	**8**
Difference	−9.9%	−9.5%	−13.9%	−9.7%	−23.5%	−3.6%	−5.0%	−9.2%	−3.8%	−7.0%	−12.6%	−3.5%	−3.8%	**0.0%**

Table 5
Running times of different methods on 20 datasets and the shortest running time in each row is highlighted in bold. **MGBOD w/o gen.** denotes the running time after removing the time for granular-ball generation.

Dataset	WDOD	ODGrCR	NC	DeepSVDD	VarE	DCROD	ECOD	ROD	WFRDA	MFGAD	ILGNI	DIF	ADkNRS	MGBOD	MGBOD w/o gen.
Ann	0.5016	77.7456	2.0040	4.5529	84.3680	0.5700	**0.0122**	4.1904	59.4405	18.4070	1014.2776	7.5593	116.8795	2.5999	0.9594
Cardio	0.1313	713.1461	0.7380	1.2154	3.5310	0.3280	**0.0061**	26.0905	2.5837	8.1752	62.8020	2.4477	22.0852	0.4735	0.4034
Cardiot	0.1137	19.9013	0.6390	1.0960	2.5630	0.2910	**0.0052**	20.4592	1.3461	7.3597	34.6431	2.1548	8.3306	0.3985	0.2736
Ecoli	0.0033	0.3092	0.0460	0.2400	0.0580	0.0290	**0.0014**	0.1613	0.0117	0.0282	0.0936	0.4640	0.0564	0.0237	0.0075
Iono	0.0088	4.3491	0.0350	0.1770	0.0180	0.0190	**0.0023**	26.3122	0.0270	0.5158	1.2366	0.5311	0.3455	0.0529	0.0384
Iris	0.0012	0.0446	0.0049	0.0940	0.0050	0.0130	0.0012	0.0156	**0.0008**	0.0029	0.0047	0.3397	0.0059	0.0049	0.0011
Mamm	1.1830	4399.8970	3.4710	6.9240	464.1551	3.1530	**0.0115**	5.1139	288.2161	59.1718	3630.4927	12.4918	1567.9200	5.4157	0.9641
Musk	2.9688	444 092.0841	5.5800	2.6400	42.7370	2.6892	**0.0769**	108 679.3556	31.4613	1962.3774	21 308.7489	4.0817	1073.1707	25.6158	23.1948
Page	0.4251	94.1743	1.4190	3.2910	143.7940	0.7540	**0.0100**	22.1369	22.6414	17.9540	595.8639	5.4788	414.4373	1.9687	0.8423
Pend	1.2371	415.3694	1.5680	4.3824	232.7470	2.8375	**0.0181**	97.5895	77.0028	115.0574	2373.1965	8.4997	6181.8974	3.6917	2.3174
Sate	2.4965	10 428.7795	4.3640	4.1359	152.1680	2.5851	**0.0313**	577.6198	77.5529	292.2335	5188.6166	7.2646	298.2433	6.8685	4.7027
Satimage	2.0234	1793.8761	3.9090	3.7780	138.8150	1.1780	**0.0258**	450.3896	57.5739	236.9135	3879.3560	7.1074	458.4165	5.3220	3.6285
Spam	0.7966	710.4044	0.9920	1.8950	36.3320	0.3880	**0.0135**	639.0015	8.8995	110.8840	528.0523	3.8105	242.2322	3.0603	1.9136
Thyroid	0.1430	19.2631	0.3390	2.4158	67.0460	1.1560	**0.0045**	1.5278	7.9752	5.3233	144.4198	4.3055	34.5228	0.7388	0.2519
Vowels	0.0528	1087.2858	0.1420	0.9910	1.3130	0.2550	**0.0036**	3.9646	0.6412	2.1712	15.7553	1.9328	53.9271	0.1894	0.0876
Waveform	0.4596	93.7647	0.6550	2.2557	45.1260	0.7310	**0.0128**	87.4332	8.3084	39.8359	429.2674	4.4545	492.4776	0.8942	0.5464
Yeast	0.0251	19.7726	0.0810	0.7800	0.7650	0.1410	**0.0042**	0.5155	0.2814	0.5068	6.3253	1.5477	1.6443	0.1229	0.0366
Aba	0.2278	21.5967	0.3460	2.6780	76.1620	0.5480	**0.0096**	7.2326	11.3265	7.3366	262.8324	4.8145	200.6570	0.9456	0.4528
Sick	0.6132	152.6067	0.8480	2.3200	47.9780	0.5280	**0.0118**	51.2061	13.5145	61.4426	739.0017	4.9505	54.0387	2.3890	1.5308
Thyroid_d	3.7622	965.1594	3.0800	12.7140	497.6691	3.6030	**0.0276**	97.9774	201.6389	359.5128	10 332.1059	10.3765	461.0387	15.9161	10.5483
Average	0.8587	23 255.4765	1.5130	2.9288	101.8675	1.0898	**0.0145**	5539.9147	43.5222	165.2605	2527.3546	4.7307	584.1163	3.8346	2.6351

