International Journal of Machine Learning and Cybernetics
https://doi.org/10.1007/s13042-024-02425-8

ORIGINAL ARTICLE

Granular-ball computing guided anomaly detection for hybrid attribute data

Xinyu Su[1] · Xiwen Wang[1] · Dezhong Peng[1,4] · Hongmei Chen[2] · Yingke Chen[3] · Zhong Yuan[1]

Received: 2 June 2024 / Accepted: 9 October 2024

Abstract
Anomaly detection is one of the important research areas in data mining or data analytics. However, most of the existing anomaly detection methods only consider homogeneous data, such as nominal or numerical attribute data, and fail to effectively deal with hybrid attribute data. Moreover, these methods also suffer from inefficiency and noise sensitivity due to their single-granularity sample-based input paradigm. In this study, we propose an unsupervised anomaly detection method based on the granular-ball fuzzy set called HGBAD. First, we define a novel granular-ball fuzzy set to deal with the uncertainty information in hybrid attribute data. Based on the novel fuzzy set, multiple granular-ball fuzzy information granules are constructed. The anomaly degrees of granular-ball fuzzy information granules are fused to calculate the anomaly factors. The anomaly factors are used to measure the anomaly degrees of samples. Based on the anomaly factors, anomalies can be detected by an anomaly determination threshold. Experimental results demonstrate the superior performance of HGBAD in detecting anomalies across various data types. The code is publicly available at https://github.com/Mxeron/HGBAD.

Keywords Granular computing · Granular-ball computing · Hybrid attribute data · Anomaly detection · Outlier detection

1 Introduction

How to mine potential anomaly objects in data is one of the important tasks in data mining. Anomalies (or outliers) are a small portion of objects whose certain characteristics or manifestations differ from the expected objects. With the growth of various industries, anomaly detection has gone beyond its traditional role of augmenting data and can now be applied across a wide range of industries, such as credit card fraud detection [1], medical diagnosis [2], and process monitoring [3].

Anomaly detection, by virtue of its importance, has attracted many scholars to carry out research, and many methods based on different theories have been proposed. According to different technical guides, existing anomaly detection methods can be broadly categorized into statistical-based, depth-based, distance-based, clustering-based, and density-based methods [4, 5].

Statistical-based methods rely on various statistical analysis principles or metrics, such as mean, median, and variance, to analyze data and identify anomalies based on the observed differences in results [6, 7]. These methods assume that the data satisfy some distribution, which is a natural

✉ Zhong Yuan
yuanzhong@scu.edu.cn

Xinyu Su
suxinyu@stu.scu.edu.cn

Xiwen Wang
wangxiwen124@stu.scu.edu.cn

Dezhong Peng
pengdz@scu.edu.cn

Hongmei Chen
hmchen@swjtu.edu.cn

Yingke Chen
yingke.chen@northumbria.ac.uk

1 College of Computer Science, Sichuan University, Chengdu 610065, China

2 School of Computing and Artificial Intelligence, Southwest Jiaotong University, Chengdu 611756, China

3 Department of Computer and Information Sciences, Northumbria University, Newcastle upon Tyne NE1 8ST, UK

4 Sichuan National Innovation New Vision UHD Video Technology Co., Ltd, 610095 Chengdu, China

Published online: 03 November 2024

assumption for continuous numerical attributes. However, when the dataset contains nominal attributes, this assumption no longer applies [7, 8]. Moreover, in high-dimensional data, the sparseness of the data may cause the statistical analysis to become inaccurate [9]. In depth-based methods, depth refers to the central tendency or positional depth of the samples in the dataset. The anomalies are usually those samples with low data depth [10]. However, these methods are not efficient enough in high-dimensional data with hybrid attributes [7, 10].

Considering the problems of the above methods, to further improve the performance of anomaly detection, distance-based methods have been proposed [11, 12]. The core idea of these methods is that a sample is considered an anomaly if its distance from most other samples exceeds a certain threshold. This type of method is easy to understand and is widely used. However, distance-based methods are still unable to deal with the sparse problem in high-dimensional data, and these methods demonstrate a high sensitivity to the selection of threshold values [8]. Additionally, clustering-based methods cannot effectively handle high-dimensional or density-varying data [9, 12].

Past methods treat anomaly detection as a binary classification task, this treatment does not take into account the anomaly degree, which should be different for complex anomalies in real scenarios. For this reason, Breuning et al. [13] first proposed a density-based anomaly detection method, whose main idea is to define the anomaly factor (or outlier factor) of a sample by comparing the density of the surroundings of the sample with the density of the surroundings of the neighbors of the sample. However, the density-based methods are sensitive to hyper-parameters [5]. Most anomaly detection methods are implemented based on Euclidean distance, which leads to the fact that these methods fail to handle nominal or hybrid attribute data effectively [14–16]. In addition, these methods almost always use a sample-based input paradigm and can only process samples one by one. This processing is often inefficient and susceptible to noisy data in the dataset. If this noisy data can be circumvented in the early data preprocessing stage, it will bring great benefits to the subsequent processing and can further improve the performance of anomaly detection.

Granular computing (GrC), a pivotal instrument in the domain of knowledge mining, addresses complex problems by emulating the inherent patterns of human cognition [17, 18], and it provides novel theoretical ideas for studying many problems in data mining. Fuzzy set theory (FST) is an essential tool in GrC for dealing with potential uncertainty and fuzziness in data, which has been applied to decision-making [19], pattern recognition [20], fuzzy clustering [21], rule induction [22], and anomaly detection [14]. FST utilizes fuzzy binary relations to characterize similarity between samples. By introducing different fuzzy relations, FST can effectively process nominal, numerical, and hybrid attribute data directly, preserving as much information as possible in the data. Integrating FST into anomaly detection enhances the ability of methods to manage uncertainty information in hybrid attribute data and to extract effective anomaly features. FST offers several advantages in the anomaly detection field, such as improving accuracy, adapting to various data types, and enhancing method interpretability [14, 23].

The input processing of traditional machine learning methods is built on single-granularity samples, which greatly reduces their efficiency and accuracy [24, 25]. To improve the performance of traditional machine learning methods, granular-ball computing (GBC) [26] is proposed as an emerging computing paradigm, which is further developed based on GrC by introducing the concept of "ball". GBC makes the processing unit further abstracted from general granules into a "ball"-shaped structure with richer geometrical and algebraic properties [24]. It attempts to provide a more intuitive and structured way to deal with complex data and processes. It innovatively uses granular-balls as method input, where granular-balls are generated in the raw data, and multiple samples are contained in one granular-ball. In the subsequent processing, these balls with different granularity are used as the smallest processing unit instead of the single-granularity samples, and thus, the efficiency and anti-noise ability of the method can be further improved. The GBC-based method visually depicts the data information with a granular-ball structure, improving the intuition and efficiency of data analysis and understanding. Since granular-balls are adaptively generated based on different data, granular-balls can be well adapted to a variety of complexly distributed data. Furthermore, its hierarchical processing capability allows problems to be analyzed and solved at different levels of abstraction, increasing the flexibility and depth of dealing with complex problems.

However, existing granular-ball generation methods are targeted at numerical attribute data and fail to handle hybrid attribute data effectively. Introducing GBC into anomaly detection and replacing samples with multi-granularity granular-balls can enhance the ability to express data, improve computational efficiency, and have better noise resistance [24, 25]. Multiple advantages of GBC have not been fully utilized in the field of anomaly detection, and the application of GBC in anomaly detection has to be further expanded [23, 27].

Based on the above discussion, to further improve the performance of unsupervised anomaly detection, we utilize the advantages of FST and GBC to construct an unsupervised anomaly detection method based on hybrid granular-ball fuzzy information granules called HGBAD. Figure 1 shows the general core idea of HGBAD. Specifically, we introduce a hybrid distance metric to extend the original granular-ball generation method to hybrid attribute data. The granular-balls

