Highlights

A Kernelized Fuzzy Approximation Fusion Model with Granular-ball Computing for Outlier Detection

Yongxiang Li, Xinyu Su, Zhong Yuan, Run Ye, Dezhong Peng, Hongmei Chen

- We pioneer the integration of MGBC into KFRS and construct a novel FRS model named MKFRS.
- We propose a novel fuzzy information fusion model using multi-granularity kernelized fuzzy approximation.
- We apply the novel fusion model to construct an unsupervised outlier detection method named KFGOD.
- Extensive experiments validate the effectiveness and robustness on 20 benchmark datasets.

A Kernelized Fuzzy Approximation Fusion Model with Granular-ball Computing for Outlier Detection

Yongxiang Li[a], Xinyu Su[a], Zhong Yuan[a,*], Run Ye[b], Dezhong Peng[a], Hongmei Chen[c]

[a]*College of Computer Science, Sichuan University, Chengdu 610065, China*
[b]*School of Automation Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China*
[c]*School of Computing and Artificial Intelligence, Southwest Jiaotong University, Chengdu 611756, China*

Abstract

Outlier detection is a fundamental task in data analytics, where fuzzy rough set-based methods have gained increasing attention for their ability to effectively model uncertainty associated with outliers in data. However, existing FRS-based methods often exhibit limitations when applied to complex scenarios. Most of these methods rely on single-granularity fusion, where all samples are processed at a uniform, fine-grained level. This restricts their ability to fuse multi-granularity information, limiting outlier discrimination and making them more susceptible to noise. Moreover, many traditional methods construct fuzzy relation matrices under linear assumptions, which fail to effectively represent the intricate, nonlinear relations commonly found in real-world data. This leads to suboptimal estimation of membership degrees and degrades the reliability of outlier detection. To address these challenges, we propose a Kernelized Fuzzy approximation fusion model with Granular-ball computing for Outlier Detection (KFGOD), which integrates multi-granularity granular-balls and kernelized fuzzy rough sets into a unified framework. KFGOD fuses multi-granularity information to capture abnormal information at different granularity levels. Simultaneously, kernel functions are employed to effectively model multi-granularity nonlinear relations, enhancing the expressive power of fuzzy relation construction. By performing information fusion across multiple kernelized fuzzy information granules associated with each granular-ball, KFGOD evaluates the outlier degrees of each ball and propagates this fused abnormality information to the corresponding samples. This hierarchical and kernelized method allows for effective outlier detection in unlabeled datasets. Extensive experiments conducted on twenty benchmark datasets confirm the effectiveness of KFGOD, which consistently outperforms several state-of-the-art baselines in terms of detection accuracy and robustness. The codes are publicly available online at https://github.com/LYXRhythm/KFGOD.

Keywords: Information fusion, Fuzzy rough sets, Multi-granularity, Granular-ball computing, Outlier detection

1. Introduction

Outlier detection involves identifying objects in data that significantly deviate from normal patterns or differ from the majority of instances. Its core objective is to recognize those distinctive instances, which may correspond to potential risk events or rare but valuable information [1–3]. Given the scarcity and uncertainty of outliers, labeling abnormal instances in data poses a challenge. As a result, unsupervised outlier detection, with its advantage of not requiring labels, has gained broader research and application.

A key challenge in unsupervised outlier detection is effectively integrating diverse data characteristics, such as multiple granularities, scales, or sources of information, to distinguish outliers from normal patterns without the aid of labeled data [4, 5]. Information fusion, which involves combining multiple data perspectives or representations into a unified representation, has long been a prominent research focus, encompassing areas such as multi-source fusion [6], multi-scale fusion [5], and multi-granularity fusion [4]. In this study, we adopt information fusion as a cornerstone of our method, integrating multi-granularity kernelized fuzzy information to enhance the detection of outliers in complex, unlabeled datasets.

Unsupervised outlier detection methods can be categorized according to their underlying assumptions about the nature of outliers. These include statistical-based methods [3], distance-based methods [3], density-based methods [7],

*Corresponding author
Email address: yuanzhong@scu.edu.cn (Zhong Yuan)

clustering-based methods [8], and rough set-based methods [2, 4, 9]. Statistical-based methods detect outliers by analyzing deviations from expected statistical patterns. However, they typically assume specific data distributions (e.g., Gaussian), which limits their applicability to real-world data with complex or unknown distributions. Distance-based methods identify outliers based on their distance to neighboring samples, with larger distances indicating a higher likelihood of being an outlier. Nevertheless, in high-dimensional or complex data, distances tend to become uniformly distributed, which significantly reduces their discriminative power. Density-based methods evaluate the local density of data points, identifying those in sparse regions as potential outliers. Despite their local focus, these methods still rely on distance calculations and therefore inherit the limitations of distance-based approaches. Clustering-based methods aim to group similar samples into clusters and label those that belong to small or isolated clusters as outliers. However, these methods often struggle with irregular or overlapping cluster structures, limiting their effectiveness in handling complex data distributions. Rough set-based methods assess outlier degrees using uncertainty measures derived from lower and upper approximations. While they are well-suited for categorical or symbolic data, most existing approaches are not designed to effectively handle numerical attributes, which constrains their practical applicability.

Rough sets, a cornerstone of Granular Computing (GrC), are widely used to manage data uncertainty and form the foundation for numerous extended models. Methods based on rough sets address the limitation of distance-based methods in handling nominal data effectively. However, they are limited to nominal data, requiring pre-discretization for numerical data, which results in information loss. To address these shortcomings, an extended rough set model called Fuzzy Rough Set (FRS) is proposed [10]. FRS directly computes fuzzy relations between samples using numerical attributes, bypassing discretization and addressing the limitations of conventional rough sets [11–13]. The ability of FRS to handle mixed data has garnered widespread attention and research, and they have been applied to attribute reduction [14, 15], classification [16], clustering [17], and outlier detection [9]. For example, Yuan et al. [14] propose a generalized unsupervised mixed attribute reduction model based on FRS. Wang et al. [18] introduce a directed FRS model, which, compared to traditional models, better captures the inherent uncertainty in sample distributions. Yuan et al. [19] propose an unsupervised outlier detection method based on fuzzy rough density. Wang et al. [20] develop an unsupervised outlier detection method based on fuzzy rough entropy. Chen et al. [21] utilize FRS to propose a consistency-based semi-supervised outlier detection method. Despite these advances, FRS-based unsupervised outlier detection methods still face significant limitations in complex data scenarios. First, most methods rely on linear assumptions to build fuzzy relation matrices, using metrics like Euclidean or Mahalanobis distance to assess sample similarity. This simplistic linear method struggles with nonlinear data, failing to accurately model complex sample relations via basic distance metrics. Moreover, these methods use single-granularity samples as the basic unit and focus solely on single-granularity fuzzy relations, ignoring multi-granularity information in the data. This not only results in lower efficiency but also makes the methods highly susceptible to noise interference.

The kernel function constitutes a fundamental paradigm in machine learning, operating on the principle of implicitly mapping nonlinearly separable samples from the original input space into a high-dimensional feature space through a nonlinear transformation, thereby rendering them linearly separable in the transformed space [22]. A notable advantage of kernel methods lies in the ability to compute inner products in the high-dimensional space indirectly via operations in the original input space, obviating the need for explicit mapping and significantly reducing computational complexity. This framework enables kernel-based models to effectively address the challenge of nonlinearity in the input space while maintaining computational tractability. In the context of integrating kernel methods with FRS, Hu et al. [23] are the first to introduce a Gaussian kernel function to construct fuzzy equivalence relations, satisfying reflexivity, symmetry, and $T_{\cos}$-transitivity. They further defined the kernelized fuzzy upper and lower approximations and proposed a supervised feature selection method based on Kernelized Fuzzy Rough Sets (KFRS). Subsequently, Wu et al. [24] developed an outlier detection approach grounded in KFRS, wherein the fuzzy approximation accuracy is employed to quantify the outlier degree of each sample. By incorporating kernel functions into the FRS framework, KFRS facilitates the effective modeling of intricate nonlinear relationships among samples, thereby extending the applicability of fuzzy rough models to more complex data domains.

Multi-granularity Granular-Ball Computing (MGBC) is a recently proposed paradigm for multi-granularity representation and computation that has been widely applied in machine learning and deep learning due to its efficiency and robustness [25, 26]. The concept of MGBC originates from the coarse-to-fine cognitive mechanism of the human brain, emphasizing the importance of multi-granularity perception. MGBC employs granular-balls with varying granularity sizes to adaptively cover and represent the original sample space. By performing simple granular-ball generation as a preprocessing step, MGBC equips downstream tasks with the ability to mine multi-granularity information from the data. The advantages of MGBC have attracted significant attention and research. For example, Yang et al. [27] propose a three-way classifier based on granular-ball neighborhood rough sets. Xia et al. [28] introduce an efficient and adaptive clustering method based on granular-balls. Qian et al. [29] propose a partial label feature selection method based on granular-balls. Cheng et al. [30] develop an outlier detection method based on

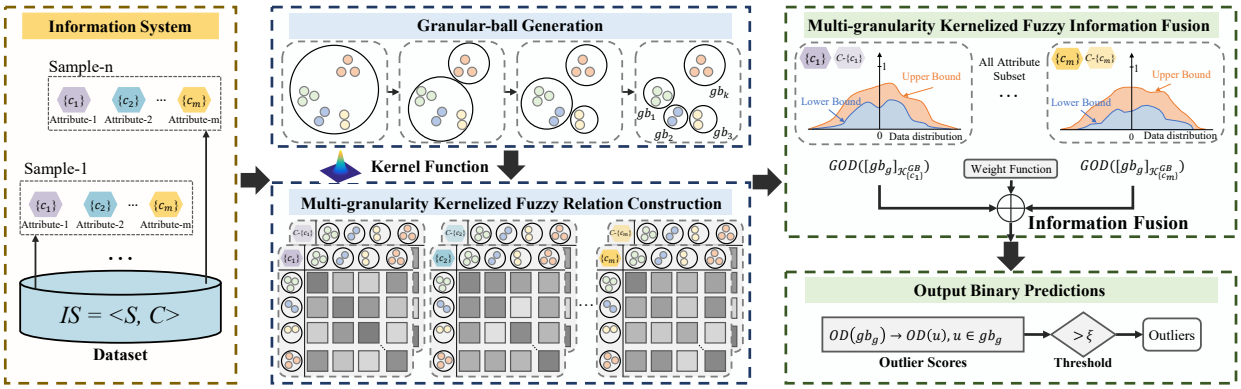


Figure 1: The overall framework of the proposed KFGOD.

granular-ball mean-shift. Su et al. [9] propose an outlier detection method based on granular-ball FRS. Su et al. [31] introduce an MGBC-guided outlier detection method for mixed attribute data. However, these MGBC-based outlier detection methods often fail to adequately account for the nonlinear characteristics of complex data, which may limit their performance in nonlinear scenarios. In [9, 31], when constructing the fuzzy relation matrix, only linear relations in the data are considered, while potential nonlinear dependencies between samples are overlooked, resulting in less accurate modeling of fuzzy relations.

To tackle the shortcomings of existing FRS-based outlier detection methods, which fail to adequately consider the nonlinear and multi-granularity characteristics of data, we integrate MGBC and KFRS to develop the Multi-granularity Kernelized Fuzzy Rough Sets (MKFRS) model for unlabeled data. This model defines multi-granularity kernelized fuzzy upper and lower approximations along with approximation accuracy. Based on MKFRS, we introduce a Kernelized Fuzzy approximation fusion model with Granular-ball computing for Outlier Detection (KFGOD). The KFGOD framework is depicted in Figure 1. In KFGOD, we begin by importing raw data into the information system, followed by granular-ball generation in the original sample space, where multi-granularity granular-balls serve as the basic processing units. Next, we integrate MGBC with KFRS to establish multi-granularity kernelized fuzzy relations, fuzzy granular structures, and corresponding information granules. Based on the KFRS model, we first define the outlier degree of multi-granularity kernelized fuzzy information granule through the multi-granularity kernelized fuzzy approximation accuracy, which reflects the outlier degree of a specific fuzzy information granule. We then fuse abnormal information of multiple multi-granularity kernelized fuzzy information granules linked to each granular-ball to determine its outlier degree. Finally, we map these outlier degrees to the samples within each granular-ball to derive outlier degrees for all samples.

The main novelties and contributions of this paper are summarized as follows:

- We pioneer the integration of MGBC into KFRS and construct MKFRS, a novel FRS model that leverages granular-balls and kernel functions to capture multi-granularity nonlinear fuzzy information in unlabeled data.
- Based on MKFRS, we propose a novel fuzzy information fusion model using multi-granularity kernelized fuzzy approximation.
- We apply the novel fusion model to outlier detection and develop KFGOD, an unsupervised outlier detection method that assesses sample abnormality by integrating fuzzy approximation accuracy information across multiple multi-granularity kernelized fuzzy information granules.
- Extensive experiments on twenty public datasets demonstrate that the proposed method outperforms several state-of-the-art outlier detection methods.

