Research Article

Zhaowen Li, Hongxuan He*, and Pei Wang*

Class-consistent technology-based outlier detection for incomplete real-valued data based on rough set theory and granular computing

https://doi.org/10.1515/jisys-2024-0347
received July 11, 2024; accepted January 09, 2025

Abstract: The goal of outlier detection is to pinpoint data points that exhibit notable deviations from the rest of the observed values. It has found successful application in numerous fields, including process inspection, anti-terrorist operations, and public security. However, the majority of existing outlier algorithms rely on methods involving filling or deleting missing data, with few directly addressing incomplete data. This article studies outlier detection for incomplete real-valued data based on class-consistent technology, rough set theory, and granular computing. First, a tolerance relation founded on class-consistent technology is presented to illustrate the similarity among information values within an incomplete real-valued information system (IRVIS). Then, the tolerance classes are established based on the tolerance relation and utilized for computing approximate accuracy and other metrics. Next, an outlier factor is defined, considering both the degree of outlierness and the weight function assigned to each object within an IRVIS, elucidating the uncertainty and degree of outlier. Finally, an outlier detection algorithm (ODIRG) for an IRVIS based on class-consistent technology, rough set theory, and granular computing is devised. Numerical experiments on seven UCI datasets are undertaken to evaluate the stability of the ODIRG algorithm. The proposed method is demonstrated to exhibit strong effectiveness and adaptability for categorical data when compared with five other algorithms. It is notable that for comprehensive comparison, precision, recall, $F1$-measure, and receiver operating characteristic curve are employed to delineate the benefits of the proposed approach.

Keywords: rough set theory, granular computing, an IRVIS, class-consistent technology, outlier detection

1 Introduction

1.1 Research and background

An outlier refers to a data point that significantly differs from “normal” observations in terms of data characteristics. Hawkins [1] gives a compelling definition of the many. “Outliers are deviations from other observations, leading to suspicions that it is produced by a different mechanism.” Since outliers can have a

*** Corresponding author: Hongxuan He,** College of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, P.R. China, e-mail: D240201014@stu.cqupt.edu.cn
*** Corresponding author: Pei Wang,** Center for Applied Mathematics of Guangxi, Yulin Normal University, Yulin, Guangxi 537000, P.R. China, e-mail: peiwang130@ylu.edu.cn
Zhaowen Li: College of Computer Science, Guangdong University of Science and Technology, Dongguan, Guangdong 523083, P.R. China, e-mail: zhaowenli@ylu.edu.cn

notable impact on the results of statistical analysis, removing them is often considered a crucial preprocessing step in data analysis models [2]. However, to remove outliers, they need to be identified first, which is the goal of outlier detection. Its applications span various fields, including intrusion detection, wireless sensor network localization, time-series sequence analysis, and process monitoring [3–6].

Outlier detection has gradually attracted the attention of a large number of researchers, who have proposed various methods for detecting outliers. Various methodologies and technical frameworks underpin outlier detection, broadly classified into statistical-based methods [7,8], cluster-based methods [9,10], depth-based methods [11], distance-based methods [12,13], and density-based methods [14]. Statistical-based methods necessitate a prior understanding of data distribution laws, rendering them inadequate for intricate and multidimensional datasets. Cluster-based methods detect outliers by tuning the parameters dictating the relationship between the detection object and the cluster. Nonetheless, the efficacy of outlier detection hinges significantly on how the parameters of the clustering algorithm are configured, resulting in subjectivity and variability. Depth-based techniques excel in lower-dimensional data but falter in high dimensions. Distance-based methods quantify the Euclidean distance between objects, while density-based methods are formulated by computing the local density of each object and its neighboring points. However, the computation involved in the latter two approaches is intricate and resource intensive. This has prompted scholars to continuously explore and improve the outlier detection algorithm (ODRIG).

To overcome the limitations of distance-based and density-based methods, new theories and paradigms related to anomaly detection have emerged. Granular computing (GrC) is a natural model that simulates the human mind when solving large-scale problems [15,16]. Zadeh [17] introduced the notion of information granularity in 1996, suggesting the potential application of the fuzzy set theory in this domain. Pawlak [10,18,19] initially introduced the rough set theory (RST), offering a concrete instance of GrC that underscores the importance of granulation. RST has gradually established itself as a dominant mathematical framework in the realm of GrC [20]. Jiang et al. introduced outlier detection approaches based on RST [21]. A previous study gave a computational framework to detect aberrations utilizing RST. Chen et al. [22] introduced a method for detecting deviations, drawing on the principle from GrC. Jiang and Chen [23], and Jiang et al. [21] delved into a hybrid approach for abnormality detection, merging RST and GrC in an algorithm that relies on approximate precision entropy. Nguyen [24] introduced a technique within RST for detecting and assessing outlying points through a layered approximate inference strategy.

Although these methods have demonstrated the practicality of RST in the field of outlier detection, it is important to note that these mathematical detection models, built upon equivalence relations, are more suited for handling nominal feature data. Prior to their application to numerical feature data, discretization is necessary, a process that not only extends the time required for data processing but also results in significant information loss.

To address the limitation of RST in only handling nominal feature data, Dubois and Prade [25] presented the concept of fuzzy rough set. Fuzzy rough set utilize fuzzy relation to characterize the similarity between objects, enabling them to directly process numerical or continuous feature data without the information loss associated with discretization. Yuan et al. [26] delved into outlier identification approaches based on fuzzy rough set.

1.2 Related work

Over the years, numerous types of anomaly detection algorithms have been developed. This section will briefly introduce some classic outlier detection techniques as well as methods based on RST.

Outlier detection methods can be broadly classified into statistical-based methods [7,8], cluster-based methods [9,10], depth-based methods [11], distance-based methods [12,13], and density-based methods [14].

Statistical-based methods assumes a distribution or probability model for the dataset to be examined. Then, discordancy tests are employed for outlier detection. When the statistical assumptions made about the data satisfy the actual constraints, it is statistically very effective. However, precisely because it requires

assuming that the data conforms to a certain distribution, it is not applicable to situations where the distribution is unknown [7,8]. Clustering-based methods categorize data into clusters. Data objects that do not belong to any cluster or small clusters with significantly fewer data points compared to other clusters are deemed as anomalies. Since clustering-based methods leverage the clustering structure of data to detect anomalies, they are highly robust unsupervised methods suitable for various types of data featuring diverse characteristics. A drawback of clustering-based methods is that their effectiveness strongly depends on the clustering algorithms employed, which may not be optimal for outlier detection. In addition, computational costs can be a bottleneck for large datasets [9,10]. The depth-based method assigns a depth value to each object and maps the object to the corresponding two-dimensional spatial layer based on this depth value, where objects located on shallower layers are likely to be outliers. It compensates for the limitations of statistical methods in certain aspects. However, the depth-based method performs better when dealing with data in two-dimensional and three-dimensional spaces, but its efficiency is relatively low when detecting high-dimensional mixed-feature data [11]. The distance-based method measures the degree of outlierness by calculating the distance between two objects, considering those that are far away from most other objects as anomalies. This method is widely used due to its ease of implementation. However, the sparsity issue in high-dimensional data is difficult to overcome. Furthermore, since it relies on two global parameters, it is extremely sensitive to the choice of these parameters. In addition, it does not take into account changes in local density [12,13]. The core concept of the density-based method lies in the fact that the density around anomalies significantly differs from the density of objects in their neighboring areas. Based on this, each object is assigned a local outlier factor to indicate its degree of outlierness. The higher the local outlier factor value of an object, the more likely it is to be an anomaly. Although the density-based method can address the issue of identifying local anomalies, it remains highly sensitive to the choice of parameters [14].

However, in practical applications, the computations required by these methods remain considerably complex, which motivates scholars to continuously research and ODIRGs. In recent years, to overcome the limitations of distance-based and density-based methods, researchers have proposed ODIRGs based on rough sets. For example, Shaari et al. [27] have explored a new approach to identify outliers using nonreduction concept within the framework of RST; Jiang et al. [28] introduced a novel definition for outlier detection based on rough sequences and delved into the definition of distance-based outlier detection metrics within the framework of RST; A previous study gave an outlier reduction method based on an outlier detection analysis system, utilizing the concepts of RST; Albanese et al. [29] used a new rough set approach to extend outlier detection to spatiotemporal data.

1.3 Motivation and contribution

However, existing methods have not considered the handling of missing real-valued data. It is noteworthy that varying degrees of missing data are prevalent across outlier detection tasks in diverse information systems, and distinct strategies for handling missing data typically yield differing experimental outcomes. One of the most frequently used preprocessing techniques involves filling in missing values to create complete datasets. Typically, missing values within an attribute are replaced with the mean, maximum, or most frequently occurring value of all available attribute values within that attribute [30]. However, these preprocessing methods can significantly alter the information structure of the original dataset, leading to increased information loss and high computational costs. In response to this challenge, the article introduces a technique for detecting outliers in incomplete real-valued data through the application of RST and GrC.

The main objective of this study is to identify anomalies in an incomplete real-valued information system (IRVIS) without necessitating the filling of missing values. From a RST standpoint, the article presents a class consistent for incomplete numerical data, aiming to account for the variability of information values across different attributes and intervals. Subsequently, the class is employed to consistently compute the tolerance class for each object. Once the tolerance class is established, this study introduces an outiler factor to evaluate the probability of a data being an outlying point, thereby facilitating outlier identification. In addition,

normalization gives the probability of an abnormal degree of the object of assessment becoming an outlier. Finally, an anomaly ODIRG is designed. The study conducts experiments on seven datasets sourced from UCI (Machine Learning Repository) and KEEL (Knowledge Extraction based on Evolutionary Learning), comparing the performance of five outlier identification algorithms (CBLOF [9], FIEOD [31], KNN [32], SEQ [28], and NOOF [33]). The experimental findings illustrate that ODIRG often surpasses other outlier detection methods in an IRVIS.

The article's contributions are summarized as follows:

(1) RST and GrC are used to directly handle an IRVIS, while considering all objects with missing attribute values when constructing information granules. This avoids operations such as data imputation or data manipulation, thus sustaining the information configuration of the initial data to a great extent, reducing the risk of information loss, and consequently lowering computational costs;
(2) We investigate the degree of abnormality for each object. It provides a systematic framework for assessing the consistency of objects within different categories or clusters. This helps identify outliers that significantly deviate from the expected behavior of their respective categories, thereby enhancing the accuracy and reliability of ODIRGs;
(3) A revolutionary outlying observation identification algorithm is put forward for an IRVIS, leveraging the abnormality degree of individual objects. Experimental findings display the superior capability and adaptability of this proposed method for an IRVIS.

1.4 Structure and organization

The subsequent sections of this article are structured as follows: Section 2 reviews some related works. Section 3 outlines the initial groundwork necessary for formulating the proposed methodology. Section 4 systematically develops the proposed anomaly factor and presents the ODIRGs. Section 5 presents the experimental findings. Section 6 performs the evaluation assessments. Section 7 wraps up the article.

The workflow of the article is depicted in Figure 1.

2 Preliminaries

In this section, the preliminary components essential for constructing the proposed methodology are delineated.