Table 6
The running times of our method on 20 datasets, where **MGBOD w/o GB** denotes the running time after replacing the original granular-balls in MGBOD with samples as the basic processing unit and **MGBOD w/o Gen.** denotes the running time after removing the time for granular-ball generation.

Method	Ann	Cardio	Cardiot	Ecoli	Iono	Iris	Mamm	Musk	Page	Pend	Sate	Satimage	Spam	Thyroid	Vowels	Waveform	Yeast	Aba	Sick	Thyroid_d	**Average**	**Difference**
MGBOD w/o GB	5.6012	2.1010	1.2889	0.0174	0.0759	0.0032	15.6411	113.3868	4.8810	16.9357	38.2122	38.8497	13.2068	1.2738	0.4445	5.0440	0.1634	2.4443	8.0686	57.5753	16.2607	+517%
MGBOD	2.5999	0.4735	0.3985	0.0237	0.0529	0.0049	5.4157	25.6158	1.9687	3.6917	6.8685	5.3220	3.0603	0.7388	0.1894	0.8942	0.1229	0.9456	2.3890	15.9161	3.8346	+46%
MGBOD w/o Gen.	**0.9594**	**0.4034**	**0.2736**	**0.0075**	**0.0384**	**0.0011**	**0.9641**	**23.1948**	**0.8423**	**2.3174**	**4.7027**	**3.6285**	**1.9136**	**0.2519**	**0.0876**	**0.5464**	**0.0366**	**0.4528**	**1.5308**	**10.5483**	**2.6351**	**0%**

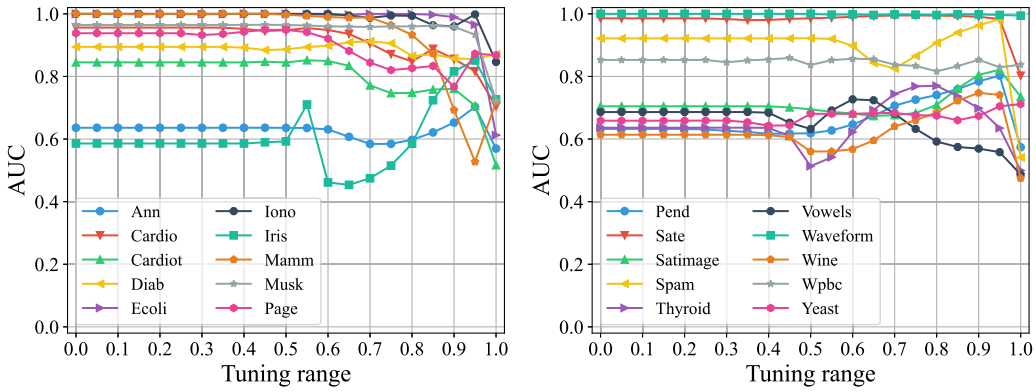
method suitable for multi-related samples. Using the Friedman test, we can obtain an F-distribution with 13 and 247 freedom degrees. At a predetermined significance level of $\alpha = 0.1$, the test statistic τ_F is computed as 4.2275, surpassing the critical value of 1.5512. The results show that there are statistically significant differences between these methods in the performance of the methods. Consequently, we need to further clarify the statistical differences between methods through post-hoc analysis.