The remaining section of this study are organized as follows. Section 2 reviews recent developments in extended rough sets-based outlier detection methods and MGBC. Section 3 reviews some basic knowledge regarding KFRS. Section 4 proposes a new FRS model called MKFRS. Section 5 introduces a new unsupervised outlier detection method based on MKFRS, named KFGOD. Section 6 validates the effectiveness of KFGOD through extensive experiments. Section 7 summarizes this study.

2. Related works

2.1. Extended rough sets based outlier detection

The limitations of traditional rough sets in handling numerical attribute data have hindered their widespread application. To address this, a series of GrC models extending rough sets have been proposed to improve their performance in processing mixed attribute data. Examples include Neighborhood Rough Sets (NRS) [32], Fuzzy Rough Sets (FRS) [10], Fuzzy Neighborhood Rough Sets (FNRS) [33], and Kernelized Fuzzy Rough Sets (KFRS) [23]. Meanwhile, these extended models have also been widely applied to outlier detection. For example, Yuan et al. [34] propose a multi-granulation relative entropy hybrid attribute outlier detection method based on NRS, utilizing outlier factors derived from multi-granulation relative entropy to measure the outlier degrees of samples. An outlier detection method based on a three-way neighborhood structure is proposed in [35], using multi-neighborhood outlier factors to reflect the outlier degrees of samples. A generalized outlier detection model based on fuzzy rough granules is introduced in [36], employing fuzzy approximation accuracy to measure the outlier degrees of samples. Chen et al. [21] propose to guide outlier detection through consistency based on FRS. Yuan et al. [19] develop an outlier detection method based on fuzzy rough density, using fuzzy information entropy to measure the importance of attributes.

Despite the notable progress these methods have achieved in the field of outlier detection, most of them rely on simple linear methods to model the fuzzy relations or neighborhood relations between samples. This may result in an inability to effectively characterize the complex relations among samples in high-dimensional data. To this end, a novel outlier detection method based on KFRS is proposed in [24], which leverages fuzzy approximation accuracy to assess the outlier degrees of samples. Although the application of KFRS in the field of outlier detection is not the first, existing methods do not account for multi-granularity information in the data, relying solely on single, fine-granularity samples as the basic processing unit. Consequently, further exploration of KFRS in the field of outlier detection is still needed. Furthermore, these methods mainly use single-granularity samples as the basic processing unit, failing to effectively leverage the potential multi-granularity information inherent in the data.

2.2. Multi-granularity granular-ball computing

Multi-granularity Granular-Ball Computing (MGBC) is an important multi-granularity computing and representation framework in GrC. It not only avoids the need to select granularity during the granulation process but also enables adaptive multi-granularity representation. In existing studies on MGBC, a granular-ball is typically defined as $\mathcal{B} = \{u_i \mid i = 1, 2, \ldots, t\}$, where u_i represents a sample in $\mathcal{B}$, and t denotes the number of samples contained in $\mathcal{B}$. Each granular-ball possesses two key attributes, its center and radius. The center of a granular-ball is usually defined as the average of all samples within it, calculated as $z = \frac{1}{t}\sum_{i=1}^{t} u_i$. The radius of a granular-ball can be defined in two distinct ways, based on the average distance or the maximum distance. The first radius is the average distance from all samples within the granular-ball to its center, computed as $r = \frac{1}{t}\sum_{i=1}^{t} \|z - u_i\|_2$. The second radius is the maximum distance from any sample within the granular-ball to its center, calculated as $r = \max_i^t(\|z - u_i\|_2)$. Given a granular-ball $\mathcal{B}_g$, its center and radius are denoted as z_g and r_g, respectively.

The adaptive multi-granularity computing and representation capabilities provided by MGBC endow many traditional methods with the ability to handle multi-granularity information in data. Consequently, MGBC has gained widespread application. For example, Xia et al. [25] propose an efficient classification method based on MGBC, achieving faster speeds while maintaining classification performance. A spectral clustering method based on MGBC is proposed to enhance the efficiency of traditional spectral clustering methods [37]. Xia et al. [38] introduce granular-ball rough sets to address the limitations of classical rough sets in handling numerical attribute data, applying it to attribute reduction with promising results. Cheng et al. [39] develop a density peak clustering method based on MGBC to improve the efficiency of the density peak clustering method. A granular-ball fuzzy support machine model combining MGBC and fuzzy support machines is proposed to enhance the efficiency of the original model [40].

Additionally, MGBC has also been applied in outlier detection. For instance, Cheng et al. [30] propose an outlier detection method based on granular-ball mean-shift. However, this method relies solely on density to identify outliers, neglecting uncertainty information in the data. Su et al. [9] present an outlier detection method based on granular-ball FRS. An outlier detection method based on MGBC, targeting mixed attribute data, is proposed in [31]. Another outlier detection method based on multi-granular-ball fuzzy information is presented in [4]. However, these methods similarly employ a simple fuzzy relation matrix construction method, making them unable to effectively capture the complex relations among samples in complex data. Gao et al. [1] propose a FRS-based multi-scale outlier detection method to identify various types of outliers. Although this method considers multi-granularity and uncertainty information in the data, it employs a simple Manhattan distance to calculate the similarity between granular-balls, failing to capture the complex relationships among samples. Jia et al. [41] detect outliers using the radius of granular-balls and the mean distance from internal samples to the center. While this method improves the quality of granular-ball

generation, it does not account for multi-granularity and uncertainty information in the data, relying solely on granular-ball properties for outlier detection, which limits its outlier detection performance. In contrast to these MGBC-based methods discussed above, we integrate classical FRS with kernel functions to effectively characterize the complex relationships among samples. Meanwhile, to leverage information from different granularity levels in the data, we further incorporate MGBC and propose a new FRS model called Multi-granularity Kernelized Fuzzy Rough Sets (MKFRS). Building on the MKFRS model, we introduce a novel and effective unsupervised outlier detection method KFGOD.

3. Preliminaries

This section presents the fundamental definitions and notations related to FRS, kernelized fuzzy relations, and multi-granularity granular-ball computing, which constitute the theoretical foundation of the proposed method. For clarity and consistency, the main symbols used throughout this paper are summarized in Table 1.

Table 1: Nomenclature.

Notations	Descriptions
S	Set of data samples (objects)
A	Set of attributes
$IS = \langle S, A\rangle$	Information system
C	Set of condition attributes
D	Set of decision attributes
B	Attribute subset used to compute fuzzy relations, $B \subseteq C$
u_i	Individual samples in the dataset
GB	Set of multi-granularity granular-balls
$\mathcal{B}_g$	A specific granular-ball
z_g	Center of granular-ball $\mathcal{B}_g$
r_g	Radius of granular-ball $\mathcal{B}_g$
σ	Kernel width parameter in the Gaussian kernel
κ_{ij}^{B}	Kernelized fuzzy similarity between u_i and u_j w.r.t. B
κ_{B}^{GB}	Multi-granularity kernelized fuzzy relation on GB w.r.t. B
$M_{\kappa_B^{GB}}$	Multi-granularity kernelized fuzzy relation matrix w.r.t. B
$G(\kappa_B^{GB})$	Multi-granularity kernelized fuzzy granular structure induced by κ_B^{GB}
$[\mathcal{B}_g]_{\kappa_Q^{GB}}$	Multi-granularity kernelized fuzzy granule of $\mathcal{B}_g$ under relation κ_Q^{GB}
$\underline{\kappa_P^{GB}}[\mathcal{B}_g]_{\kappa_Q^{GB}}$	Multi-granularity kernelized fuzzy lower approximation of $[\mathcal{B}_g]_{\kappa_Q^{GB}}$ w.r.t. κ_P^{GB}
$\overline{\kappa_P^{GB}}[\mathcal{B}_g]_{\kappa_Q^{GB}}$	Multi-granularity kernelized fuzzy upper approximation of $[\mathcal{B}_g]_{\kappa_Q^{GB}}$ w.r.t. κ_P^{GB}
$\mathcal{A}(\kappa_P^{GB}, [\mathcal{B}_g]_{\kappa_Q^{GB}})$	Multi-granularity kernelized fuzzy approximation accuracy of $[\mathcal{B}_g]_{\kappa_Q^{GB}}$ w.r.t. κ_P^{GB}
$GOD([\mathcal{B}_g]_{\kappa_Q^{GB}})$	Outlier degree of $[\mathcal{B}_g]_{\kappa_Q^{GB}}$
$OD(\mathcal{B}_g)$	Outlier degree of granular-ball $\mathcal{B}_g$
$OD(u_i)$	Outlier degree of sample $u_i \in \mathcal{B}_g$

Before applying KFRS, the dataset needs to be imported into an Information System (IS), which can be denoted as $IS = \langle S, A\rangle$, where $S = \{u_1, u_2, \ldots, u_n\}$ is a non-empty set of samples, and $A = \{a_1, a_2, \ldots, a_m\}$ is a non-empty set of attributes. When $A = C \cup D$ and $C \cap D = \emptyset$, the IS is referred to as a Decision System (DS), where $C = \{c_1, c_2, \ldots, c_m\}$ is the set of condition attributes, and D is the set of decision attributes. For any $u_i \in S$ and $c_j \in C$, $c_j(u_i)$ represents the value of the c_j attribute for the sample u_i. In this study, we focus on unsupervised outlier detection; thus, we only consider the case where $A = C$. Consequently, the IS can be expressed as $IS = \langle S, C\rangle$.

Given $IS = \langle S, C\rangle$, a real-valued function $\kappa : S \times S \to \mathbb{R}$ is said to be kernel if it satisfies semipositivity and symmetry for any $u \in S$ [23]. For any kernel $\kappa : S \times S \to [0, 1]$ with $\kappa(u, u) = 1$ at least satisfies $T_{\cos}$-transitive [23], where

$$T_{\cos}(p, q) = \max\left\{pq - \sqrt{1 - p^2}\sqrt{1 - q^2}, 0\right\}. \tag{1}$$

Rational quadratic kernel, Spherical kernel, and Gaussian kernel are all commonly used kernel functions [23], which simultaneously satisfy reflexivity, symmetry, and $T_{\cos}$-transitivity. Consequently, the fuzzy relations calculated from these kernel functions are fuzzy T-equivalence relations. By replacing the fuzzy relations in the original FRS model with these kernel functions, the KFRS model can be constructed.

Given $IS = \langle S, C\rangle$ and for any $B \subseteq C$, the kernelized fuzzy relation κ_B w.r.t. B on S is a fuzzy set on $S \times S$. Similar to FRS, the kernelized fuzzy relation can also be denoted as a kernelized fuzzy relation matrix, i.e., $M_{\kappa_B} = (\kappa_{ij}^B)_{n\times n}$, where $\kappa_{ij}^B = \kappa_B(u_i, u_j)$ denotes the kernelized fuzzy relation between samples u_i and u_j w.r.t. B. Each row vector in M_{κ_B} represents a kernelized fuzzy set.

Definition 1 *Given IS* $= \langle S, C \rangle$ *and for any* $B \subseteq C$*, the kernelized fuzzy relation* κ_B *can be used to granulate the sample set* S *into multiple kernelized fuzzy information granules, referred to as the kernelized fuzzy granular structure* $G(\kappa_B)$*, which is defined as*

$$G(\kappa_B) = \{[u_1]_{\kappa_B}, [u_2]_{\kappa_B}, \ldots, [u_n]_{\kappa_B}\}, \tag{2}$$

where $[u_i]_{\kappa_B} = (\kappa^B_{i1}/u_1) + (\kappa^B_{i2}/u_2) + \cdots + (\kappa^B_{in}/u_n) = (\kappa^B_{i1}, \kappa^B_{i2}, \ldots, \kappa^B_{in})$ *denotes the kernelized fuzzy information granule induced by the kernelized fuzzy relation* κ_B*.*

Clearly, $[u_i]_{\kappa_B}$ is a fuzzy set on κ_B, where $[u_i]_{\kappa_B}(u_j) = \kappa_B(u_i, u_j) = \kappa^B_{ij}$. When $\kappa^B_{ij} = 0$, it indicates that u_j definitely does not belong to $[u_i]_{\kappa_B}$; when $\kappa^B_{ij} = 1$, it indicates that u_j definitely belongs to $[u_i]_{\kappa_B}$. When $\kappa^B_{ij} \in (0, 1)$, it suggests that u_j belongs to $[u_i]_{\kappa_B}$ to a certain extent. This non-absolute relations can effectively handle uncertainty information in the data. The cardinality of $[u_i]_{\kappa_B}$ is calculated as $|[u_i]_{\kappa_B}| = \sum_{u_j \in S} \kappa_B(u_i, u_j)$. Obviously, $1 \leq |[u_i]_{\kappa_B}| \leq n$. The cardinality of $[u_i]_{\kappa_B}$ reflects the similarity between u_i and all samples under κ_B.

It is not difficult to see that the kernelized fuzzy relation is a special type of fuzzy set, which also possesses the properties of fuzzy sets and operations such as intersection, union, and complement. To accommodate the applicability to different data, we can extend fuzzy set operations to triangular norm (t-norm), triangular conorm (s-norm or t-conorm), and negator [23]. Next, we use $T_{\cos}$, $S_{\cos}$, and N to denote the t-norm, s-norm, and the negator, respectively.