Throughout this article, $O = \{o_1, o_2, \ldots, o_n\}$, $A = \{a_1, a_2, \ldots, a_m\}$ denote two nonempty finite sets, 2^O the power set of O and $|X|$ the cardinality of $X \in 2^O$.

2.1 An IRVIS

Definition 2.1. [18] Let O be an object set and A an attribute set. Suppose that O and A are finite sets. Then the pair (O, A) is called an information system (IS), if each attribute $a \in A$ determines an information function $a : O \to V_a$, where $V_a = \{a(o) : o \in O\}$.

Definition 2.2. [18] Let (O, A) be an IS. If there is $a \in A$ such that $* \in V_a$, here $*$ means a null or unknown value, then (O, A) is called an incomplete information system (IIS).

For each $a \in A$, denote

$$V_a^* = V_a - \{a(o) : a(o) = *\}.$$

Then, V_a^* means the set of all nonmissing information values with respect to the attribute a.

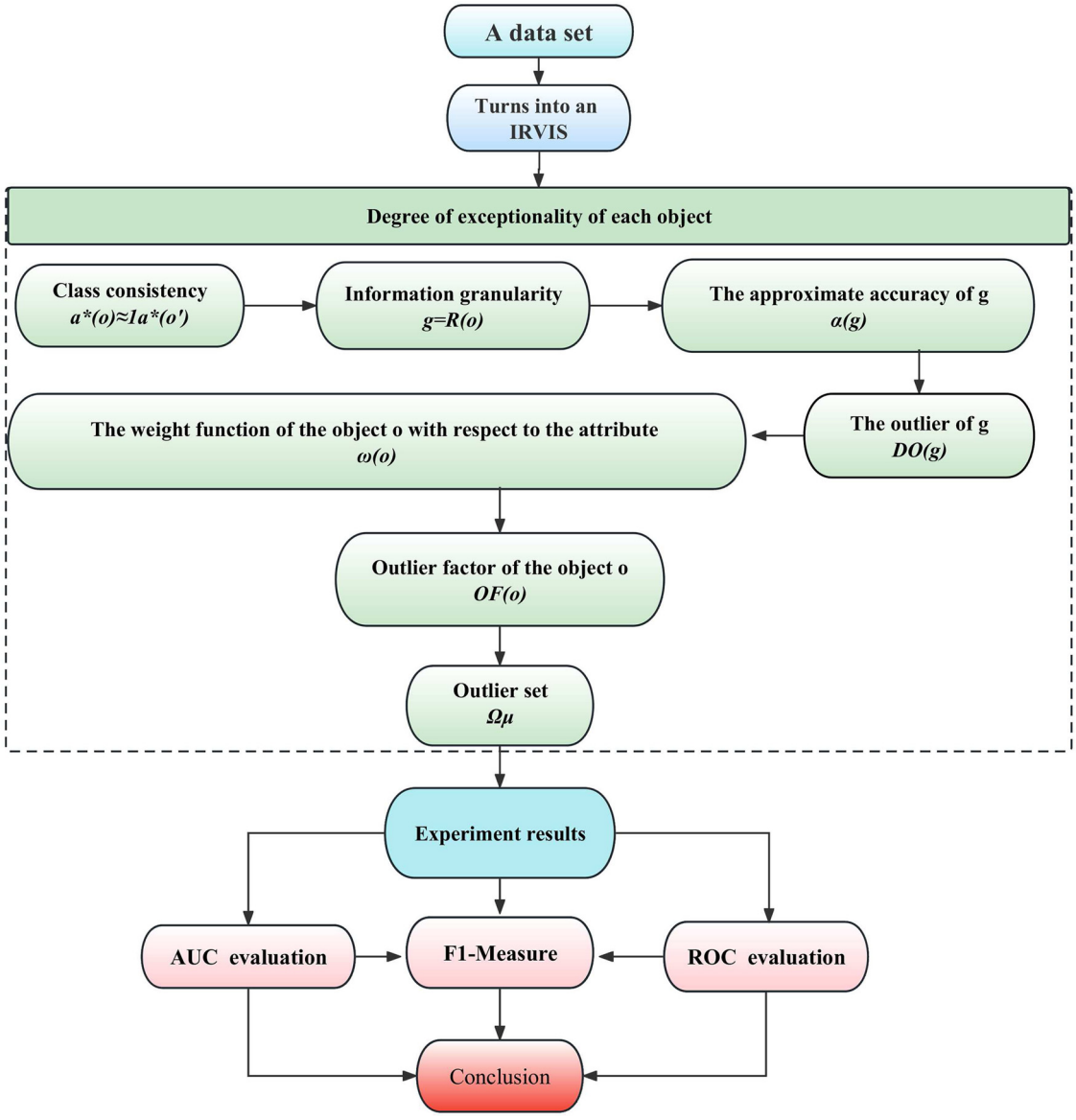

Figure 1: The workflow of this article. (The image was created by the authors.).

Definition 2.3. [34] Suppose that (O, A) is an IIS. Then (O, A) is referred to as an IRVIS, if for any $a \in A$ and $o \in O$, $a(o)$ is a real number.

If $P \subseteq A$, then (O, P) is referred to as the subsystem of (O, A).

Example 2.4. Table 1 expresses an IRVIS (O, A), where $O = \{o_1, o_2, \ldots, o_7\}$ is an object set and $A = \{a_1, a_2, a_3, a_4\}$ is a set of attributes.

$$V_{a_1}^* = \{1, 3, 6, 9\}, V_{a_2}^* = V_{a_2} = \{0, 1\},$$
$$V_{a_3}^* = \{10, 20, 30\}, V_{a_4}^* = \{10, 40, 80\}.$$

Table 1: An IRVIS

O	a_1	a_2	a_3	a_4
o_1	6	0	20	*
o_2	1	1	*	10
o_3	*	0	*	80
o_4	3	1	20	10
o_5	6	0	30	*
o_6	*	1	10	40
o_7	9	1	*	80

* unknown.

2.2 Class-consistent technology

In an IRVIS (O, A), there is either $a(o) = *$ or $a(o)$ is a real number for any $a \in A$ and $o \in O$. So (O, A) can be dealt with in a discrete way.

Definition 2.5. [35] Let (O, A) be an IRVIS. $\forall a \in A$ and $o \in O$, $a(o)$ is a real number. The following equation is used to transform the value of $a(o)$ to [0,1]:

$$a^*(o) = \begin{cases} \dfrac{a(o) - \min V_a^*}{\max V_a^* - \min V_a^*}, & a(o) \neq * \\ 0, & a(o) = * \end{cases}$$

Then a^* determines a fuzzy set on O:

$$a^* = \frac{a^*(o_1)}{o_1} + \frac{a^*(o_2)}{o_2} + \cdots + \frac{a^*(o_n)}{o_n}.$$

The membership degrees of objects o and o' to the fuzzy set a^* are denoted as $a(o)$ and $a(o')$, respectively.

In RST, information particles are usually constructed with an equivalence relation, but for an IRVIS, achieving $a^*(o) = a^*(o')$ is highly challenging. Hence, the introduction of the following class-consistent relation is necessary.

Definition 2.6. [36] Let $k \in N, m, n \in [0, 1]$. If $m = n = 0$ or $m, n \in (0, \frac{1}{10^k})$ or $m = n = \frac{1}{10^k}$ or ... or $m = n = \frac{10^k - 1}{10^k}$ or $m, n \in (\frac{10^k - 1}{10^k}, 1)$ or $m = n = 1$, then m and n are said to be class consistent, denote it by $m \approx_k n$, where k is said to be a threshold value.

In this article, we pick $k = 1$.

Definition 2.7. [36] Given $m, n \in [0, 1]$. If the following condition hold:

$$m = n \quad \text{when } m, n \in \{0, 0.1, \ldots, 0.9, 1\}$$

or

$$n \in \left(\frac{i-1}{10}, \frac{i}{10}\right) \quad \text{when } m \in \left(\frac{i-1}{10}, \frac{i}{10}\right) (k = 1, 2, \ldots, 9, 10).$$

Then $m \approx_1 n$.

Definition 2.7 is grounded on the concept of partitioning $[0, 1]$, wherein $[0, 1]$ is partitioned into 21 segments: 11 endpoints and 10 open intervals. If two numbers must strictly match across the 11 endpoints or fall within the same open interval among the 10 open intervals, they are deemed class consistent.

Example 2.8. Pick $a = 0.79$, $b = 0.71$ and $c = 0.81$. Then $a \approx_1 b$, $a \not\approx_1 c$. Here, $|a - b| = 0.08 > 0.02 = |a - c|$.

While the usual metric for object similarity involves assessing the distance or proximity between them, example 2.8 highlights that Definition 2.7 does not consider the proximity between two numbers within $[0, 1]$. Consequently, Definition 2.7 offers an alternative perspective on describing the similarity between two objects.

This article terms the manipulation of numbers within $[0, 1]$ as the class consistency technique, which leads to the establishment of the subsequent class consistency relation.

2.3 A tolerance relation based on class-consistent technology

Definition 2.9. [18] Let R be a binary relation on O whenever $R \subseteq O \times O$. R is called
(1) reflexive, if $(o, o') \in R$ for any $o \in O$;
(2) symmetric, if $(o, o') \in R$ implies $(o', o) \in R$;
(3) transitive, if $(o, o') \in R$ and $(o', o'') \in R$ imply $(o, o'') \in R$.

R is said to be an equivalence relation on O, if R is reflexive, symmetric and transitive; R is called a tolerance relation on O, if R is reflexive and symmetric.

Definition 2.10. [34] Let (O, A) be an IRVIS. Given $P \subseteq A$. Define

$$(o, o') \in R_P \Leftrightarrow \forall a \in P, a^*(o) \approx_1 a^*(o') \text{ or } a^*(o) = * \text{ or } a^*(o') = *.$$

Then R_P is called the binary relation induced by the subspace (O, P)

Clearly, R_P is a tolerance relation on O.

Definition 2.11. [34] Let (O, A) be an IRVIS. Given $P \subseteq A$. Define

$$R_P(o) = \{o' \in O : (o, o') \in R_P\}.$$

Then $R_P(o)$ is referred as to the tolerance class of the object o under R_P. Denote

$$O/R_P = \{R_P(o) : o \in O\}.$$

Proposition 2.12. *Let (O, A) be an IRVIS. If $P_1 \subseteq P_2 \subseteq A$, then $\forall o \in O$,*

$$R_{P_2}(o) \subseteq R_{P_1}(o).$$

Proof. Obviously. □

Proposition 3.12 indicates that $R_P(o)$ is monotonically increasing with respect to P.

Definition 2.13. [34] Let (O, A) be an IRVIS. Given $P \subseteq A$ and $X \in 2^O$. Define

$$\underline{R}_P(X) = \{o \in O : R_P(o) \subseteq X\};$$
$$\overline{R}_P(X) = \{o \in O : R_P(o) \cap X \neq \varnothing\}.$$

Then $\underline{R}_P(X)$ and $\overline{R}_P(X)$ are called the lower approximation and upper approximation of X, respectively.

Moreover, if $\underline{R}_P(X) = \overline{R}_P(X)$, then X is called an exact set with respect to R_P; otherwise, X is called a rough set with respect to R_P.

