Information Sciences 678 (2024) 121016

Contents lists available at ScienceDirect

Information Sciences

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

Detecting anomalies with granular-ball fuzzy rough sets

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

[a] College of Computer Science, Sichuan University, Chengdu 610065, China
[b] School of Computing and Artificial Intelligence, Southwest Jiaotong University, Chengdu 611756, China
[c] Department of Computer and Information Sciences, Northumbria University, Newcastle upon Tyne NE1 8ST, UK
[d] Sichuan Newstrong UHD Video Technology Co., Ltd., Chengdu 610095, China

ARTICLE INFO

Keywords:
Granular computing
Fuzzy rough sets
Granular-ball
Anomaly detection
Outlier detection

ABSTRACT

Most of the existing anomaly detection methods are based on a single and fine granularity input pattern, which is susceptible to noisy data and inefficient for detecting anomalies. Granular-ball computing, as a novel multi-granularity representation and computation method, can effectively compensate for these shortcomings. We utilize the fuzzy rough sets to mine the potential uncertainty information in the data efficiently. The combination of granular-ball computing and fuzzy rough sets takes into account the benefits of both methods, providing great application and research value. However, this novel combination still needs to be explored, especially for unsupervised anomaly detection. In this study, we first propose the granular-ball fuzzy rough set model, and the relevant definitions in the model are given. Subsequently, we pioneeringly present an unsupervised anomaly detection method based on granular-ball fuzzy rough sets called granular-ball fuzzy rough sets-based anomaly detection (GBFRD). Our method introduces the granular-ball fuzzy rough granules-based outlier factor to characterize the outlier degree of an object effectively. The experimental results demonstrate that GBFRD exhibits superior performance compared to the state-of-the-art methods. The code is publicly available at https://github.com/Mxeron/GBFRD.

1. Introduction

Anomaly detection, a pivotal branch in machine learning and data analysis, is used to identify objects or sample points significantly deviating from most data patterns. During data processing in many models, anomalies are usually regarded as a hindrance to improving model performance, and it is desirable to exclude as many anomalies present in the data as possible when performing data preprocessing. However, in some domains, discovering anomalies present in the data is often more valuable than analyzing normal objects therein, such as fraud detection [32], network security [38], and electricity theft detection [42]. Therefore, the study of anomaly detection methods has significant research and practical value.

Based on different rules for mining anomalous objects, existing anomaly detection methods can be broadly classified into statistical-based, distance-based, density-based, clustering-based, and rough set theory-based methods. Among them, the statistics-based methods map each object to a certain probability distribution and consider objects distributed in a low-probability region to be

* Corresponding author.
E-mail addresses: suxinyu@stu.scu.edu.cn (X. Su), yuanzhong@scu.edu.cn (Z. Yuan), farstars@qq.com (B. Chen), pengdz@scu.edu.cn (D. Peng), hmchen@swjtu.edu.cn (H. Chen), yingke.chen@northumbria.ac.uk (Y. Chen).

https://doi.org/10.1016/j.ins.2024.121016
Received 13 March 2024; Received in revised form 6 June 2024; Accepted 9 June 2024

Available online 12 June 2024

anomalies. Distance-based methods use the distance between an object and its nearest neighbors as the basis for anomaly detection, which assumes that an object has a higher probability of being an anomaly if it is farther away from its nearest neighbors [4,41]. Although distance-based anomaly detection methods are very intuitive, in high-dimensional data, these methods are often unable to effectively handle high-dimensional data, where the “curse of dimensionality” exists [29]. The “curse of dimensionality” causes the data to become very sparse and the sample points more dispersed and isolated, which is an excellent challenge for distance-based anomaly detection methods. The density-based anomaly detection method is based on the different density regions in which the objects are located to identify the anomalies, and the most representative method is the local outlier factor (LOF) [3], which was proposed by Breunig et al. In density-based methods, the density estimation is non-parametric and has great interpretability but has a higher computation complexity [27] compared to statistical-based methods. Clustering-based methods are unsupervised, which divide objects into clusters utilizing different clustering algorithms, where the smaller the size of the clusters, the more likely they are to be considered anomalies [6,23,25]. Rough set theory-based methods emphasize the concept of granularity, mining anomalous features through multi-granularity level information in the data to achieve anomaly detection [47]. Moreover, many extended methods of anomaly detection based on rough set theory have been proposed [28,46,47].

Several methods above usually take a single object or a sample point as the input. The finest granularity-based input pattern is susceptible to noise points and is not efficient enough. Moreover, these methods do not consider the granularity of the data and fail to mine the complex information in the data. The performances of these methods can be further improved.

Granular-ball computing (GBC) is an efficient, robust, and highly interpretable multi-granularity representation and computation method [36], which has now gained widespread attention. GB is a novel way of data representation, and each GB can contain some sample points. In the subsequent processing, it is no longer dealing with individual sample points, but GBs and the coarse granularity nature of GBs can reduce the influence of noise [34]. Moreover, the number of GBs generated in the original dataset is much smaller than the number of sample points, thus improving the processing efficiency of the model. Most of the current researches on GBC are mainly divided into two fields, the GB generation methods [34,39] and the application of GBs, such as clustering [5], classification [35], attribute reduction [37] and anomaly detection [2]. Bai et al. [2] proposed an anomaly detection method based on the GB clustering for the first time, which first generates GBs in the original unlabeled data, and then the anomaly detection is achieved based on the number of internal sample points, overlapping relations, and domain changes in each GB. Although this method constructs an effective anomaly detection method based on GBs, it does not consider the uncertainty information, such as fuzziness in the data. Moreover, no research has combined GBC with uncertainty information in unlabeled data to construct an unsupervised anomaly detection method.

Fuzzy rough sets (FRS), an essential granular computing (GrC) method, is often used to deal with the uncertainty information and knowledge representation. FRS combines the advantages of rough set [20] and fuzzy set [7]. FRS simulates the natural patterns in human thinking [45], which emphasizes the different knowledge granularities and helps to solve complex problems in the real world. Soon after FRS was proposed, it received significant attention, and many scholars improved and extended it to propose many new FRS models [26,43,46]. However, to our knowledge, no study has been proposed combining FRS with GBC for anomaly detection. The natural multi-granularity characteristic of GBC enables it to be naturally combined with FRS. Many excellent characteristics of GBC can be applied to anomaly detection, which can further improve the efficiency and robustness of anomaly detection. However, many of the advantages of GBC have yet to be applied to anomaly detection, and applying this method in anomaly detection requires further research.

Based on the above discussion, strengths in GBC and FRS are utilized to construct a novel granular-ball fuzzy rough set (GBFR) model. We innovatively propose a novel unsupervised anomaly detection method based on the model. Subsequently, based on the GB fuzzy approximation accuracy proposed in this study, the outlier degree of a granular-ball fuzzy rough granule (GBG) is calculated to describe the outlier degree of this granule. The outlier factor of an object is calculated based on the outlier degrees of the GBGs containing that object. For a GBG that may contain an anomalous object, the GB fuzzy approximation accuracy between this GBG and multiple fuzzy relations is usually low, and the outlier degree is high. Then, the anomalies are detected by threshold judgments. Specifically, the contributions of this study include

(1) To the best of our knowledge, we are the first to propose the unsupervised anomaly detection method based on granular-ball fuzzy rough sets.
(2) A novel GBFR model is proposed, and a novel anomaly detection method is proposed based on this model.
(3) The outlier factor of an object is constructed and is used to characterize the outlier degree of an object.
(4) The experimental results show that the proposed method outperforms most SOTA methods.

The remainder of this study is organized as follows. Section 2 briefly reviews anomaly detection methods based on rough set theory and granular-ball computing. Section 3 introduces the preliminary knowledge on fuzzy rough sets and granular-ball computing. In Section 4, generalized fuzzy rough sets are constructed based on granular-ball, and the related definitions are given. In Section 5, we build a novel anomaly detection method based on granular-ball fuzzy rough sets and present the corresponding anomaly detection algorithm. The results of our experiments are shown in Section 6. Finally, Section 7 summarizes this study.

2. Related works

This section reviews recent related works on rough set theory-based anomaly detection and granular-ball computing.

2.1. Rough set theory-based anomaly detection

In order to make up for the fact that distance-based and density-based anomaly detection methods cannot effectively deal with nominal attribute data, methods based on rough set theory have been widely studied and proposed. For example, Nguyen et al. [22] proposed a novel anomaly detection method using multi-level approximate reasoning schemes in rough set theory. Xue et al. [40] detected outliers with the help of some labeled samples and rough C-means clustering. Albanese et al. [1] proposed a method for anomaly detection using rough set approximations. Jiang et al. [10] detected sequence-based outliers based on rough set theory. Jiang et al. [11] built an anomaly detection method using information entropy in rough sets. Macia et al. [18] proposed an efficient method for anomaly detection in large volumes of information. Jiang et al. [12] built a novel anomaly detection method based on approximation accuracy entropy.

However, the classical rough set is only suitable for dealing with nominal attribute data and cannot deal with numerical attribute data. FRS characterizes the similarity between objects through fuzzy relations and can directly deal with numerical attribute data without discretization, which greatly preserves the information in the data. Inspired by the idea of FRS, some methods based on FRS have been proposed [31,46,48].

Elhoussaine et al. [8] proposed a highly scalable and fuzzy neighborhood rough set-based anomaly detection method in large-scale data. Yuan et al. [47] defined a new fuzzy relation for mixed attribute data and characterized the outlier degree of an object by fusing the information of multi-fuzzy rough granules to which the object belongs. Mazarbhuiya et al. [21] proposed an intuitionistic fuzzy rough set-based anomaly detection method. In [48], an anomaly detection method was proposed based on a fuzzy rough computing model using multi-fuzzy granules. Wang et al. [31] proposed a novel method for anomaly detection by distance-based fuzzy rough entropy. Yuan et al. [46] proposed an anomaly detection method based on the weighted fuzzy rough density.