Definition 2 *Given IS* $= \langle S, C \rangle$*, for any* $P, Q \subseteq C$ *and for any* $u_i \in S$*.* κ_P *and* κ_Q *are two kernelized fuzzy relations.* $G(\kappa_Q) = \{[u_1]_{\kappa_Q}, [u_2]_{\kappa_Q}, \ldots, [u_n]_{\kappa_Q}\}$ *is the kernelized fuzzy granular structure induced by* κ_Q*. For any* $[u_i]_{\kappa_Q} \in G(\kappa_Q)$*, the kernelized fuzzy upper and lower approximations of* $[u_i]_{\kappa_Q}$ *w.r.t.* κ_P *are a pair of fuzzy sets on* S*, whose membership functions are defined as follows*

$$\overline{\kappa_P}[u_i]_{\kappa_Q}(u) = \sup_{u_j \in S} T_{\cos}\left(\kappa_P(u, u_j), [u_i]_{\kappa_Q}(u_j)\right), \tag{3}$$

$$\underline{\kappa_P}[u_i]_{\kappa_Q}(u) = \inf_{u_j \in S} S_{\cos}\left(N(\kappa_P(u, u_j)), [u_i]_{\kappa_Q}(u_j)\right), \tag{4}$$

where

$$T_{\cos}(p, q) = \max\left\{pq - \sqrt{1 - p^2}\sqrt{1 - q^2}, 0\right\}, \tag{5}$$

$$S_{\cos}(p, q) = \min\left\{p + q - pq - \sqrt{2p - p^2}\sqrt{2q - q^2}, 1\right\}. \tag{6}$$

4. Multi-granularity kernelized fuzzy rough sets

In this section, we overcome the limitations of traditional FRS models by combining MGBC with kernel functions, introducing a novel model, MKFRS. By embedding multi-granularity granular-balls into KFRS, MKFRS effectively utilizes multi-granularity data information, improving both efficiency and robustness. We begin by presenting the granular-ball generation in MGBC, then combine MGBC with KFRS to develop MKFRS.

4.1. Granular-ball generation

The granular-ball generation can adaptively endow different methods with the ability to process multi-granularity information, serving as a crucial step in the application of MGBC. The quality of granular-ball generation significantly affects the effectiveness of subsequent task computations. The process of granular-ball generation is a perception process that transitions from coarse granularity to fine granularity in the sample space. This study adopts an efficient and widely used granular-ball generation method in [37], which employs a top-down method to progressively refine the granular-balls.

Specifically, the method first considers the entire dataset as a single large granular-ball, then determines whether the granular-ball meets the conditions for splitting. If the conditions are met, the granular-ball is split. The granular-ball generation process ends when all granular-balls can no longer be split. This method uses a metric called Distribution Measure (DM) as the criterion for determining granular-ball splitting, which is calculated as

$$DM_g = \frac{1}{|\mathcal{B}_g|} r_g, \tag{7}$$

where $|\mathcal{B}_g|$ denotes the number of samples in $\mathcal{B}_g$.

Algorithm 1: Granular-ball generation

Input: $IS = \langle S, C \rangle$.
Output: A set of granular-balls GB.

$GB \leftarrow \{S\}$;
for $\mathcal{B}_g$ *in* GB **do**
  calculate DM_w and DM_g by Eqs. (7) and (8);
  if $DM_w > DM_g$ **then**
    Split $\mathcal{B}_g$ into $\mathcal{B}_{g_1}$ and $\mathcal{B}_{g_2}$;
    Remove $\mathcal{B}_g$ from GB;
    $GB \leftarrow GB \cup \{\mathcal{B}_{g_1}, \mathcal{B}_{g_2}\}$;
  end
end
Calculate mean(r) and median(r) in GB;
for $\mathcal{B}_g$ *in* GB **do**
  if $r_g > 2 \times \max(\text{mean}(r), \text{median}(r))$ **then**
    Split $\mathcal{B}_g$ into $\mathcal{B}_{g_1}$ and $\mathcal{B}_{g_2}$;
    Remove $\mathcal{B}_g$ from GB;
    $GB \leftarrow GB \cup \{\mathcal{B}_{g_1}, \mathcal{B}_{g_2}\}$;
  end
end
return GB.

After treating the entire dataset as a single large granular-ball $\mathcal{B}_A$, we identify the sample p_1 that is farthest from the center z_A, as well as the sample p_2 that is farthest from p_1. Next, we calculate the midpoints p_1^* and p_2^* between z_A and p_1, and between z_A and p_2, respectively. p_1^* and p_2^* are designated as the initial clustering centers, and we then assign the samples to either p_1^* or p_2^* based on their distances to these two centers, thereby forming two granular-balls $\mathcal{B}_{A_1}$ and $\mathcal{B}_{A_2}$. Before splitting, we also need to calculate the weighted DM value to determine whether $\mathcal{B}_A$ meets the conditions for division. The weight DM value is defined as

$$DM_w = \frac{|\mathcal{B}_{A_1}|}{|\mathcal{B}_A|} DM_{\mathcal{B}_{A_1}} + \frac{|\mathcal{B}_{A_2}|}{|\mathcal{B}_A|} DM_{\mathcal{B}_{A_2}}, \tag{8}$$

where $|\mathcal{B}_A|$, $|\mathcal{B}_{A_1}|$, and $|\mathcal{B}_{A_2}|$ denote the number of samples in each granular-ball. If $DM_w > DM_A$, then $\mathcal{B}_A$ will be split into $\mathcal{B}_{A_1}$ and $\mathcal{B}_{A_2}$, otherwise no split will be performed [37]. In addition, this method also takes into account that some boundary points and noise points will make the radius of some granular-balls too large. Therefore, after completing the above granular-ball generation, this method also needs to determine the radius of each granular-ball. If $r_g > 2 \times \max(\text{mean}(r), \text{median}(r))$, $\mathcal{B}_g$ needs to be further split, where mean(r) and median(r) represent the mean and median of the radii of all granular-balls. The whole granular-ball generation process is shown in Algorithm 1. The time complexity of this process is $O(|S| \log |S|)$ [37]. After completing the granular-ball generation, the original sample space will be covered and represented by multi-granularity granular-balls. Since the number of generated granular-balls is significantly smaller than the number of original samples, and subsequent operations use multi-granularity granular-balls as the basic processing units, MGBC can effectively improve the efficiency. Moreover, the multi-granularity nature of the granular-balls allows them to encompass noise within a large granular-ball, where the noise characteristics are smoothed out by the remaining normal samples, thus providing good robustness against noise [25, 26].

4.2. FRS model with multi-granularity and kernelization

We first define the kernelized fuzzy relation in the classic KFRS model and further generalize it to multi-granularity kernelized fuzzy relation to construct MKFRS. The selection of kernel function is the key to constructing KFRS. In this study, we choose the widely used Gaussian kernel function. Given $IS = \langle S, C \rangle$, for any $u_i, u_j \in S$ and $B \subseteq C$, the kernelized fuzzy similarity between u_i and u_j induced by B is defined as

$$\kappa_B(u_i, u_j) = \kappa_{ij}^B = \exp\left(-\frac{1}{\sigma} \sum_{b \in B} (b(u_i) - b(u_j))^2\right), \tag{9}$$

where σ denotes the Gaussian kernel parameter. The kernelized fuzzy relation defined above can simultaneously calculate the fuzzy relations between samples under single attributes and multiple attributes. However, the kernelized fuzzy relation does not consider the multi-granularity information in the data and does not have the ability to utilize multi-granularity information. To this end, we improve the original kernelized fuzzy relation and combine it with multi-granularity granular-balls to help process multi-granularity information.

Specifically, based on MGBC, we propose the multi-granularity kernelized fuzzy relation, which is defined as follows.

Definition 3 *Given $IS = \langle S, C\rangle$ and GB is the set of granular-balls generated on IS. For any $B \subseteq C$ and any $\mathcal{B}_g, \mathcal{B}_h \in GB$, the multi-granularity kernelized fuzzy relation κ_B^{GB} w.r.t. B on GB is defined as*

$$\kappa_B^{GB}(\mathcal{B}_g, \mathcal{B}_h) = \kappa_B(z_g, z_h) = \exp\left(-\frac{1}{\sigma}\sum_{b\in B}(b(z_g) - b(z_h))^2\right), \tag{10}$$

where z_g denotes the center of $\mathcal{B}_g$ and $b(z_g)$ denotes the value of the b attribute for the center of granular-ball $\mathcal{B}_g$.

Similarly, the multi-granularity kernelized fuzzy relation κ_B^{GB} w.r.t. B on GB is a fuzzy set on $GB \times GB$ and κ_B^{GB} can be denoted as a multi-granularity kernelized fuzzy relation matrix, i.e., $M_{\kappa_B^{GB}} = (\kappa_B^{GB}(\mathcal{B}_g, \mathcal{B}_h))_{|GB|\times|GB|}$, where $g, h = \{1, 2, \ldots, |GB|\}$. Each row vector in $M_{\kappa_B^{GB}}$ denotes a multi-granularity kernelized fuzzy set.

Definition 4 *Given $IS = \langle S, C\rangle$ and GB is the set of granular-balls generated on IS. For any $B \subseteq C$, κ_B^{GB} can be used to granulate GB into multi-granularity kernelized fuzzy information granules, referred to as the multi-granularity kernelized fuzzy granular structure $G(\kappa_B^{GB})$, which is defined as*

$$G(\kappa_B^{GB}) = \left\{[\mathcal{B}_1]_{\kappa_B^{GB}}, [\mathcal{B}_2]_{\kappa_B^{GB}}, \ldots, [\mathcal{B}_k]_{\kappa_B^{GB}}\right\}, \tag{11}$$

where $[\mathcal{B}_g]_{\kappa_B^{GB}} = (\kappa_B^{GB}(\mathcal{B}_g, \mathcal{B}_1), \ldots, \kappa_B^{GB}(\mathcal{B}_g, \mathcal{B}_k))$ denotes the multi-granularity kernelized fuzzy information granule induced by the multi-granularity kernelized fuzzy relation κ_B^{GB}.

For any $[\mathcal{B}_g]_{\kappa_B^{GB}} \in G(\kappa_B^{GB})$, $[\mathcal{B}_g]_{\kappa_B^{GB}}$ is a fuzzy set on κ_B^{GB}, where $[\mathcal{B}_g]_{\kappa_B^{GB}}(\mathcal{B}_h) = \kappa_B^{GB}(\mathcal{B}_g, \mathcal{B}_h)$, indicating the degree to which the granular-ball $\mathcal{B}_h$ belongs to $[\mathcal{B}_g]_{\kappa_B^{GB}}$ under κ_B^{GB}. The cardinality of $[\mathcal{B}_g]_{\kappa_B^{GB}}$ is calculated as $|[\mathcal{B}_g]_{\kappa_B^{GB}}| = \sum_{\mathcal{B}_h\in GB} \kappa_B^{GB}(\mathcal{B}_g, \mathcal{B}_h)$ and $1 \le |[\mathcal{B}_g]_{\kappa_B^{GB}}| \le k$, where k is the number of granular-balls generated on IS.

Next, we further give the definition of multi-granularity kernelized fuzzy approximations.

Definition 5 *Given $IS = \langle S, C\rangle$ and GB is the set of granular-balls generated on IS. For any $P, Q \subseteq C$ and for any $\mathcal{B}_g \in GB$. κ_P^{GB} and κ_Q^{GB} are two multi-granularity kernelized fuzzy relations. $G(\kappa_Q^{GB}) = \{[\mathcal{B}_1]_{\kappa_Q^{GB}}, [\mathcal{B}_2]_{\kappa_Q^{GB}}, \ldots, [\mathcal{B}_g]_{\kappa_Q^{GB}}\}$ is the multi-granularity kernelized fuzzy granular structure induced by κ_Q^{GB}. For any $[\mathcal{B}_k]_{\kappa_Q^{GB}} \in G(\kappa_Q^{GB})$, the multi-granularity kernelized fuzzy upper and lower approximations of $[\mathcal{B}_k]_{\kappa_Q^{GB}}$ w.r.t. κ_P^{GB} are a pair of fuzzy sets on GB, whose membership functions are defined as follows*

$$\overline{\kappa_P^{GB}}[\mathcal{B}_g]_{\kappa_Q^{GB}}(\mathcal{B}) = \sup_{\mathcal{B}_h\in GB} T_{\cos}\left(\kappa_P^{GB}(\mathcal{B}, \mathcal{B}_h), [\mathcal{B}_g]_{\kappa_Q^{GB}}(\mathcal{B}_h)\right), \tag{12}$$

$$\underline{\kappa_P^{GB}}[\mathcal{B}_g]_{\kappa_Q^{GB}}(\mathcal{B}) = \inf_{\mathcal{B}_h\in GB} S_{\cos}\left(N(\kappa_P^{GB}(\mathcal{B}, \mathcal{B}_h)), [\mathcal{B}_g]_{\kappa_Q^{GB}}(\mathcal{B}_h)\right). \tag{13}$$

The upper approximation $\overline{\kappa_P^{GB}}[\mathcal{B}_g]_{\kappa_Q^{GB}}(\mathcal{B})$ describes the maximum extent to which the granular ball $\mathcal{B}$ may belong to $[\mathcal{B}_g]_{\kappa_Q^{GB}}$ from the perspective of κ_P^{GB}. The lower approximation $\underline{\kappa_P^{GB}}[\mathcal{B}_g]_{\kappa_Q^{GB}}(\mathcal{B})$ describes the minimum extent to which the granular ball $\mathcal{B}$ necessarily belongs to $[\mathcal{B}_g]_{\kappa_Q^{GB}}$ from the perspective of κ_P^{GB}.