Theorem 2.14. *Let (O, A) be an IRVIS.*
(1) $\overline{R}_P(\varnothing) = \underline{R}_P(\varnothing) = \varnothing$, $\underline{R}_P(O) = \overline{R}_P(O) = O$.
(2) $\underline{R}_P(X) \subseteq X \subseteq \overline{R}_P(X)$.

(3) $X \subseteq Y \Rightarrow \underline{R}_P(X) \subseteq \underline{R}_P(Y), \overline{R}_P(X) \subseteq \overline{R}_P(Y)$.
(4) *If* $P_1 \subseteq P_2 \subseteq A$, *then* $\forall X \in 2^O$,

$$\underline{R}_{P_1}(X) \subseteq \underline{R}_{P_2}(X), \overline{R}_{P_2}(X) \subseteq \overline{R}_{P_1}(X).$$

(6) $\underline{R}_P(X \cap Y) = \underline{R}_P(X) \cap \underline{R}_P(Y)$; $\overline{R}_P(X \cap Y) = \overline{R}_P(X) \cup \overline{R}_P(Y)$.
(7) $\underline{R}_P(O - X) = O - \overline{R}_P(X)$; $\overline{R}_P(O - X) = O - \underline{R}_P(X)$.

Proof. According to Definition 2.10, Proposition 2.12, and Definition 2.13, the conclusion is obvious. □

3 Outliers for incomplete real-valued data based on RST and GrC

3.1 The outlier detection method

In an IRVIS (O, A), for any $o \in O$ and a set of class consistency relation, we can obtain a granule g containing o with respect to each of these relations.

To calculate the degree of outliers for the given granule, we use the approximate precision of the RST to define the degree of outliers for the granule.

Definition 3.1. Let (O, A) be an IRVIS. Given $P \subseteq A$. Put $C = A - P = \{a_{k_1}, \ldots, a_{k_s}\}$. Suppose $|C| \geq 2$ and $E \subseteq C$. Then E-accuracy of approximation of the information granule $g \in O/R_P$ is defined as follows:

$$\alpha_E(g) = |\underline{R}_E(g)|/|\overline{R}_E(g)|. \tag{3.1}$$

For the given granule g, we calculate the approximate accuracy of g for a set of class consistency relations. If the approximate progress of g to these relations is always low, then g behavior can be considered to be deviant from normal, and g is highly outlierable.

Conventional outlier detection techniques usually emphasize categorizing outliers into binary classifications, deciding whether an object falls into the outlier category or not. Nevertheless, in many cases, it could be more beneficial to assign a precise degree of outlierness to each object. In this article, an outlier factor is created based on the accuracy of approximation to quantify the degree of outlierness for objects. Objects with a higher outlier factor are more probable to be outliers. To accomplish this, we initially establish the degree of outlierness for each object concerning the attribute subset $P \subseteq A$.

Definition 3.2. Let (O, A) be an IRVIS. Given $P \subseteq A$. Put $C = A - P = \{a_{k_1}, \ldots, a_{k_s}\}$. Suppose $|C| \geq 2$. Then degree of outlierness of the information granule $g \in O/R_P$ with with respect to P is defined as follows:

$$DO_P(g) = 1 - |g|\frac{\alpha_C(g) + \sum_{i=1}^{s}(\alpha_{C-\{a_{k_i}\}}(g) + 1)/2}{n(s+1)}, \tag{3.2}$$

where $\alpha_C(g)$ and $\alpha_{C-\{a_{k_i}\}}(g)$, respectively, denote the accuracies of approximation of g with respect to relations R_C and $R_{C-\{a_{k_i}\}}$, $1 \leq k_i \leq n$.

Denote

$$DO_a(g) = DO_{\{a\}}(g).$$

$DO_P(g)$ It quantifies the abnormality of granule g, and the abnormality of g is described by the uncertainty of g. Because the accuracy of approximation of g can be used to measure the uncertainty of g, we calculate $DO_P(g)$ by using the accuracy of approximation of g.

It is crucial to highlight that the calculation of $DO_P(g)$ relies on $P \subseteq A$. Nonetheless, considering that A encompasses $2^{|A|}$ subsets, it becomes impractical to compute $DO_P(g)$ for every object across all attribute

subsets. This would lead to a substantial escalation in the time complexity of the associated algorithm. Therefore, we just calculate the accuracies of approximation of g with respect to relations R_C and $R_{C-a_{k_i}}$.

When creating an outlier factor of o, we also consider the weight function linked to o.

Definition 3.3. Let (O, A) be an IRVIS. Then weight function of $a \in A$ is defined as follows:

$$\omega_a(o) = 1 - \sqrt[3]{\frac{|R_a(o)|}{n}}, \tag{3.3}$$

where $\omega_a(o)$ is the weight of object o, and the $\omega_a(o)$ tells us from another perspective whether o belongs to a minority or majority class.

Definition 3.4. Let (O, A) be an IRVIS with $m \geq 3$. Then outlier factor of $o \in O$ is defined as follows:

$$OF(o) = \frac{\sum_{j=1}^{m} \omega_{a_j}(o) DO_{a_j}(R_{a_j}(o))}{m}. \tag{3.4}$$

Then $OF(o)$ is called the outlier factor of the object O in the IRVIS (O, A). On the basis of the definition of an outlier, we aim to assign a higher outlier factor to object o if it is significantly distant from the majority of other objects in O. When $|R_P(o)|$ is small, indicating that o is part of the minority of objects, we assign a low weight to o to amplify the degree of outlierness in o.

The concepts of $DO_P(g)$ and $OF(o)$ stem from Hawkins' definition of outliers, which states that any object o within a set O that possesses characteristics differing significantly from the rest of the objects in O can be deemed an outlier. In the context of $DO_P(g)$ and $OF(o)$, uncertainty is considered a distinctive characteristic indicating abnormality. If $DO_P(g)$ consistently remains at a high level, it may indicate that o is functioning abnormally, resulting in a high outlier score for o. Therefore, in $OF(o)$, the outlier score of o is directly related to the value of $DO_P(g)$.

Definition 3.5. Let (O, A) be an IRVIS with $m \geq 3$. Given $\mu \in [0,1]$. Then $o^* \in O$ is called μ-outlier in (O, A), if $OF(o^*) > \mu$.

In this article, the set of all μ-outlier in an IRVIS is denoted as Ω_μ.

We propose the generalized outlier detection model, named ODIRG. The process of building this model encompasses five key components: (1) the RST and GrC framework; (2) assessment of approximation accuracy; (3) outlier detection; (4) outlier factor; and (5) the actual outlier detection procedure. Next, we delve into a detailed analysis of a specific ODIRG method.

Example 3.6. Given an IRVIS (O, A) in Table 1. Then the tolerance class for each object is shown in Table 2.

Let $g_1 = R_{a_1}(o_1)$, $g_2 = R_{a_2}(o_1)$, $g_3 = R_{a_3}(o_1)$, $g_4 = R_{a_4}(o_1)$, and from Table 2, we have $g_1 = \{o_1, o_3, o_5, o_6\}$, $g_2 = \{o_1, o_3, o_5\}$, $g_3 = \{o_1, o_2, o_3, o_4, o_7\}$, $g_4 = \{o_1, o_2, o_3, o_4, o_5, o_6, o_7\}$.

Table 2: The tolerance class of each object

	$R_{a_1}(o)$	$R_{a_2}(o)$	$R_{a_3}(o)$	$R_{a_4}(o)$
o_1	$\{o_1, o_3, o_5, o_6\}$	$\{o_1, o_3, o_5\}$	$\{o_1, o_2, o_3, o_4, o_7\}$	$\{o_1, o_2, o_3, o_4, o_5, o_6, o_7\}$
o_2	$\{o_2, o_3, o_6\}$	$\{o_2, o_4, o_6, o_7\}$	$\{o_1, o_2, o_3, o_4, o_5, o_6, o_7\}$	$\{o_1, o_2, o_4, o_5\}$
o_3	$\{o_1, o_2, o_3, o_4, o_5, o_6, o_7\}$	$\{o_1, o_3, o_5\}$	$\{o_1, o_2, o_3, o_4, o_5, o_6, o_7\}$	$\{o_1, o_3, o_5, o_7\}$
o_4	$\{o_3, o_4, o_6\}$	$\{o_2, o_4, o_6, o_7\}$	$\{o_1, o_2, o_3, o_4, o_7\}$	$\{o_1, o_2, o_4, o_5\}$
o_5	$\{o_1, o_3, o_5, o_6\}$	$\{o_1, o_3, o_5\}$	$\{o_2, o_3, o_5, o_7\}$	$\{o_1, o_2, o_3, o_4, o_5, o_6, o_7\}$
o_6	$\{o_1, o_2, o_3, o_4, o_5, o_6, o_7\}$	$\{o_2, o_4, o_6, o_7\}$	$\{o_2, o_3, o_6, o_7\}$	$\{o_1, o_5, o_6\}$
o_7	$\{o_3, o_6, o_7\}$	$\{o_2, o_4, o_6, o_7\}$	$\{o_1, o_2, o_3, o_4, o_5, o_6, o_7\}$	$\{o_1, o_3, o_5, o_7\}$

(1) For $g_i(i = 1, 2, 3, 4)$, the E-accuracy of approximation is calculated as follows:

$$\alpha_{\{a_2,a_3,a_4\}}(g_1) = \frac{|\underline{R}_{\{a_2,a_3,a_4\}}(g_1)|}{|\overline{R}_{\{a_2,a_3,a_4\}}(g_1)|} = \frac{|\{o_1, o_3, o_5, o_6\}|}{|\{o_1, o_3, o_5, o_6\}|} = 1,$$
$$\alpha_{\{a_2,a_3\}}(g_1) = \frac{|\underline{R}_{\{a_2,a_3\}}(g_1)|}{|\overline{R}_{\{a_2,a_3\}}(g_1)|} = \frac{|\{o_1, o_3, o_5\}|}{|\{o_1, o_2, o_3, o_5, o_6, o_7\}|} = 0.5,$$
$$\alpha_{\{a_2,a_4\}}(g_1) = \frac{|\underline{R}_{\{a_2,a_4\}}(g_1)|}{|\overline{R}_{\{a_2,a_4\}}(g_1)|} = \frac{|\{o_1, o_3, o_5, o_6\}|}{|\{o_1, o_3, o_5, o_6\}|} = 1,$$
$$\alpha_{\{a_3,a_4\}}(g_1) = \frac{|\underline{R}_{\{a_3,a_4\}}(g_1)|}{|\overline{R}_{\{a_3,a_4\}}(g_1)|} = \frac{|\{o_6\}|}{|\{o_1, o_2, o_3, o_4, o_5, o_6, o_7\}|} \approx 0.1429.$$