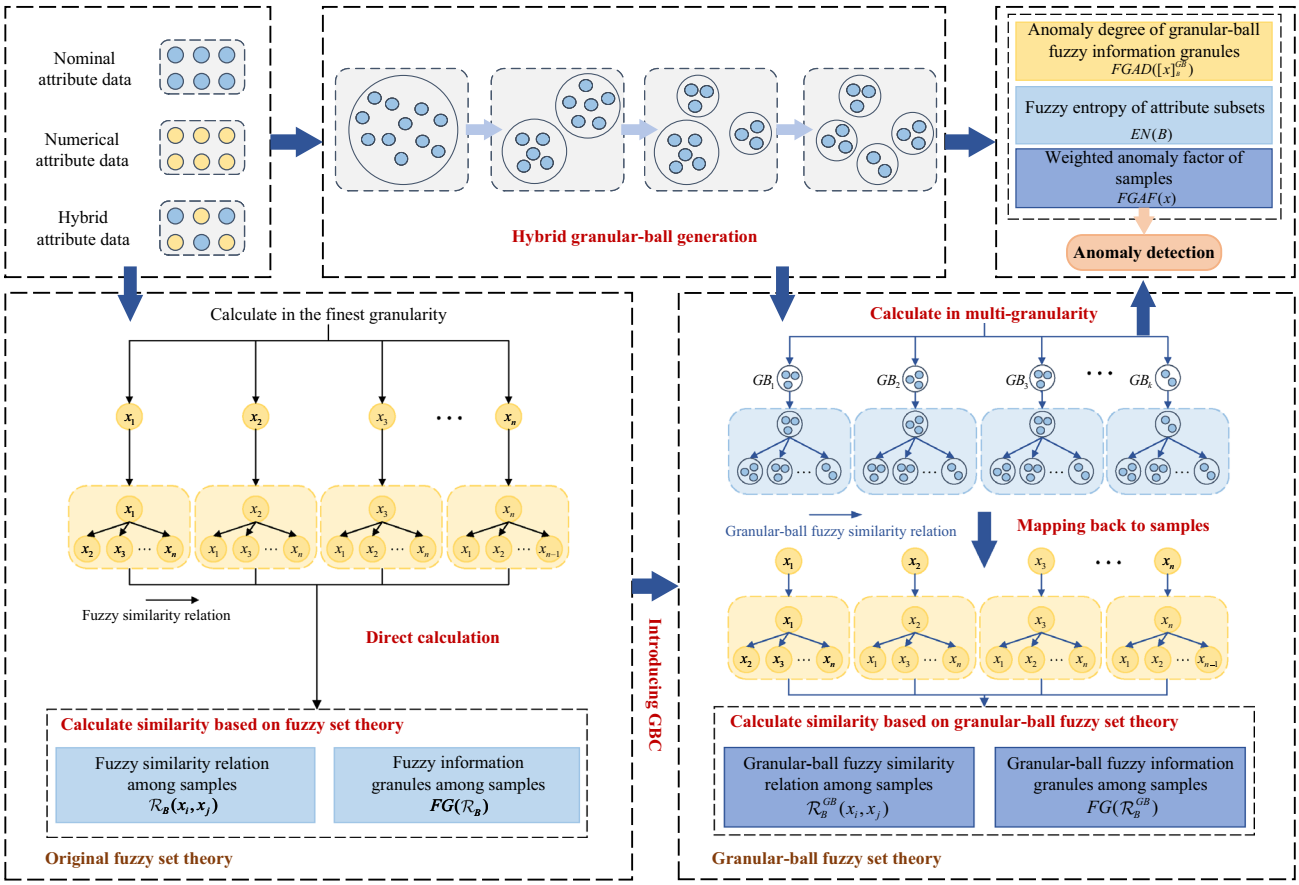


Fig. 1 The framework diagram of HGBAD

are generated on the input dataset by the hybrid granular-ball generation method. Subsequently, we calculate the fuzzy relations between granular-balls and then map these fuzzy relations to the samples to obtain the granular-ball fuzzy relations between the samples. Finally, based on the fuzzy relations, we define the anomaly factors to measure the anomaly degree of the samples through the hybrid granular-ball fuzzy information granules to which they belonged. In addition, we introduce the granular-ball fuzzy entropy to set weights for different attributes. The lower left corner shows the original FST, which is directly calculated in the finest granularity. With the introduction of GBC, a multi-granularity representation and computing method is realized, which leads to a superior anomaly detection method.

Overall, the contribution of this study consists of the following aspects.

1. To apply granular-ball computing to hybrid attribute data, a novel granular-ball generation method is proposed. Then, a novel granular-ball fuzzy set is proposed based on hybrid granular-balls.
2. To improve outlier detection performance in hybrid attribute data, a novel unsupervised anomaly detection method based on the granular-ball fuzzy set called HGBAD is proposed.
3. To quantify the anomaly degrees of each sample, the fuzzy granular-ball information granules-based anomaly factor is proposed.

The remaining sections of this study are organized as follows. Section 2 reviews related works on fuzzy set theory-based anomaly detection methods and granular-ball computing. Section 3 reviews some knowledge about the fuzzy set theory and granular-ball computing. Section 4 proposes the novel granular-ball fuzzy set in hybrid attribute data. Section 5 details the proposed method and gives the corresponding pseudo-code. Section 6 demonstrates the experimental results of our method. Section 7 summarizes this study. To facilitate reading and understanding of this study, the descriptions of different abbreviations are shown in Table 1.

Table 1 The descriptions of different abbreviations

Abbr.	Descriptions
HGBAD	Hybrid granular-ball fuzzy information granules-based anomaly detection
GrC	Granular computing
FST	Fuzzy set theory
GBC	Granular-ball computing
IS	Information system
FG	Family of fuzzy information granules
HD	Hybrid distance
FGAD	Anomaly degree of hybrid granular-ball fuzzy information granule
GBFE	Granular-ball fuzzy entropy
FGAF	Fuzzy granular-ball information granules-based anomaly factor
ROC	Receiver operating characteristic
AUC	Area under the curve
CD	Critical difference

2 Related works

In this section, we review some related works about granular-ball computing and fuzzy set theory-based anomaly detection methods.

2.1 Granular-ball computing

In early applications of GBC, a granular-ball is defined in the following form: $GB_i = \{x_j | j = 1, 2, \ldots, g\}$, where x_j denotes a sample in the granular-ball, g denotes the number of samples in the granular-ball and $|GB_i| = g$. The center and radius are two important properties of granular-ball. The center of GB_i is defined as $c = \frac{1}{|GB_i|} \sum_{x_j \in GB_i} x_j$. There are two ways to define the radius, the first way to define it is the average distance, i.e., $r = \frac{1}{|GB_i|} \sum_{x_j \in GB_i} \|x_j - c\|$. Another way to define it is by the maximum distance, i.e., $r = \max_{x_j \in GB_i} \|x_j - c\|$. The granular-balls generated with the average distance give a better fit to the data distribution, resulting in clearer decision boundaries than the maximum distance and less susceptible to noise [25]. Adopting the maximum distance as a radius provides better coverage of the raw data compared to the average distance. The choices of the radii vary according to different application scenarios. The choice and size of the radius affect the ability of balls to cover and represent the raw data. In classification tasks, the average distance is often chosen, and not all samples need to have a granular-ball to which they belong so that a clearer decision boundary can be generated [24]. Whereas in granular-ball clustering, the radius is defined by the maximum radius to achieve full coverage of all samples [25].

Granular-balls can be generated by different generation methods. The quality of the granular-balls greatly affects the efficiency and performance of downstream task execution. Therefore, the quality needs to be judged during the generation process to ensure that all the balls are of superior quality. In supervised tasks, such as classification, the purity of granular-balls is used to assess their quality. The purity is defined as the proportion of samples from the category with the highest percentage in the ball [24], which only applies to supervised tasks. Figure 2 shows the generation process of the granular-balls. To simulate the global priority in human brain cognition, the whole dataset is regarded as a big ball, which is the coarsest granularity and of the worst quality and does not meet the requirements. Subsequently, the k-means or k-division is used to divide the big ball [25]. Based on the definition of quality under different tasks, the quality of the newly generated balls is calculated and evaluated to see if the quality meets the set threshold. If the current ball already meets the requirements, the division is stopped, otherwise, the previous steps need to be continued. The whole generation process is an iterative updating process to achieve a better decomposition and representation of the raw data.

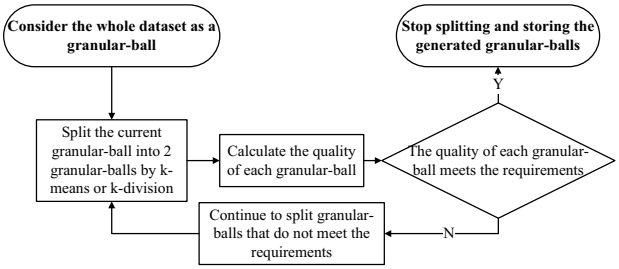
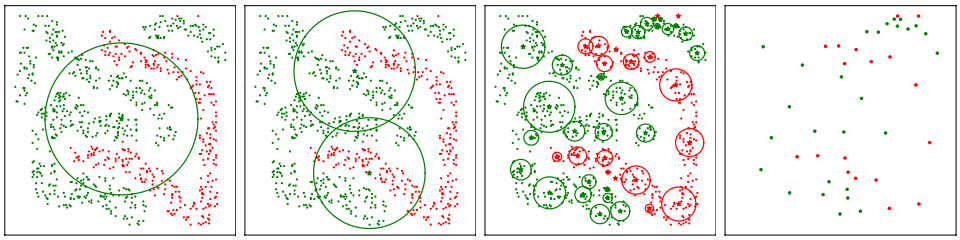


Fig. 2 The flowchart of the granular-ball generation

Taking the classification task as an example, Fig. 3 shows the process of granular-ball generation. At first, the whole dataset is considered as a big ball, as shown in Fig. 3a Subsequently, the ball is gradually generated by k-means or k-division. Figure 3c shows the coverage of all the samples by multiple balls generated in the final



(a) First step (b) Second step (c) Final step (d)Granular-ball centers

Fig. 3 The process of granular-ball generation

process. Figure 3d shows the centers of all the balls generated in the raw data. The generated balls constitute a multi-granularity coverage of the samples and the number of balls is much smaller than the number of samples. Note that the time complexity of granular-ball generation depends on the generation of the largest granular-ball among them since the number of generated granular-balls can be regarded as small constants. k-means and k-division can be viewed as approximating linear time complexity in existing work [25]. The process of calculating the corresponding properties of multiple balls is approximately negligible, so the time complexity of the granular-ball generation is also approximately linear.