From the above definitions, it can be seen that the multi-granularity kernelized fuzzy approximation we defined is fundamentally different from existing fuzzy approximations. First, we use kernel functions to more accurately capture the complex relations between samples. Second, we employ MGBC to effectively extract multi-granularity information from the data.

Property 1 *Given $IS = \langle S, C\rangle$ and GB is the set of granular-balls generated on IS. For any $P, Q \subseteq C$, κ_P^{GB} and κ_Q^{GB} are two multi-granularity kernelized fuzzy relations on GB. For any $[\mathcal{B}_g]_{\kappa_Q^{GB}} \in G(\kappa_Q^{GB})$, $\overline{\kappa_P^{GB}}[\mathcal{B}_g]_{\kappa_Q^{GB}}$ and $\underline{\kappa_P^{GB}}[\mathcal{B}_g]_{\kappa_Q^{GB}}$ satisfies the following properties.*

- $\underline{\kappa_P^{GB}}[\mathcal{B}_g]_{\kappa_Q^{GB}} \subseteq [\mathcal{B}_g]_{\kappa_Q^{GB}} \subseteq \overline{\kappa_P^{GB}}[\mathcal{B}_g]_{\kappa_Q^{GB}}$;
- $\underline{\kappa_P^{GB}}(\underline{\kappa_P^{GB}}[\mathcal{B}_g]_{\kappa_Q^{GB}}) = \underline{\kappa_P^{GB}}[\mathcal{B}_g]_{\kappa_Q^{GB}}$; $\overline{\kappa_P^{GB}}(\overline{\kappa_P^{GB}}[\mathcal{B}_g]_{\kappa_Q^{GB}}) = \overline{\kappa_P^{GB}}[\mathcal{B}_g]_{\kappa_Q^{GB}}$;
- *For any* $Y \subseteq C$, *if* $\kappa_Y^{GB} \subseteq \kappa_P^{GB}$, *then* $\underline{\kappa_P^{GB}}[\mathcal{B}_g]_{\kappa_Q^{GB}} \subseteq \underline{\kappa_Y^{GB}}[\mathcal{B}_g]_{\kappa_Q^{GB}} \subseteq [\mathcal{B}_g]_{\kappa_Q^{GB}} \subseteq \overline{\kappa_Y^{GB}}[\mathcal{B}_g]_{\kappa_Q^{GB}} \subseteq \overline{\kappa_P^{GB}}[\mathcal{B}_g]_{\kappa_Q^{GB}}$.

In summary, by integrating MGBC and kernel functions, we propose a new FRS model, namely MKFRS. In the next section, we develop a new unsupervised outlier detection method based on MKFRS.

5. The proposed outlier detection

In this section, unlike simple binary classification methods, we assign each sample a value within the [0,1] interval to characterize its outlier degree. Specifically, we propose a Kernelized Fuzzy approximation fusion model with Granular-ball computing for Outlier Detection (KFGOD). In KFGOD, we first define the outlier degree of multi-granularity kernelized fuzzy information granules. Then, the abnormal information of multiple fuzzy information granules associated with a granular-ball is weighted and fused to characterize the outlier degree of the granular-ball. Finally, we map the outlier degrees of each granular-ball to the samples within that granular-ball to obtain the outlier degrees of the samples.

5.1. Detection method

In MKFRS, we define the corresponding fuzzy approximation, which describes how a multi-granularity kernelized fuzzy set (e.g., $[\mathcal{B}_g]_{\kappa_Q^{GB}}$) is approximated by another multi-granularity kernelized fuzzy relation (e.g., κ_P^{GB}). This fuzzy approximation not only integrates multi-granularity kernelized fuzzy information but also provides a theoretical foundation for subsequent outlier detection. Next, we define a novel fuzzy approximation accuracy, which can be measured by the difference between the fuzzy upper and lower approximations, which reflects the degree of uncertainty in the granular-ball's membership. Specifically, if the upper approximation significantly exceeds the lower approximation, it indicates a high level of uncertainty for the granular-ball, suggesting a potentially higher outlier degree; conversely, if the two are close, it implies that the granular-ball's membership is relatively clear, typically corresponding to normal data. By quantifying this uncertainty, we can further derive outlier degrees for granular-balls and map them to the sample points within, thus achieving efficient and robust outlier detection. First, we provide the definition of the multi-granularity kernelized fuzzy approximation accuracy by fusing corresponding approximation information.

Definition 6 *Given* $IS = \langle S, C\rangle$ *and* GB *is the set of granular-balls generated on* IS. *For any* $P, Q \subseteq C$, κ_P^{GB} *and* κ_Q^{GB} *are two multi-granularity kernelized fuzzy relations on* GB. *For any* $[\mathcal{B}_g]_{\kappa_Q^{GB}} \in G(\kappa_Q^{GB})$, *the multi-granularity kernelized fuzzy approximation accuracy of* $[\mathcal{B}_g]_{\kappa_Q^{GB}}$ *w.r.t.* κ_P^{GB} *is defined as*

$$\mathcal{A}(\kappa_P^{GB}, [\mathcal{B}_g]_{\kappa_Q^{GB}}) = \frac{|\underline{\kappa_P^{GB}}[\mathcal{B}_g]_{\kappa_Q^{GB}}|}{|\overline{\kappa_P^{GB}}[\mathcal{B}_g]_{\kappa_Q^{GB}}|}. \tag{14}$$

$\mathcal{A}(\kappa_P^{GB}, [\mathcal{B}_g]_{\kappa_Q^{GB}})$ reflects the difficulty of approximating $[\mathcal{B}_g]_{\kappa_Q^{GB}}$ by κ_P^{GB}. The closer its value is to 0, the greater the difference between the upper and lower approximations, indicating a higher corresponding uncertainty. Thus, it can be used to reflect the outlier degree of $[\mathcal{B}_g]_{\kappa_Q^{GB}}$. Based on this, we next provide the definition of the outlier degree of multi-granularity kernelized fuzzy information granules.

Definition 7 *Given* $IS = \langle S, C\rangle$ *and* GB *is the set of granular-balls generated on* IS. *For any* $Q \subseteq C$ *and* $P = C - Q$, κ_P^{GB} *and* κ_Q^{GB} *are two multi-granularity kernelized fuzzy relations on* GB. *The outlier degree of* $[\mathcal{B}_g]_{\kappa_Q^{GB}} \in G(\kappa_Q^{GB})$ *is defined as*

$$GOD([\mathcal{B}_g]_{\kappa_Q^{GB}}) = 1 - W([\mathcal{B}_g]_{\kappa_Q^{GB}}) \cdot \mathcal{A}(\kappa_P^{GB}, [\mathcal{B}_g]_{\kappa_Q^{GB}}), \tag{15}$$

where $W([\mathcal{B}_g]_{\kappa_Q^{GB}}) = |[\mathcal{B}_g]_{\kappa_Q^{GB}}|/|GB|$, $|[\mathcal{B}_g]_{\kappa_Q^{GB}}|$ *denotes the cardinality of* $[\mathcal{B}_g]_{\kappa_Q^{GB}}$, *and* $|GB|$ *denotes the number of granular-balls.*

In $GOD([\mathcal{B}_g]_{\kappa_Q^{GB}})$, we utilize the uncertainty reflected by $\mathcal{A}(\kappa_P^{GB}, [\mathcal{B}_g]_{\kappa_Q^{GB}})$ to measure the outlier degree of $[\mathcal{B}_g]_{\kappa_Q^{GB}}$. Additionally, in $W([\mathcal{B}_g]_{\kappa_Q^{GB}})$, we use the cardinality of $[\mathcal{B}_g]_{\kappa_Q^{GB}}$ to calculate its weight in measuring the outlier degree. $|[\mathcal{B}_g]_{\kappa_Q^{GB}}|$ denotes the cardinality of $[\mathcal{B}_g]_{\kappa_Q^{GB}}$ and reflects the sum of similarities between $\mathcal{B}_g$ and other granular-balls under κ_Q^{GB}. A smaller value indicates that the granule $[\mathcal{B}_g]_{\kappa_Q^{GB}}$ significantly differs from most granular-balls, thus suggesting a higher likelihood of being abnormal.

In $IS = \langle S, C\rangle$, each subset of C can be used to construct a fuzzy relation matrix. Consequently, there would be $2^{|C|}$ fuzzy relation matrices to be constructed, which is clearly impractical for real-world applications. Therefore, following existing research [2, 4, 9, 19], we only consider individual attributes in C when calculating $GOD([\mathcal{B}_g]_{\kappa_Q^{GB}})$. Based on this, we introduce the outlier degree of a granular-ball, which is defined as follows.

Definition 8 *Given $IS = \langle S, C\rangle$ and GB is the set of granular-balls generated on IS. For any $\mathcal{B}_g \in GB$, the outlier degree of $\mathcal{B}_g$ is defined as*

$$OD(\mathcal{B}_g) = \frac{1}{|C|}\sum_{c\in C} W(c, \mathcal{B}_g)\cdot GOD([\mathcal{B}_g]_{\kappa_c^{GB}}), \tag{16}$$

where $W(c, \mathcal{B}_g) = 1 - \sqrt[3]{|[\mathcal{B}_g]_{\kappa_c^{GB}}|/|GB|}$, $|GB|$ denotes the number of granular-balls, and $|C|$ denotes the number of attributes.

In the above definition, we measure the outlier degree of a granular-ball by weighted fusion of the abnormal information from multiple granules associated with the granular-ball. In designing the weight $W(c, \mathcal{B}_g)$, we take into account that outliers should be a minority of objects in the data. For any $\mathcal{B}_g \in GB$ and $c \in C$, the cardinality of $[\mathcal{B}_g]_{\kappa_c^{GB}}$ reflects the similarity between $\mathcal{B}_g$ and other granular-balls under κ_c^{GB}. If the cardinality of $[\mathcal{B}_g]_{\kappa_c^{GB}}$ is smaller than that of other granules, $\mathcal{B}_g$ may belong to a minority class in the data, and thus it should be assigned a higher weight when calculating the outlier degree. KFGOD's multi-granularity fusion aggregates information across different granularity levels, allowing the model to distinguish between noise-induced variations and true outliers by leveraging broader contextual patterns. Specifically, larger granular-balls smooth out local noise by encapsulating more sample points, while smaller granular-balls preserve fine-grained details, ensuring that outliers with subtle deviations are not overlooked. This balance enhances KFGOD's robustness to noise.

Through the above calculations, we can obtain the outlier degrees for each granular-ball in the granular-ball set. Subsequently, we map the outlier degrees of the granular-balls to their internal samples to derive the outlier degrees of the samples. For any $\mathcal{B}_g \in GB$ and $u \in \mathcal{B}_g$, $OD(u) = OD(\mathcal{B}_g)$. Finally, we can set an outlier threshold ξ for outlier detection. For any $u \in S$, if $OD(u) > \xi$ then u is detected as an outlier.

5.2. Detection algorithm

Based on the KFGOD method mentioned above, we provide the corresponding pseudocode and time complexity analysis below.

As shown in Algorithm 2, we first generate the granular-ball set GB on IS. Then, we calculate the fuzzy relation matrix $M_{\kappa_c^{GB}}$ under a single attribute $c \in C$, laying the foundation for the subsequent calculation of fuzzy approximation accuracy. Subsequently, for any $\mathcal{B} \in GB$ and $c \in C$, we calculate the fuzzy relation matrix $M_{\kappa_{C-\{c\}}^{GB}}$ and calculate the fuzzy approximation accuracy $\mathcal{A}(\kappa_{C-\{c\}}^{GB}, [\mathcal{B}]_{\kappa_c^{GB}})$ of $[\mathcal{B}]_{\kappa_c^{GB}}$ being approximated by $\kappa_{C-\{c\}}^{GB}$. Based on this, we further calculate the outlier degree $GOD([\mathcal{B}]_{\kappa_c^{GB}})$ of $[\mathcal{B}]_{\kappa_c^{GB}}$. Then, we fuse the abnormal information of multiple fuzzy information granules associated with $\mathcal{B}$ to measure the outlier degree $OD(\mathcal{B})$ of $\mathcal{B}$. Finally, we map $OD(\mathcal{B})$ to the samples $u \in \mathcal{B}$ to obtain the outlier degrees of the samples. During the algorithm execution, the time complexity of the corresponding granular-ball generation algorithm in Step 2 is $O(|S|\log|S|)$. The time complexity of Steps 3-5 is $O(|C||GB|^2)$. The time complexity of Steps 6-18 is $O(|GB||C|)$. Therefore, the final time complexity of KFGOD is $O(|C||GB|^2)$, where $|C|$ denotes the number of attributes and $|GB|$ denotes the number of granular-balls generated on IS.