$$\alpha_{\{a_1,a_3,a_4\}}(g_2) = \frac{|\underline{R}_{\{a_1,a_3,a_4\}}(g_2)|}{|\overline{R}_{\{a_1,a_3,a_4\}}(g_2)|} = \frac{|\{o_1, o_5\}|}{|\{o_1, o_3, o_5, o_7\}|} = 0.5,$$
$$\alpha_{\{a_1,a_3\}}(g_2) = \frac{|\underline{R}_{\{a_1,a_3\}}(g_2)|}{|\overline{R}_{\{a_1,a_3\}}(g_2)|} = \frac{|\{o_1, o_5\}|}{|\{o_1, o_2, o_3, o_4, o_5, o_6, o_7\}|} \approx 0.2857,$$
$$\alpha_{\{a_1,a_4\}}(g_2) = \frac{|\underline{R}_{\{a_1,a_4\}}(g_2)|}{|\overline{R}_{\{a_1,a_4\}}(g_2)|} = \frac{|\varnothing|}{|\{o_1, o_3, o_5, o_6, o_7\}|} = 0,$$
$$\alpha_{\{a_3,a_4\}}(g_2) = \frac{|\underline{R}_{\{a_3,a_4\}}(g_2)|}{|\overline{R}_{\{a_3,a_4\}}(g_2)|} = \frac{|\varnothing|}{|\{o_1, o_2, o_3, o_4, o_5, o_7\}|} = 0.$$

$$\alpha_{\{a_1,a_2,a_4\}}(g_3) = \frac{|\underline{R}_{\{a_1,a_2,a_4\}}(g_3)|}{|\overline{R}_{\{a_1,a_2,a_4\}}(g_3)|} = \frac{|\{o_2, o_4, o_7\}|}{|\{o_1, o_2, o_3, o_4, o_5, o_7\}|} = 0.5,$$
$$\alpha_{\{a_1,a_2\}}(g_3) = \frac{|\underline{R}_{\{a_1,a_2\}}(g_3)|}{|\overline{R}_{\{a_1,a_2\}}(g_3)|} = \frac{|\varnothing|}{|\{o_1, o_2, o_3, o_4, o_5, o_6, o_7\}|} = 0,$$
$$\alpha_{\{a_1,a_4\}}(g_3) = \frac{|\underline{R}_{\{a_1,a_4\}}(g_3)|}{|\overline{R}_{\{a_1,a_4\}}(g_3)|} = \frac{|\{o_2, o_4, o_7\}|}{|\{o_1, o_2, o_3, o_4, o_5, o_6, o_7\}|} \approx 0.4286,$$
$$\alpha_{\{a_2,a_4\}}(g_3) = \frac{|\underline{R}_{\{a_2,a_4\}}(g_3)|}{|\overline{R}_{\{a_2,a_4\}}(g_3)|} = \frac{|\{o_2, o_4, o_7\}|}{|\{o_1, o_2, o_3, o_4, o_5, o_7\}|} = 0.5.$$

$$\alpha_{\{a_1,a_2,a_3\}}(g_4) = \frac{|\underline{R}_{\{a_1,a_2,a_3\}}(g_4)|}{|\overline{R}_{\{a_1,a_2,a_3\}}(g_4)|} = \frac{|\{o_1, o_2, o_3, o_4, o_5, o_6, o_7\}|}{|\{o_1, o_2, o_3, o_4, o_5, o_6, o_7\}|} = 1,$$
$$\alpha_{\{a_1,a_2\}}(g_4) = \frac{|\underline{R}_{\{a_1,a_2\}}(g_4)|}{|\overline{R}_{\{a_1,a_2\}}(g_4)|} = \frac{|\{o_1, o_2, o_3, o_4, o_5, o_6, o_7\}|}{|\{o_1, o_2, o_3, o_4, o_5, o_6, o_7\}|} = 1,$$
$$\alpha_{\{a_1,a_3\}}(g_4) = \frac{|\underline{R}_{\{a_1,a_3\}}(g_4)|}{|\overline{R}_{\{a_1,a_3\}}(g_4)|} = \frac{|\{o_1, o_2, o_3, o_4, o_5, o_6, o_7\}|}{|\{o_1, o_2, o_3, o_4, o_5, o_6, o_7\}|} = 1,$$
$$\alpha_{\{a_2,a_3\}}(g_4) = \frac{|\underline{R}_{\{a_2,a_3\}}(g_4)|}{|\overline{R}_{\{a_2,a_3\}}(g_4)|} = \frac{|\{o_1, o_2, o_3, o_4, o_5, o_6, o_7\}|}{|\{o_1, o_2, o_3, o_4, o_5, o_6, o_7\}|} = 1.$$

(2) The corresponding P-degree of outlierness could be calculated accordingly:

$$DO_{\{a_1\}}(g_1) = 1 - |g_1|\frac{\alpha_{\{a_2,a_3,a_4\}}(g_1) + (\alpha_{\{a_2,a_3\}}(g_1) + \alpha_{\{a_2,a_4\}}(g_1) + \alpha_{\{a_3,a_4\}}(g_1) + 3)/2}{n(s+1)}$$
$$= 1 - |\{o_1, o_3, o_5, o_6\}|\frac{1 + (0.5 + 1 + 0.1429 + 3)/2}{7(3+1)}$$
$$\approx 0.5255,$$

$$DO_{\{a_2\}}(g_2) = 1 - |g_2|\frac{\alpha_{\{a_1,a_3,a_4\}}(g_2) + (\alpha_{\{a_1,a_3\}}(g_2) + \alpha_{\{a_1,a_4\}}(g_2) + \alpha_{\{a_3,a_4\}}(g_2) + 3)/2}{n(s+1)}$$
$$= 1 - |\{o_1, o_3, o_5\}|\frac{0.5 + (0.2857 + 0 + 0 + 3)/2}{7(3+1)}$$
$$\approx 0.7704,$$

$$DO_{\{a_3\}}(g_3) = 1 - |g_3|\frac{\alpha_{\{a_1,a_2,a_4\}}(g_3) + (\alpha_{\{a_1,a_2\}}(g_3) + \alpha_{\{a_1,a_4\}}(g_3) + \alpha_{\{a_2,a_4\}}(g_3) + 3)/2}{n(s+1)}$$
$$= 1 - |\{o_1, o_2, o_3, o_4, o_7\}|\frac{0.5 + (0 + 0.4286 + 0.5 + 3)/2}{7(3+1)}$$
$$\approx 0.5599.$$
$$DO_{\{a_4\}}(g_4) = 1 - |g_4|\frac{\alpha_{\{a_1,a_2,a_3\}}(g_4) + (\alpha_{\{a_1,a_2\}}(g_4) + \alpha_{\{a_1,a_3\}}(g_4) + \alpha_{\{a_2,a_3\}}(g_4) + 3)/2}{n(s+1)}$$
$$= 1 - |\{o_1, o_2, o_3, o_4, o_5, o_6, o_7\}|\frac{1 + (1 + 1 + 1 + 3)/2}{7(3+1)}$$
$$= 0.$$

(3) The weight function by Definition 3.3 is calculated as follows:

$$\omega_{a_1}(o_1) = 1 - \sqrt[3]{\frac{|R_{a_1}(o_1)|}{n}} = 1 - \sqrt[3]{\frac{|\{o_1, o_3, o_5, o_6\}|}{n}} = 1 - \sqrt[3]{\frac{4}{7}} \approx 0.1702,$$
$$\omega_{a_2}(o_1) = 1 - \sqrt[3]{\frac{|R_{a_2}(o_1)|}{n}} = 1 - \sqrt[3]{\frac{|\{o_1, o_3, o_5\}|}{n}} = 1 - \sqrt[3]{\frac{3}{7}} \approx 0.2461,$$
$$\omega_{a_3}(o_1) = 1 - \sqrt[3]{\frac{|R_{a_3}(o_1)|}{n}} = 1 - \sqrt[3]{\frac{|\{o_1, o_2, o_3, o_4, o_7\}|}{n}} = 1 - \sqrt[3]{\frac{5}{7}} \approx 0.1061,$$
$$\omega_{a_4}(o_1) = 1 - \sqrt[3]{\frac{|R_{a_4}(o_1)|}{n}} = 1 - \sqrt[3]{\frac{|\{o_1, o_2, o_3, o_4, o_5, o_6, o_7\}|}{n}} = 1 - \sqrt[3]{\frac{7}{7}} = 0.$$

(4) Finally, the outlier factor proposed in Definition 3.4 can be calculated as follows:

$$OF(o_1) = \frac{\sum_{j=1}^{m}\omega_{a_j}(o_1)DO_{a_j}(R_{a_j}(o_1))}{m} = \frac{0.1702 \times 0.5255 + 0.2461 \times 0.7704 + 0.1061 \times 0.5599 + 0 \times 0}{4} \quad (3.5)$$
$$\approx 0.0846.$$

(5) This way we can obtain all θ-outlier factors of every instances:

$OF(o_1) = 0.0846$, $OF(o_2) = 0.1104$, $OF(o_3) = 0.0732$, $OF(o_4) = 0.1253$, $OF(o_5) = 0.0992$, $OF(o_6) = 0.1077$, $OF(o_7) = 0.1013$.

Given $\mu = 0.12$, we have

$$OF(o_3) < OF(o_1) < OF(o_5) < OF(o_7) < OF(o_6) < OF(o_2) < \mu < OF(o_4),$$

Only o_4 has an OF higher than the threshold μ, and is thus taken it as an outlier according to Definition 3.5.

3.2 Outlier detection algorithms

Building upon the deviation detection methodology, this section presents an algorithm, denoted as ODIRG, tailored to perform outlier detection tasks and thoroughly examines its time complexity. Given an IRVIS (O, A), our approach portrayed as Algorithms 1 and 2. Considering an incomplete real-valued data with n objects, m attributes. The number of tolerance classes denotes as r.

Algorithm 1: Calculating $R_P(o)$

Input: An IRVIS (O, A), $P \subseteq A$ and $o \in O$.
Output: The tolerance class $R_P(o)$.
1: **for** $a \in P$ **do**
2: **for** $o \in O$ **do**
3: $a^*(o) = \frac{a(o) - \min V_a^*}{\max V_a^* - \min V_a^*}$ or $a^*(o) = 0$
4: **end for**
5: **end for**
6: $R_p(o) \leftarrow \{o' \in O : \forall a \in P, a^*(o) \approx_1 a^*(o')$ or $a^*(o) = *$ or $a^*(o') = *\}$;
7: return $R_P(o)$.

The computational complexity for Algorithm 1 is in the manner described. In steps 1–5, the computational complexity for figuring out $a^*(o)$ is $O(mn)$. In step 6, we compute the $R_p(o)$, which spends $O(n)$. Therefore, the total time complexity is $O(mn)$.

Algorithm 2: Outlier detection algorithm for an IRVIS based on class-consistent technology, rough set theory, and granular computing (ODIRG)

Input: An IRVIS (O, A).
Output: Calculating Ω_μ.
1: $\Omega_\mu \leftarrow \varnothing$
2: **for** $o \in O$ **do**
3: **for** $a \in A$ **do**
4: Obtain tolerance class $R_a(O)$ by Algorithm 1;
5: Calculate $\alpha_{A-\{a\}}(R_a(o))$;
6: **for** $b \in A - \{a\}$ **do**
7: Calculate $\alpha_{A-\{a,b\}}(R_a(o))$;
8: **end for**
9: Calculate $DO_a(R_a(o))$;
10: Calculate $\omega_a(o)$;
11: **end for**
12: Calculate $OF(o)$;
13: **if** $OF(o) > \mu$
14: $\Omega_\mu \leftarrow \Omega_\mu \cup \{o\}$;
15: **end if**
16: **end for**
17: return Ω_μ.