As above, we choose $\alpha = 0.1$, which in turn gives us critical distance (CD) $CD_{0.1} = 4.1274$. To provide a visual interpretation of the results of the Nemenyi test, we plot a Nemenyi test figure that shows the average ordinal rankings of the compared methods. Within this figure, each method is represented by a single point that corresponds to its average ordinal value. Adjacent to this point, a centered horizontal line segment, emanating from the point, is utilized to denote the size of the CD. The presence of connecting horizontal line segments between a set of methods indicates a lack of significant differences in their performance, suggesting that these methods perform similarly in the context of the test conducted.

As can be seen in Fig. 5, there are statistically significant differences between MGBOD and most of the methods. For example, in Fig. 5, there is no line coverage or connection between MGBOD and MFGAD, ROD, ODGrCR, DeepSVDD, WDOD, ILGNI, NC, and VarE methods, showing that MGBOD is statistically significantly different from these methods. However, there is line coverage between MGBOD, DIF, ADkNRS, DCROD, WFRDA, and ECOD, i.e., the available evidence does not provide a clear indication of statistical differences among these methods. Of course, this is only a statistical analysis of the variability between the methods and does not indicate that the core ideas of these methods are the same.

5.5. Ablation experiments

In this section, we further validate the effectiveness of MGBOD through ablation experiments, demonstrating the promoting effect of



(a) Datasets 1-10 (b) Datasets 11-20

Fig. 4. The variation curves of AUC on hyperparameter σ with step size 0.05.

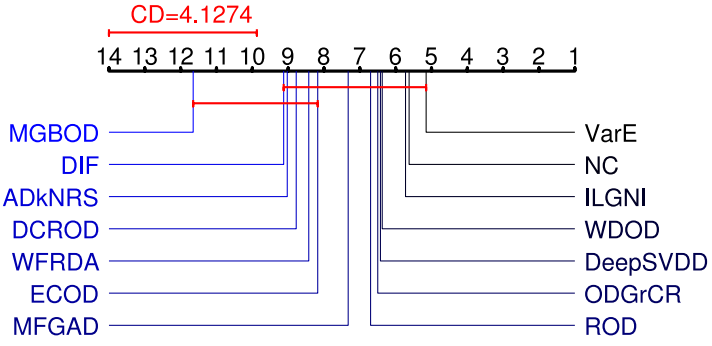


Fig. 5. The figure of Nemenyi test with respect to compared methods on AUC.

Table 7
The AUC results of ablation experiments on 20 datasets, where **MGBOD w/o seq.** denotes MGBOD after removing the two attribute sequences.

Method	Ann	Cardio	Cardiot	Ecoli	Iono	Iris	Mamm	Musk	Page	Pend	Sate	Satimage	Spam	Thyroid	Vowels	Waveform	Yeast	Aba	Sick	Thyroid_d	**Average**	**Difference**
MGBOD	**0.703**	**0.956**	**0.852**	**0.912**	**1.000**	**1.000**	**0.851**	**1.000**	**0.965**	**0.949**	**0.803**	**0.995**	**0.822**	**0.984**	**0.770**	**0.727**	**1.000**	**0.748**	0.859	**0.712**	**0.880**	**0.00%**
MGBOD w/o seq.	0.669	0.944	0.810	0.873	0.997	**1.000**	0.816	**1.000**	0.954	0.942	0.798	0.966	0.763	0.977	0.594	0.688	0.998	0.633	**0.869**	0.664	0.848	−3.70%