6. Experiments

In this section, we validate the effectiveness of KFGOD through a series of experiments. We begin by outlining the experimental setup, including the datasets and baseline methods used. Next, we analyze the experimental results based on Receiver Operating Characteristic (ROC) curves, Area Under the ROC Curve (AUC), Average Precision (AP), and geometric mean (g-mean) metrics. Furthermore, we investigate the impact of multi-granularity granular-balls and different kernel functions on KFGOD's performance, and the sensitivity of KFGOD to its hyperparameters. Finally, we conduct a statistical analysis to determine whether there are statistically significant differences between the methods.

Algorithm 2: KFGOD

Input: $IS = \langle S, C \rangle, \sigma$.

Output: A set of outlier degrees OD.

Initialize a set of outlier degrees $OD \leftarrow \emptyset$;
Generate granular-balls GB on IS by Algorithm 1;
for $c \in C$ **do**
    Calculate the multi-granularity kernelized fuzzy relation matrix $M_{\kappa_c^{GB}}$ w.r.t. c by Eq. (10);
end
for $\mathcal{B} \in GB$ **do**
    for $c \in C$ **do**
        Calculate the multi-granularity kernelized fuzzy relation matrix $M_{\kappa_{C-\{c\}}^{GB}}$ w.r.t. $C - \{c\}$ by Eq. (10);
        Calculate the multi-granularity kernelized fuzzy approximation accuracy $\mathcal{A}(\kappa_{C-\{c\}}^{GB}, [\mathcal{B}]_{\kappa_c^{GB}})$ of $[\mathcal{B}]_{\kappa_c^{GB}}$ w.r.t. $\kappa_{C-\{c\}}^{GB}$ by Eq. (14);
        Calculate the outlier degree $GOD([\mathcal{B}]_{\kappa_c^{GB}})$ of $[\mathcal{B}]_{\kappa_c^{GB}}$ by Eq. (15);
        Calculate the weight coefficient $W(c, \mathcal{B})$;
    end
    Calculate the outlier degree $OD(\mathcal{B})$ of $\mathcal{B}$ by Eq. (16);
    for $u \in \mathcal{B}$ **do**
        Calculate the outlier degree $OD(u) = OD(\mathcal{B})$ of u;
        Remove $OD(\mathcal{B})$ from OD;
    end
end
return OD.

Table 2: The basic information of 20 datasets in our experiments.

No.	Datasets	Abbr.	# Samples	# Attributes	# Outliers (% Ratios)	Domains
1	cardio	Cardio	1831	21	176 (9.60%)	Healthcare
2	cardiotocography_2and3_33_variant1	Cardiot	1688	21	33 (2.00%)	Healthcare
3	diabetes_tested_positive_26_variant1	Diab	526	8	26 (4.90%)	Healthcare
4	ecoli	Ecoli	336	7	9 (2.70%)	Biology
5	ionosphere_b_24_variant1	Iono	249	34	24 (9.60%)	Oryctognosy
6	iris_Irisvirginica_11_variant1	Iris	111	4	11 (9.90%)	Biology
7	pima_TRUE_55_variant1	Pima	555	9	55 (9.90%)	Healthcare
8	sonar_M_10_variant1	Sonar	107	60	10 (9.30%)	Physics and Chemistry
9	wbc_malignant_39_variant1	Wbc	483	9	90 (18.60%)	Healthcare
10	wdbc_M_39_variant1	Wdbc	396	31	39 (9.80%)	Healthcare
11	yeast_ERL_5_variant1	Yeast	1141	8	5 (0.40%)	Biology
12	annealing_variant1	Ann	798	38	42 (5.30%)	Physics and Chemistry
13	arrhythmia_variant1	Arr	452	279	66 (14.60%)	Healthcare
14	bands_band_27_variant1	Bands_27	339	39	27 (7.96%)	Physics and Chemistry
15	bands_band_42_variant1	Bands_42	354	39	42 (11.90%)	Physics and Chemistry
16	creditA_plus_42_variant1	Credit	425	15	42 (9.90%)	Business
17	german_1_14_variant1	German	714	20	14 (2.00%)	Business
18	heart270_2_16_variant1	Heart	166	13	16 (9.60%)	Healthcare
19	horse_1_12_variant1	Horse	256	27	12 (4.70%)	Biology
20	sick_sick_35_variant1	Sick	3576	29	35 (0.63%)	Healthcare

6.1. Experimental setups

We select 20 commonly used datasets from publicly available datasets[1][2] for our experiments, with their basic information presented in Table 2. These datasets cover real-world application domains including Healthcare, Biology, Oryctognosy, Physics and Chemistry, and Business. We choose 14 baseline methods, encompassing both classic and state-of-the-art methods. Specifically, these methods include: the Local Distance-based Outlier Factor (LDOF) [42], the Local Outlier Probabilities-based outlier detection (LoOP) [43], the Outlier Detection based on Granular Computing and Rough set theory (ODGrCR) [44], the reverse uNreaChability-based outlier detection (NC) [45], the provable Self-Representation based Outlier detection (SRO) [46], the Deep Support Vector Data Description-based

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

outlier detection (DeepSVDD) [47], the Variance structural scorE-based outlier detection (VarE) [48], the Rotation-based Outlier Detection (ROD) [49], the Directed density ratio Changing Rate-based Outlier Detection (DCROD) [50], the Deep Isolation Forest for outlier detection (DIF) [51], the Incomplete Local and Global Neighborhood Information-based outlier detection (ILGNI) [52], the Kernelized Fuzzy-Rough Anomaly Detection (KFRAD) [24], the Anomaly Detection based improved k-nearest Neighbor Rough Sets (ADkNRS) [53], and the Multi-Granular-Ball fuzzy information granules-based Outlier Detection (MGBOD) [4]. The relevant information and hyperparameter settings for these methods are shown in Table 3.

Table 3: The main ideas and hyperparameter settings of all methods in the experiment, where ✖ denotes that the method does not require hyperparameter selection or step size.

Methods (Years)	**Main ideas**	**Hyperparameter tuning ranges**	**Step sizes**
KFGOD (Ours)	Multi-granularity kernelized fuzzy approximation fusion	$\sigma \in [0.02, 0.8]$	0.02
MGBOD (2025) [4]	Multi-granular-ball fuzzy information granules	$\sigma \in [0, 1]$	0.05
ADkNRS (2025) [53]	Improved k-nearest neighbor rough sets	$k \in [1, 60]$	1
KFRAD (2024) [24]	Kernelized fuzzy rough sets	$\delta \in [0.02, 0.8]$	0.02
ILGNI (2023) [52]	Incomplete local and global neighborhood information	✖	✖
DIF (2023) [51]	Deep isolation forest	$r = 50, t = 6, \psi = 256$	✖
DCROD (2022) [50]	Directed density ratio changing rate	$k \in [1, 60]$	1
ROD (2020) [49]	Rotation information	✖	✖
VarE (2018) [48]	Structural scores in a high-dimensional space	$\lambda \in \{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3\}$	✖
DeepSVDD (2018) [47]	Deep support vector data description	✖	✖
SRO (2017) [46]	Provable self-representation	$\alpha \in [2, 20]$	1
NC (2016) [45]	Reverse unreachability	$k \in [1, 60]$	1
ODGrCR (2015) [44]	Granular computing and rough set theory	✖	✖
LoOP (2009) [43]	Local outlier probability	$k \in [1, 60]$	1
LDOF (2009) [42]	Local distance	$k \in [1, 60]$	1

6.2. Evaluation protocol

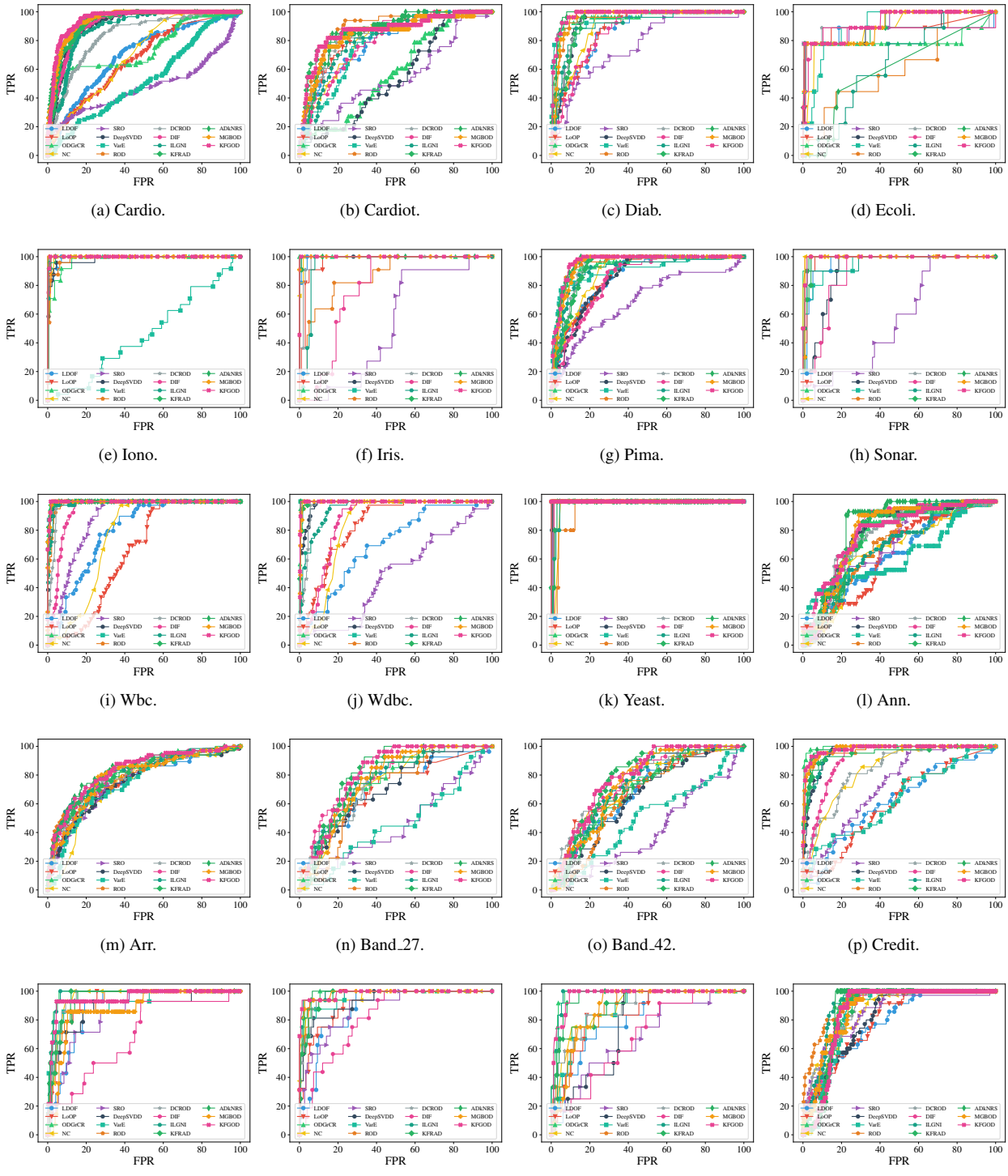
We analyze the experimental results using four widely adopted evaluation metrics: ROC, AUC, AP, and g-mean. The ROC curve illustrates the trade-off between the True Positive Rate (TPR) and False Positive Rate (FPR) across different thresholds, while the AUC provides a single scalar value summarizing the overall discriminative ability of the model. AP reflects the area under the precision-recall curve, offering a reliable measure for performance under class imbalance by emphasizing the precision of outlier predictions. The g-mean metric is defined as the geometric mean of the TPR and True Negative Rate (TNR). It evaluates the balance between correctly detecting both outliers and inliers. This metric is particularly useful in imbalanced scenarios, as it penalizes methods that perform well on only one class while neglecting the other, thereby providing a more comprehensive assessment of detection robustness.

6.3. Compared with the state-of-the-arts

To comprehensively evaluate the performance of KFGOD and baseline methods, we first present the ROC curves of 15 representative outlier detection methods across 20 public benchmark datasets. The ROC curve illustrates the trade-off between TPR and FPR, with the TPR plotted on the y-axis and the FPR on the x-axis. A curve that approaches the top-left corner indicates a high TPR and low FPR, thus reflecting better detection capability.

As shown in Figure 2, the ROC curves of KFGOD on the *Cardio*, *Diab*, *Ecoli*, *Iris*, *Pima*, *Wbc*, *Yeast*, *Band_42*, and *Heart* datasets are closer to the top-left corner of the axes, indicating superior performance. Next, we calculate the AUC results for each method across different datasets to compare their AUC results. As shown in Table 4, KFGOD achieves the highest rank in 9 out of the 20 datasets and boasts the best average AUC result of 0.931. The AP results of the 15 methods across 20 datasets are presented in Table 5, where KFGOD again achieves the best average AP result, surpassing the second-best method by 0.043. Finally, regarding the g-mean results, as shown in Table 6, KFGOD also achieves the best average g-mean result. From these experimental results, it is evident that KFGOD consistently demonstrates superior performance across these metrics, validating the effectiveness of KFGOD.