For Algorithm 2, the computational cost for steps 1–11 is $O(mn^2)$ according to Definitions 3.1–3.3. Steps 12 costs $O(1)$ according to Definition 3.4. Thus, the entire algorithmic complexity of algorithm 2 is equivalent to $O(mn^2)$.

4 The experimental result and analyses

This section presents the exploratory findings and conducts analyses.

4.1 Experimental setup

In the experiments, the designed ODIRG in this article is compared with cluster-based local outliers factor (CBLOF) algorithms, fuzzy information entropy-based outlier detection (FIEOD) algorithms, k-nearest neighbors (KNN) algorithms, sequence-based (SEQ) algorithms, and neighborhood outlier detection (NOOF) algorithms. We will assess the efficacy of all external detection algorithms in this section using the metrics introduced by Aggarwal and Yu [37].

Generally speaking, outliers have a higher probability of occurring in the minority class. Therefore, we assume that outliers are distributed within the minority class of the dataset. For each algorithm, we initially compute the abnormal values of each object in the dataset and reshuffle them in dropping order. Objects with higher outlier values after rearrangement are prone to be deviations. Then, we select objects from the ascending portion of the reordered dataset to determine how many of them belong to the rare class. A higher probability of chosen objects in the minority class indicates better performance.

The designed outlying observation identification algorithm employs a threshold, denoted as μ, to classify objects as outliers. This threshold is established relying on the quantity of objects in the uncommon class. Specifically, the algorithm computes outliers for each object, resulting in a sequence arranged in descending order. The μ threshold is then set as the value of the mth outlier in this sequence, where m represents the amount of objects in the rare class within the dataset.

The experiments in this section utilize a computer equipped with an Intel Core i5-8250U processor running at a frequency of 1.60 GHz and 4 GB of memory. The operating system employed is Windows 10, and the experiments are conducted using Python 3.9. The Python Integrated Development Environment (IDE) utilized is PyCharm 2022.2.1.

4.2 Datasets and experiment results

The experiment employs seven datasets sourced from UCI Machine Learning Repository and KEEL: Mammographic Masses, Breast Cancer Wisconsin (Original), Hayes, Breast Cancer Wisconsin (Diagnose), Newthyriod, Credit Approval, and Pima. In general, a small subset of objects in each dataset is considered as outliers. It is usually defined on a case by case based on a specific dataset. For instance, dataset Breast Cancer Wisconsin (Diagnose), malignant objects are considered as outliers.

Public datasets are rarely directly applicable to outlier detection tasks. Preprocessing is necessary for outlier detection tasks with an IRVIS dataset. This involves transforming the dataset into a table representing an information system, with each row representing an object and each column representing an attribute. In addition, it involves introducing random 1% data loss for all information values in each information system table. Due to the balanced distribution of decision classes in most of the chosen seven datasets, they are unsuitable for outlier detection. Consequently, the approach adopted in this article leverages the methodologies proposed in [38,39] to induce an imbalanced distribution within the decision class datasets for outlier detection. Outliers are generated by randomly downsampling a specific class, while all objects from the other classes are retained to construct a dataset. Table 3 supplies an overview of the datasets utilized in the study.

Mammographic masses dataset The mammographic masses dataset from UCI have 961 objects and six attributes. The decision class in the last column can be divided into two parts: “benign(0)” and “malignant(1)” which contain 516 and 445 objects, respectively. Hence, preprocessing the dataset is necessary to enhance its suitability for outlier detection tasks. The experimental technique randomly downsampled some objects in the

Table 3: Statistics of the public benchmark datasets

	#Attributes	#Objects	#Preprocessing	# Outlier ratio
Mammographic masses	6	611	Downsampling class “malignant(1)” to 95 objects	15.55%
Breast Cancer Wisconsin (Original)	10	699	Class “Malignant(4)” is treat as outliers	34.50%
Hayes	5	132	Class “3” is treated as outliers	22.73%
Breast Cancer Wisconsin (Diagnose)	32	569	Class “M” is treat as outliers	37.26%
Newthyriod	6	215	Class “positive” is treated as outliers	16.28%
Credit approval	16	449	Downsampling class “+” to 66 objects	14.70%
Pima	9	768	Class “positive” is treat as outliers	34.90%

“1” (malignant) class to 95, making this class rare. The final experimental subjects were 611, and the proportion of rare classes was 15.55% (Table 3).

(1) The breakpoint μ, obtained from the deviation ratio, stands at 0.2257, as illustrated in Figure 2.

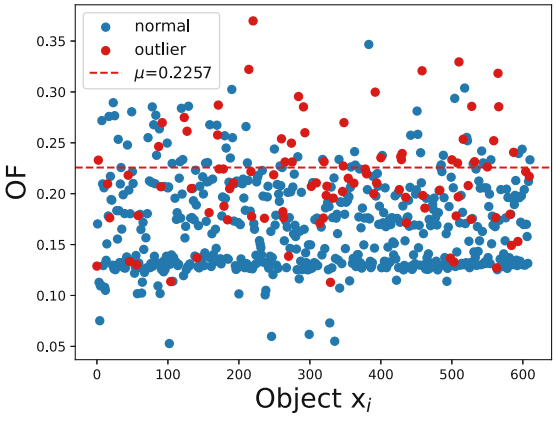

Figure 2: Object distribution for mammographic masses dataset. (The image was created by the authors.).

(2) Table 4 displays that of the top 20 objects most probable to be true deviations, the ODIRG algorithm find out 11 true outliers. Compared with, CBLOF algorithm figures out four genuine outliers, FIEOD algorithm figures out zero true anomalies, KNN algorithm figures out zero authentic outliers, SEQ algorithm figures out five legitimate deviations, and NOOF algorithm figures out two actual anomalies. ODIRG algorithm has always been proved that has the best outlier detection rate in this article in subsequent comparisons.

Table 4: The experimental result in Mammographic Masses dataset

Top ratio (%) (number of objects)	Number of rare classes included (coverage (%))					
	ODIRG	CBLOF	FIEOD	KNN	SEQ	NOOF
0.03 (20)	**11 (11.5789)**	4 (4.2105)	0 (0.0)	0 (0.0)	5 (5.2632)	2 (2.1053)
0.07 (40)	**17 (17.8947)**	5 (5.2632)	4 (4.2105)	5 (5.2632)	14 (14.7368)	7 (7.3684)
0.1 (60)	**24 (25.2632)**	8 (8.4211)	8 (8.4211)	9 (9.4737)	23 (24.2105)	7 (7.3684)
0.13 (80)	**33 (34.7368)**	12 (12.6316)	9 (9.4737)	10 (10.5263)	31 (32.6316)	7 (7.3684)
0.16 (100)	**38 (40.0)**	14 (14.7368)	10 (10.5263)	11 (11.5789)	37 (38.9474)	8 (8.4211)
0.2 (120)	**45 (47.3684)**	19 (20.0)	13 (13.6842)	14 (14.7368)	44 (46.3158)	20 (21.0526)

Bold values represent the optimal experimental results and play a prominent role.

This article presents scatter charts and detection rate comparison charts for six datasets simultaneously.

Breast Cancer Wisconsin (Original) dataset The Breast Cancer Wisconsin (Original) dataset from UCI comprises 699 instances with 10 features, categorized into two classes: “benign(2)” (selected as inliers) and “malignant(4)” (chosen as outliers). The “malignant(4)” accounts for 34.48% of the entirety dataset (Table 3). Based on the underrepresented class percentage and the limit μ mathematical approach, the cutoff μ for this dataset can be determined, as depicted in Figure 3(a). As illustrated in Table 5, SEQ algorithm exhibits the best performance, particularly during the initial and middle detection stages. The SEQ algorithm has a unique advantage in capturing the time series characteristics of data, and its outstanding performance during the initial and middle detection stages may be attributed to the prominent time series features present in the dataset. The ODIRG algorithm and the NOOF algorithm demonstrate similar performance. The NOOF algorithm excels at capturing local features of data. The proximity in performance between the ODIRG algorithm and the NOOF algorithm suggests that the ODIRG algorithm possesses excellent performance in local search. In addition, the ODIRG algorithm outperforms the NOOF algorithm in global search. It is evident that the ODIRG algorithm exhibits superior performance in anomaly detection for an IRVIS.

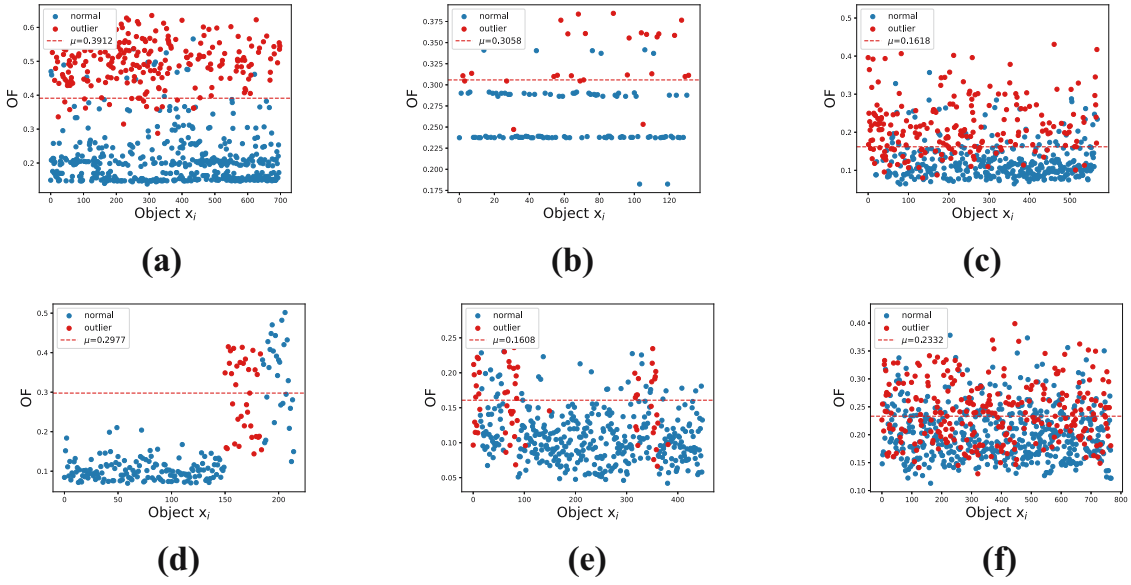

Figure 3: Object distribution scatter chart results. (These image were created by the authors.) (a) Breast Cancer Wisconsin (Original), (b) Hayes, (c) Breast Cancer Wisconsin (Diagnose), (d) Newthyriod, (e) credit approval, and (f) Pima.