the two attribute sequences designed in MGBOD on outlier detection performance. In the ablation experiments, instead of using two attribute sequences, we only utilize each single attribute to compute outlier scores. The experiments are carried out in the same experimental environment and with the same parameter configuration. The experimental results are shown in Table 7, where **MGBOD w/o seq.** denotes the AUC results of MGBOD after removing the two attribute sequences. As can be seen from the table, after removing the two attribute sequences in MGBOD, the AUC results of MGBOD decreased by 3.70% compared with the original results, which demonstrates that the two attribute sequences have a promoting effect on the performance of MGBOD.

Although our method has achieved satisfactory performance in the experiments, MGBOD still has some limitations. First, the performance of MGBOD is partly dependent on the quality of the generated granular-balls. The limitations of the granular-ball generation method are inherited by MGBOD, such as its inability to effectively handle nominal attribute data. Second, MGBOD requires hyperparameter tuning. Although it is relatively insensitive to hyperparameter selection, the choice of hyperparameters can still have some impact on its performance. Moreover, setting hyperparameters may not be user-friendly for non-experts. To address these limitations, future work could focus on optimizing the granular-ball generation method and developing adaptive outlier detection techniques to further enhance detection performance.

6. Conclusions

How to quickly and effectively uncover potential outliers in data is a challenging and worthwhile research topic. In this work, we provide a novel feasible solution for outlier detection by utilizing granular-balls and fuzzy information granulation. Our method, based on multi-granular-ball fuzzy binary granular structures constructed on two attribute sequences, achieves effective detection of outliers in data. Our method not only extends the theory of GBC and information granulation but also extends the application scenarios of GBC. However, existing granular-ball generation methods are still unable to handle nominal attribute data efficiently, and the presence of hyperparameter in our methods leads to user-unfriendliness. In future work, we consider improving the granular-ball generation process and hyperparameter to propose more efficient and hyperparameter-free outlier detection methods.

CRediT authorship contribution statement

Xinyu Su: Writing – review & editing, Writing – original draft, Visualization, Validation. **Shitong Cheng:** Visualization, Validation, Resources. **Dezhong Peng:** Funding acquisition, Data curation, Conceptualization. **Hongmei Chen:** Resources, Project administration, Investigation. **Zhong Yuan:** Writing – review & editing, Writing – original draft, Validation, Supervision, Resources, Methodology.

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors thank both the editors and reviewers for their valuable suggestions, which substantially improve this paper. This work was supported by the National Natural Science Foundation of China (62306196, 62372315, and 62376230), the Sichuan Science and Technology Program (2024NSFSC0443, 2024YFHZ0089, 2024NSFTD0049, and 2024YFHZ0144), the Chengdu Science and Technology Project (Grant no. 2023-XT00-00004-GX), and the Fundamental Research Funds for the Central Universities, China (YJ202245).

Data availability

The datasets can be found at https://github.com/BELLoney/Outlier-detection.