To evaluate KFGOD's performance in detecting minority-class outliers, we select the *Cardiot* and *Yeast* datasets, generating datasets with outlier ratios of 2%, 4%, 6%, 8%, and 10% using the widely-used SMOTE technique [54]. As shown in Table 7, on both the *Cardiot* and *Yeast* datasets, the detection performance of DCROD, DIF, KFRAD, and MGBOD declines as the outlier ratio increases, whereas KFGOD maintains satisfactory outlier detection performance. This further demonstrates that KFGOD exhibits strong robustness in imbalanced datasets.



(a) Cardio. (b) Cardiot. (c) Diab. (d) Ecoli.

(e) Iono. (f) Iris. (g) Pima. (h) Sonar.

(i) Wbc. (j) Wdbc. (k) Yeast. (l) Ann.

(m) Arr. (n) Band_27. (o) Band_42. (p) Credit.

(q) German. (r) Heart. (s) Horse. (t) Sick.

Figure 2: ROC curves of 15 methods and 20 datasets.

The superior performance of KFGOD can be attributed to several key advantages: 1) the use of multi-granularity granular-balls enables robust and noise-tolerant data representation; 2) kernelized fuzzy relations allow modeling of complex nonlinear patterns; 3) fuzzy approximation accuracy provides a principled measure of outlier uncertainty, and 4) multi-view weighted fusion captures abnormal information from different attribute perspectives. These components work jointly to enhance the discriminative capability of KFGOD in unsupervised outlier detection scenarios.

Table 4: AUC results comparison of 15 outlier detection methods across 20 datasets. The best AUC result in each row is **bolded**, while the second highest performance is underlined.

Datasets	LDOF	LoOP	ODGrCR	NC	SRO	DeepSVDD	VarE	ROD	DCROD	DIF	ILGNI	KFRAD	ADkNRS	MGBOD	KFGOD
Cardio	0.698	0.667	0.748	0.703	0.478	0.909	0.526	0.932	0.834	0.918	0.887	0.934	0.933	0.941	**0.950**
Cardiot	0.790	0.833	0.581	0.806	0.549	0.549	0.780	**0.893**	0.834	0.796	0.811	0.880	0.863	0.846	0.869
Diab	0.863	0.870	0.958	0.902	0.754	0.929	0.965	0.954	0.935	0.861	0.905	0.931	0.975	0.960	**0.978**
Ecoli	0.863	0.886	0.799	0.881	0.922	0.879	0.899	0.613	0.875	0.864	0.638	0.910	0.602	0.909	**0.926**
Iono	0.988	0.993	0.976	0.993	**1.000**	0.984	0.456	0.987	**1.000**	0.999	0.998	0.998	**1.000**	**1.000**	0.999
Iris	0.995	0.969	**1.000**	0.996	0.534	0.999	**1.000**	0.866	0.983	0.779	0.951	0.999	**1.000**	**1.000**	**1.000**
Pima	0.855	0.860	0.938	0.888	0.685	0.849	0.904	0.947	0.928	0.839	0.901	0.920	0.956	0.947	**0.962**
Sonar	0.964	0.981	**0.998**	**0.998**	0.562	0.901	0.955	0.990	0.989	0.882	0.965	0.982	0.992	0.993	0.993
Wbc	0.792	0.645	0.995	0.756	0.870	0.993	**0.997**	0.985	0.976	0.939	0.988	**0.997**	0.996	**0.997**	**0.997**
Wdbc	0.702	0.839	**0.998**	0.834	0.489	0.981	0.997	0.995	0.968	0.871	0.951	0.993	0.997	0.994	0.997
Yeast	0.987	0.988	0.999	0.970	0.995	0.997	**1.000**	0.949	0.990	0.987	0.981	0.999	0.984	**1.000**	**1.000**
Ann	0.627	0.623	0.734	0.653	0.693	0.787	0.619	0.712	0.755	0.792	0.686	0.783	**0.816**	0.791	0.784
Arr	0.749	0.771	0.813	0.734	0.757	0.750	0.739	0.801	0.796	0.809	0.816	0.792	**0.831**	0.789	0.828
Bands_27	0.715	0.696	0.755	0.739	0.456	0.695	0.459	0.705	0.759	0.796	0.759	**0.813**	0.764	0.766	0.808
Bands_42	0.683	0.710	0.696	0.732	0.408	0.650	0.496	0.661	0.756	0.752	0.743	0.781	0.686	0.751	**0.784**
Credit	0.616	0.598	**0.994**	0.837	0.724	0.959	0.625	0.974	0.841	0.910	0.964	0.950	0.979	0.980	0.981
German	0.889	0.925	0.979	0.955	0.852	0.882	0.919	0.927	0.955	0.686	**0.980**	0.950	0.956	0.874	0.951
Heart	0.880	0.923	0.985	0.969	0.880	0.930	0.962	0.975	0.970	0.817	0.983	0.976	0.976	**0.986**	**0.986**
Horse	0.803	0.838	**0.980**	0.885	0.685	0.761	0.869	0.880	0.878	0.683	0.975	0.903	0.954	0.883	0.971
Sick	0.779	0.788	0.870	0.853	0.840	0.815	0.842	**0.927**	0.875	0.879	0.902	0.852	0.865	0.875	0.858
Average	0.812	0.820	0.890	0.854	0.707	0.860	0.800	0.884	0.895	0.843	0.889	0.917	0.906	0.914	**0.931**

Table 5: AP results comparison of 15 outlier detection methods across 20 datasets. The best AP result in each row is **bolded**, while the second highest performance is underlined.

Datasets	LDOF	LoOP	ODGrCR	NC	SRO	DeepSVDD	VarE	ROD	DCROD	DIF	ILGNI	KFRAD	ADkNRS	MGBOD	KFGOD
Cardio	0.209	0.192	0.487	0.190	0.189	0.500	0.105	0.554	0.366	0.551	0.439	0.627	0.541	0.647	**0.656**
Cardiot	0.100	0.148	0.067	0.091	0.053	0.038	0.058	0.122	0.129	0.081	0.062	0.147	0.150	0.129	**0.158**
Diab	0.254	0.266	0.610	0.276	0.151	0.321	**0.704**	0.481	0.440	0.215	0.331	0.392	0.575	0.447	0.656
Ecoli	0.587	0.561	0.585	0.258	**0.712**	0.537	0.336	0.052	0.638	0.471	0.041	0.658	0.056	0.680	0.646
Iono	0.880	0.947	0.814	0.919	**1.000**	0.919	0.089	0.907	**1.000**	0.989	0.982	0.984	0.997	**1.000**	0.993
Iris	0.967	0.751	**1.000**	0.971	0.106	0.992	**1.000**	0.429	0.864	0.200	0.531	0.992	**1.000**	**1.000**	**1.000**
Pima	0.359	0.376	0.573	0.395	0.235	0.347	0.644	0.591	0.523	0.304	0.471	0.485	0.618	0.560	**0.680**
Sonar	0.717	0.804	**0.983**	0.981	0.126	0.425	0.774	0.887	0.904	0.329	0.861	0.815	0.921	0.911	0.930
Wbc	0.185	0.104	0.930	0.144	0.251	0.925	**0.970**	0.825	0.683	0.412	0.889	0.965	0.960	0.964	0.965
Wdbc	0.187	0.294	**0.976**	0.241	0.140	0.805	0.951	0.894	0.651	0.298	0.675	0.890	0.940	0.894	0.940
Yeast	0.206	0.171	0.759	0.076	0.315	0.464	**1.000**	0.055	0.223	0.159	0.152	0.796	0.157	**1.000**	**1.000**
Ann	0.126	0.069	0.114	0.080	0.097	0.147	0.147	0.112	0.107	0.120	0.093	0.221	0.181	0.196	**0.230**
Arr	0.393	0.427	0.494	0.269	0.337	0.405	0.288	**0.514**	0.449	0.457	0.494	0.447	0.513	0.419	0.502
Bands_27	0.245	0.256	0.177	0.218	0.131	0.148	0.088	0.131	**0.286**	0.212	0.174	0.186	0.282	0.173	0.196
Bands_42	0.197	0.237	0.200	0.247	0.118	0.172	0.118	0.161	**0.275**	0.234	0.223	0.233	0.224	0.250	0.258
Credit	0.145	0.157	**0.956**	0.339	0.195	0.696	0.229	0.866	0.409	0.444	0.786	0.776	0.845	0.846	0.882
German	0.165	0.197	0.435	0.318	0.121	0.209	0.418	0.287	0.300	0.062	**0.502**	0.264	0.342	0.100	0.406
Heart	0.379	0.600	0.895	0.849	0.486	0.675	0.841	0.811	0.813	0.268	0.863	0.863	0.811	0.920	**0.935**
Horse	0.149	0.171	**0.708**	0.282	0.111	0.154	0.205	0.371	0.340	0.118	0.666	0.418	0.552	0.198	0.590
Sick	0.046	0.039	0.040	0.053	0.072	0.067	0.035	**0.206**	0.098	0.063	0.057	0.031	0.033	0.080	0.033
Average	0.325	0.338	0.590	0.360	0.247	0.447	0.450	0.463	0.475	0.299	0.465	0.560	0.535	0.571	**0.633**

6.4. Ablation study

To further investigate the impact of multi-granularity granular-balls and different kernel functions on KFGOD's performance, we compare the AUC results of single-granularity-based KFGOD and KFGOD with its Gaussian kernel replaced by Polynomial and Sigmoid kernels. As shown in Table 8, **KFGOD** uses multi-granularity granular-balls as the processing unit and employs a Gaussian kernel to measure the fuzzy similarity between granular-balls. **Single-granularity** refers to modifying KFGOD to process samples at a single granularity while still using the Gaussian kernel. **Polynomial kernel** and **Sigmoid kernel** indicate replacing the Gaussian kernel in KFGOD with Polynomial and Sigmoid kernels, respectively.

Table 8 reveals that multi-granularity granular-balls and the Gaussian kernel contribute to improving KFGOD's detection performance. KFGOD adaptively groups data into multiple granular-balls, with each granular-ball forming local structures of varying sizes in the data, effectively handling multi-granularity information. The Gaussian kernel in KFGOD, with its kernel values decaying rapidly with distance, significantly affects only locally adjacent granular-ball pairs. This results in a more reasonable measure of fuzzy similarity between granular-balls, enabling better capture of abnormal deviations between granular-balls and thus enhancing outlier detection performance. Although KFGOD with the Gaussian kernel achieves promising results on these datasets, kernel selection is context-dependent. Datasets with explicit polynomial relationships or global structures might benefit from polynomial kernels, while sigmoid kernels could be explored in scenarios resembling neural network applications. We recommend analyzing dataset

Table 6: G-mean results comparison of 15 outlier detection methods across 20 datasets. The best g-mean result in each row is **bolded**, while the second highest performance is underlined.

Datasets	LDOF	LoOP	ODGrCR	NC	SRO	DeepSVDD	VarE	ROD	DCROD	DIF	ILGNI	KFRAD	ADkNRS	MGBOD	KFGOD
Cardio	0.670	0.621	0.734	0.640	0.512	0.850	0.508	0.878	0.778	0.857	0.836	0.863	0.880	0.880	**0.890**
Cardiot	0.720	0.787	0.564	0.777	0.558	0.527	0.745	**0.846**	0.812	0.730	0.771	0.826	0.841	0.810	0.840
Diab	0.828	0.819	0.930	0.879	0.694	0.908	0.920	0.914	0.911	0.807	0.877	0.911	**0.965**	0.932	0.937
Ecoli	0.879	0.882	0.879	0.864	**0.909**	0.878	0.844	0.607	0.899	0.892	0.658	0.878	0.605	0.879	0.878
Iono	0.982	0.991	0.936	0.984	**1.000**	0.957	0.486	0.948	**1.000**	0.998	0.993	0.980	0.998	**1.000**	0.996
Iris	0.980	0.938	**1.000**	0.990	0.654	0.995	**1.000**	0.819	0.975	0.794	0.959	0.995	**1.000**	**1.000**	**1.000**
Pima	0.786	0.794	0.897	0.843	0.645	0.778	0.867	0.902	0.898	0.799	0.863	0.891	**0.927**	0.920	0.926
Sonar	0.925	0.969	0.990	**0.995**	0.609	0.902	0.898	0.979	0.979	0.879	0.939	0.979	0.990	**0.995**	0.990
Wbc	0.788	0.663	0.985	0.777	0.846	0.982	0.989	0.957	0.974	0.926	0.973	0.990	0.984	0.990	**0.992**
Wdbc	0.671	0.792	0.996	0.837	0.548	0.953	**0.997**	0.986	0.959	0.860	0.906	0.979	**0.997**	0.993	**0.997**
Yeast	0.993	0.994	0.999	0.980	0.996	0.997	**1.000**	0.936	0.993	0.992	0.978	0.999	0.985	**1.000**	**1.000**
Ann	0.604	0.658	0.751	0.693	0.684	0.770	0.613	0.675	0.756	0.786	0.651	0.768	**0.851**	0.802	0.775
Arr	0.704	0.711	0.742	0.701	0.715	0.699	0.700	0.737	0.745	0.742	0.756	0.727	**0.779**	0.763	0.765
Bands_27	0.711	0.688	0.717	0.716	0.471	0.663	0.524	0.714	0.722	0.736	0.751	**0.784**	0.744	0.730	0.780
Bands_42	0.646	0.667	0.656	0.703	0.461	0.631	0.554	0.678	0.707	0.711	0.690	**0.754**	0.670	0.713	0.749
Credit	0.627	0.603	**0.968**	0.781	0.668	0.921	0.602	0.934	0.783	0.876	0.917	0.908	0.953	0.934	0.942
German	0.870	0.892	**0.970**	0.937	0.797	0.841	0.887	0.866	0.946	0.693	0.967	0.937	0.919	0.877	0.941
Heart	0.841	0.876	0.955	0.939	0.862	0.868	0.894	0.942	0.935	0.754	0.959	0.919	**0.966**	0.955	0.962
Horse	0.784	0.835	0.965	0.822	0.656	0.789	0.826	0.815	0.837	0.651	**0.969**	0.819	0.930	0.854	0.962
Sick	0.737	0.776	0.871	0.788	0.801	0.768	0.849	0.865	0.847	0.849	0.879	0.850	**0.912**	0.835	0.874
Average	0.787	0.798	0.875	0.832	0.704	0.834	0.785	0.850	0.873	0.817	0.865	0.888	0.895	0.893	**0.910**

Table 7: AUC results comparison of 5 outlier detection methods across different outlier ratios on the *Cardiot* and *Yeast* datasets. The best AUC result in each row is **bolded**, while the second highest performance is underlined.