The multiple advantages of GBC have gained a lot of attention in a short period of time and have been applied to many scenarios of artificial intelligence. Considering that the coarse granularity nature of granular-balls makes them less susceptible to noise, Peng et al. [28] proposed a novel GBC model by combining neighborhood rough set with granular-balls, and applied it to attribute reduction and classification with label noise, and achieved excellent detection results. Cheng et al. [29] combined GBC with the density peaks clustering algorithm to define the density of the granular-ball by the attributes of granular-balls, and the whole process has no parameters. Finally, the density of granular-balls and the $\sigma-$distance are implemented to cluster the granular-balls, and then the clustering results are expanded to the samples. In addition to classification and clustering, granular-balls have also been applied to sampling. Xia et al. [30] proposed a general sampling method based on granular-balls called granular-ball sampling (GBS). The method not only reduces the data size but also improves the quality of the data in noisy label environments. Qian et al. [31] proposed an efficient GBC model by combining neighborhood rough set and granular-ball. The model first clusters the multi-label data into multiple granules, and then proposes a label enhancement method to convert the original labels into label distribution. After that, based on the generated label distribution data, a novel feature selection method is proposed to realize the feature selection with multi-label data.

In summary, GBC has relatively desirable migration, and it can be easily applied to many fields. However, the research on the combination of GBC and FST is less, and the application of GBC in anomaly detection is still to be further explored and researched. In addition, existing granular-ball generation methods fail to deal with nominal or hybrid attribute data, so the granular-ball generation methods need to be further researched to better adapt to the complex data.

2.2 Fuzzy set theory-based anomaly detection

Compared with existing anomaly detection methods, the FST-based methods take into account the uncertainty information in the data and allow a more detailed characterization of the anomaly degree in the data. The characterization of objects is enriched to better uncover potential outliers. Xue et al. [32] proposed a novel anomaly detection method, which achieves semi-supervised detection of anomalies with only partially labeled sample data and fuzzy rough c-means clustering. Liu et al. [15] calculated the dissimilarity between samples and further used it to define the distances between granules. The distances are used to construct the state transition matrix. Anomalies are detected by the stationary distribution generated by iterative calculations. Jin et al. [33] proposed a fuzzy constraint-based anomaly detection method, which characterizes the object prior knowledge by nearness measure theory and reduces the objects by fuzzy constraints. Finally, the anomaly is detected by searching the sparse subspace in the reduced dataset. Yuan et al. [14] calculated the aggregation degrees of the samples by introducing the fuzzy rough density. The fuzzy entropy is used to calculate the weights of different attributes. Finally, the anomaly factors are calculated based on the fuzziness of the samples and the corresponding fuzzy rough density.

Although these methods have achieved desirable results in experiments, they require the calculation of sample-to-sample fuzzy relations. In addition, the input of these methods is a single sample with the finest and single granularity, which makes these methods susceptible to noise, lack of robustness, and low efficiency in anomaly detection. The detection efficiency and performance of these methods can be further improved.

3 Preliminaries

FST serves as a practical tool for managing uncertainty information in numerical, nominal, and hybrid attribute data. When applying FST to a data table, the data table without decisions is denoted as a 2-tuple $IS = \langle U, C \rangle$, where $U = \{x_1, x_2, \ldots, x_n\}$ is a nonempty finite set of samples; C is a nonempty finite set of condition attributes, storing all condition attributes for each sample; for any $x \in U$ and any $a \in C$, x^a denotes the value of the a attribute of sample x.

Definition 1 Given a nonempty finite set $U = \{x_1, x_2, \ldots, x_n\}$, if $\mathcal{R}$ is a map from U to [0, 1], i.e., $\mathcal{R} : U \to [0, 1]$, then $\mathcal{R}$ is a fuzzy set on U.

For any $x_i \in U$, $\mathcal{R}(x_i)$ denotes the membership of x_i for $\mathcal{R}$, or the membership function of $\mathcal{R}$. The set of all fuzzy sets on the universe is denoted as $\mathcal{F}(U)$. The fuzzy set $\mathcal{R}$ can be denoted as $\mathcal{R} = (\mathcal{R}(x_1), \mathcal{R}(x_2), \ldots, \mathcal{R}(x_n))$ or $\mathcal{R} = \sum_{i=1}^{n} \mathcal{R}(x_i)/x_i$.

Definition 2 Given a nonempty finite set $U = \{x_1, x_2, \ldots, x_n\}$, a fuzzy relation $\mathcal{R}$ on U is defined as $\mathcal{R} : U \times U \to [0, 1]$. $\mathcal{R}$ refers to a fuzzy set on $U \times U$.

Similar to the traditional operations between binary relations, some commonly used operations for fuzzy relations are given below.

(1) $\mathcal{R}_1 = \mathcal{R}_2 \Leftrightarrow \forall (s,t) \in U \times U, \mathcal{R}_1(s,t) = \mathcal{R}_2(s,t)$;
(2) $\mathcal{R}_1 \subseteq \mathcal{R}_2 \Leftrightarrow \forall (s,t) \in U \times U, \mathcal{R}_1(s,t) \leq \mathcal{R}_2(s,t)$;
(3) $\forall (s,t) \in U \times U, (\mathcal{R}_1 \cap \mathcal{R}_2)(s,t) = \mathcal{R}_1(s,t) \wedge \mathcal{R}_2(s,t) = \min\{\mathcal{R}_1(s,t), \mathcal{R}_2(s,t)\}$;
(4) $\forall (s,t) \in U \times U, (\mathcal{R}_1 \cup \mathcal{R}_2)(s,t) = \mathcal{R}_1(s,t) \vee \mathcal{R}_2(s,t) = \max\{\mathcal{R}_1(s,t), \mathcal{R}_2(s,t)\}$;

For any $x_i, x_j \in U \times U$, $\mathcal{R}(x_i, x_j)$ indicates the degree to which x_i has a relation $\mathcal{R}$ with x_j. A fuzzy relation $\mathcal{R}$ on U can be denoted as a fuzzy relation matrix, i.e., $M_{\mathcal{R}} = [r_{ij}]_{n \times n}$, where $r_{ij} = \mathcal{R}(x_i, x_j)$, and each row vector denotes a fuzzy set.

Let $\mathcal{R}$ be a fuzzy relation on U. For any $x, y, z \in U$, if $\mathcal{R}$ holds the following properties

(1) $\mathcal{R}(x,x) = 1 \Leftrightarrow \mathcal{R}$ is reflexive;
(2) $\mathcal{R}(x,y) = \mathcal{R}(y,x) \Leftrightarrow \mathcal{R}$ is symmetric;
(3) $\mathcal{R}(x,z) \geq \min\{\mathcal{R}(x,y), \mathcal{R}(y,z)\} \Leftrightarrow \mathcal{R}$ is transitive.

Then $\mathcal{R}$ is said to be a fuzzy equivalence relation; if $\mathcal{R}$ only satisfies items (1) and (2), then $\mathcal{R}$ is called a fuzzy similarity relation.

An information granule is a collective of elements aggregated based on principles of similarity or specific functions. The concept is initially introduced by Zadeh in 1975 through rough set theory. Expanding from FST, fuzzy information granules adeptly manage uncertainty information present within data. When multiple information granules are generated from data, they constitute a family of information granules.

For any $B \subseteq C$ and the fuzzy relation $\mathcal{R}_B$ induced by B, the family of fuzzy information granules $FG(\mathcal{R}_B)$ induced by $\mathcal{R}_B$ with respect to $x_i \in U$ is defined as

$$FG(\mathcal{R}_B) = \left\{[x_1]_{\mathcal{R}_B}, [x_2]_{\mathcal{R}_B}, \ldots, [x_n]_{\mathcal{R}_B}\right\}, \tag{1}$$

where $[x_i]_{\mathcal{R}_B} = (r_{i1}^B, r_{i2}^B, \ldots, r_{in}^B)$. $[x_i]_{\mathcal{R}_B}$ is a fuzzy information granule in $FG(\mathcal{R}_B)$ induced by $\mathcal{R}_B$. Clearly, $[x_i]_{\mathcal{R}_B}$ is a fuzzy set on $\mathcal{R}_B$. $[x_i]_{\mathcal{R}_B}(x_j) = r_{ij}^B = \mathcal{R}_B(x_i, x_j)$ quantifies the fuzzy similarity degree between x_i and x_j on $\mathcal{R}_B$. For simplicity, B is used to replace $\mathcal{R}_B$ in this study. Thus, $[x_i]_{\mathcal{R}_B}(x_j)$ can also be denoted as $[x_i]_B(x_j)$.

4 Granular-ball fuzzy set in hybrid attribute data

In this study, we introduce a hybrid distance metric to handle numerical, nominal, and hybrid attribute data efficiently. After that, a novel granular-ball generation method for hybrid attribute data is proposed. Finally, a novel granular-ball fuzzy set is proposed.