References

[1] A. Smiti, A critical overview of outlier detection methods, Comput. Sci. Rev. 38 (2020) 100306.
[2] K. Li, X. Gao, S. Fu, X. Diao, P. Ye, B. Xue, J. Yu, Z. Huang, Robust outlier detection based on the changing rate of directed density ratio, Expert Syst. Appl. 207 (2022) 117988.
[3] G. Giacinto, R. Perdisci, M. Del Rio, F. Roli, Intrusion detection in computer networks by a modular ensemble of one-class classifiers, Inf. Fusion 9 (1) (2008) 69–82.
[4] K. Huang, H. Wen, C. Yang, W. Gui, S. Hu, Outlier detection for process monitoring in industrial cyber-physical systems, IEEE Trans. Autom. Sci. Eng. 19 (3) (2021) 2487–2498.
[5] H.K. Khanuja, D. Adane, To monitor and detect suspicious transactions in a financial transaction system through database forensic audit and rule-based outlier detection model, in: Organizational Auditing and Assurance in the Digital Age, IGI Global, 2019, pp. 224–255.
[6] L. Li, W. Ding, L. Huang, X. Zhuang, V. Grau, Multi-modality cardiac image computing: A survey, Med. Image Anal. (2023) 102869.
[7] H. Xin, Z. Hao, Z. Sun, R. Wang, Z. Miao, F. Nie, Multi-view and multi-order graph clustering via constrained l1, 2-norm, Inf. Fusion (2024) 102483.
[8] W. Wei, J.Y. Liang, Information fusion in rough set theory: An overview, Inf. Fusion 48 (2019) 107–118.
[9] P. Zhang, T. Li, G. Wang, C. Luo, H. Chen, J. Zhang, D. Wang, Z. Yu, Multi-source information fusion based on rough set theory: A review, Inf. Fusion 68 (2021) 85–117.
[10] J.L. Garrido-Labrador, A. Serrano-Mamolar, J. Maudes-Raedo, J.J. Rodríguez, C. García-Osorio, Ensemble methods and semi-supervised learning for information fusion: A review and future research directions, Inf. Fusion (2024) 102310.
[11] Z. Yuan, H. Chen, C. Luo, D. Peng, MFGAD: Multi-fuzzy granules anomaly detection, Inf. Fusion 95 (2023) 17–25.
[12] J. Xie, L. Jiang, S. Xia, X. Xiang, G. Wang, An adaptive density clustering approach with multi-granularity fusion, Inf. Fusion (2024) 102273.
[13] M.S. Reis, Multiscale and multi-granularity process analytics: A review, Processes 7 (2) (2019) 61.
[14] X. Yang, Y. Li, T. Li, A review of sequential three-way decision and multi-granularity learning, Internat. J. Approx. Reason. 152 (2023) 414–433.
[15] J.T. Yao, A.V. Vasilakos, W. Pedrycz, Granular computing: perspectives and challenges, IEEE Trans. Cybern. 43 (6) (2013) 1977–1989.
[16] Y. Qian, J. Liang, Z.W. Wei-zhi, C. Dang, Information granularity in fuzzy binary GrC model, IEEE Trans. Fuzzy Syst. 19 (2) (2010) 253–264.
[17] C. Mencar, A.M. Fanelli, Interpretability constraints for fuzzy information granulation, Inform. Sci. 178 (24) (2008) 4585–4618.
[18] Y. Qian, Y. Li, J. Liang, G. Lin, C. Dang, Fuzzy granular structure distance, IEEE Trans. Fuzzy Syst. 23 (6) (2015) 2245–2259.
[19] G. Wang, J. Yang, J. Xu, Granular computing: from granularity optimization to multi-granularity joint problem solving, Granul. Comput. 2 (2017) 105–120.
[20] L.A. Zadeh, Fuzzy sets and information granularity, Fuzzy Sets Fuzzy Log. Fuzzy Syst.: Sel. Pap. (1979) 433–448.
[21] Z. Yuan, P. Hu, H. Chen, Y. Chen, Q. Li, DFNO: Detecting fuzzy neighborhood outliers, IEEE Trans. Knowl. Data Eng. 37 (1) (2025) 200–209.
[22] Z. Yuan, B.Y. Chen, J. Liu, H.M. Chen, D.Z. Peng, P.L. Li, Anomaly detection based on weighted fuzzy-rough density, Appl. Soft Comput. 134 (2023) 109995.
[23] Z. Yuan, H. Chen, T. Li, B. Sang, S. Wang, Outlier detection based on fuzzy rough granules in mixed attribute data, IEEE Trans. Cybern. 52 (8) (2021) 8399–8412.
[24] J. Ye, J. Zhan, B. Sun, A three-way decision method based on fuzzy rough set models under incomplete environments, Inform. Sci. 577 (2021) 22–48.
[25] Z. Yuan, H. Chen, P. Xie, P. Zhang, J. Liu, T. Li, Attribute reduction methods in fuzzy rough set theory: An overview, comparative experiments, and new directions, Appl. Soft Comput. 107 (2021) 107353.
[26] Z. Wang, H. Chen, X. Yang, J. Wan, T. Li, C. Luo, Fuzzy rough dimensionality reduction: a feature set partition-based approach, Inform. Sci. 644 (2023) 119266.
[27] S. Xia, Y. Liu, X. Ding, G. Wang, H. Yu, Y. Luo, Granular ball computing classifiers for efficient, scalable and robust learning, Inform. Sci. 483 (2019) 136–152.
[28] S. Xia, X. Dai, G. Wang, X. Gao, E. Giem, An efficient and adaptive granular-ball generation method in classification problem, IEEE Trans. Neural Netw. Learn. Syst. 35 (4) (2024) 5319–5331.