Datasets (% Outlier ratios)	DCROD	DIF	KFRAD	MGBOD	KFGOD
Cardiot (2%)	0.834	0.796	**0.880**	0.846	0.869
Cardiot (4%)	0.823	0.792	0.857	0.893	**0.899**
Cardiot (6%)	0.745	0.726	0.863	0.895	**0.899**
Cardiot (8%)	0.679	0.648	0.841	**0.875**	**0.875**
Cardiot (10%)	0.601	0.626	0.862	0.862	**0.868**
Yeast (2%)	0.985	0.970	0.999	0.999	**1.000**
Yeast (4%)	0.968	0.889	0.998	0.991	**0.999**
Yeast (6%)	0.834	0.752	0.997	0.975	**0.999**
Yeast (8%)	0.470	0.748	0.997	0.970	**0.999**
Yeast (10%)	0.525	0.645	0.997	0.963	**0.998**
Average	0.746	0.759	0.929	0.927	**0.941**

characteristics, such as non-linearity or feature interactions, to guide kernel selection and exploring hybrid kernel approaches to further enhance KFGOD's adaptability to diverse scenarios.

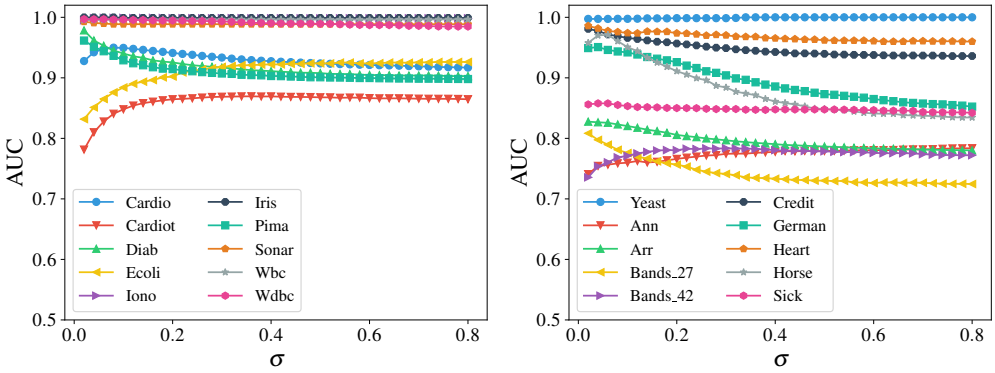
6.5. Hyperparameter sensitivity analysis

In KFGOD, it is necessary to determine the hyperparameter σ for computing the multi-granularity kernelized fuzzy relation. In this section, we investigate the impact of σ on the outlier detection performance of KFGOD, analyzing KFGOD's sensitivity to this hyperparameter. We plot the variation curves of KFGOD's AUC and g-mean results with different values of the hyperparameter σ across 20 datasets in Figure 3 and Figure 4.

As observed from Figure 3, the AUC variation curves for most datasets are relatively smooth and exhibit no abrupt changes, indicating the overall robustness of KFGOD with respect to the kernel parameter σ. For certain datasets, such as *Cardio*, *Cardiot*, *Ecoli*, and *Band_42*, the AUC tends to increase as σ becomes larger, and subsequently stabilizes. This phenomenon occurs primarily because, when the hyperparameter σ is small, only closely located granular-ball pairs exhibit relatively high fuzzy similarity. This causes KFGOD to focus solely on a small subset of neighboring granular-balls during uncertainty measurement, leading to suboptimal performance. As σ increases, the range of fuzzy similarity extends to more distant granular-ball pairs, enabling KFGOD to incorporate a broader range of neighborhood information in its uncertainty measurement. This expansion allows KFGOD to more comprehensively capture both local and global data structures, thereby enhancing its ability to detect outliers. When σ exceeds a certain threshold, the neighborhood information considered by KFGOD during uncertainty measurement becomes relatively stable, resulting in a plateau in AUC performance. In contrast, for other datasets including *Diab*, *Pima*, *Horse*, *Arr*, *Credit*, *German*, and *Heart*, the AUC decreases with increasing σ, before reaching a stable plateau. These patterns suggest that KFGOD becomes less sensitive to the kernel parameter once σ exceeds a certain threshold.

Table 8: AUC results comparison for KFGOD using Gaussian, Polynomial, and Sigmoid kernels and single-granularity-based KFGOD. The best AUC result in each row is **bolded**, while the second highest performance is underlined.

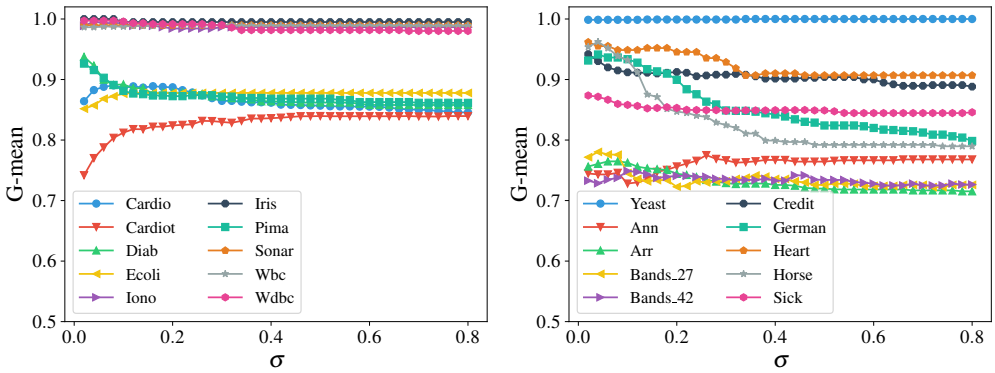
Datasets	KFGOD (Gaussian kernel)	Single-granularity	Polynomial kernel	Sigmoid kernel
Cardio	**0.950**	0.924	0.434	0.293
Cardiot	0.869	**0.879**	0.795	0.639
Diab	**0.978**	0.927	0.674	0.923
Ecoli	**0.926**	0.919	0.675	0.646
Iono	**0.999**	0.998	0.907	0.169
Iris	**1.000**	0.999	0.848	0.997
Pima	**0.962**	0.913	0.647	0.843
Sonar	**0.993**	0.981	0.922	0.922
Wbc	**0.997**	**0.997**	0.995	0.995
Wdbc	0.997	0.990	0.875	**1.000**
Yeast	**1.000**	**1.000**	0.981	0.994
Ann	0.784	**0.793**	0.609	0.665
Arr	**0.828**	0.781	0.595	0.815
Bands_27	**0.808**	0.801	0.562	0.307
Bands_42	**0.784**	0.777	0.497	0.359
Credit	**0.981**	0.940	0.667	0.291
German	**0.951**	0.930	0.860	0.792
Heart	**0.986**	0.965	0.682	0.895
Horse	**0.971**	0.869	0.641	0.831
Sick	**0.858**	0.850	0.672	0.807
Average	**0.931**	0.912	0.727	0.709



(a) Datasets: from Card to Wdbc.

(b) Datasets: from Yeast to Horse.

Figure 3: The AUC result curves for KFGOD under different hyperparameter σ settings.



(a) Datasets: from Card to Wdbc.

(b) Datasets: from Yeast to Horse.

Figure 4: The g-mean result curves for KFGOD under different hyperparameter σ settings.

In addition, we plot the g-mean result curves for KFGOD under different hyperparameter σ settings in Figure 4. It can be observed that the trend of the g-mean curves with respect to σ is consistent with that of the AUC curves, aligning with the conclusions of the above analysis. Overall, KFGOD exhibits low sensitivity to σ, making it relatively easy to tune in practice. To facilitate consistent application and experimental comparison, we recommend setting $\sigma = 0.2$, which yields favorable performance across most datasets.

The kernel parameter σ is varied within the range $[0.02, 0.8]$ to balance local discrimination and global generalization across diverse datasets. This range is selected based on empirical observations and the normalization of attributes to the interval $[0, 1]$. Extremely small values tend to produce sparse and unstable fuzzy relations, whereas excessively large values lead to oversmoothing, thereby diminishing the model's capacity to distinguish outliers. Within the chosen range, KFGOD maintains stable and competitive performance, as demonstrated in the empirical results. Furthermore, the variation in sensitivity across datasets can be attributed to several underlying factors, such as data dimensionality, attribute distribution, and intrinsic structural complexity. In high-dimensional datasets, larger values of σ generally produce more stable kernelized fuzzy relations. In contrast, for low-dimensional or well-clustered datasets, smaller σ values are often more effective in preserving local structure. Therefore, it is advisable to determine σ based on pairwise distance statistics or conduct empirical sensitivity analysis on a validation subset. This adaptive strategy is especially important in unsupervised scenarios, where labeled data are unavailable.

6.6. Statistical analysis

The Friedman test and the Nemenyi post-hoc test are two commonly used methods for statistical analysis between outlier detection methods. The Friedman test is employed to determine whether there are statistically significant differences among the methods in an experiment, though it cannot identify significant differences between any specific pair of methods. Once statistically significant differences are confirmed among the methods, the Nemenyi post-hoc test is applied to further determine the significant differences between specific pairs of methods.

In the Friedman test, the AUC results of each method across all datasets are ranked from lowest to highest, and corresponding ordinal values are assigned. Based on these ordinal values, the Friedman test determines whether there are statistically significant differences among the methods. Let M denote the number of methods and N denote the number of datasets, the Friedman test statistic is calculated as

$$\tau_F = \frac{(N-1)\tau_{\chi^2}}{N(M-1)-\tau_{\chi^2}}, \tag{17}$$

$$\tau_{\chi^2} = \frac{12N}{M(M+1)}\sum_{i=1}^{M} V_i^2 - \frac{1}{4}M(M+1)^2, \tag{18}$$

where V_i is the average ordinal value of the ith method, and τ_F follows an F distribution with $M-1$ and $(M-1)(N-1)$ degrees of freedom. In the Nemenyi post-hoc test, a statistic called the Critical Difference (CD) is calculated as

$$CD_\alpha = q_\alpha \sqrt{\frac{1}{6N}M(M+1)}, \tag{19}$$

where q_α is the critical value of Tukey's distribution. The Nemenyi test figure is typically used to visually illustrate the statistically significant differences between any two methods. Specifically, in this figure, the average ordinal values of the methods are marked as points on a number line, and a horizontal line segment representing the CD value extends symmetrically around each point. If multiple methods' points are covered by the same CD line segment, it indicates that there are no statistically significant differences among those methods.

As shown in Table 2, there are 15 methods and 20 datasets. Thus, the degrees of freedom for the τ_F distribution are calculated as 14 and 266. In the Friedman test, we set $\alpha = 0.1$, and the value of τ_F is 13.2378, which exceeds the critical value of 1.5303. This indicates that there are significant differences among the methods, necessitating the use of the Nemenyi post-hoc test for further differentiation. When $\alpha = 0.1$, the critical distance is calculated as $CD_{0.1} = 4.4675$. We plot the Nemenyi test figure in Figure 5. From the figure, it can be observed that KFGOD shares line segment overlap only with MGBOD, ADkNRS, KFRAD, ODGrCR, and DCROD, indicating that there are no statistically significant differences between KFGOD and these five methods. However, KFGOD exhibits significant differences with the remaining nine methods, suggesting that KFGOD statistically outperforms methods such as ILGNI, ROD, VarE, NC, and others.