Table 5: The experimental result in Breast Cancer Wisconsin (Original) dataset

Top ratio (%) (number of objects)	Number of rare classes included (coverage (%))					
	ODIRG	**CBLOF**	**FIEOD**	**KNN**	**SEQ**	**NOOF**
0.01 (10)	**10 (4.1494)**	8 (3.3195)	**10 (4.1494)**	9 (3.7344)	**10 (4.1494)**	**10 (4.1494)**
0.03 (20)	**20 (8.2988)**	17 (7.0539)	19 (7.8838)	18 (7.4689)	19 (7.8838)	19 (7.8838)
0.04 (30)	**29 (12.0332)**	27 (11.2033)	**29 (12.0332)**	28 (11.6183)	**29 (12.0332)**	**29 (12.0332)**
0.06 (40)	**39 (16.1826)**	36 (14.9378)	38 (15.7676)	37 (15.3527)	**39 (16.1826)**	**39 (16.1826)**
0.07 (50)	48 (19.917)	44 (18.2573)	48 (19.917)	47 (19.5021)	**49 (20.332)**	**49 (20.332)**
0.09 (60)	58 (24.0664)	53 (21.9917)	58 (24.0664)	56 (23.2365)	**59 (24.4813)**	58 (24.0664)

Bold values represent the optimal experimental results and play a prominent role.

Hayes dataset The Hayes dataset from KEEL comprises 132 objects categorized into three classes based on five attributes. Among these classes, "3" is considered a rare class due to its relatively small proportion in the dataset, accounting for 22.73% (Table 3). Consequently, the boundary μ is determined as the outlierness ranking at 22.73% in a decreasing order (Figure 3(b)). The distribution of Hayes dataset in the figure shows that the detection ratio is very well. As Table 6 shows that the performance of four algorithms are well. ODIRG algorithm is tied for best with FIEOD algorithm, KNN algorithm and NOOF algorithm, next is CBLOF algorithm, the SEQ algorithm is the worst. The calculation process of the FIEOD algorithm may be relatively complex, and its performance depends on the setting of parameters. The performance of the KNN algorithm and the NOOF algorithm is easily affected by the data distribution. Therefore, ODIRG is slightly better. From the data results, the ODIRG algorithm is slightly better.

Table 6: The experimental result in Hayes dataset

Top ratio (%) (number of objects)	Number of rare classes included (coverage (%))					
	ODIRG	CBLOF	FIEOD	KNN	SEQ	NOOF
0.61 (80)	**30 (100.0)**	29 (96.6667)	**30 (100.0)**	**30 (100.0)**	28 (93.3333)	**30 (100.0)**
0.68 (90)	**30 (100.0)**	29 (96.6667)	**30 (100.0)**	**30 (100.0)**	28 (93.3333)	**30 (100.0)**
0.76 (100)	**30 (100.0)**	**30 (100.0)**	**30 (100.0)**	**30 (100.0)**	28 (93.3333)	**30 (100.0)**
0.83 (110)	**30 (100.0)**	**30 (100.0)**	**30 (100.0)**	**30 (100.0)**	28 (93.3333)	**30 (100.0)**
0.91 (120)	**30 (100.0)**	**30 (100.0)**	**30 (100.0)**	**30 (100.0)**	28 (93.3333)	**30 (100.0)**
0.98 (130)	**30 (100.0)**	**30 (100.0)**	**30 (100.0)**	**30 (100.0)**	**30 (100.0)**	**30 (100.0)**

Bold values represent the optimal experimental results and play a prominent role.

Breast Cancer Wisconsin (Diagnose) dataset The Breast Cancer Wisconsin (Diagnose) dataset from UCI contains 569 objects with 32 attributes and can be classified as two classes: "M" and "B." "M" means the cancer is diagnosed as malignant, and "B" means the cancer is diagnosed as benign. According to the actual means and the proportion of the Breast Cancer Wisconsin (Diagnose) dataset, "M" is considered a rare class, with a ratio of 37.26% (Table 3). Therefore, the outlier with 37.26% in a declining order was selected as the point of demarcation μ = 0.1618 (Figure 3(c)). As shown in Table 7, the ODIRG algorithm performs exceptionally well, and the SEQ algorithm also shows good performance. However, the average performance of the other algorithms is relatively poor. This indicates that the performance of these other algorithms fluctuates significantly due to the incompleteness of the data, while our proposed algorithm is more robust.

Table 7: The experimental result in breast cancer wisconsin (diagnose) dataset

Top ratio (%)	Number of rare classes included (coverage(%))					
(Number of objects)	ODIRG	CBLOF	FIEOD	KNN	SEQ	NOOF
0.26(150)	**123(58.0189)**	102(48.1132)	119(56.1321)	99(46.6981)	**123(58.0189)**	115(54.2453)
0.28(160)	**130(61.3208)**	112(52.8302)	126(59.434)	103(48.5849)	129(60.8491)	122(57.5472)
0.3(170)	**136(64.1509)**	120(56.6038)	131(61.7925)	106(50.0)	135(63.6792)	124(58.4906)
0.32(180)	**142(66.9811)**	124(58.4906)	137(64.6226)	115(54.2453)	141(66.5094)	129(60.8491)
0.33(190)	**148(69.8113)**	125(58.9623)	141(66.5094)	117(55.1887)	147(69.3396)	133(62.7358)
0.35(200)	**155(73.1132)**	127(59.9057)	143(67.4528)	122(57.5472)	153(72.1698)	140(66.0377)

Bold values represent the optimal experimental results and play a prominent role.

Newthyriod dataset In the Newthyriod dataset from KEEL, there are 215 objects with 6 attributes. The decision column, divided into "positive" and other parts, designates "positive" as the outlier fraction. By

considering the proportion of the outlier fraction and applying the threshold μ calculation method, the breakpoint μ for the Newthyriod dataset is determined (Figure 3(d)). Analyzing Table 8, the ODIRG algorithm demonstrates high detection ratios, followed by gradual improvement in performance by the FIEOD algorithm. However, towards the end, the ODIRG, the FIEOD, and KNN algorithms show significant enhancement. It is evident that the ODIRG algorithm exhibits the most robust outlier detection capability in the Newthyriod dataset.

Table 8: The experimental result in Newthyriod dataset

Top ratio (%) (number of objects)	Number of rare classes included (coverage (%))					
	ODIRG	**CBLOF**	**FIEOD**	**KNN**	**SEQ**	**NOOF**
0.23 (50)	**25 (71.4286)**	**25 (71.4286)**	**25 (71.4286)**	20 (57.1429)	23 (65.7143)	23 (65.7143)
0.28 (60)	**29 (82.8571)**	28 (80.0)	**29 (82.8571)**	25 (71.4286)	28 (80.0)	26 (74.2857)
0.33 (70)	**34 (97.1429)**	28 (80.0)	31 (88.5714)	29 (82.8571)	31 (88.5714)	30 (85.7143)
0.37 (80)	**35 (100.0)**	28 (80.0)	**35 (100.0)**	34 (97.1429)	31 (88.5714)	33 (94.2857)
0.42 (90)	**35 (100.0)**	28 (80.0)	**35 (100.0)**	**35 (100.0)**	33 (94.2857)	33 (94.2857)
0.47 (100)	**35 (100.0)**	28 (80.0)	**35 (100.0)**	**35 (100.0)**	33 (94.2857)	33 (94.2857)

Bold values represent the optimal experimental results and play a prominent role.

Credit approval dataset The Credit Approval dataset from UCI have 690 objects and 16 attributes. The decision class in the last column has only two parts "+" and "−" , where the objects of "+" are 307 and the objects of "−" are 383. The Credit Approval dataset is too balanced to deal with the outlier detected mission. Therefore, this article adopted the method that is randomly remove some objects of the less part. Table 3 illustrates that among the final trial objects totaling 449, the rare class constitutes 14.70%. The μ of Credit Approval dataset can be clearly seen in (Figure 3(e)). Table 9 indicates that the ODIRG algorithm consistently delivered strong performance alongside the SEQ and NOOF algorithms. However, it is evident that the SEQ and NOOF algorithms slightly lag behind the KNN algorithm in effectiveness.

Table 9: The experimental result in credit approval dataset

Top ratio (%) (number of objects)	Number of rare classes included (coverage (%))					
	ODIRG	**CBLOF**	**FIEOD**	**KNN**	**SEQ**	**NOOF**
0.13 (60)	**30 (45.4545)**	13 (19.697)	18 (27.2727)	13 (19.697)	28 (42.4242)	**30 (45.4545)**
0.22 (100)	**41 (62.1212)**	22 (33.3333)	26 (39.3939)	25 (37.8788)	**41 (62.1212)**	39 (59.0909)
0.31 (140)	**47 (71.2121)**	26 (39.3939)	35 (53.0303)	32 (48.4848)	46 (69.697)	**47 (71.2121)**
0.4 (180)	51 (77.2727)	40 (60.6061)	48 (72.7273)	44 (66.6667)	**52 (78.7879)**	49 (74.2424)
0.49 (220)	56 (84.8485)	48 (72.7273)	52 (78.7879)	**59 (89.3939)**	56 (84.8485)	51 (77.2727)
0.58 (260)	56 (84.8485)	57 (86.3636)	56 (84.8485)	**63 (95.4545)**	56 (84.8485)	56 (84.8485)

Bold values represent the optimal experimental results and play a prominent role.

Pima dataset The Pima dataset from KEEL is composed by 768 objects and 9 attributes. The last column is decision that can be classified into two classes. "positive" is regarded as outlier part, because the positive class is accounting for only 34.90% (Table 3). The threshold μ in the Pima dataset can be computed through the computing way of threshold μ (Figure 3(f)). From the Table 10, ODIRG algorithm and SEQ algorithm are performed well in the whole process, but as we can see, the outlier detection ability of ODIRG algorithm is superior than the SEQ algorithm in Pima dataset.

A concise comparison between the six outlier detection methods is presented in Table 11.

Table 10: The experimental result in Pima dataset

Top ratio (%) (number of objects)	Number of rare classes included (coverage (%))					
	ODIRG	**CBLOF**	**FIEOD**	**KNN**	**SEQ**	**NOOF**
0.01 (10)	**7 (2.6119)**	2 (0.7463)	5 (1.8657)	3 (1.1194)	**7 (2.6119)**	3 (1.1194)
0.03 (20)	14 (5.2239)	8 (2.9851)	6 (2.2388)	6 (2.2388)	**15 (5.597)**	8 (2.9851)
0.04 (30)	**22 (8.209)**	10 (3.7313)	12 (4.4776)	13 (4.8507)	21 (7.8358)	14 (5.2239)
0.05 (40)	**28 (10.4478)**	16 (5.9701)	18 (6.7164)	17 (6.3433)	**28 (10.4478)**	19 (7.0896)
0.07 (50)	**35 (13.0597)**	20 (7.4627)	24 (8.9552)	22 (8.209)	**35 (13.0597)**	23 (8.5821)
0.08 (60)	42 (15.6716)	23 (8.5821)	31 (11.5672)	23 (8.5821)	**43 (16.0448)**	29 (10.8209)

Bold values represent the optimal experimental results and play a prominent role.