[29] Q. Xie, Q. Zhang, S. Xia, F. Zhao, C. Wu, G. Wang, W. Ding, GBG++: A fast and stable granular ball generation method for classification, IEEE Trans. Emerg. Top. Comput. Intell. 8 (2) (2024) 2022–2036.
[30] Y. Xue, Y. Shao, S. Xia, G. Wang, The dual model of support vector machine based on granular ball computing, in: 2021 3rd International Conference on Applied Machine Learning, ICAML, IEEE, 2021, pp. 43–47.
[31] S. Xia, S. Zheng, G. Wang, X. Gao, B. Wang, Granular ball sampling for noisy label classification or imbalanced classification, IEEE Trans. Neural Netw. Learn. Syst. 34 (4) (2021) 2144–2155.
[32] X. Peng, P. Wang, S. Xia, C. Wang, W. Chen, VPGB: A granular-ball based model for attribute reduction and classification with label noise, Inform. Sci. 611 (2022) 504–521.
[33] J. Xie, W. Kong, S. Xia, G. Wang, X. Gao, An efficient spectral clustering algorithm based on granular-ball, IEEE Trans. Knowl. Data Eng. 35 (9) (2023) 9743–9753.
[34] D. Cheng, Y. Li, S. Xia, G. Wang, J. Huang, S. Zhang, A fast granular-ball-based density peaks clustering algorithm for large-scale data, IEEE Trans. Neural Netw. Learn. Syst. (2023) 1–14.
[35] Z. Yuan, X.Y. Zhang, S. Feng, Hybrid data-driven outlier detection based on neighborhood information entropy and its developmental measures, Expert Syst. Appl. 112 (2018) 243–257.
[36] Y. Wang, Y.P. Li, Outlier detection based on weighted neighbourhood information network for mixed-valued datasets, Inform. Sci. 564 (2021) 396–415.
[37] C. Liu, Z. Yuan, B. Chen, H. Chen, D. Peng, Fuzzy granular anomaly detection using Markov random walk, Inform. Sci. 646 (2023) 119400.
[38] B. Chen, Z. Yuan, D. Peng, X. Chen, H. Chen, Consistency-guided semi-supervised outlier detection in heterogeneous data using fuzzy rough sets, Appl. Soft Comput. 165 (2024) 112070.
[39] X. Chen, Z. Yuan, S. Feng, Anomaly detection based on improved k-nearest neighbor rough sets, Internat. J. Approx. Reason. 176 (2025) 109323.
[40] Y. Xue, Y. Shao, H. Xia, GBFSVM: A robust classification learning method, Sci. J. Intell. Syst. Res. Vol. 4 (1) (2022) 1–8.
[41] Q. Zhang, C. Wu, S. Xia, F. Zhao, M. Gao, Y. Cheng, G. Wang, Incremental learning based on granular ball rough sets for classification in dynamic mixed-type decision system, IEEE Trans. Knowl. Data Eng. 35 (9) (2023) 9319–9332.
[42] S. Xia, H. Zhang, W. Li, G. Wang, E. Giem, Z. Chen, GBNRS: A novel rough set algorithm for fast adaptive attribute reduction in classification, IEEE Trans. Knowl. Data Eng. 34 (3) (2020) 1231–1242.
[43] D. Cheng, C. Zhang, Y. Li, S. Xia, G. Wang, J. Huang, S. Zhang, J. Xie, GB-DBSCAN: A fast granular-ball based DBSCAN clustering algorithm, Inform. Sci. (2024) 120731.
[44] F. Jiang, Y. Sui, C. Cao, Outlier detection using rough set theory, in: International Workshop on Rough Sets, Fuzzy Sets, Data Mining, and Granular-Soft Computing, Springer, 2005, pp. 79–87.
[45] F. Jiang, Y.F. Sui, C.G. Cao, A rough set approach to outlier detection, Int. J. Gen. Syst. 37 (5) (2008) 519–536.
[46] F. Jiang, Y.-M. Chen, Outlier detection based on granular computing and rough set theory, Appl. Intell. 42 (2015) 303–322.
[47] F. Jiang, H. Zhao, J. Du, Y. Xue, Y. Peng, Outlier detection based on approximation accuracy entropy, Int. J. Mach. Learn. Cybern. 10 (2019) 2483–2499.
[48] M. Singh, R. Pamula, An outlier detection approach in large-scale data stream using rough set, Neural Comput. Appl. 32 (13) (2020) 9113–9127.
[49] Z. Yuan, H.M. Chen, T.R. Li, X.Y. Zhang, B.B. Sang, Multigranulation relative entropy-based mixed attribute outlier detection in neighborhood systems, IEEE Trans. Syst. Man Cybern.: Syst. 52 (8) (2022) 5175–5187.
[50] S. Xia, J. Xie, G. Wang, GBC: An efficient and adaptive clustering algorithm based on granular-ball, 2022, arXiv preprint arXiv:2205.14592.
[51] Q.H. Hu, D.R. Yu, Z.X. Xie, J.F. Liu, Fuzzy probabilistic approximation spaces and their information measures, IEEE Trans. Fuzzy Syst. 14 (2) (2006) 191–201.
[52] H. Xu, G. Pang, Y. Wang, Y. Wang, Deep isolation forest for anomaly detection, IEEE Trans. Knowl. Data Eng. 35 (12) (2023) 12591–12604.
[53] R. Li, H. Chen, S. Liu, X. Li, Y. Li, B. Wang, Incomplete mixed data-driven outlier detection based on local–global neighborhood information, Inform. Sci. 633 (2023) 204–225.