Nevertheless, the above methods process input objects individually, and the computation of fuzzy relations between objects is needed. The finest granularity-based computation results in an intolerable increase in time complexity. Furthermore, this computation renders the performance of the methods susceptible to the influence of noise points. Consequently, the detection efficiency and performance of the method can be further improved.

2.2. Granular-ball computing

Existing research on GBC can be broadly divided into two fields: optimizing the GB generation process and applying GBs in multiple tasks. The 2-means clustering algorithm achieves the earliest GB generation. However, the classic generation method needs to manually set the stopping threshold according to different data distributions, and the threshold setting values will affect the quality of the generated GBs [35]. The efficiency and quality of GB generation play a pivotal role in subsequent model learning. In order to accelerate the GB generation and reduce the non-essential computation overheads, Xia et al. [34] proposed an efficient GB generation method by using the k-division algorithm to accelerate the GB generation process while ensuring classification performance. They proposed another adaptive GB generation method, which significantly improves the efficiency of the GB generation. Previous GB generation methods are implemented based on k-means or k-division, resulting in randomness in the generation process. To avoid being affected by randomness, in [39], a set of sample points that have not been divided into GBs is used to divide the GBs. In addition, they introduced anomaly detection for GBs, which improved the efficiency and robustness of GB generation.

Based on efficient GB generation methods, the GB generation process is no longer a bottleneck in constructing an efficient GB learning model. The GB's multi-granularity characteristic and robustness shine in many application fields. Xia et al. [35] proposed an efficient and robust GBC-based classification framework with superior time and space complexity. Based on this, the mathematical models of GBSVM and GBkNN are constructed, which are efficient and perform better than traditional classification methods. An efficient and non-parameter clustering method was proposed in [5], which combined GBs with density peak (DP) clustering algorithm, constructed the density of GBs based on the information of the center and radius, and performed the clustering using the DP algorithm. Furthermore, an effective GBC model called rough GBC was proposed in [24] and applied to multi-label feature selection, which had superior classification performance compared with the SOTA methods. Furthermore, Xia et al. [37] proposed a novel neighborhood rough set based on GBs, which adaptively generated different neighborhoods for each object with linear time complexity.

As a novel method in GrC, GBC has many subfields that deserve further exploration and research. Furthermore, the combination of GBC and FRS for anomaly detection still needs to be explored.

3. Preliminaries

3.1. Fuzzy rough sets

FRS is an effective way to deal with the fuzziness information in data. To facilitate data processing, the information is stored in a table. The data table is also called the information system (IS), which is defined as follows

Definition 1. Given an IS, which is denoted as a 4-tuple $IS = \langle U, A, V, f \rangle$, where $U = \{o_1, o_2, \ldots, o_n\}$ denotes a non-empty and finite set of objects; $A = \{a_1, a_2, \ldots, a_m\}$ denotes a non-empty and finite set of attributes; $V = \bigcup_{a \in A} V_a$ denotes the value domains of all attributes, where V_a denotes the value domain of the attribute a; for any $o_i \in U$ and any $a \in A$, o_i^a denotes the value of o_i with respect to attribute a; $o_i^a \in V_a$.

In this study, we discuss the unsupervised anomaly detection method, where A in IS contains only the set of condition attributes C, and IS can also be represented as $IS = \langle U, C, V, f \rangle$.

Let O denote a map from the universe U to $[0,1]$, which is $O : U \to [0,1]$, then O is a fuzzy set on U. $O(o_i)$ denotes the membership degree of o_i to the fuzzy set O. The set of all fuzzy sets on U is denoted as $\mathcal{F}(U)$. The fuzzy set O is denoted as $O = (O(o_1), O(o_2), \ldots, O(o_n))$ or $O = \sum_{i=1}^{n} \frac{O(o_i)}{o_i}$, where the summation notation is only borrowed.

The fuzzy relation $\mathcal{R}$ on U is defined as $\mathcal{R} : U \times U \to [0,1]$. The set of all fuzzy relations on U is denoted as $\mathcal{F}(U \times U)$. For any $(o_i, o_j) \in U \times U$, the membership degree $\mathcal{R}(o_i, o_j)$ indicates the degree to which o_i has a relation $\mathcal{R}$ with o_j. A fuzzy relation $\mathcal{R}$ on U is represented by a fuzzy relation matrix, $M_{\mathcal{R}} = [r_{ij}]_{n \times n}$, where $r_{ij} = \mathcal{R}(o_i, o_j)$, and each row represents a fuzzy set.

For any $x, y, z \in U$, if $\mathcal{R}$ satisfies

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

then $\mathcal{R}$ is said to be a fuzzy equivalence relation on U. If $\mathcal{R}$ satisfies only (1) and (2), then $\mathcal{R}$ is said to be a fuzzy similarity relation on U.

To better utilize the uncertainty information present in mixed attribute data for knowledge analysis and discovery, Dubois and Prade introduced the concept of upper and lower approximation in fuzzy rough sets [7], which is defined as follows.

Definition 2. Let $\mathcal{R}$ be a fuzzy equivalence relation on U. For any $O \in \mathcal{F}(U)$ and any $x, y \in U$, the membership function of the upper approximation $\overline{\mathcal{R}}O$ and the lower approximation $\underline{\mathcal{R}}O$ of O are defined as

$$\begin{aligned} \overline{\mathcal{R}}O(x) &= \sup_{y \in U} \min\{\mathcal{R}(x,y), O(y)\}, \\ \underline{\mathcal{R}}O(x) &= \inf_{y \in U} \max\{1 - \mathcal{R}(x,y), O(y)\}. \end{aligned} \tag{1}$$

3.2. Granular-ball computing

GBC, as a novel, robust, and efficient GrC method, replaces the previous single granularity sample points with GBs to provide multi-granularity descriptions of the original data rather than considering only the finest granularity sample points [34]. During model learning, the GBs have coarse granularity characteristics, which makes the GBs more robust and interpretable. Moreover, the number of generated GBs is much smaller than the number of objects so that the model can be more efficient.

Due to various factors in the real world, the collected data often contain both noise and anomalies, and it is essential to note that the concepts of anomalies and noise are different. Most of the data are also called observations, which consist of a combination of accurate data and noise. Noise is a random variable in the data that does not conform to the implied pattern or structure of the data, usually introduced when measuring, sampling, or recording data, and is often difficult to analyze and model. Anomalies are data significantly different from the overall distribution of most of the data. This may be due to errors, malfunctions, fraud, unusual events, or other causes that may have recognizable patterns. In analyzing and detecting anomalies, the data may be affected by noise. In contrast, when the GBs are used as input for model learning, the naturally coarse granularity nature of the GB makes it useful for improving robustness in the subsequent model learning process. It reduces the impact of noise on the data.

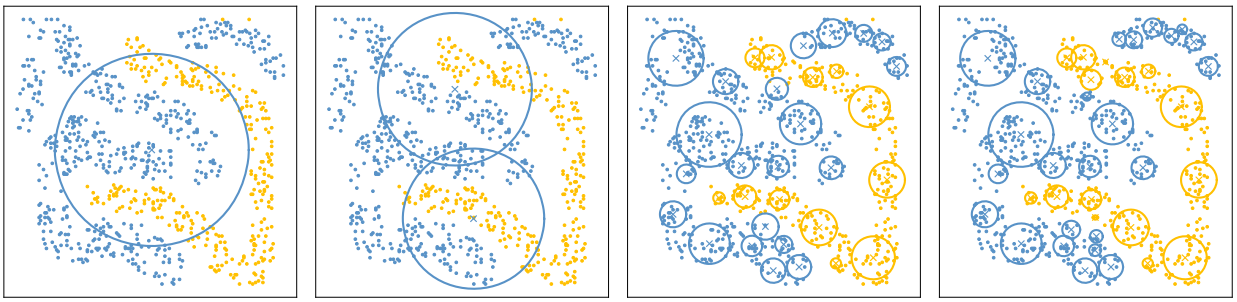
Definition 3. Each granular-ball $GB = \{o_i, i = 1, 2, \ldots, N\}$ can be denoted by two features [44], i.e., its center c and radius r. The relevant representations are as follows.

$$c = \frac{1}{N} \sum_{i=1}^{N} o_i, r = \frac{1}{N} \sum_{i=1}^{N} \delta(o_i, c), \tag{2}$$

where N denotes the number of objects in the GB and δ denotes the distance function. In this study, Euclidean distance is used as the distance function.

From the above definition, it can be seen that the center of a GB is equal to the center of gravity of all the objects inside that GB, and the radius is equal to the average distance from all the objects in the GB to its center. In addition, the radius of the GB can also be defined in terms of the maximum distance, in which case the GBs can cover the original data more comprehensively. Of the two radius calculation methods mentioned above, the average distance is often used in classification tasks because it allows for clear decision boundaries. At the same time, maximum distance is often used in clustering tasks to cover all objects more comprehensively [34].

In the process of GB generation, the 2-means clustering algorithm is a commonly used method for GB division, and this method is very efficient in the process of division [35]. In the GB division, the quality of the GB needs to be assessed by the purity in classification tasks. The quality of the GB is the basis for whether the GB can be further divided. When the quality of the GB reaches a certain threshold, the quality of the GB has already met the requirements, and the GB is not involved in the subsequent division. Instead, the current GB will be divided into two new sub-GBs.



(a) (b) (c) (d)

Fig. 1. The process of GB generation.

In classification tasks, the purity of a GB is used to indicate its quality, with higher purity indicating higher quality. The purity measures are based on the ratio of majority label [19]. The quality of a GB is calculated from the labels of the objects inside the GB, which is undesirable for the unsupervised anomaly detection method in this study because the label data cannot be obtained. Therefore, in the subsequent sections, we apply GB generation to unlabeled data based on existing research.