Definition 3 For any $B \subseteq C$ and any $x_i, x_j \in U$, the hybrid distance metric between x_i and x_j on condition attribute subset B is calculated by

$$HD_B(x_i, x_j) = \sqrt{\sum_{c_S \in S} \left(x_i^{c_S} - x_j^{c_S}\right)^2} + \sum_{c_N \in N} P\left(x_i^{c_N}, x_j^{c_N}\right), \tag{2}$$

where S and N denote the set of numerical attributes and the set of nominal attributes on B, respectively; $B = S \cup N$ and $S \cap N = \emptyset$; $P(x_i^{c_N}, x_j^{c_N}) = 0$ if $x_i^{c_N} = x_j^{c_N}$; $P(x_i^{c_N}, x_j^{c_N}) = 1$ if $x_i^{c_N} \neq x_j^{c_N}$.

As a novel hybrid distance metric, HD_B needs to be proved to satisfy non-negativity, symmetry, and triangle inequality. For any $x_i, x_j \in U$, $HD_B(x_i, x_j) \geq 0$ and if and only if $x_i = x_j$, $HD_B(x_i, x_j) = 0$, so HD_B satisfies non-negativity. Next, $HD_B(x_i, x_j) = HD_B(x_j, x_i)$, so HD_B satisfies symmetry. Finally, the triangle inequality, HD_B consists of a numerical attribute part and a nominal attribute part. In the numerical attribute part, $\sqrt{\sum_{c_S \in S}(x_i^{c_S} - x_j^{c_S})^2}$ as the Euclidean distance satisfies the triangle inequality. In the nominal attribute part, $\sum_{c_N \in N} P(x_i^{c_N}, x_j^{c_N})$ as the Hamming distance also satisfies the triangle inequality. So, for any

$x_i, x_j, x_k \in U$, $HD_B(x_i, x_j) + HD_B(x_j, x_k) \geq HD_B(x_i, x_k)$. So, HD_B satisfies the triangle inequality. Therefore, HD_B is a valid distance metric.

4.1 Hybrid granular-ball generation

The original granular-ball generation methods use *k*-means [24] or *k*-division [26]. The granular-ball quality is defined by the labels of the samples inside the granular-ball. The use of *k*-means or *k*-division leads to the fact that granular-balls can only be constructed in numerical attribute data and fail to handle nominal or hybrid attribute data. Therefore, there is a need to give a novel granular-ball generation method under hybrid attribute data. In addition, we need to give a novel definition of the granular-ball quality in unsupervised tasks.

To apply the *k*-means to hybrid attribute data, we use the hybrid distance metric in Eq. (2) to replace the original Euclidean distance. In particular, we calculate the distances of the samples under nominal and numerical attributes separately and add these two distances to get the hybrid distance. Then we use the hybrid distance as the basis for *k*-means, which in turn enables the generation of granular-balls under hybrid attribute data.

Furthermore, the center of clusters needs to be calculated in the clustering process. The original *k*-means is realized by directly solving for the mean value, which does not apply to nominal and hybrid attribute data. Therefore, we calculate the center of numerical attributes and the center of nominal attributes separately. Specifically, for numerical attributes, the mean is directly solved as the center as in the original *k*-means; for nominal attributes, the category with the highest number of occurrences is taken as the center, i.e., mode. Similarly, when assigning samples to clusters, the hybrid distance is calculated. Additionally, there may be empty clusters during granular-ball generation, and if a cluster is empty, the sample farthest from the cluster will be chosen as the centroid. Based on the above hybrid distance metric, the granular-ball generation under hybrid attribute data can be realized. We refer to the above process as the hybrid granular-ball generation. The related algorithm implementation is shown in Algorithm 1. First of all, the input dataset is considered as a coarse ball. Then, determine if the coarse ball needs to be divided. If it exceeds a specified threshold, it is divided by Algorithm 1. Then, the two sub-balls after division are stored, and the parent ball is deleted. If the ball does not exceed the threshold, no division is required, and it is stored directly. Then, iterates over the newly generated balls until all balls meet the threshold.

Algorithm 1 Hybrid granular-ball generation

Input: $IS = \langle U, C \rangle$;
Output: K;
Initializing: a set of granular-balls $K = \emptyset$, a granular-ball $GB = \emptyset$, a queue $Q = \emptyset$;
$GB = \{U\}$ and put GB into Q;
while $Q \neq \emptyset$ **do**
  Get the top granular-ball GB in Q and remove GB in Q;
  if $|GB| \geq \sqrt{|U|}$ **then**
    Divide GB into GB_1 and GB_2 by 2-means with the hybrid distance metric in Eq. (2);
    Put GB_1 and GB_2 into the tail of Q;
  else
    $K = K \cup \{GB\}$;
  end if
end while
return K;

The time complexity of the classic granular-ball generation algorithm is approximately linear [24]. In this study, our proposed hybrid granular-ball generation algorithm is similar to the previous ones, except that the relevant distance metric is replaced with the hybrid distance metric, so the time complexity of Algorithm 1 can also be approximated as linear $O(|U|)$.

In addition, the classic definitions of the center and radius of the granular-balls are based on numerical attribute data and could not be applied to the nominal and hybrid attribute data. Therefore, we redefine the center and radius based on the different attribute types of data. First is the center, which is defined in the same way as the cluster centers defined above, dealing with numerical and nominal attribute data, respectively. Specifically, the previous method of calculating the mean values is still taken for numerical attributes, and the mode is taken as the radius for nominal attributes. Next is the radius, for any $GB \in K_B$ generated on $B \subseteq C$, its radius is defined as $r = \frac{1}{|GB|} \sum_{x_i \in GB} HD_B(x_i, c)$. Both nominal and numeric attributes can be included in B.

In this study, granular-balls serve as a key element for anomaly detection. Since labeled data is not available in unsupervised tasks, the original granular-ball generation

method needs to be improved to be applicable in unsupervised anomaly detection. Based on the relevant research [23, 29], we utilize the number of samples contained within each granular-ball as a metric for evaluating its quality. The termination threshold for granular-ball generation is set to $\sqrt{|U|}$. When the number of samples within a granular-ball exceeds or equals $\sqrt{|U|}$, it suggests inadequate compactness, prompting further subdivision. Conversely, if the sample count is less than $\sqrt{|U|}$, the current granular-ball is considered sufficiently compact, and division is stopped. This adaptive approach eliminates the need for manual threshold adjustments typically required in traditional granular-ball generation methods.

4.2 Granular-ball fuzzy set

Based on the hybrid distance metric in Eq. (2), the fuzzy relation $\mathcal{R}_B(x_i, x_j)$ between x_i and x_j on the condition attribute subset $B \subseteq C$ is calculated as

$$\mathcal{R}_B(x_i, x_j) = \begin{cases} 0, & 1 - \frac{HD_B(x_i, x_j)}{|B|} < \sigma; \\ 1 - \frac{HD_B(x_i, x_j)}{|B|}, & \text{Other cases.} \end{cases} \tag{3}$$

Clearly, the above fuzzy relation is reflexive and symmetric. Therefore, it is a fuzzy similarity relation. σ is a hyperparameter that sets fuzzy relation below this value to zero. In the above equation, the hybrid distance metric is used to calculate the fuzzy relations between samples. In addition, the numerical attributes are normalized before calculation, so that the value domain of $HD_B(x_i, x_j)$ uniformly falls within the interval [0, 1].

Combining GBC with FST allows for the construction of efficient and robust methods. Unlike the traditional FST, the granular-ball fuzzy set considers the granular-ball to which the sample belongs rather than a particular sample alone. The granular-ball fuzzy set is proposed in this study and its related definitions are given below.

Definition 4 For any $B \subseteq C (B \neq \emptyset)$, $K_B = \left\{GB_1^B, GB_2^B, \ldots, GB_k^B\right\}$ is a set of granular-balls generated on condition attribute subset B. For any $x_i, x_j \in U$, the granular-ball fuzzy relation between x_i and x_j induced by K_B is defined as

$$\mathcal{R}_B^{GB}(x_i, x_j) = \mathcal{R}_B(T(x_i), T(x_j)), \tag{4}$$

where $T(\cdot)$ denotes a mapping function that returns the center of the granular-ball to which the input sample belongs. The granular-ball fuzzy relations between the samples are calculated by mapping the fuzzy relations between granular-balls to the samples within these balls.

Obviously, after mapping, $\mathcal{R}_B^{GB}$ is reflexive and symmetric, i.e., it is a fuzzy similarity relation. The family of hybrid granular-ball fuzzy information granules $FG(\mathcal{R}_B^{GB})$ with respect to U is defined as

$$FG(\mathcal{R}_B^{GB}) = \left\{[x_1]_B^{GB}, [x_2]_B^{GB}, \ldots, [x_n]_B^{GB}\right\}. \tag{5}$$

The cardinality of the hybrid granular-ball fuzzy information granule $[x_i]_B^{GB}$ is calculated as $|[x_i]_B^{GB}| = \sum_{j=1}^{n} \mathcal{R}_B^{GB}(x_i, x_j)$. If the cardinality of $[x_i]_B^{GB}$ is smaller than the other granules in $FG(\mathcal{R}_B^{GB})$, then x_i is regarded as the minority class in U. Similarly, $M_{\mathcal{R}_B^{GB}}$ denotes a fuzzy relation matrix.