Table 11: Outlier detection method comparison

Method	Superiority	Inferiority
ODIRG	High efficiency	High time complexity
CBLOF	Handle local outliers	Low performance
FIEOD	Strong adaptability	Complex computation
KNN	Adaptability for classification	High time complexity
NOOF	Relatively simple	Under use of data information
SEQ	Deal with sequence	Discretization pretreatment for numeric data

5 Evaluation analyses

This section uses two quantitative criteria for assessing experimental results: (1) precision (P), recall (R), and $F1$ measure ($F1$); (2) receiver operating characteristic (ROC) curves and area under the curve (AUC).

5.1 P, R and $F1$

Every algorithm has the capability to assign an outlier factor to every object present in a dataset. This factor essentially represents the probability of an object being an outlier. By arranging all the objects in the dataset in a descending order based on their outlier factors, we can easily identify the first t objects as outliers when given a threshold value t.

The groups you aim to identify as outliers (all ground truth outliers) are regarded as the positive class, whereas the remaining categories are seen as the negative class. When the t objects are chosen, there are four potential scenarios, as outlined in Table 12.

Table 12: Confusion matrix for predicting outliers

	Predicted outlier	**Predicted inlier**
Actual outlier	TP	FN
Actual inlier	FP	TN

Table 13: Comparison results with precision, recall and $F1$-measure

Data sets		ODIRG			CBLOF			FIEOD			KNN			NOOF			SEQ		
	t	$P(t)$	$R(t)$	$F1$	$P(t)$	$R(t)$	$F1$	$P(t)$	$R(t)$	$F1$	$P(t)$	$R(t)$	$F1$	$P(t)$	$R(t)$	$F1$	$P(t)$	$R(t)$	$F1$
Mammographic masses	120	0.375	0.4737	0.4186	0.1583	0.2	0.1767	0.1083	0.1368	0.1209	0.1167	0.1474	0.1302	0.3667	0.4632	0.4093	0.1667	0.2105	0.186
	180	0.3167	0.6	0.4145	0.2167	0.4105	0.2836	0.1722	0.3263	0.2255	0.1333	0.2526	0.1745	0.3333	0.6316	0.4364	0.2444	0.4632	0.32
	240	0.2792	0.7053	0.4	0.2042	0.5158	0.2925	0.2208	0.5579	0.3164	0.1625	0.4105	0.2328	0.2917	0.7368	0.4179	0.2458	0.6211	0.3522
	300	0.2667	0.8421	0.4051	0.17	0.5368	0.2582	0.24	0.7579	0.3646	0.1867	0.5895	0.2835	0.27	0.8526	0.4101	0.2267	0.7158	0.3443
	360	0.2306	0.8737	0.3648	0.1444	0.5474	0.2286	0.2417	0.9158	0.3824	0.2278	0.8632	0.3604	0.2333	0.8842	0.3692	0.225	0.8526	0.356
	420	0.2071	0.9158	0.3379	0.1238	0.5474	0.2019	0.2167	0.9579	0.3534	0.2119	0.9368	0.3456	0.2071	0.9158	0.3379	0.2119	0.9368	0.3456
	Average	0.2792	0.7351	0.3902	0.1696	0.4597	0.2402	0.2	0.6088	0.2939	0.1732	0.5333	0.2545	0.2837	0.7474	**0.3968**	0.2201	0.6333	0.3174
Breast Cancer Wisconsin (Original)	120	0.9667	0.4813	0.6427	0.8417	0.4191	0.5596	0.9833	0.4896	0.6537	0.9333	0.4647	0.6205	0.975	0.4855	0.6482	0.9	0.4481	0.5983
	180	0.9389	0.7012	0.8029	0.8444	0.6307	0.7221	0.9722	0.7261	0.8314	0.9167	0.6846	0.7838	0.9556	0.7137	0.8171	0.8111	0.6058	0.6936
	240	0.925	0.9212	0.9231	0.8542	0.8506	0.8524	0.9125	0.9087	0.9106	0.8958	0.8921	0.894	0.925	0.9212	0.9231	0.8375	0.834	0.8358
	300	0.8	0.9959	0.8872	0.7933	0.9876	0.8799	0.7967	0.9917	0.8835	0.8033	1.0	0.8909	0.8033	1.0	0.8909	0.7667	0.9544	0.8503
	360	0.6694	1.0	0.802	0.6667	0.9959	0.7987	0.6694	1.0	0.802	0.6694	1.0	0.802	0.6694	1.0	0.802	0.6694	1.0	0.802
	420	0.5738	1.0	0.7292	0.5714	0.9959	0.7262	0.5738	1.0	0.7292	0.5738	1.0	0.7292	0.5738	1.0	0.7292	0.5738	1.0	0.7292
	Average	0.8123	0.8499	0.7978	0.762	0.8133	0.7565	0.818	0.8527	0.8017	0.7987	0.8402	0.7867	0.817	0.8534	**0.8018**	0.7598	0.807	0.7515
Hayes	120	0.25	1.0	0.4	0.25	1.0	0.4	0.25	1.0	0.4	0.25	1.0	0.4	0.2333	0.9333	0.3733	0.25	1.0	0.4
	180	0.2273	1.0	0.3704	0.2273	1.0	0.3704	0.2273	1.0	0.3704	0.2273	1.0	0.3704	0.2273	1.0	0.3704	0.2273	1.0	0.3704
	240	0.2273	1.0	0.3704	0.2273	1.0	0.3704	0.2273	1.0	0.3704	0.2273	1.0	0.3704	0.2273	1.0	0.3704	0.2273	1.0	0.3704
	300	0.2273	1.0	0.3704	0.2273	1.0	0.3704	0.2273	1.0	0.3704	0.2273	1.0	0.3704	0.2273	1.0	0.3704	0.2273	1.0	0.3704
	360	0.2273	1.0	0.3704	0.2273	1.0	0.3704	0.2273	1.0	0.3704	0.2273	1.0	0.3704	0.2273	1.0	0.3704	0.2273	1.0	0.3704
	420	0.2273	1.0	0.3704	0.2273	1.0	0.3704	0.2273	1.0	0.3704	0.2273	1.0	0.3704	0.2273	1.0	0.3704	0.2273	1.0	0.3704
	Average	0.2311	1.0	**0.3753**	0.2311	1.0	**0.3753**	0.2311	1.0	**0.3753**	0.2311	1.0	**0.3753**	0.2283	0.9889	0.3709	0.2311	1.0	**0.3753**
Breast cancer wisconsin (diagnose)	120	0.825	0.467	0.5964	0.625	0.3538	0.4518	0.85	0.4811	0.6145	0.7083	0.4009	0.512	0.825	0.467	0.5964	0.775	0.4387	0.5602
	180	0.7889	0.6698	0.7245	0.6889	0.5849	0.6327	0.7611	0.6462	0.699	0.6389	0.5425	0.5867	0.7833	0.6651	0.7194	0.7167	0.6085	0.6582
	240	0.7292	0.8255	0.7743	0.5917	0.6698	0.6283	0.6542	0.7406	0.6947	0.5667	0.6415	0.6018	0.7083	0.8019	0.7522	0.6417	0.7264	0.6814
	300	0.6467	0.9151	0.7578	0.5833	0.8255	0.6836	0.5867	0.8302	0.6875	0.5533	0.783	0.6484	0.6433	0.9104	0.7539	0.57	0.8066	0.668
	360	0.5611	0.9528	0.7063	0.5611	0.9528	0.7063	0.5306	0.9009	0.6678	0.5333	0.9057	0.6713	0.5611	0.9528	0.7063	0.5139	0.8726	0.6469
	420	0.4952	0.9811	0.6582	0.4857	0.9623	0.6456	0.4833	0.9575	0.6424	0.4952	0.9811	0.6582	0.4952	0.9811	0.6582	0.4619	0.9151	0.6139
	Average	0.6744	0.8019	**0.7029**	0.5893	0.7248	0.6247	0.6443	0.7594	0.6677	0.5826	0.7091	0.6131	0.6694	0.7964	0.6977	0.6132	0.728	0.6381
Newthyriod	120	0.2917	1.0	0.4516	0.2333	0.8	0.3613	0.2917	1.0	0.4516	0.2917	1.0	0.4516	0.2917	1.0	0.4516	0.2917	1.0	0.4516
	180	0.1944	1.0	0.3256	0.1556	0.8	0.2605	0.1944	1.0	0.3256	0.1944	1.0	0.3256	0.1944	1.0	0.3256	0.1944	1.0	0.3256
	240	0.1628	1.0	0.28	0.1628	1.0	0.28	0.1628	1.0	0.28	0.1628	1.0	0.28	0.1628	1.0	0.28	0.1628	1.0	0.28

(*Continued*)

Table 13: *Continued*

Data sets		ODIRG			CBLOF			FIEOD			KNN			NOOF			SEQ		
	t	$P(t)$	$R(t)$	$F1$	$P(t)$	$R(t)$	$F1$	$P(t)$	$R(t)$	$F1$	$P(t)$	$R(t)$	$F1$	$P(t)$	$R(t)$	$F1$	$P(t)$	$R(t)$	$F1$
	300	0.1628	1.0	0.28	0.1628	1.0	0.28	0.1628	1.0	0.28	0.1628	1.0	0.28	0.1628	1.0	0.28	0.1628	1.0	0.28
	360	0.1628	1.0	0.28	0.1628	1.0	0.28	0.1628	1.0	0.28	0.1628	1.0	0.28	0.1628	1.0	0.28	0.1628	1.0	0.28
	420	0.1628	1.0	0.28	0.1628	1.0	0.28	0.1628	1.0	0.28	0.1628	1.0	0.28	0.1628	1.0	0.28	0.1628	1.0	0.28
	Average	0.1896	1.0	**0.3162**	0.1734	0.9333	0.2903	0.1896	1.0	**0.3162**	0.1896	1.0	**0.3162**	0.1896	1.0	**0.3162**	0.1896	1.0	**0.3162**
Credit approval	120	0.3583	0.6515	0.4624	0.2	0.3636	0.2581	0.2667	0.4848	0.3441	0.2417	0.4394	0.3118	0.3667	0.6667	0.4731	0.375	0.6818	0.4839
	180	0.2833	0.7727	0.4146	0.2222	0.6061	0.3252	0.2667	0.7273	0.3902	0.2444	0.6667	0.3577	0.2889	0.7879	0.4228	0.2722	0.7424	0.3984
	240	0.2333	0.8485	0.366	0.2167	0.7879	0.3399	0.225	0.8182	0.3529	0.2625	0.9545	0.4118	0.2333	0.8485	0.366	0.225	0.8182	0.3529
	300	0.2033	0.9242	0.3333	0.2167	0.9848	0.3552	0.2	0.9091	0.3279	0.2167	0.9848	0.3552	0.1967	0.8939	0.3224	0.1967	0.8939	0.3224
	360	0.175	0.9545	0.2958	0.1806	0.9848	0.3052	0.1778	0.9697	0.3005	0.1806	0.9848	0.3052	0.1722	0.9394	0.2911	0.1722	0.9394	0.2911
	420	0.1571	1.0	0.2716	0.1571	1.0	0.2716	0.1571	1.0	0.2716	0.1571	1.0	0.2716	0.15	0.9545	0.2593	0.1548	0.9848	0.2675
	Average	0.235	0.8586	**0.3573**	0.1989	0.7879	0.3092	0.2156	0.8182	0.3312	0.2172	0.8384	0.3355	0.2346	0.8485	0.3558	0.2326	0.8434	0.3527
Pima	120	0.6417	0.2873	0.3969	0.4583	0.2052	0.2835	0.575	0.2575	0.3557	0.5167	0.2313	0.3196	0.7333	0.3284	0.4536	0.4917	0.2201	0.3041
	180	0.6167	0.4142	0.4955	0.5056	0.3396	0.4062	0.5944	0.3993	0.4777	0.5111	0.3433	0.4107	0.6556	0.4403	0.5268	0.5111	0.3433	0.4107
	240	0.5792	0.5187	0.5472	0.5125	0.459	0.4843	0.5542	0.4963	0.5236	0.525	0.4701	0.4961	0.5917	0.5299	0.5591	0.4875	0.4366	0.4606
	300	0.52	0.5821	0.5493	0.5033	0.5634	0.5317	0.5367	0.6007	0.5669	0.54	0.6045	0.5704	0.5367	0.6007	0.5669	0.4433	0.4963	0.4683
	360	0.4944	0.6642	0.5669	0.4778	0.6418	0.5478	0.5139	0.6903	0.5892	0.5056	0.6791	0.5796	0.4944	0.6642	0.5669	0.4417	0.5933	0.5064
	420	0.4714	0.7388	0.5756	0.4476	0.7015	0.5465	0.481	0.7537	0.5872	0.4857	0.7612	0.593	0.4643	0.7276	0.5669	0.419	0.6567	0.5116
	Average	0.5539	0.5342	0.5219	0.4842	0.4851	0.4667	0.5425	0.533	0.5167	0.514	0.5149	0.4949	0.5793	0.5485	**0.54**	0.4657	0.4577	0.4436