Fig. 1 demonstrates the process of GB generation in the classification task. At the beginning, all the data is treated as one GB, and then the 2-means clustering algorithm will divide each GB until the quality of all GBs meets the given threshold. After a few rounds of iteration, the number of GBs gradually increases as the division proceeds and can better cover the data than the beginning. As can be seen from the figure, the final generated GBs are multi-granularity, a GB can characterize the typical characteristics of multiple internal sample points, and the number of generated GBs is much less than the number of original objects, using GBs to replace the original sample points as model inputs, can build efficient models.

4. Granular-ball fuzzy rough sets

Definition 4. For any $B \subseteq C$ is a subset of condition attributes. For any $o_i, o_j \in U$, the fuzzy relation $\mathcal{R}_B$ between o_i and o_j induced by B is calculated as

$$\mathcal{R}_B(o_i, o_j) = \begin{cases} F_B(o_i, o_j), & \text{if } F_B(o_i, o_j) \geq \sigma; \\ 0, & \text{if } F_B(o_i, o_j) < \sigma; \end{cases} \tag{3}$$

where $F_B(o_i, o_j) = 1 - \frac{1}{|B|}\sqrt{\sum_{c_k \in B} |o_i^{c_k} - o_j^{c_k}|^2}$ and σ is an adjustable parameter or a hyper-parameter, where $o_i^{c_k}$ and $o_j^{c_k}$ are the values of the attribute c_k of the objects o_i and o_j, which are in the range $[0, 1]$ because of the normalization of the data before anomaly detection. Then, the ranges of $|o_i^{c_k} - o_j^{c_k}|^2$ and $1 - \frac{1}{|B|}\sqrt{\sum_{c_k \in B} |o_i^{c_k} - o_j^{c_k}|^2}$ are also in $[0, 1]$. Therefore, the inequality $0 \leq F_B(o_i, o_j) \leq 1$ holds. Obviously, the fuzzy relation $\mathcal{R}_B$ is reflexive and symmetric, i.e., it is a fuzzy similarity relation. The family of fuzzy information granules $G(\mathcal{R}_B)$ induced by $\mathcal{R}_B$ with respect to any $o_i \in U$ is defined as

$$G(\mathcal{R}_B) = \{[o_1]_{\mathcal{R}_B}, [o_2]_{\mathcal{R}_B}, \ldots, [o_n]_{\mathcal{R}_B}\}, \tag{4}$$

where $[o_i]_{\mathcal{R}_B} = \{(o_1, r_{i1}^B), (o_2, r_{i2}^B), \ldots, (o_n, r_{in}^B)\} = \sum_{j=1}^{n} \frac{r_{ij}^B}{o_j}$. $[o_i]_{\mathcal{R}_B}$ is a fuzzy information granule in $G(\mathcal{R}_B)$ induced by $\mathcal{R}_B$. Each $[o_i]_{\mathcal{R}_B}$ is an n-tuple, which contains the fuzzy similarity degree between the object o_i on $\mathcal{R}_B$ and all other objects (including itself). $r_{ij}^B = \mathcal{R}_B(o_i, o_j)$ denotes the fuzzy similarity degree of o_i and o_j on $\mathcal{R}_B$. The cardinality of $[o_i]_{\mathcal{R}_B}$ is calculated as $|[o_i]_{\mathcal{R}_B}| = \sum_{j=1}^{n} \mathcal{R}_B(o_i, o_j)$. Without causing confusion, B is used to replace $\mathcal{R}_B$ in this study. $\mathcal{R}_B$ is defined as a fuzzy relation matrix $M_{\mathcal{R}_B} = [r_{ij}^B]_{n \times n}$.

The GB has a natural multi-granularity characteristic, and the GBs generated on different subsets of attributes can produce different coverage of the original dataset. Moreover, a GB can characterize the common properties of the objects in the GB, and when constructing the outlier factor of an object, a single object is no longer considered but the GB to which the object belongs. The GB fuzzy relation and related concepts in GBFR will be introduced below. Considering fuzzy similarity relations between all objects in one GB and all in another GB in all GBs may give better results but introduces a significant time overhead. In order to simplify the calculation of the above equation and improve the efficiency of the calculation of the GB fuzzy relation, we replace the GB with its corresponding center.

Definition 5. For any $B \subseteq C$ is a condition attribute subset, then $GB_B = \{GB_1^B, GB_2^B, \ldots, GB_k^B\}$ is a GB set generated on B, where k denotes the number of GBs contained in GB_B. For any $o_i, o_j \in U$, $\widetilde{GB_i^B}$ and $\widetilde{GB_j^B}$ are the nearest GBs generated on B to o_i and o_j respectively. The GB fuzzy relation $\mathcal{R}_B^{GB}$ on U induced by GB_B is defined as

$$\mathcal{R}_B^{GB}(o_i,o_j)=\mathcal{R}_B^{GB}(\widetilde{GB_i^B},\widetilde{GB_j^B})=\frac{1}{|\widetilde{GB_i^B}||\widetilde{GB_j^B}|}\sum_{o_e\in\widetilde{GB_i^B}}\sum_{o_f\in\widetilde{GB_j^B}}\mathcal{R}_B(o_e,o_f), \tag{5}$$

where $|\cdot|$ denotes the size of a GB, i.e., the number of objects in that GB. The nearest GB to which an object belongs is the GB with the closest Euclidean distance to that object.

Considering fuzzy similarity relations between all objects in one GB and all in another GB in all GBs may give better results but introduces a significant time overhead. In order to simplify the calculation of the above equation and improve the efficiency of the calculation of the GB fuzzy relation, we replace the GB with its corresponding center. Let c_i and c_j be the centers of $\widetilde{GB_i^B}$ and $\widetilde{GB_j^B}$ respectively, then $\mathcal{R}_B^{GB}(o_i,o_j)$ is calculated as $\mathcal{R}_B(c_i,c_j)$. Obviously, after substitution, $\mathcal{R}_B^{GB}$ is reflexive and symmetric, i.e., it is a fuzzy similarity relation. The family of GB fuzzy information granules $G(\mathcal{R}_B^{GB})$ with respect to U is defined as

$$G(\mathcal{R}_B^{GB})=\{[o_1]_B^{GB},[o_2]_B^{GB},\dots,[o_n]_B^{GB}\}. \tag{6}$$

The cardinality of $[o_i]_B^{GB}$ is calculated as $|[o_i]_B^{GB}|=\sum_{j=1}^{n}\mathcal{R}_B^{GB}(o_i,o_j)$. Similarly, $M_{\mathcal{R}_B^{GB}}$ is a fuzzy relation matrix.

Definition 6. For any $S,B\subseteq C$ and any $[o_i]_B^{GB}\in G(\mathcal{R}_B^{GB})$, the upper approximation $\overline{\mathcal{R}_S^{GB}}[o_i]_B^{GB}$ and the lower approximation $\underline{\mathcal{R}_S^{GB}}[o_i]_B^{GB}$ of $[o_i]_B^{GB}$ with respect to $\mathcal{R}_S^{GB}$ are two fuzzy sets on U, and their membership functions are

$$\begin{aligned}\overline{\mathcal{R}_S^{GB}}[o_i]_B^{GB}(o)&=\sup_{p\in U}\min\{\mathcal{R}_S^{GB}(o,p),[o_i]_B^{GB}(p)\},\\ \underline{\mathcal{R}_S^{GB}}[o_i]_B^{GB}(o)&=\inf_{p\in U}\max\{1-\mathcal{R}_S^{GB}(o,p),[o_i]_B^{GB}(p)\}.\end{aligned} \tag{7}$$

In rough sets, the approximation accuracy only applies to nominal attribute data and cannot handle numerical attribute data. In order to handle numerical attribute data directly, the fuzzy approximation accuracy is introduced. Similarly, we define the GB fuzzy approximation accuracy based on the condition attributes.

Definition 7. For any $B\subseteq C$, where $|C-B|\geq 2$. For any $[o_i]_B^{GB}\in G(\mathcal{R}_B^{GB})$ and $S\subseteq C-B$, the GB fuzzy approximation accuracy of $[o_i]_B^{GB}$ with respect to $\mathcal{R}_S^{GB}$ is defined as

$$\alpha_{\mathcal{R}_S^{GB}}([o_i]_B^{GB})=\frac{|\underline{\mathcal{R}_S^{GB}}[o_i]_B^{GB}|}{|\overline{\mathcal{R}_S^{GB}}[o_i]_B^{GB}|}. \tag{8}$$

With the above definitions, we construct a new GBFR model, which gives the definitions of the GB fuzzy relations between objects and the GB fuzzy approximation accuracy. The GB fuzzy relation is based on multi-granularity GBs instead of the finest granularity of object-to-object calculation, which is robust to noise and improves the model performance. The next step is to propose a novel unsupervised anomaly detection method called GBFRD based on the above-constructed GBFR model.

5. GBFR-based anomaly detection

This section proposes an anomaly detection method called GBFRD based on GBFR. First, the unsupervised GB generation algorithm is introduced, and the corresponding pseudo-code is given in this study. Subsequently, the outlier degrees of the GBGs are constructed based on GBFR, and the object-specific outlier factors are further constructed. We also give specific anomaly detection examples to help understand the logic of the method and the detection process. The framework diagram of GBFRD is shown in Fig. 2.

The framework diagram demonstrates the whole process of anomaly detection with GBFRD. First, the GBs are generated, and the GB fuzzy relation matrices are constructed. Based on the GB fuzzy relation matrices, the outlier degree GBOD of the GBG and the corresponding weight information are calculated. The outlier factors of each object are calculated by fusing the outlier degree and the corresponding weight information of multiple GBGs to which the object belongs. Finally, the outlier factors are compared with the threshold to determine the specific anomalies and achieve anomaly detection.