In the above definition, we define a novel granular-ball fuzzy set for hybrid attribute data. We introduce a hybrid distance metric to calculate the granular-ball fuzzy relations between samples. In addition, we give the definition of the corresponding hybrid granular-ball fuzzy information granules. We utilize the advantages in GBC and FST, which enable the inter-sample fuzzy relations to be calculated more efficiently and robustly, and next, based on this novel FST, we construct an efficient unsupervised anomaly detection method.

5 Granular-ball computing guided anomaly detection for hybrid attribute data

In this section, an unsupervised anomaly detection method called HGBAD is proposed based on hybrid granular-ball fuzzy information granules.

5.1 Anomaly detection

For a given dataset, multiple hybrid granular-ball fuzzy information granules are generated under different subsets of attributes. Each granule contains the fuzzy relations between each sample and the rest of the samples, and its cardinality reveals the similarity degree between the samples and the different samples under a certain granular-ball fuzzy relation. If the cardinality is small, indicating a low similarity degree, then the granule should be more anomalous.

Definition 5 For any $B \subseteq C$ and any $x \in U$, the anomaly degree of the hybrid granular-ball fuzzy information granule $[x]_B^{GB}$ is defined as

$$FGAD([x]_B^{GB}) = 1 - \frac{1}{|U|}\left|[x]_B^{GB}\right|. \tag{6}$$

If the value of $FGAD([x]_B^{GB})$ is large, it means that the fuzzy similarity degree between the rest of the samples and x under $\mathcal{R}_B^{GB}$ is low, and the anomaly degree of this granule

will be large. The above definition considers the anomaly degree of the hybrid granular-ball fuzzy information granules under different subsets of attributes.

In real-world scenarios, it is essential to acknowledge that various subsets of attributes have distinct impacts on the overall dataset. With this understanding. The concept of fuzzy entropy has been used to quantify the levels of uncertainty and fuzziness present within the data [34]. Uncertainty in the data can be considered as a representation of anomalies, The higher the degree of uncertainty, the more likely it is to be associated with anomalies. Uncertainty can be considered as a representation of anomalies, so fuzzy entropy can be applied to anomaly feature mining to construct efficient anomaly detection methods.

In this study, we also introduce fuzzy entropy to construct anomaly features. The granular-ball fuzzy entropy calculated based on the hybrid granular-ball fuzzy information granules is given below.

Definition 6 For any $B \subseteq C$, the granular-ball fuzzy entropy of the condition attribute subset B is defined as

$$GBFE(B) = -\frac{1}{|U|}\log_2 \sum_{x\in U}\frac{1}{|U|}\left|[x]_B^{GB}\right|. \tag{7}$$

For any $a \in C$, if the value of $GBFE(a)$ is larger, it means that the distribution of the value domain of a is more chaotic or disordered. Combining the granular-ball fuzzy entropy with the anomaly degrees of the hybrid granular-ball fuzzy information granules defined previously, a more realistic weighted anomaly factor can be constructed.

Definition 7 For any $x \in U$, the fuzzy granular-ball information granules-based anomaly factor of x is defined as

$$FGAF(x) = \frac{1}{|C|}\sum_{a\in C} FGAD([x]_a^{GB}) \cdot W(a), \tag{8}$$

where the weights of each attribute $W(a) = \sqrt[3]{\frac{GBFE(a)}{\sum_{c\in C} GBFE(c)}} \in [0, 1]$.

We calculate the weight $W(a)$ by the granular-ball fuzzy entropy of each attribute. If an attribute is unordered distributed, the calculation of the anomaly factors should focus on the impact and influence that the attribute produces. In addition, each attribute has a weight between [0, 1], and the sum of all weights is equal to 1. After min-max normalization for the numerical attributes values, the domain of values of *FGAD* for each hybrid granular-ball fuzzy information granule is between [0, 1]. Therefore, the *FGAF* for each sample is also between [0, 1].

It should be noted that the calculation of *FGAF* is based on attribute-by-attribute. This is because, if different attribute subsets are considered according to the previous definitions, an exponential amount of computation will be generated, which is unrealistic for practical computation. Therefore, to improve the detection efficiency, we use attribute-by-attribute instead.

Based on the above thinking, the anomaly factors of each sample are calculated, which determine whether a sample is anomaly or not by a threshold. The anomaly factors are calculated from the multiple hybrid granular-ball fuzzy information granules. Suppose the anomaly degree of multiple hybrid granular-ball fuzzy information granules of a sample is high. In that case, the anomaly factor of the sample should also be large, i.e., the probability that the sample is an anomaly is high.

Definition 8 Given an anomaly determination threshold τ. For any $x \in U$, if $FGAF(x) > \tau$, x is considered an anomaly detected by HGBAD.

5.2 Algorithm implementation

In this section, we give the pseudo-code implementation of the HGBAD algorithm. From Algorithm 2, it can be seen that the whole algorithm flow is relatively straightforward. First, an IS containing only the condition attributes C and a hyper-parameter σ is input. Iterating over different attribute subsets $B \subseteq C$, K_B is generated on condition attribute subsets B, and then the corresponding granular-ball fuzzy relation $\mathcal{R}_B^{GB}$ and corresponding matrices $M_{\mathcal{R}_B^{GB}}$ is calculated. Then, according to Eq. (6), the anomaly degree $FGAD([x_i]_{\mathcal{R}_B^{GB}})$ and the corresponding weight $W_B(x_i)$ are calculated. Finally, the anomaly factor $FGAF(x_i)$ of x_i is calculated according to Eq. (8).

Algorithm 2 HGBAD

Input: $IS = \langle U, C\rangle$, σ;
Output: $FGAF$;
1: Initializing: $FGAF = \emptyset$;
2: **for** $B \subseteq C$ **do**
3: Generate granular-ball set K_B on B by Algorithm 1;
4: Calculate $\mathcal{R}_B^{GB}$ and $M_{\mathcal{R}_B^{GB}}$ by Eq. (3);
5: Calculate $W(B)$ by Eq. (7);
6: **for** $x \in U$ **do**
7: Calculate $FGAD([x]_B^{GB})$ by Eq. (6);
8: **end for**
9: **end for**
10: **for** $x \in U$ **do**
11: Calculate $FGAF(x)$ by Eq. (8);
12: **end for**
13: **return** $FGAF$;

In the given pseudo-code, granular-balls are first generated under different attribute subsets. From the above analysis, the time complexity of Algorithm 1 is $O(|U|)$. However, an

Table 2 Experimental datasets

ID	Datasets	Abbr.	Condition attributes		Samples	Anomalies	Anomaly ratios (%)
			Numerical	Nominal			
1	audiology_variant1	Audio	0	69	226	53	23.5
2	lymphography	Lymph	0	8	148	6	4.1
3	monks_0_4_variant1	Monks_4	0	6	232	4	1.7
4	tic_tac_toe_negative_12_variant1	Tic_12	0	9	638	12	1.9
5	tic_tac_toe_negative_26_variant1	Tic_26	0	9	652	26	4.0
6	tic_tac_toe_negative_32_variant1	Tic_32	0	9	658	32	4.9
7	tic_tac_toe_negative_69_variant1	Tic_69	0	9	695	69	9.9
8	annthyroid	Annthyroid	6	0	7200	534	7.4
9	cardiotocography_2and3_33_variant1	Card	21	0	1688	33	2.0
10	glass	Glass	9	0	214	9	4.2
11	iris_Irisvirginica_11_variant1	Iris	4	0	111	11	9.9
12	musk	Musk	166	0	3062	97	3.2
13	pima_TRUE_55_variant1	Pima	9	0	555	55	9.9
14	waveform_0_100_variant1	Wave	21	0	3443	100	2.9
15	wine	Wine	13	0	129	10	7.8
16	wpbc_variant1	Wpbc	33	0	198	47	23.7
17	arrhythmia_variant1	Arrhy	206	73	452	66	14.6
18	german_1_14_variant1	German	7	13	714	14	2.0
19	heart270_2_16_variant1	Heart	6	7	166	16	9.6
20	hepatitis_2_9_variant1	Hepatitis	6	13	94	9	9.6
21	horse_1_12_variant1	Horse	8	19	256	12	4.7
22	sick_sick_35_variant1	Sick_35	7	22	3576	35	1.0
23	sick_sick_72_variant1	Sick_72	7	22	3613	72	2.0
24	thyroid_disease_variant1	Thyroid_d	7	21	9172	74	0.8

exponential subset of attributes will greatly increase the time complexity of the whole algorithm, so we replace the subset of attributes with a single attribute. Thus, the time complexity of the whole algorithm is approximated as $O(|C|(|U| + |K_B|^2))$. We do not calculate the fuzzy relations between samples directly, but map the fuzzy relations between granular-balls to the samples. In the mapping process, we only need to index according to the subscripts, so the time required for mapping can be approximately ignored, which greatly improves the efficiency of anomaly detection.

6 Experiments

This section demonstrates several experiments conducted in this study and describes the relevant aspects of the experiments, such as the datasets and compared methods. The results of the experiments are also visualized, explained, and discussed.