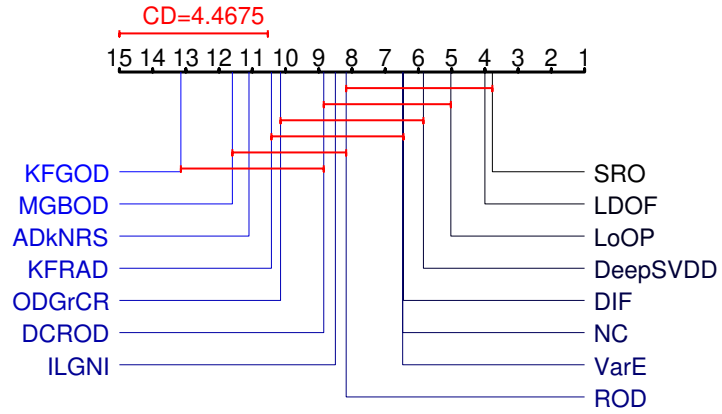


Figure 5: Nemenyi test figure on AUC results of 15 methods and 20 datasets.

7. Conclusions

In this study, we introduce multi-granularity granular-balls into the classical kernelized fuzzy rough sets, proposing a new FRS model called MKFRS. Based on MKFRS, we present an unsupervised outlier detection method named KFGOD. This method addresses the limitations of existing FRS-based outlier detection methods, which fail to effectively leverage multi-granularity information in the data and cannot adequately model the complex membership relations between samples. In KFGOD, we select appropriate kernel functions to construct multi-granularity kernelized fuzzy relation matrices under different attribute subsets, which are then used to build a multi-granularity kernelized fuzzy approximation space. Finally, by fusing abnormal information from multiple multi-granularity kernelized fuzzy information granules, we reflect the outlier degrees of the samples. We validate the effectiveness of KFGOD by comparing it with 14 baseline methods on 20 publicly available datasets.

Although KFGOD achieves promising results across multiple datasets, several limitations remain. First, the computational cost increases with the number of attributes and granular-balls, potentially affecting scalability. Second, the kernel function's performance relies on manually tuned parameters, which poses challenges in unsupervised scenarios. Third, a key limitation is that the model currently only incorporates information fusion at the single-attribute level, exploring multi-attribute fusion is an important direction for future research. Lastly, the interpretability of the fused outlier scores warrants further improvement. Addressing these issues will be a primary focus of our future work.

References

[1] C. Gao, X. Tan, J. Zhou, W. Ding, W. Pedrycz, Fuzzy granule density-based outlier detection with multi-scale granular balls, IEEE Transactions on Knowledge and Data Engineering 37 (3) (2025) 1182–1197.
[2] Z. Yuan, P. Hu, H. Chen, Y. Chen, Q. Li, Dfno: Detecting fuzzy neighborhood outliers, IEEE Transactions on Knowledge and Data Engineering 37 (1) (2025) 200–209.
[3] A. Smiti, A critical overview of outlier detection methods, Computer Science Review 38 (2020) 100306.
[4] X. Su, S. Cheng, D. Peng, H. Chen, Z. Yuan, Fusing multi-granular-ball fuzzy information to detect outliers, Applied Soft Computing (2025) 113045.
[5] B. Chen, Y. Li, D. Peng, H. Chen, Z. Yuan, Fusing multi-scale fuzzy information to detect outliers, Information Fusion 103 (2024) 102133.
[6] R. Liu, J. Zhang, H. Li, Hierarchical multi-source cues fusion for mono-to-binaural based audio deepfake detection, Information Fusion (2025) 103097.
[7] H. Liu, S. Zhang, Z. Wu, X. Li, Outlier detection using local density and global structure, Pattern Recognition 157 (2025) 110947.
[8] M. A. Samara, I. Bennis, A. Abouaissa, P. Lorenz, A survey of outlier detection techniques in iot: Review and classification, Journal of Sensor and Actuator Networks 11 (1) (2022) 4.
[9] X. Su, Z. Yuan, B. Chen, D. Peng, H. Chen, Y. Chen, Detecting anomalies with granular-ball fuzzy rough sets, Information Sciences 678 (2024) 121016.
[10] D. Dubois, H. Prade, Rough fuzzy sets and fuzzy rough sets, International Journal of General System 17 (2-3) (1990) 191–209.
[11] W. Ding, M. Abdel-Basset, H. Hawash, W. Pedrycz, Multimodal infant brain segmentation by fuzzy-informed deep learning, IEEE Transactions on Fuzzy Systems 30 (4) (2021) 1088–1101.
[12] W. Ding, S. Geng, H. Wang, J. Huang, T. Zhou, Fdiff-fusion: Denoising diffusion fusion network based on fuzzy learning for 3d medical image segmentation, Information Fusion 112 (2024) 102540.
[13] W. Ding, H. Wang, J. Huang, H. Ju, Y. Geng, C.-T. Lin, W. Pedrycz, Ftranscnn: Fusing transformer and a cnn based on fuzzy logic for uncertain medical image segmentation, Information Fusion 99 (2023) 101880.
[14] Z. Yuan, H. Chen, T. Li, Z. Yu, B. Sang, C. Luo, Unsupervised attribute reduction for mixed data based on fuzzy rough sets, Information Sciences 572 (2021) 67–87.
[15] W. Ding, C.-T. Lin, Z. Cao, Deep neuro-cognitive co-evolution for fuzzy attribute reduction by quantum leaping pso with nearest-neighbor memeplexes, IEEE transactions on cybernetics 49 (7) (2018) 2744–2757.
[16] R. Jensen, C. Cornelis, Fuzzy-rough nearest neighbour classification and prediction, Theoretical Computer Science 412 (42) (2011) 5871–5884.
[17] B. Yu, Z. Zheng, M. Cai, W. Pedrycz, W. Ding, Frcm: A fuzzy rough c-means clustering method, Fuzzy Sets and Systems 480 (2024) 108860.
[18] C. Wang, C. Wang, S. An, W. Ding, Y. Qian, Feature selection and classification based on directed fuzzy rough sets, IEEE Transactions on Systems, Man, and Cybernetics: Systems 55 (1) (2025) 699–711.

[19] Z. Yuan, B. Chen, J. Liu, H. Chen, D. Peng, P. Li, Anomaly detection based on weighted fuzzy-rough density, Applied Soft Computing 134 (2023) 109995.
[20] S. Wang, Z. Yuan, C. Luo, H. Chen, D. Peng, Exploiting fuzzy rough entropy to detect anomalies, International Journal of Approximate Reasoning 165 (2024) 109087.
[21] B. Chen, Z. Yuan, Z. Liu, D. Peng, Y. Li, C. Liu, G. Duan, Outlier detection in mixed-attribute data: a semi-supervised approach with fuzzy approximations and relative entropy, International Journal of Approximate Reasoning (2025) 109373.
[22] S. Bergman, M. Schiffer, Kernel functions and conformal mapping, Compositio Mathematica 8 (1951) 205–249.
[23] Q. Hu, D. Yu, W. Pedrycz, D. Chen, Kernelized fuzzy rough sets and their applications, IEEE Transactions on Knowledge and Data Engineering 23 (11) (2010) 1649–1667.
[24] Y. Wu, S. Wang, H. Chen, D. Peng, Z. Yuan, Kernelized fuzzy-rough anomaly detection, IEEE Transactions on Fuzzy Systems (2024) 1–12.
[25] S. Xia, Y. Liu, X. Ding, G. Wang, H. Yu, Y. Luo, Granular ball computing classifiers for efficient, scalable and robust learning, Information Sciences 483 (2019) 136–152.
[26] S. Xia, X. Dai, G. Wang, X. Gao, E. Giem, An efficient and adaptive granular-ball generation method in classification problem, IEEE Transactions on Neural Networks and Learning Systems 35 (4) (2022) 5319–5331.
[27] J. Yang, Z. Liu, S. Xia, G. Wang, Q. Zhang, S. Li, T. Xu, 3wc-gbnrs++: A novel three-way classifier with granular-ball neighborhood rough sets based on uncertainty, IEEE Transactions on Fuzzy Systems 32 (8) (2024) 4376–4387.
[28] S. Xia, B. Shi, Y. Wang, J. Xie, G. Wang, X. Gao, Gbct: Efficient and adaptive clustering via granular-ball computing for complex data, IEEE Transactions on Neural Networks and Learning Systems (2024) 1–14.
[29] W. Qian, J. Li, X. Cai, J. Huang, W. Ding, Granular ball-based partial label feature selection via fuzzy correlation and redundancy, Information Sciences (2025) 122047.
[30] S. Cheng, X. Su, B. Chen, H. Chen, D. Peng, Z. Yuan, Gbmod: A granular-ball mean-shift outlier detector, Pattern Recognition 159 (2025) 111115.
[31] X. Su, X. Wang, D. Peng, H. Chen, Y. Chen, Z. Yuan, Granular-ball computing guided anomaly detection for hybrid attribute data, International Journal of Machine Learning and Cybernetics (2024) 1–16.
[32] Q. Hu, D. Yu, J. Liu, C. Wu, Neighborhood rough set based heterogeneous feature subset selection, Information sciences 178 (18) (2008) 3577–3594.
[33] L. Yang, K. Qin, B. Sang, W. Xu, Dynamic fuzzy neighborhood rough set approach for interval-valued information systems with fuzzy decision, Applied Soft Computing 111 (2021) 107679.
[34] 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 Transactions on Systems, Man, and Cybernetics: Systems 52 (8) (2022) 5175–5187.
[35] X. Zhang, Z. Yuan, D. Miao, Outlier detection using three-way neighborhood characteristic regions and corresponding fusion measurement, IEEE Transactions on Knowledge and Data Engineering 36 (5) (2023) 2082–2095.
[36] Z. Yuan, H. Chen, T. Li, B. Sang, S. Wang, Outlier detection based on fuzzy rough granules in mixed attribute data, IEEE Transactions on Cybernetics 52 (8) (2021) 8399–8412.
[37] J. Xie, W. Kong, S. Xia, G. Wang, X. Gao, An efficient spectral clustering algorithm based on granular-ball, IEEE Transactions on Knowledge and Data Engineering 35 (9) (2023) 9743–9753.
[38] S. Xia, C. Wang, G. Wang, X. Gao, W. Ding, J. Yu, Y. Zhai, Z. Chen, Gbrs: A unified granular-ball learning model of pawlak rough set and neighborhood rough set, IEEE Transactions on Neural Networks and Learning Systems (2023) 1–15.
[39] 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 Transactions on Neural Networks and Learning Systems (2023) 1–14.
[40] S. Xia, X. Lian, G. Wang, X. Gao, Q. Hu, Y. Shao, Granular-ball fuzzy set and its implement in svm, IEEE Transactions on Knowledge and Data Engineering 36 (11) (2024) 6293–6304.
[41] Z. Jia, Z. Zhang, W. Pedrycz, Generation of granular-balls for clustering based on the principle of justifiable granularity, IEEE Transactions on Cybernetics 55 (4) (2025) 1687–1700.
[42] K. Zhang, M. Hutter, H. Jin, A new local distance-based outlier detection approach for scattered real-world data, in: Advances in Knowledge Discovery and Data Mining: 13th Pacific-Asia Conference, PAKDD 2009 Bangkok, Thailand, April 27-30, 2009 Proceedings 13, Springer, 2009, pp. 813–822.
[43] H.-P. Kriegel, P. Kröger, E. Schubert, A. Zimek, Loop: local outlier probabilities, in: Proceedings of the 18th ACM conference on Information and knowledge management, 2009, pp. 1649–1652.
[44] F. Jiang, Y.-M. Chen, Outlier detection based on granular computing and rough set theory, Applied intelligence 42 (2015) 303–322.
[45] X. Li, J. Lv, Z. Yi, An efficient representation-based method for boundary point and outlier detection, IEEE transactions on neural networks and learning systems 29 (1) (2016) 51–62.
[46] C. You, D. P. Robinson, R. Vidal, Provable self-representation based outlier detection in a union of subspaces, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 3395–3404.
[47] 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.
[48] X. Li, J. Lv, Z. Yi, Outlier detection using structural scores in a high-dimensional space, IEEE transactions on cybernetics 50 (5) (2018) 2302–2310.
[49] Y. Almardeny, N. Boujnah, F. Cleary, A novel outlier detection method for multivariate data, IEEE Transactions on Knowledge and Data Engineering 34 (9) (2020) 4052–4062.
[50] 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 Systems with Applications 207 (2022) 117988.
[51] H. Xu, G. Pang, Y. Wang, Y. Wang, Deep isolation forest for anomaly detection, IEEE Transactions on Knowledge and Data Engineering 35 (12) (2023) 12591–12604.
[52] R. Li, H. Chen, S. Liu, X. Li, Y. Li, B. Wang, Incomplete mixed data-driven outlier detection based on local–global neighborhood information, Information Sciences 633 (2023) 204–225.
[53] X. Chen, Z. Yuan, S. Feng, Anomaly detection based on improved k-nearest neighbor rough sets, International Journal of Approximate Reasoning 176 (2025) 109323.
[54] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, Smote: synthetic minority over-sampling technique, Journal of artificial intelligence research 16 (2002) 321–357.