5.1. Unsupervised granular-ball generation method

The quality of GB is an integral part of GB generation. In the process of GB generation, it is necessary to calculate the quality of the GBs to determine whether they need to be further divided. Until the quality of all the pellets meets the requirements, the entire process of GB generation is completed. The quality of GBs dramatically affects the performance of subsequent processing [34]. The original quality of a GB is defined using the labels of the objects in that GB in classification tasks [35]. Therefore, the original GB generation method is supervised, which is ineffective for learning in unlabeled data. In this study, GBs are introduced to build

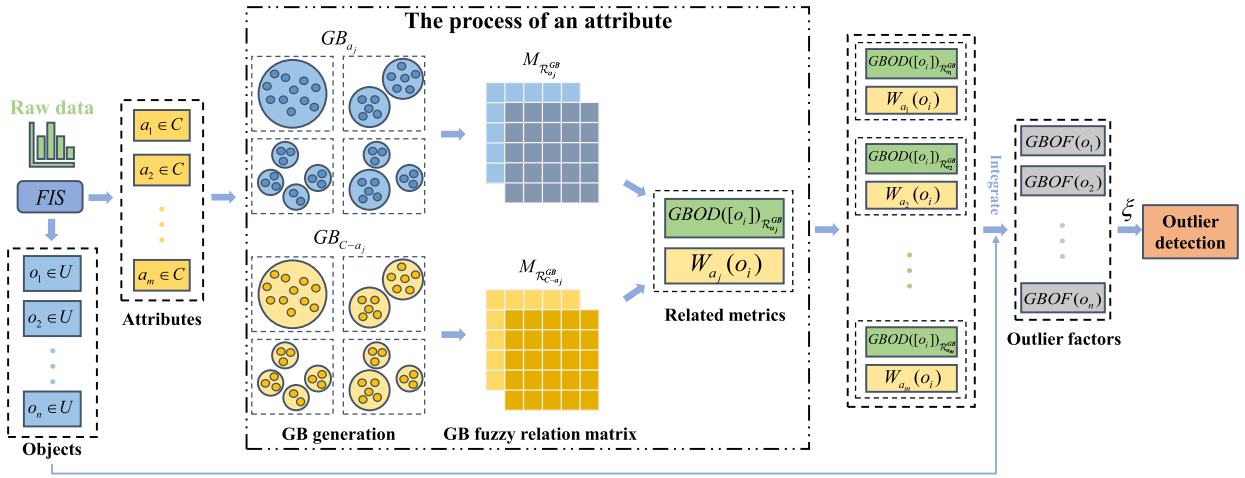


Fig. 2. The framework diagram of GBFRD.

Algorithm 1: GB generation.

Input: A dataset $O = \{o_1, o_2, \ldots, o_n\}$, a set of GBs G
Output: G
Initializing: $T \leftarrow \sqrt{|O|}$, $gb \leftarrow O$, $G \leftarrow \emptyset$;
Initialize a queue $Q \leftarrow \emptyset$, put gb into Q;
while $Q \neq \emptyset$ **do**
  Get the top gb from Q and remove this gb from Q;
  if $|gb| > T$ **then**
    Divide gb into gb_{sub_1} and gb_{sub_2} by 2-means algorithm;
    Put gb_{sub_1} and gb_{sub_2} into the tail of Q;
  end
  if $|gb| \leq T$ **then**
    $G \leftarrow G \cup gb$;
  end
end
return G.

an unsupervised detection method, so the labels in objects fail to represent the quality of GBs. GBs need to be generated using an unsupervised method.

Cheng et al. [5] argued that having fewer objects within a GB results in more GBs, leading to broader coverage of all objects and increased similarity among the objects in a GB. In this study, similar to the quality calculation method in [5], to apply GBs in unsupervised learning, the size of a GB is taken as the quality, i.e., the number of objects within each GB should not exceed $\sqrt{n}$.

We give the pseudo-code for the GB generation algorithm in this study. As shown in Algorithm 1, the 2-means algorithm is a particular case of the k-means algorithm where k is taken as 2. The time complexity of the k-means algorithm is $O(ktn)$, where k is the number of clusters. Since the 2-means clustering algorithm has a fast convergence speed, it can be considered as an approximately linear algorithm [35]. Therefore, the time complexity of generating the largest GB is also linear, i.e., the time complexity of Algorithm 1 can be considered to be approximately $O(n)$ [35].

5.2. Anomaly detection method

The GB fuzzy approximation accuracy characterizes the approximation accuracy between a GBG and a set of fuzzy relations, which can be used to measure the uncertainty of GBGs. We characterize the outlier degree of a GBG by the GB fuzzy approximation accuracy. Considering that the outlier degree of an object is often more valuable than determining whether the object is an anomaly in many cases, the outlier factor based on the GBGs is introduced to calculate the outlier degree of the object.

Definition 8. For any $B \subseteq C$, let $S = C - B$. For any $o_i \in U$, the outlier degree GBOD of $[o_i]_B^{GB}$ is defined as

$$GBOD([o_i]_B^{GB}) = 1 - \frac{1}{|U|}|[o_i]_B^{GB}| \cdot \alpha_{\mathcal{R}_S^{GB}}([o_i]_B^{GB}). \tag{9}$$

The outlier degree of $[o_i]_B^{GB}$ is calculated from its approximation accuracy. Given a set of fuzzy relations on U, if the approximation accuracy of $[o_i]_B^{GB}$ on these relations is always low, then it is assumed that o_i is anomalous and that the value of $GBOD([o_i]_B^{GB})$ will be high.

The GBGs to which an object belongs can characterize the features of the object from multi-granularity levels. For any $o_i \in U$ and $B \subseteq C$, the outlier factor of o_i is defined by the outlier degree of the GBGs containing o_i. Next, we calculate the outlier factor of the object by the GB fuzzy approximation accuracy of the GBGs to describe the outlier degree of the object.

Definition 9. Let $C = \{a_1, a_2, \ldots, a_m\}$. For any $o_i \in U$, the outlier factor GBOF of o_i is defined as

$$GBOF(o_i) = \frac{1}{|C|} \sum_{k=1}^{m} (GBOD([o_i]_{a_k}^{GB}) \cdot W_{a_k}(o_i)), \tag{10}$$

where $W_{a_k}(o_i) = 1 - \sqrt{\frac{|[o_i]_{a_k}^{GB}|}{|U|}}$.

The outlier factor of an object o_i is calculated by the outlier degree of the multiple GBGs to which it belongs. The weight of an object under an attribute $W_{a_k}(o_i)$ is calculated by the cardinality of the GBG under the GB fuzzy relation induced by the attribute to which it belongs. The cardinality of a GBG reflects the sum of similarities between the object and the rest of the objects under the current fuzzy relation; the smaller the cardinality indicates that the object is more likely to be a minority class, which means that the object is more likely to be an anomaly. So, a greater weight is assigned to it in the calculation.

Theoretically, any $B \subseteq C$ can determine a fuzzy relation, and each fuzzy relation can obtain a GBG containing o_i. A total of $2^{|C|}$ GBGs containing o_i are obtained. These exponentially possible cases are unrealistic for practical computation. To reduce the computation complexity, in the computation of $GBOF(o_i)$, the m GBGs composed attribute by attribute in C are replaced by the previous theoretical $2^{|C|}$ GBGs.

When calculating the outlier factor of an object, the outlier degrees of GBGs containing that object are fused with their corresponding weight. The smaller cardinality of a GBG indicates that the object is more likely to be a minority class in this GB fuzzy relation, which means that the object is more likely to be an anomaly. Consequently, assigning greater weight to these GBGs during the calculation process is prudent. A higher outlier factor indicates a higher outlier degree for the given object. A predefined threshold can be set to demarcate anomalies from normal objects if there is a need to distinguish certain anomalies from normal objects specifically.

Definition 10. Given an anomaly determination threshold ξ. For any $o \in U$, if $GBOF(o) > \xi$, o is said to be an anomaly in U.

5.3. Anomaly detection algorithm

Algorithm 2: GBFRD.

Input: $IS = \langle U, C, V, f \rangle, \sigma$, outlier factors $GBOF$
Output: $GBOF$

```
Initializing: GBOF ← ∅
for j ← 1 to |C| do
    Generate GB_{a_j} and GB_{C−a_j} by Algorithm 1;
    Calculate M_{R^{GB}_{a_j}} and M_{R^{GB}_{C−a_j}} by Eq. (5);
    for i ← 1 to |U| do
        Calculate α_{R^{GB}_{C−a_j}}([o_i]^{GB}_{a_j}) by Eq. (8);
        Calculate GBOD([o_i]^{GB}_{a_j}) and W_{a_j}(o_i) by Eq. (9);
    end
end
for i ← 1 to |U| do
    Calculate GBOF(o_i) by Eq. (10);
end
return GBOF.
```

Algorithm 2 shows the implementation of the GBFRD algorithm based on the GBGs. First, an IS containing only the condition attributes C and an adjustable parameter σ is input. Iterating over each attribute a_j in C, GB_{a_j} and GB_{C-a_j} are generated according to Algorithm 1, and then the corresponding GB fuzzy relation matrices $M_{R_{a_j}^{GB}}$ and $M_{R_{C-a_j}^{GB}}$ are calculated according to Eq. (5). Then the approximation accuracy $\alpha_{R_{C-a_j}^{GB}}([o_i]_{a_j}^{GB})$ is calculated according to Eq. (8). Then, according to Eq. (9), the outlier degree of different GBGs $GBOD([o_i]_{a_j}^{GB})$ and the corresponding weight $W_{a_j}(o_i)$ are calculated based on the corresponding outlier degree of $GBOD([o_i])$. Finally, the outlier factor $GBOF(o_i)$ of o_i is calculated according to Eq. (10).

Table 1
A data table.