6.1 Experimental settings

We conduct experiments on publicly available datasets.[1, 2] Rich datasets, including nominal, numerical, and hybrid attribute datasets with large samples and attributes, are selected for our experiments. As shown in Table 2, this table shows the original name, abbreviation, number of numerical attributes, number of nominal attributes, number of samples, number of anomalies, and percentage of anomalies for each dataset. These datasets contain numerical attribute data, nominal attribute data, and hybrid attribute data, which simulate applications in a variety of scenarios so that the effectiveness of the methods in this study can be verified realistically. The maximum number of attributes is 279, the maximum number of samples is 9172, and the maximum percentage of anomalies is 23.7%.

We conduct a comparison of our method with several well-known and efficient anomaly detection methods, as shown in Table 3. Table 3 contains a variety of methods

[1] https://github.com/BELLoney/Outlier-detection.

[2] https://odds.cs.stonybrook.edu/.

Table 3 The descriptions of compared methods and hyper-parameter adjustment ranges, where ✘ denotes that this method has no hyper-parameter adjustment range or step size

No	Methods	Descriptions	Hyper-parameter adjustment ranges	Step sizes
1	HGBAD (Ours)	Hybrid granular-ball fuzzy information granules-based anomaly detection	[0, 1]	0.05
2	WFRDA [14]	Weighted fuzzy rough density-based anomaly detection	[0.1, 2]	0.1
3	FGAS [15]	Fuzzy granules based-outlier detection	[0, 1]	0.05
4	WNINOD [35]	Outlier detection based on weighted neighbourhood information network	[1, 10]	1
5	VarE [36]	Outlier detection using variance structural scores	$\{10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}\}$	✘
6	ApproE [36]	Outlier detection using variance structural scores	$\{10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}\}$	✘
7	ODGrCR [37]	Granular computing and rough set theory-based outlier detection	✘	✘
8	WDOD [38]	Weighted density-based outlier detection	✘	✘
9	ITB [39]	Information-theoretic-based outlier detection	✘	✘
10	LDOF [40]	Local distance-based outlier detection	[1, 60]	1
11	SEQ [41]	Sequence-based outlier detection	✘	✘
12	ODIN [42]	Indegree number-based outlier detection	[1, 60]	1

that have been well-received in the field for their effectiveness and innovative techniques. We aim to demonstrate the effectiveness of HGBAD in various scenarios by comparing it with these methods.

As shown in Table 3, we set various hyper-parameter adjustment ranges for different methods to ensure that the performance of each method can be best demonstrated, and try to ensure the fairness of the experiment. We normalize the numerical attribute values for the input samples and replace the nominal attribute values with the same integers. During the prediction process, each anomaly factors of samples factor is calculated to show its anomaly degree.

6.2 Experimental results

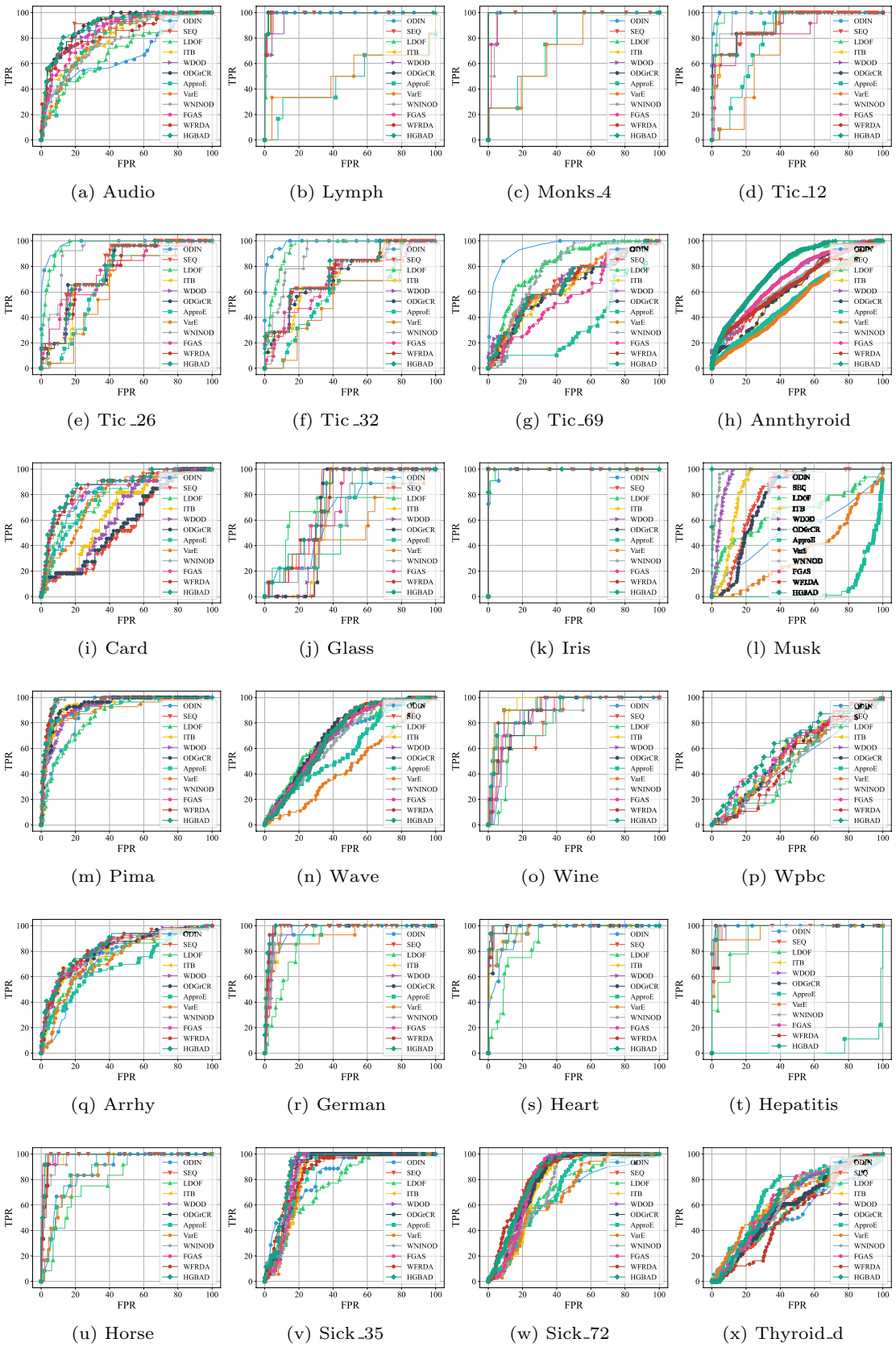
The receiver operating characteristic (ROC) curve and the area under the curve (AUC) metric constitute fundamental tools for evaluating the performance of different methods [14, 23]. In this study, we also use the ROC curve and AUC metric to evaluate the performance of different methods. For the ROC curve, the closer the curve is to the upper left corner of the axis, the better the performance of the method. For the AUC metric, it takes a value between 0 and 1, and the closer the value is to 1, the better the performance of the method.

Figure 4 shows the ROC curves of the 12 methods on 24 datasets, where HGBAD corresponds to the army green curve. From the figure, it can be seen that HGBAD performs best on datasets such as Audio, Lymph, Monks_4, Card, etc. However, due to many comparison methods, there are overlapping curves in the figure. Since there are some overlaps in the ROC curves that are not easy to compare, the following will compare the average AUC of the different methods across all datasets.

In Table 4, we underline the best AUC value under each dataset to emphasize it. The last row in the table counts the number of times the different methods achieved the best scores on all datasets. It is easy to see from the table that HGBAD achieved first place in 15 out of 24 datasets with an average AUC of 0.868, which demonstrates that HGBAD has the best performance and achieved better results in numerical, nominal, and hybrid attribute datasets, validating the anomaly detection capability of HGBAD with hybrid attribute data.

6.3 Hyper-parameter sensitivity analysis

When calculating fuzzy relations in HGBAD, the hyper-parameter σ is needed. This section analyzes the effect of σ on the performance of HGBAD. Since σ does not affect the results in nominal attribute data. Therefore, only numerical attribute data and hybrid attribute data are analyzed here. The adjustment range of σ is [0, 1] with a step size of 0.05, and the AUC curves of HGBAD under different datasets with different σ are shown in Fig. 5. From the figure, it can be seen that HGBAD is affected by hyper-parameters within acceptable limits under most datasets. Especially in the hybrid attribute data, the change in the AUC of HGBAD under different σ is extremely small, and the curves are smoother under different datasets. However, it is worth noting that there is a sudden drop when the value of σ is close to 1.0 on datasets such as Musk, Glass, Wave, and Card. This is because when $\sigma = 1.0$, after Eq. 3, the fuzzy relations between the samples are all set to zero and fail to be used to mine anomaly features in the data. In summary, in order