[54] Y. Almardeny, N. Boujnah, F. Cleary, A novel outlier detection method for multivariate data, IEEE Trans. Knowl. Data Eng. 34 (9) (2020) 4052–4062.
[55] Z. Li, Y. Zhao, X. Hu, N. Botta, C. Ionescu, G.H. Chen, Ecod: Unsupervised outlier detection using empirical cumulative distribution functions, IEEE Trans. Knowl. Data Eng. 35 (12) (2022) 12181–12193.
[56] X. Li, J. Lv, Z. Yi, Outlier detection using structural scores in a high-dimensional space, IEEE Trans. Cybern. 50 (5) (2018) 2302–2310.
[57] L. Ruff, R. Vandermeulen, N. Goernitz, L. Deecke, S.A. Siddiqui, A. Binder, E. Müller, M. Kloft, Deep one-class classification, in: International Conference on Machine Learning, PMLR, 2018, pp. 4393–4402.
[58] X. Li, J. Lv, Z. Yi, An efficient representation-based method for boundary point and outlier detection, IEEE Trans. Neural Netw. Learn. Syst. 29 (1) (2018) 51–62.
[59] X. Zhao, J. Liang, F. Cao, A simple and effective outlier detection algorithm for categorical data, Int. J. Mach. Learn. Cybern. 5 (2014) 469–477.