U	a_1	a_2	a_3
o_1	0.3	0.2	3
o_2	0.1	0.5	2
o_3	0.1	0.6	5
o_4	0.2	0.3	6

In the "for" loop from Steps 2 to 9, the GBs are first generated according to Algorithm 1, whose time complexity is $O(|U|)$. When calculating the fuzzy relation in Step 4, the fuzzy relation between GBs is first calculated. Assuming that the number of GBs in these two GB sets generated in Step 3 is $|GB_1|$ and $|GB_2|$, the number of iterations of Step 4 is $|C|(|GB_1||GB_1| + |GB_2||GB_2|)$. When calculating the fuzzy relation between the objects, retrieving the fuzzy relation between two GBs belonging to two objects is only necessary. The retrieval time can be considered to be approximately ignored. In Steps 6 and 7, the number of iterations is $|C||U|$. The number of iterations in Step 11 is $|U|$. The final time complexity of Algorithm 2 can be approximately considered to be $O(|C|(|U| + |GB_1||GB_1| + |GB_2||GB_2|))$. The sizes of $|GB_1|$ and $|GB_2|$ may vary in the actual computation, but they are always much smaller than the total number of objects $|U|$.

5.4. Anomaly detection example

Given a data table, as shown in Table 1, where $U = \{o_1, o_2, o_3, o_4\}$ and $C = \{a_1, a_2, a_3\}$. Each object contains three attributes. In data processing, there are usually differences in the order of magnitude; to ensure the reliability and consistency of the processing results, it is necessary first to standardize the numerical attribute data, and min-max normalization is used for preprocessing.

For $C = \{a_1, a_2, a_3\}$ and let $\sigma = 0.5$. According to Eq. (6), the fuzzy relation matrices on each attribute in C are calculated as

$$M_{\mathcal{R}^{GB}_{a_1}} = \begin{bmatrix} 1 & 0 & 0 & 0.50 \\ 0 & 1 & 1 & 0.50 \\ 0 & 1 & 1 & 0.50 \\ 0.50 & 0.50 & 0.50 & 1 \end{bmatrix}, M_{\mathcal{R}^{GB}_{a_2}} = \begin{bmatrix} 1 & 0 & 0 & 0.75 \\ 0 & 1 & 0.75 & 0.50 \\ 0 & 0.75 & 1 & 0 \\ 0.75 & 0.50 & 0 & 1 \end{bmatrix}, M_{\mathcal{R}^{GB}_{a_3}} = \begin{bmatrix} 1 & 0.75 & 0.50 & 0 \\ 0.75 & 1 & 0 & 0 \\ 0.50 & 0 & 1 & 0.75 \\ 0 & 0 & 0.75 & 1 \end{bmatrix}.$$

For any $[o_i]^{GB}_{a_1} \in G(\mathcal{R}^{GB}_{a_1})$, its GB fuzzy approximation accuracy with respect to $\mathcal{R}^{GB}_{C-a_1}$ is calculated as $\alpha_{\mathcal{R}^{GB}_{C-a_1}}([o_1]^{GB}_{a_1}) \approx 0.2918$, $\alpha_{\mathcal{R}^{GB}_{C-a_1}}([o_2]^{GB}_{a_1}) = \alpha_{\mathcal{R}^{GB}_{C-a_1}}([o_3]^{GB}_{a_1}) \approx 0.4021$, $\alpha_{\mathcal{R}^{GB}_{C-a_1}}([o_4]^{GB}_{a_1}) \approx 0.7381$.

According to Eq. (9), the outlier degrees of $[o_i]^{GB}_{a_1}$ is calculated as $GBOD([o_1]^{GB}_{a_1}) = 1 - \frac{1}{|U|}|[o_1]^{GB}_{a_1}| \cdot \alpha_{\mathcal{R}^{GB}_{C-a_1}}([o_1]^{GB}_{a_1}) = 1 - \frac{1}{4} \times 1.50 \times 0.2918 \approx 0.8906$, $GBOD([o_2]^{GB}_{a_1}) = GBOD([o_3]^{GB}_{a_1}) \approx 0.7487$, $GBOD([o_4]^{GB}_{a_1}) \approx 0.5387$. Similarly, the outlier degrees of $[o_i]^{GB}_{a_2}$ and $[o_i]^{GB}_{a_3}$ are calculated. $GBOD([o_1]^{GB}_{a_2}) = GBOD([o_3]^{GB}_{a_2}) \approx 0.8177$, $GBOD([o_2]^{GB}_{a_2}) \approx 0.6780$, $GBOD([o_4]^{GB}_{a_2}) \approx 0.7340$. $GBOD([o_1]^{GB}_{a_3}) \approx 0.8107$, $GBOD([o_2]^{GB}_{a_3}) \approx 0.9471$, $GBOD([o_3]^{GB}_{a_3}) \approx 0.8355$, $GBOD([o_3]^{GB}_{a_4}) \approx 0.9451$.

Based on the above existing outlier degrees of multiple GBGs, according to Eq. (10), the outlier factor of o_1 is calculated as $GBOF(o_1) = \frac{1}{|C|}\sum_{k=1}^{m}(GBOD([o_1]^{GB}_{a_k}) \cdot W_{a_k}(o_1)) = \frac{1}{3} \times (0.8906 \times (1 - \sqrt{\frac{1.5}{4}}) + 0.8177 \times (1 - \sqrt{\frac{1.75}{4}}) + 0.8107 \times (1 - \sqrt{\frac{2.25}{4}})) \approx 0.2749$. Similarly, the outlier factors of o_2, o_3 and o_4 are calculated as $GBOF(o_2) \approx 0.2157$, $GBOF(o_3) \approx 0.2142$, $GBOF(o_4) \approx 0.2054$.

The outlier factor describes the outlier degree of an object. Suppose we set the threshold $\xi = 0.25$; in the above calculation result, only $GBOF(o_1) \approx 0.2749 > 0.25$. It is said that o_1 is an anomaly based on the GBGs.

6. Experiments

This section begins by describing the relevant settings for the experiments. Subsequently, the performance among the methods is visually compared using the ROC curves and boxplots, and we give a table of the AUC values for all methods. In addition, statistical analyses are performed to test statistical differences. Eventually, parameter sensitivity analyses are performed to test the robustness of the proposed method.

6.1. Experimental settings

To validate the effectiveness of our method, we conduct several experiments on several publicly available datasets.[1,2] For these datasets, the number of attributes ranges from 4 to 279, the number of samples ranges from 111 to 9172, and the number of anomalies ranges from 5 to 176. The datasets used are shown in Table 2.

We compare our method with several SOTA anomaly detection methods, including multi-fuzzy granules anomaly detection (MFGAD, 2023) [48], weighted fuzzy-rough density-based anomaly (WFRDA, 2023) [46], empirical cumulative distribution-based outlier detection (ECOD, 2022) [15], weighted neighborhood information network-based outlier detection (WNINOD, 2021) [33],

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

Table 2
Experimental datasets.

No.	Datasets	Abbr.	Number of attributes	Number of samples	Number of anomalies	Anomaly ratio
1	Audiology_variant1	Audio	69	226	53	23.5%
2	Lymphography	Lymp	8	148	6	4.1%
3	Cardio	Card	21	1831	176	9.6%
4	Cardiotocography_2and3_33_variant1	Cardio	21	1688	33	2.0%
5	Diabetes_tested_positive_26_variant1	Diab	8	526	26	4.9%
6	Iris_Irisvirginica_11_variant1	Iris	4	111	11	9.9%
7	Pima_TRUE_55_variant1	Pima	9	555	55	9.9%
8	Thyroid	Thyro	6	3772	93	2.5%
9	Wbc_malignant_39_variant1	Wbc	9	483	39	8.1%
10	Wdbc_M_39_variant1	Wdbc	31	396	39	9.9%
11	Yeast_ERL_5_variant1	Yeast	8	1141	5	0.4%
12	Arrhythmia_variant1	Arr	279	452	66	14.6%
13	CreditA_plus_42_variant1	Cred	15	425	42	9.9%
14	Horse_1_12_variant1	Horse	27	256	12	4.7%
15	Sick_sick_35_variant1	Sick	29	3576	35	1.0%
16	Thyroid_disease_variant1	Thyr	28	9172	74	0.8%

variance structural score (VarE, 2020) [13], copula-based outlier detection (COPOD, 2020) [14], virtual outlier score (VOS, 2019) [30], local projection-based outlier detection (LPOD, 2018) [17], natural outlier factor (NOF, 2016) [9] and isolation forest (IForest, 2012) [16].

In the above methods, MFGAD and WFRDA mainly involve the parameter σ. In order to determine the optimal parameters on different datasets, we set the adjustable range of σ to $[0.1, 2]$ with a step size of 0.1. VOS and LPOD are affected by their internal parameter k. The adjustable range of k to $[1, 60]$ with a step size of 1. The adjustable range of the parameter for WNINOD is $[1, 10]$ with a step size of 1. For VarE, we set the candidate values of λ to $\{1000, 100, 10, 1, 0.1, 0.01, 0.001\}$. For IForest, which is mainly affected by the number of base estimators, we set this number to 100. For GBFRD, there is an adjustable parameter σ in the calculation of fuzzy similarity; we set its adjustable range to $[0, 1]$ with a step size of 0.05. We consider the comparison between the different methods in our experiments to be fair. However, it is not absolutely fair. Some methods have an infinite range of adjustable parameters. Therefore, we can only approximate the optimal case as much as possible.

The ROC curve and AUC index are often used to evaluate the performance of different methods [46]. They are straightforward, intuitive, and easy to compare. The ROC curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR), providing insight into how well the model distinguishes between classes. This is crucial in applications where false positives and false negatives have different costs or implications. The AUC provides a single aggregate performance measure across all possible classification thresholds. A higher AUC value indicates a better model performance, which is easy to interpret and compare.

6.2. Experimental results

Fig. 3 shows the ROC curves of 11 methods on 16 datasets, where the orange color is the curve of GBFRD, and it can be seen from the figure that GBFRD outperforms the other methods on Audio, Lymp, Cardio, Yeast, Arr, Horse and Sick. However, the high number of ROC curves plotted in a single figure makes the distinction between some methods unclear. So, below, we compare the AUC values of the 11 methods on different datasets.