(a) Audio (b) Lymph (c) Monks_4 (d) Tic_12

(e) Tic_26 (f) Tic_32 (g) Tic_69 (h) Annthyroid

(i) Card (j) Glass (k) Iris (l) Musk

(m) Pima (n) Wave (o) Wine (p) Wpbc

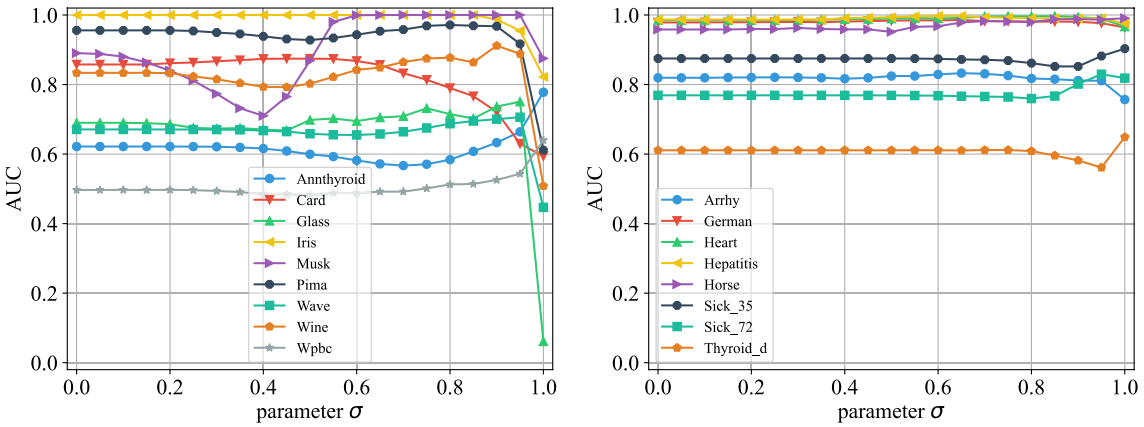
(q) Arrhy (r) German (s) Heart (t) Hepatitis

(u) Horse (v) Sick_35 (w) Sick_72 (x) Thyroid_d

Fig. 4 ROC curves for different methods on all datasets

Table 4 AUC of different methods on all datasets, where the method with the highest AUC in each dataset is underlined

Dataset	ODIN	SEQ	LDOF	ITB	WDOD	ODGrCR	ApproE	VarE	WNINOD	FGAS	WFRDA	**HGBAD**
Audio	0.672	0.896	0.701	0.815	0.868	0.903	0.796	0.792	0.778	0.846	0.834	0.903
Lymph	0.994	0.991	0.992	0.991	0.974	0.995	0.477	0.507	0.992	0.989	0.993	0.996
Monks_4	1.000	0.987	1.000	1.000	0.987	1.000	0.772	0.727	0.978	0.978	1.000	1.000
Tic_12	0.996	0.900	0.981	0.903	0.908	0.912	0.798	0.733	0.967	0.856	0.913	0.914
Tic_26	0.981	0.784	0.963	0.757	0.775	0.776	0.685	0.643	0.906	0.760	0.764	0.779
Tic_32	0.988	0.754	0.944	0.729	0.753	0.733	0.592	0.563	0.894	0.701	0.758	0.763
Tic_69	0.950	0.669	0.812	0.638	0.686	0.634	0.356	0.668	0.775	0.574	0.660	0.668
Annthyroid	0.696	0.649	0.772	0.628	0.655	0.625	0.547	0.523	0.667	0.707	0.658	0.778
Card	0.824	0.560	0.790	0.673	0.641	0.581	0.795	0.780	0.851	0.850	0.865	0.874
Glass	0.650	0.694	0.726	0.663	0.679	0.678	0.629	0.538	0.674	0.704	0.746	0.751
Iris	0.993	1.000	0.995	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Musk	0.516	0.794	0.663	0.881	0.950	0.767	0.072	0.346	0.983	1.000	1.000	1.000
Pima	0.889	0.929	0.855	0.939	0.919	0.938	0.920	0.904	0.963	0.971	0.974	0.971
Wave	0.683	0.704	0.711	0.725	0.707	0.726	0.595	0.473	0.671	0.699	0.704	0.706
Wine	0.887	0.833	0.816	0.933	0.900	0.896	0.939	0.941	0.873	0.885	0.915	0.912
Wpbc	0.507	0.548	0.504	0.560	0.560	0.553	0.557	0.552	0.506	0.594	0.528	0.640
Arrhy	0.737	0.814	0.749	0.801	0.811	0.813	0.691	0.739	0.815	0.824	0.826	0.833
German	0.952	0.976	0.889	0.959	0.953	0.979	0.946	0.919	0.965	0.975	0.984	0.986
Heart	0.949	0.991	0.880	0.983	0.987	0.985	0.965	0.962	0.992	0.997	0.995	0.997
Hepatitis	0.997	0.987	0.922	0.988	0.986	0.988	0.033	0.962	0.992	0.997	0.995	0.997
Horse	0.870	0.983	0.803	0.981	0.985	0.980	0.883	0.869	0.966	0.987	0.982	0.991
Sick_35	0.837	0.890	0.779	0.859	0.880	0.870	0.848	0.842	0.864	0.869	0.868	0.903
Sick_72	0.682	0.820	0.739	0.777	0.808	0.794	0.722	0.691	0.754	0.804	0.837	0.830
Thyroid_d	0.543	0.591	0.572	0.649	0.635	0.604	0.705	0.650	0.640	0.657	0.531	0.649
Average	0.825	0.823	0.815	0.826	0.834	0.822	0.680	0.722	0.853	0.843	0.847	0.868
1st order	6	1	1	2	1	4	2	2	1	4	5	15



(a)Numerical attribute datasets (b) Hybrid attribute datasets

Fig. 5 The variation curves of AUC on hyper-parameter σ with different datasets

to show the best performance of HGBAD, the best hyper-parameters need to be selected based on different datasets.

6.4 Statistical analysis

In addition to the above experiments, referring to the existing research [14, 23], the Friedman test and Nemenyi

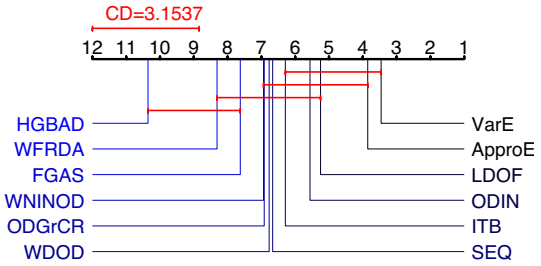


Fig. 6 Nemenyi test figure on AUC

test are conducted for statistical analysis of all methods. As can be seen in Tables 2 and 3, there are 12 methods and 24 datasets used during the experiment. The F distributions with freedom degrees of 11 and 253 can be obtained. In the Friedman test, when the significance level $\alpha = 0.1$, $\tau_F = 8.5064$, it is greater than the critical value 1.5962. This result indicates a statistically significant difference among the methods under investigation. To this end, we proceed to conduct a post-hoc analysis to elucidate and distinguish the specific differences between these methods further.

In the Nemenyi test, when the significance level $\alpha = 0.1$, the corresponding critical distance $CD_{0.1} = 3.1537$. To illustrate the results of the Nemenyi test, we plot the average ordinal values of all evaluated methods along with a line segment representing the critical difference on a single axis. This figure allows for an intuitive comparison of the performances of different methods, highlighting the statistical significance of differences observed in their rankings as determined by the Nemenyi test.

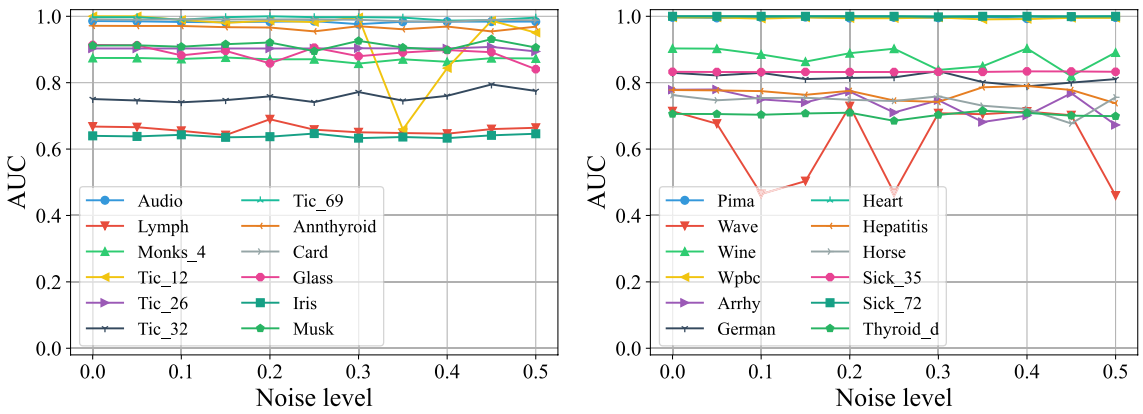
We show the results of the Nemenyi test in Fig. 6, each red line segment in the figure is the critical distance line segment. If more than one method is covered by a single critical distance line segment on the axes, it means that there is no significant difference between these methods. In the figure, HGBAD only overlaps with WFRDA and FGAS, indicating that there is no statistically significant difference between them. It should be noted that this does not mean that HGBAD is similar to these two methods.