The AUC values of different methods are shown in Table 3, where the best AUC values in each dataset are bolded. Through this table, we can compare the performance of different methods more clearly and intuitively. Among them, GBFRD performs best on most datasets, e.g., Audio, Lymp, Cardio, Iris, Wbc, Yeast, etc. However, on Card, Diab, Pima, etc., the AUC values of GBFRD are slightly lower than those of the other better methods. From the point of view of the average AUC values, the value of GBFRD is 0.942, which is significantly better than other methods, verifying the effectiveness of our method.

Moreover, as illustrated in Fig. 4, the AUC boxplots for various methods across all datasets are presented. These boxplots serve as valuable tools for visualizing the distribution of AUC values for different methods when applied to diverse datasets, thereby offering insights into the stability of these methods. A careful examination of the figure reveals that the boxplot representing GBFRD exhibits a notably more compact and upward form. This observation implies that GBFRD demonstrates superior performance and showcases commendable stability, underlining its efficacy as a robust anomaly detection method.

The running times of the different methods are shown in Table 4, from which it can be seen that the time overhead of GBFRD is slightly higher. The reasons for this are as follows. To improve the method performance, we utilize the GB fuzzy approximation accuracy to characterize the outlier degree of the GBG, which achieves desirable results, but the computation process is time-consuming. Two fuzzy relation matrices $M_{R_{a_j}^{GB}}$ and $M_{R_{C-a_j}^{GB}}$ need to be calculated in each iteration. From the previous analysis, the theoretical time complexity of GBFRD is $O(|C|(|U| + |GB_1||GB_1| + |GB_2||GB_2|))$, which neglects the GB fuzzy relations mapped back to the objects and has a time overhead. However, there is still a discrepancy between this theoretical analysis and the actual running time. When the dataset is large, this part of the mapping time is also non-negligible, leading to a higher overall running time.

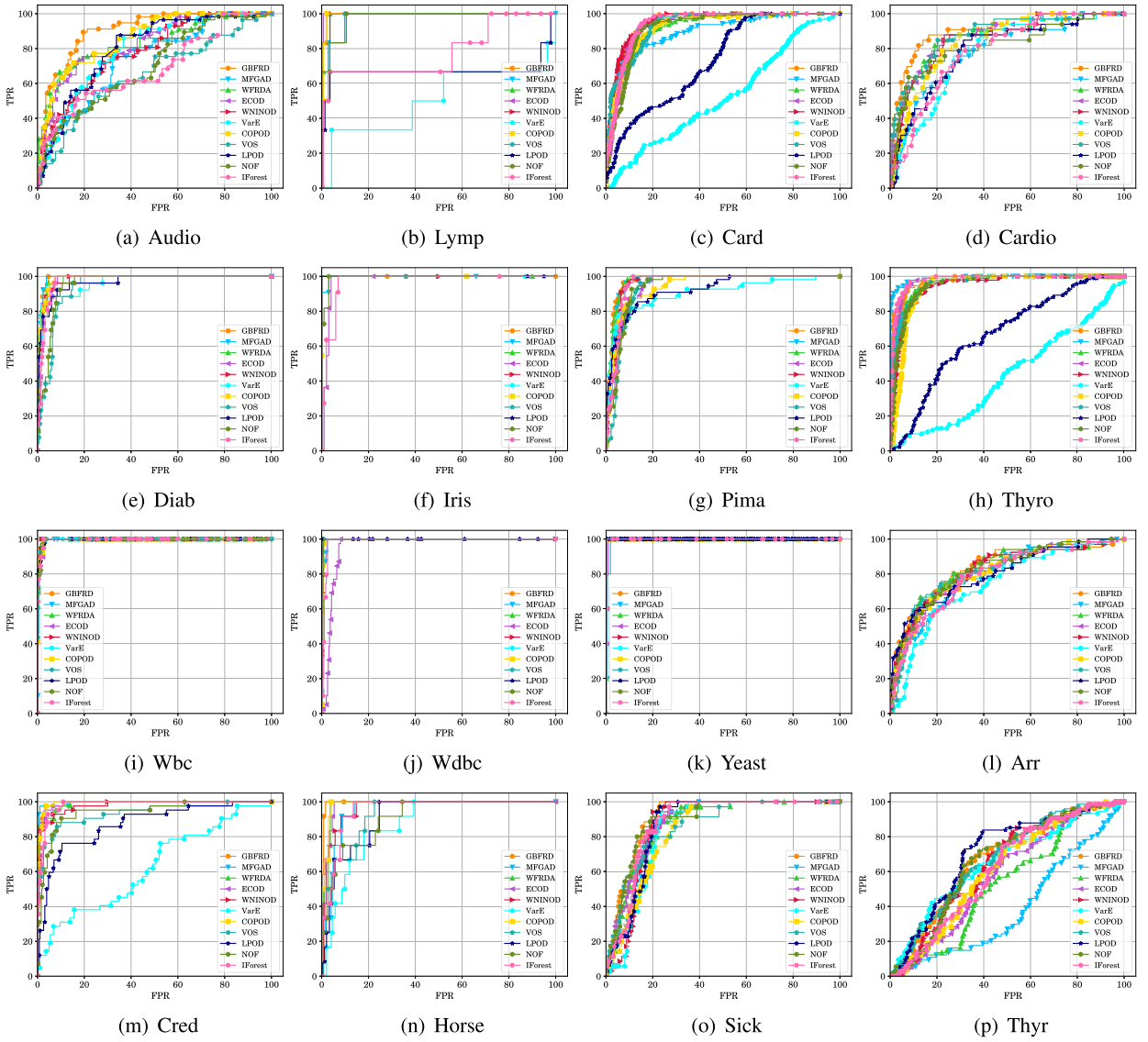


Fig. 3. Experimental comparison results on ROC.

6.3. Statistical analysis

Following the previous works [46–48], we conduct the statistical analysis of the above experimental results. First, the Friedman test is used to determine whether different methods achieve the same performance. After that, the Nemenyi test further discriminates the methods, and the Nemenyi test figure visualizes the differences between the methods.

As seen in Table 3, 11 methods and 16 datasets are utilized in our experiment. The F distributions with freedom degrees of 10 and 150 can be obtained. According to the Friedman test, when $\alpha = 0.05$, the value of $\tau_F = 4.8688$ is greater than the critical value 1.8943. Therefore, the null hypothesis that 'all the methods have the same performance' does not hold. In other words, the methods are significantly different, and a post-hoc test is needed to further distinguish between them.

According to the Nemenyi test, when the significance level $\alpha = 0.05$, the corresponding critical distance $CD_{0.05} = 3.7746$. By plotting the average ordinal value of all methods and the line segment of CD length on a single axis, the Nemenyi test figure can be obtained. As shown in Fig. 5, when there is a red horizontal line segment CD covering between some methods in the figure, it is assumed that there are no statistical differences between these methods. It can be seen that there is no horizontal line segment coverage between GBFRD, ECOD, COPOD, VOS, LPOD, NOF, MFGAD, IForest, VarE, and it is considered that there are significant statistical differences between GBFRD and these methods. However, there is no consistent evidence of statistical differences between GBFRD, WFRDA, and WNINOD.

Table 3
Experimental results on AUC.

Datasets	GBFRD (Ours)	MFGAD	WFRDA	ECOD	WNINOD	VarE	COPOD	VOS	LPOD	NOF	IForest
Audio	**0.903**	0.735	0.834	0.833	0.778	0.792	0.860	0.651	0.799	0.694	0.685
Lymp	**0.996**	0.978	0.993	**0.996**	0.992	0.507	0.994	0.974	0.673	0.977	0.781
Card	0.948	0.897	0.922	0.935	**0.949**	0.526	0.922	0.933	0.732	0.920	0.941
Cardio	**0.903**	0.784	0.865	0.871	0.851	0.780	0.844	0.882	0.791	0.817	0.788
Diab	0.987	**0.989**	0.984	0.979	0.981	0.965	0.986	0.942	0.966	0.954	0.976
Iris	**1.000**	0.997	**1.000**	0.977	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	0.994	0.971
Pima	0.972	0.971	**0.974**	0.947	0.963	0.904	0.942	0.935	0.920	0.942	0.957
Thyro	0.977	**0.989**	0.957	0.977	0.943	0.432	0.939	0.959	0.668	0.949	0.980
Wbc	**0.998**	0.994	0.997	0.995	0.997	0.997	0.995	0.995	0.997	0.997	0.996
Wdbc	0.998	0.996	0.999	0.959	0.996	0.997	0.996	**1.000**	**1.000**	0.995	0.987
Yeast	**1.000**	0.992	0.998	0.995	0.998	**1.000**	0.997	**1.000**	**1.000**	**1.000**	0.997
Arr	**0.829**	0.790	0.826	0.807	0.815	0.739	0.805	0.813	0.799	0.816	0.788
Cred	0.994	**0.995**	0.986	0.990	0.979	0.625	0.992	0.934	0.873	0.941	0.983
Horse	**0.995**	0.951	0.982	0.981	0.966	0.869	0.986	0.929	0.911	0.908	0.954
Sick	**0.907**	0.860	0.868	0.883	0.864	0.842	0.839	0.862	0.867	0.894	0.881
Thyr	0.663	0.388	0.531	0.581	0.640	0.650	0.611	0.671	**0.708**	0.662	0.614
Average	**0.942**	0.894	0.920	0.919	0.920	0.789	0.919	0.905	0.856	0.904	0.892

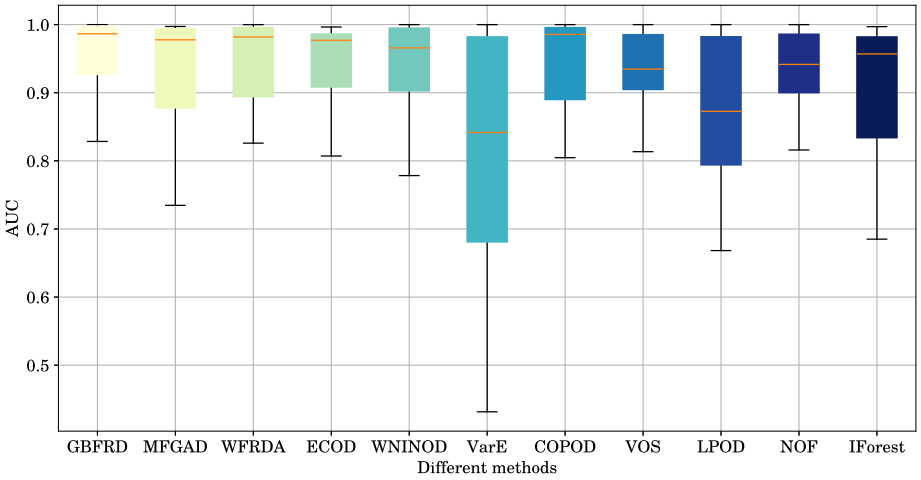


Fig. 4. The AUC boxplots for each method across 16 datasets.

Table 4
Running times of different methods (unit: seconds).

Datasets	GBFRD (Ours)	MFGAD	WFRDA	ECOD	WNINOD	VarE	COPOD	VOS	LPOD	NOF	IForest
Audio	1.328	0.216	0.784	1.055	0.079	0.584	0.218	0.349	0.015	1.598	0.104
Lymp	0.153	0.023	0.090	0.004	0.010	0.004	0.004	0.147	0.016	0.338	0.078
Card	42.281	13.071	32.919	0.004	13.236	3.531	0.005	33.251	0.700	39.004	0.132
Cardio	34.390	11.380	27.560	0.004	10.257	2.563	0.004	22.037	0.263	36.841	1.300
Diab	1.224	0.237	1.046	0.001	0.109	0.005	0.001	1.863	0.006	3.228	0.070
Iris	0.022	0.006	0.024	0.001	0.002	0.184	0.002	0.084	0.049	0.147	0.095
Pima	1.574	0.317	1.311	0.947	0.149	0.904	0.942	6.866	0.920	4.199	0.957
Thyro	78.381	7.299	40.804	0.006	60.720	67.046	0.004	99.782	0.664	376.420	0.164
Wbc	0.858	0.203	0.918	0.001	0.088	0.118	0.001	1.547	0.071	11.054	0.086
Wdbc	2.801	1.495	2.321	0.002	0.134	0.084	0.002	1.048	0.058	2.987	0.088
Yeast	5.728	0.861	4.756	0.001	1.642	0.765	0.001	8.628	0.117	30.324	0.111
Arr	26.384	120.382	22.561	0.017	1.918	0.319	0.017	3.744	5.070	31.814	0.137
Cred	1.183	0.370	0.855	0.002	0.084	0.398	0.001	2.896	0.122	2.614	0.092
Horse	0.664	0.309	0.498	0.001	0.036	0.054	0.001	1.503	0.125	1.066	0.083
Sick	196.408	85.560	105.574	0.011	245.511	47.978	0.012	211.760	0.844	636.024	0.193
Thyr	1380.512	555.978	663.063	0.029	4059.247	497.669	0.030	1240.336	5.372	6720.517	0.507
Average	110.868	49.857	56.568	0.130	274.576	38.888	0.078	102.240	0.901	493.636	0.262

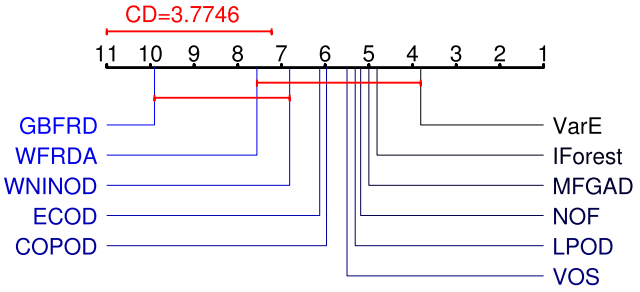
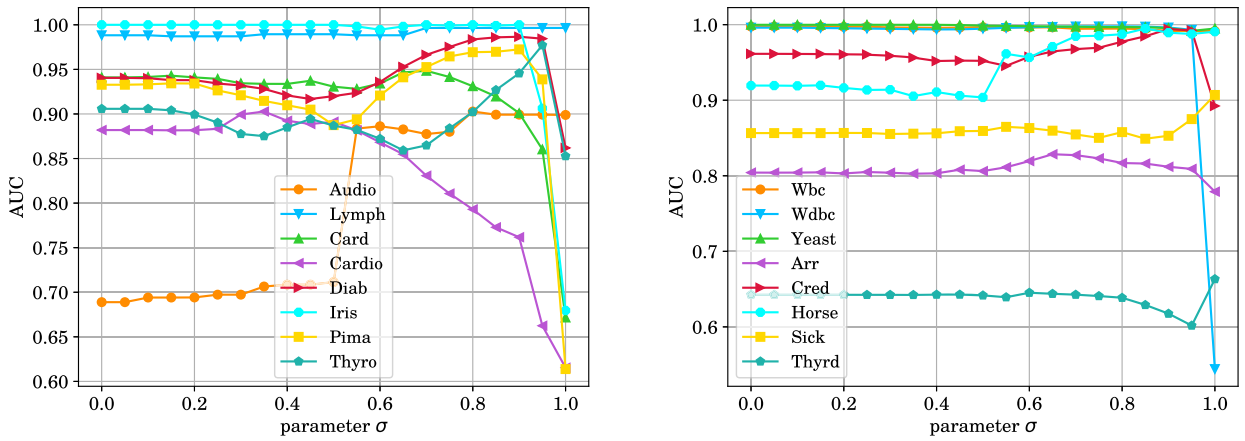


Fig. 5. Nemenyi's test figure on AUC.



(a) Datasets 1-8 (b) Datasets 9-16

Fig. 6. The variation curves of AUC on parameter σ.

6.4. Parameter sensitivity analysis

There is only one adjustable parameter σ in the proposed method. Fig. 6 demonstrates the variation curves of different σ. This figure sets the adjustable range of σ to $[0, 1]$ and the step size to 0.05.

The depicted results demonstrate the robustness of GBFRD on the majority of datasets, where variations in σ have minimal impact on the AUC values. However, it is noteworthy that within a small subset of datasets, a particular pattern emerges, wherein an increase in σ initially leads to a rise in the AUC value, followed by a sharp decline. This is because when σ is close to 1.0, according to Eq. (3), the GB fuzzy relations may be too small to be set to 0 during calculation, resulting in no valid information being provided.

The reasons why GBFRD can improve anomaly detection performance can be summarized as follows. GBC, a novel GrC method, adopts coarse granularity GBs as the input to circumvent the influence of noisy data and improve its efficiency. We propose a novel FRS model named GBFR. GBFR utilizes the strengths of GBC and FRS to mine anomalous features efficiently with GBs while dealing with uncertainty in the data. GBC and FRS are essential tools in GrC, and both can mine the information in data at multiple levels of granularity rather than considering only a small amount of information at a single granularity. GBFRD is a method built on the GBFR model for anomaly detection. Specifically, the outlier factors of each object are characterized by calculating the outlier degrees of multiple GBGs. The outlier degree of a GBG is determined by the GB fuzzy approximation accuracy for a set of GB fuzzy relations since the GB fuzzy approximation accuracy of a GBG can be used to measure the uncertainty of that GBG. If the GB fuzzy approximation accuracy of a GBG about these relations is always low, it indicates that the GBG has significant abnormal behavior. Different GBGs can characterize objects from multiple GB fuzzy relations, incorporating diversified feature information rather than relying on only one. In addition, we consider that multiple GBGs contain different anomalous features. Their anomalous features have varying degrees of importance. Therefore, we have set corresponding weights for each object under different attributes. Each weight is determined by the cardinality of the GBG under the GB fuzzy relation induced by the attribute to which it belongs. The cardinality of a GBG reflects the sum of similarities between the object and the rest of the objects under the current fuzzy relation; the smaller the cardinality indicates that the object is more likely to be a minority class, which means that the object is more likely to be an anomaly. So, a greater weight is assigned to it in the calculation.

7. Conclusion

In this study, we propose an effective unsupervised anomaly detection method called GBFRD based on GBC and FRS. The experimental results show that GBFRD outperforms most existing methods and has noteworthy performance. In the experiments section, we analyze the reasons for the excellent performance of GBFRD. Although our method achieves promising detection results, some things could be improved. Specifically, since the granular-ball generation is based on the k-means algorithm, it is not applicable to data with nominal attributes. In addition, our method has a hyper-parameter, which often requires multiple tunings to determine the optimal parameter for new datasets, and this tuning process is time-consuming. Therefore, our method needs to be further improved in detecting mixed-attribute data in future work, and the tuning process can be avoided by further investigating adaptive methods.

CRediT authorship contribution statement

Xinyu Su: Writing – review & editing, Writing – original draft. **Zhong Yuan:** Data curation, Conceptualization, Funding acquisition, Writing – review & editing. **Baiyang Chen:** Resources, Methodology. **Dezhong Peng:** Software, Resources, Project administration. **Hongmei Chen:** Investigation, Funding acquisition. **Yingke Chen:** Data curation, Conceptualization.

Declaration of competing interest

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

Data availability

We have shared the link to my data/code in the article.

Acknowledgements

The authors thank both the editors and reviewers for their valuable suggestions, which substantially improve this paper. This work was supported by the National Natural Science Foundation of China (62306196, 62372315, and 62376230), Sichuan Science and Technology Program (2023YFQ0020, 2023ZYD0143, 24ZDZX0007, 2024YFHZ0144, 2024YFHZ0089, and 2024NSFSC0443), and the Fundamental Research Funds for the Central Universities (YJ202245).