6.5 Attribute noise sensitivity analysis

To evaluate the sensitivity of our method to noise, we adopt the strategy used in existing studies [15, 43]. We randomly select $\lceil |U| \times \beta \rceil$ samples, where β denotes the noise level. Then, a random attribute is selected from the attributes of these samples, and these attribute values are replaced with random values between the maximum and minimum values of the attribute. The effect of different noise levels on AUC is shown in Fig. 7. From the figure, it can be seen that the AUC variation curves of HGBAD under most of the datasets do not fluctuate too much under different noise levels, and there is only a significant fluctuation under Tic_12 and Wave, which indicates that the anti-noise ability of HGBAD is relatively good on the majority of nominal, numerical, and hybrid attribute datasets.

7 Conclusion

In this study, we provide an effective method for anomaly detection in hybrid attribute data, called HGBAD. It uses the hybrid granular-ball fuzzy information granules belonging to the samples to quantify the anomaly degree of the samples and assign reasonable weights to different attributes through granular-ball fuzzy entropy. We validate



(a) Datasets 1-12 (b) Datasets 13-24

Fig. 7 The variation curves of AUC with different noise levels

the effectiveness of HGBAD through rich experiments. Experimental results demonstrate that HGBAD achieves first place on 15 of 24 datasets and outperforms existing methods. There are statistically significant differences between HGBAD and most of the methods. In addition, he experimental results also demonstrates that HGBAD is insensitive to hyper-parameter σ and robust to the noise in the data. In future work, we will continue to improve the existing methods and propose parameter-free and efficient anomaly detection methods for hybrid attribute data.

Acknowledgements This work was supported by the National Natural Science Foundation of China (62306196, 62372315, and 62376230), Sichuan Science and Technology Program (2024NSFTD0049, 2023YFQ0020 and 2024NSFSC0443), and the Fundamental Research Funds for the Central Universities (YJ202245).

Data availability The data used in this study are available at https://github.com/Mxeron/HGBAD.

References

1. Pawar R, Kathuria H, Joe P (2023) Credit card fraud detection and analysis. In: 2023 14th international conference on computing communication and networking technologies (ICCCNT). IEEE, pp. 1–5
2. Matsushima Y, Noma H, Yamada T, Furukawa TA (2020) Influence diagnostics and outlier detection for meta-analysis of diagnostic test accuracy. Res Synth Methods 11(2):237–247
3. Huang K, Wen H, Yang C, Gui W, Hu S (2021) Outlier detection for process monitoring in industrial cyber-physical systems. IEEE Trans Autom Sci Eng 19(3):2487–2498
4. Hussain I (2020) Outlier detection using nonparametric depth-based techniques in hydrology. Int J Adv Comput Sci Appl. https://doi.org/10.14569/IJACSA.2020.0110954
5. Smiti A (2020) A critical overview of outlier detection methods. Comput Sci Rev 38:100306
6. Domański PD (2020) Study on statistical outlier detection and labelling. Int J Autom Comput 17(6):788–811
7. Suboh S, Aziz IA, Shaharudin SM, Ismail SA, Mahdin H (2023) A systematic review of anomaly detection within high dimensional and multivariate data. JOIV Int J Inform Visual 7(1):122–130
8. Samariya D, Thakkar A (2023) A comprehensive survey of anomaly detection algorithms. Ann Data Sci 10(3):829–850
9. Boukerche A, Zheng L, Alfandi O (2020) Outlier detection: methods, models, and classification. ACM Comput Surv (CSUR) 53(3):1–37
10. Staerman G, Adjakossa E, Mozharovskyi P, Hofer V, Sen Gupta J, Clémençon S (2023) Functional anomaly detection: a benchmark study. Int J Data Sci Anal 16(1):101–117
11. Tayeh T, Aburakhia S, Myers R, Shami A (2020) Distance-based anomaly detection for industrial surfaces using triplet networks. In: 2020 11th IEEE annual information technology, electronics and mobile communication conference (IEMCON). IEEE, pp 0372–0377
12. Sikder MNK, Batarseh FA (2023) Outlier detection using AI: a survey. AI Assur. https://doi.org/10.48550/arXiv.2112.00588
13. Breuni, MM, Kriegel H-P, Ng RT, Sander J (2000) LOF: identifying density-based local outliers. In: Proceedings of the 2000 ACM SIGMOD international conference on management of data, pp 93–104
14. Yuan Z, Chen B, Liu J, Chen H, Peng D, Li P (2023) Anomaly detection based on weighted fuzzy-rough density. Appl Soft Comput 134:109995
15. Liu C, Yuan Z, Chen B, Chen H, Peng D (2023) Fuzzy granular anomaly detection using Markov random walk. Inf Sci 646:119400
16. Aydın F (2023) Boundary-aware local density-based outlier detection. Inf Sci 647:119520
17. Yao Y (2008) Granular computing: past, present and future. In: 2008 IEEE international conference on granular computing. IEEE, pp 80–85
18. Yao JT, Vasilakos AV, Pedrycz W (2013) Granular computing: perspectives and challenges. IEEE Trans Cybern 43(6):1977–1989
19. Rani P, Chen S-M, Mishra AR (2023) Multiple attribute decision making based on Mairca, standard deviation-based method, and Pythagorean fuzzy sets. Inf Sci 644:119274
20. Demir I (2023) Novel correlation coefficients for interval-valued Fermatean hesitant fuzzy sets with pattern recognition application. Turk J Math 47(1):213–233
21. Tang Y, Huang J, Pedrycz W, Li B, Ren F (2023) A fuzzy clustering validity index induced by triple center relation. IEEE Trans Cybern 53(8):5024–5036
22. Zhao S, Dai Z, Wang X, Ni P, Luo H, Chen H, Li C (2021) An accelerator for rule induction in fuzzy rough theory. IEEE Trans Fuzzy Syst 29(12):3635–3649
23. Su X, Yuan Z, Chen B, Peng D, Chen H, Chen Y (2024) Detecting anomalies with granular-ball fuzzy rough sets. Inf Sci 678:121016
24. Xia S, Liu Y, Ding X, Wang G, Yu H, Luo Y (2019) Granular ball computing classifiers for efficient, scalable and robust learning. Inf Sci 483:136–152
25. Xia S, Wang G, Gao X (2023) Granular ball computing: an efficient, robust, and interpretable adaptive multi-granularity representation and computation method. arXiv preprint arXiv:2304.11171
26. Xia S, Dai X, Wang G, Gao X, Giem E (2022) An efficient and adaptive granular-ball generation method in classification problem. IEEE Trans Neural Netw Learn Syst 35(4):5319–5331
27. Bai H, Shen F, Kong W, Feng J (2023) Granular-ball clustering based neighbourhood outliers detection method. In: 2023 6th International conference on electronics technology (ICET). IEEE, pp 1306–1312
28. Peng X, Wang P, Xia S, Wang C, Chen W (2022) VPGB: a granular-ball based model for attribute reduction and classification with label noise. Inf Sci 611:504–521
29. Cheng D, Li Y, Xia S, Wang G, Huang J, Zhang S (2023) A fast granular-ball-based density peaks clustering algorithm for large-scale data. IEEE Trans Neural Netw Learn Syst. https://doi.org/10.1109/TNNLS.2023.3300916
30. Xia S, Zheng S, Wang G, Gao X, Wang B (2023) Granular ball sampling for noisy label classification or imbalanced classification. IEEE Trans Neural Netw Learn Syst 34(4):2144–2155
31. Qian W, Xu F, Qian J, Shu W, Ding W (2023) Multi-label feature selection based on rough granular-ball and label distribution. Inf Sci 650:119698
32. Xue Z, Shang Y, Feng A (2010) Semi-supervised outlier detection based on fuzzy rough c-means clustering. Math Comput Simul 80(9):1911–1921
33. Jin L, Chen J, Zhang X (2019) An outlier fuzzy detection method using fuzzy set theory. IEEE Access 7:59321–59332
34. Hu Q, Yu D, Xie Z, Liu J (2006) Fuzzy probabilistic approximation spaces and their information measures. IEEE Trans Fuzzy Syst 14(2):191–201

35. Wang Y, Li Y (2021) Outlier detection based on weighted neighbourhood information network for mixed-valued datasets. Inf Sci 564:396–415
36. Li X, Lv J, Yi Z (2018) Outlier detection using structural scores in a high-dimensional space. IEEE Trans Cybern 50(5):2302–2310
37. Jiang F, Chen Y-M (2015) Outlier detection based on granular computing and rough set theory. Appl Intell 42:303–322
38. Huang J, Zhu Q, Yang L, Feng J (2016) A non-parameter outlier detection algorithm based on natural neighbor. Knowl Based Syst 92:71–77
39. Wu S, Wang S (2013) Information-theoretic outlier detection for large-scale categorical data. IEEE Trans Knowl Data Eng 25(3):589–602
40. Zhang K, Hutter M, Jin H (2009) 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, pp 813–822
41. Jiang F, Sui Y, Cao C (2009) Some issues about outlier detection in rough set theory. Expert Syst Appl 36(3):4680–4687
42. Hautamaki V, Karkkainen I, Franti P (2004) Outlier detection using k-nearest neighbour graph. In: Proceedings of the 17th international conference on pattern recognition, 2004. ICPR 2004, vol 3. IEEE, pp 430–433
43. Zhu X, Wu X (2004) Class noise vs. attribute noise: a quantitative study. Artif Intell Rev 22:177–210