References

[1] A. Albanese, S.K. Pal, A. Petrosino, Rough sets, kernel set, and spatiotemporal outlier detection, IEEE Trans. Knowl. Data Eng. 26 (1) (2012) 194–207.
[2] H. Bai, F. Shen, W. Kong, J. Feng, Granular-ball clustering based neighbourhood outliers detection method, in: 2023 6th International Conference on Electronics Technology (ICET), IEEE, 2023, pp. 1306–1312.
[3] M.M. Breunig, H.-P. Kriegel, R.T. Ng, J. Sander, Lof: identifying density-based local outliers, in: Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, 2000, pp. 93–104.
[4] E. Cabana, R.E. Lillo, H. Laniado, Multivariate outlier detection based on a robust Mahalanobis distance with shrinkage estimators, Stat. Pap. 62 (2021) 1583–1609.
[5] D. Cheng, Y. Li, S. Xia, G. Wang, J. Huang, S. Zhang, A fast granular-ball-based density peaks clustering algorithm for large-scale data, IEEE Trans. Neural Netw. Learn. Syst. (2023) 1–14.
[6] A. Degirmenci, O. Karal, Efficient density and cluster based incremental outlier detection in data streams, Inf. Sci. 607 (2022) 901–920.
[7] D. Dubois, H. Prade, Rough fuzzy sets and fuzzy rough sets, Int. J. Gen. Syst. 17 (2–3) (1990) 191–209.
[8] Z. Elhoussaine, et al., A fuzzy neighborhood rough set method for anomaly detection in large scale data, IAES Int. J. Artif. Intell. 9 (1) (2020) 1–10.
[9] J. Huang, Q. Zhu, L. Yang, J. Feng, A non-parameter outlier detection algorithm based on natural neighbor, Knowl.-Based Syst. 92 (2016) 71–77.
[10] F. Jiang, Y. Sui, C. Cao, Some issues about outlier detection in rough set theory, Expert Syst. Appl. 36 (3) (2009) 4680–4687.
[11] F. Jiang, Y. Sui, C. Cao, An information entropy-based approach to outlier detection in rough sets, Expert Syst. Appl. 37 (9) (2010) 6338–6344.
[12] F. Jiang, H. Zhao, J. Du, Y. Xue, Y. Peng, Outlier detection based on approximation accuracy entropy, Int. J. Mach. Learn. Cybern. 10 (2019) 2483–2499.
[13] X. Li, J. Lv, Z. Yi, Outlier detection using structural scores in a high-dimensional space, IEEE Trans. Cybern. 50 (5) (2018) 2302–2310.
[14] Z. Li, Y. Zhao, N. Botta, C. Ionescu, X. Hu, Copod: copula-based outlier detection, in: 2020 IEEE International Conference on Data Mining (ICDM), IEEE, 2020, pp. 1118–1123.
[15] Z. Li, Y. Zhao, X. Hu, N. Botta, C. Ionescu, G.H. Chen, Ecod: unsupervised outlier detection using empirical cumulative distribution functions, IEEE Trans. Knowl. Data Eng. 35 (12) (2023) 12181–12193.
[16] F.T. Liu, K.M. Ting, Z.-H. Zhou, Isolation-based anomaly detection, ACM Trans. Knowl. Discov. Data 6 (1) (2012) 1–39.
[17] H. Liu, X. Li, J. Li, S. Zhang, Efficient outlier detection for high-dimensional data, IEEE Trans. Syst. Man Cybern. Syst. 48 (12) (2017) 2451–2461.
[18] F. Maciá-Pérez, J.V. Berna-Martinez, A.F. Oliva, M.A.A. Ortega, Algorithm for the detection of outliers based on the theory of rough sets, Decis. Support Syst. 75 (2015) 63–75.
[19] A. Mani, Algebraic, topological, and mereological foundations of existential granules, in: International Joint Conference on Rough Sets, Springer, 2023, pp. 185–200.
[20] A. Mani, Granular knowledge and rational approximation in general rough sets – I, J. Appl. Non-Class. Log. 34 (2–3) (2024) 294–329.
[21] F.A. Mazarbhuiya, M. Shenify, An intuitionistic fuzzy-rough set-based classification for anomaly detection, Appl. Sci. 13 (9) (2023) 5578.
[22] T.T. Nguyen, Outlier detection: an approximate reasoning approach, in: Rough Sets and Intelligent Systems Paradigms: International Conference, RSEISP 2007, Proceedings 1, Warsaw, Poland, June 28–30, 2007, Springer, 2007, pp. 495–504.
[23] Y. Peng, Y. Yang, Y. Xu, Y. Xue, R. Song, J. Kang, H. Zhao, Electricity theft detection in ami based on clustering and local outlier factor, IEEE Access 9 (2021) 107250–107259.

[24] W. Qian, F. Xu, J. Qian, W. Shu, W. Ding, Multi-label feature selection based on rough granular-ball and label distribution, Inf. Sci. 650 (2023) 119698.
[25] S. Sandosh, V. Govindasamy, G. Akila, Enhanced intrusion detection system via agent clustering and classification based on outlier detection, Peer-to-Peer Netw. Appl. 13 (3) (2020) 1038–1045.
[26] B. Sang, W. Xu, H. Chen, T. Li, Active antinoise fuzzy dominance rough feature selection using adaptive k-nearest neighbors, IEEE Trans. Fuzzy Syst. 31 (11) (2023) 3944–3958.
[27] A. Smiti, A critical overview of outlier detection methods, Comput. Sci. Rev. 38 (2020) 100306.
[28] Y. Song, H. Lin, Z. Li, Outlier detection in a multiset-valued information system based on rough set theory and granular computing, Inf. Sci. 657 (2024) 119950.
[29] S. Thudumu, P. Branch, J. Jin, J. Singh, A comprehensive survey of anomaly detection techniques for high dimensional big data, J. Big Data 7 (2020) 1–30.
[30] C. Wang, Z. Liu, H. Gao, Y. Fu, Vos: a new outlier detection model using virtual graph, Knowl.-Based Syst. 185 (2019) 104907.
[31] S. Wang, Z. Yuan, C. Luo, H. Chen, D. Peng, Exploiting fuzzy rough entropy to detect anomalies, Int. J. Approx. Reason. 165 (2024) 109087.
[32] X. Wang, Z. Liu, J. Liu, J. Liu, Fraud detection on multi-relation graphs via imbalanced and interactive learning, Inf. Sci. 642 (2023) 119153.
[33] Y. Wang, Y. Li, Outlier detection based on weighted neighbourhood information network for mixed-valued datasets, Inf. Sci. 564 (2021) 396–415.
[34] S. Xia, X. Dai, G. Wang, X. Gao, E. Giem, An efficient and adaptive granular-ball generation method in classification problem, IEEE Trans. Neural Netw. Learn. Syst. 35 (4) (2024) 5319–5331.
[35] S. Xia, Y. Liu, X. Ding, G. Wang, H. Yu, Y. Luo, Granular ball computing classifiers for efficient, scalable and robust learning, Inf. Sci. 483 (2019) 136–152.
[36] S. Xia, G. Wang, X. Gao, Granular ball computing: an efficient, robust, and interpretable adaptive multi-granularity representation and computation method, arXiv preprint arXiv:2304.11171, 2023.
[37] S. Xia, H. Zhang, W. Li, G. Wang, E. Giem, Z. Chen, Gbnrs: a novel rough set algorithm for fast adaptive attribute reduction in classification, IEEE Trans. Knowl. Data Eng. 34 (3) (2020) 1231–1242.
[38] W. Xiaolan, M.M. Ahmed, M.N. Husen, Z. Qian, S.B. Belhaouari, Evolving anomaly detection for network streaming data, Inf. Sci. 608 (2022) 757–777.
[39] Q. Xie, Q. Zhang, S. Xia, F. Zhao, C. Wu, G. Wang, W. Ding, Gbg++: a fast and stable granular ball generation method for classification, arXiv preprint arXiv:2305.18450, 2023.
[40] Z. Xue, S. Liu, Rough-based semi-supervised outlier detection, in: 2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery, vol. 1, IEEE, 2009, pp. 520–523.
[41] J. Yang, X. Tan, S. Rahardja, Outlier detection: how to select k for k-nearest-neighbors-based outlier detectors, Pattern Recognit. Lett. 174 (2023) 112–117.
[42] K. Yang, W. Chen, J. Bi, M. Wang, F. Luo, Multi-view broad learning system for electricity theft detection, Appl. Energy 352 (2023) 121914.
[43] X. Yang, H. Chen, T. Li, Y. Yao, Geodesic fuzzy rough sets for discriminant feature extraction, IEEE Trans. Fuzzy Syst. 32 (3) (2024) 778–791.
[44] X. Yang, Y. Li, S. Xia, X. Lian, G. Wang, T. Li, Granular-ball three-way decision, in: International Joint Conference on Rough Sets, Springer, 2023, pp. 283–295.
[45] J.T. Yao, A.V. Vasilakos, W. Pedrycz, Granular computing: perspectives and challenges, IEEE Trans. Cybern. 43 (6) (2013) 1977–1989.
[46] Z. Yuan, B. Chen, J. Liu, H. Chen, D. Peng, P. Li, Anomaly detection based on weighted fuzzy-rough density, Appl. Soft Comput. 134 (2023) 109995.
[47] Z. Yuan, H. Chen, T. Li, B. Sang, S. Wang, Outlier detection based on fuzzy rough granules in mixed attribute data, IEEE Trans. Cybern. 52 (8) (2021) 8399–8412.
[48] Z. Yuan, H. Chen, C. Luo, D. Peng, Mfgad: multi-fuzzy granules anomaly detection, Inf. Fusion 95 (2023) 17–25.