Bold values represent the optimal experimental results and play a prominent role.

Table 12 provides a summary of the confusion matrix for the ROC curve, consisting of four possible outcomes in outlier forecasting: when an outlier is correctly identified as an outlier (true positive, TP), when an outlier is mistakenly classified as an inlier (false negative, FN), when an inlier is incorrectly classified as an outlier (false positive, FP), and when an inlier is correctly identified as an inlier (true negative, TN).

Then, P, R, and $F1$ can be calculated as follows:

$$P = \frac{\text{TP}}{\text{TP} + \text{FP}},$$
$$R = \frac{TP}{\text{TP} + \text{FN}},$$
$$F1 = \frac{2 \times P \times R}{P + R},$$

where TP, FP, and FN are the identical means as talked about earlier. The $F1$ metric is the balance between P and R and is typically used to evaluate the performance of an algorithm.

When given a value of t, P represents the proportion of actual detected outliers to the total number of selected objects, while R denotes the proportion of actual outliers detected at a specific t to the total number of outliers in a dataset. It is evident that for a given t, the algorithm's outlier detection performance improves as both P and R increase. Furthermore, when comparing algorithms with the same P or R, a smaller t indicates stronger performance. In addition, $F1$ is the harmonic mean of P and R and is used to assess the detection effectiveness by combining P and R. If $P = 0$ and $R = 0$, we define $F1 = 0$ to indicate that the algorithm's detection efficiency is not promising.

Table 13 displays the comparison results of P, R, and $F1$ in seven datasets.

In the experiment, we set the threshold t for the six algorithms within an interval from the number of actual outliers to twice that number. This interval, spanning from the total number of actual outliers to double that count, was evenly divided into five segments. Subsequently, we assigned the six subinterval endpoint values to t in succession to calculate the P, R, and $F1$ of each algorithm, followed by a comparison of their respective average values. For example, the number of outliers in the preprocessed Mammographic Masses dataset is 95, and accordingly, the threshold t was set to 95, 105, 115, 125, 135, and 145. In addition, when the R reaches 1 for the first time, it means that all real outliers in a dataset have been detected. At this time, the smaller the threshold t (denoted as t') corresponding to each algorithm, the better the algorithm performance.

As shown in Figure 4, when t is equal to the number of all actual outliers in a dataset, the P of the ODIRG algorithm is higher than that of other algorithms. For example, if $t = 100$ in Mammographic Masses dataset, P of the ODIRG algorithm is 0.38, while P of the CBLOF, FIEOD, KNN, SEQ and NOOF algorithm is 0.14, 0.10, 0.11, 0.37, and 0.08, respectively.

The comparison chart of detection rate outcomes for the six datasets are shown in Figure 5 simultaneously.

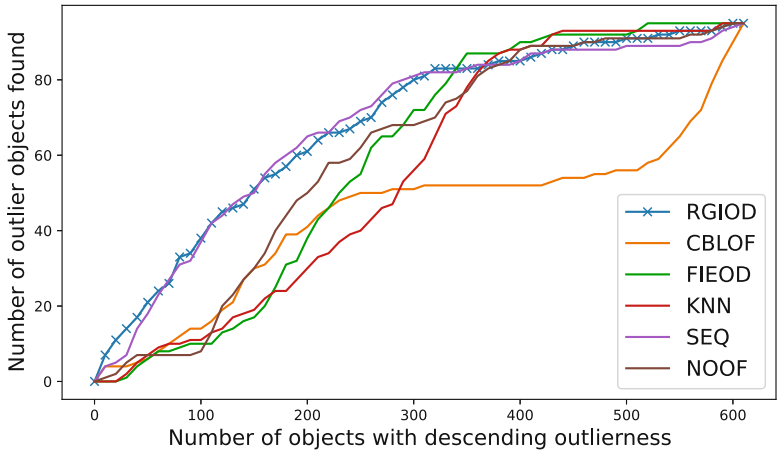

Figure 4: t' value in Mammographic Masses dataset. (The image was created by the authors.).

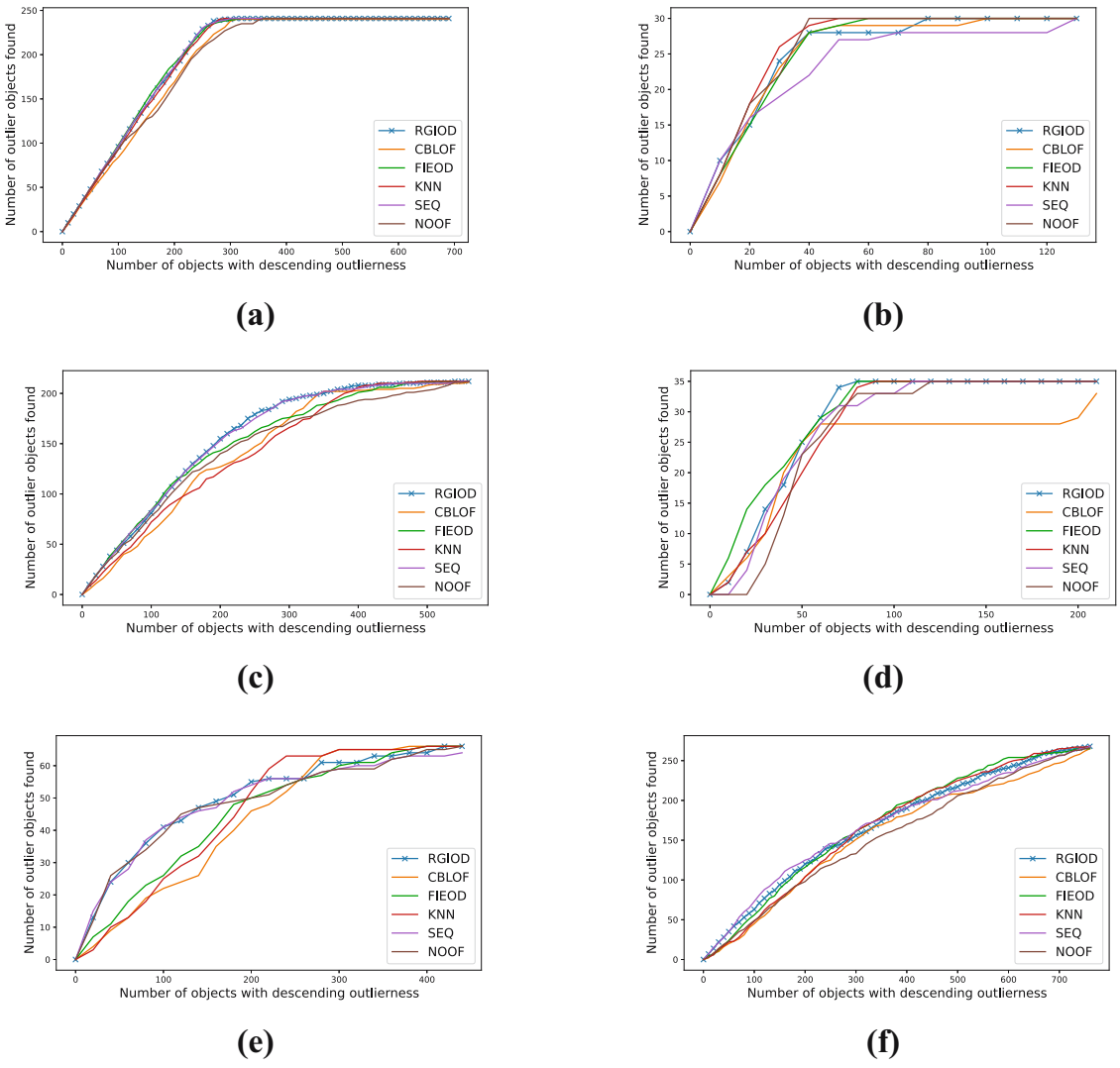

Figure 5: Comparison chart of detection rate outcomes. (These image were created by the authors.) (a) Breast Cancer Wisconsin (original), (b) Hayes, (c) Breast Cancer Wisconsin (Diagnose), (d) newthyroid, (e) credit approval, and (f) Pima.

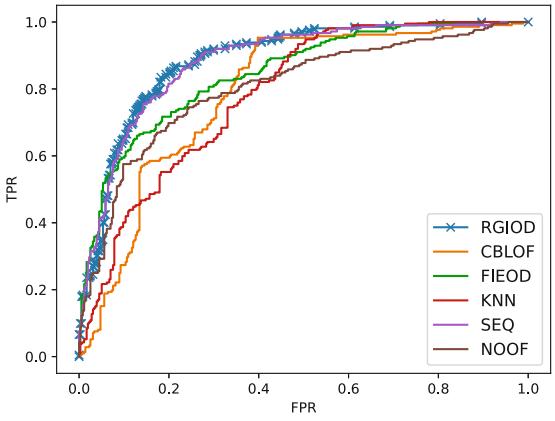

Figure 6: ROC for Breast Cancer Wisconsin (Diagnose). (The image was created by the authors.).

5.2 ROC and AUC

When the dataset has an imbalance between positive and negative instances, the ROC curve and AUC metric can serve as effective categorical evaluation indices. ROC is a comprehensive metric that captures the trade-off between sensitivity and specificity for continuous variables. It is a graph with the false positive rate (FPR) on the x-axis and the true positive rate (TPR) on the y-axis. The FPR and TPR are calculated as follows:

$$\text{TPR} = \text{TP}/(\text{TP} + \text{FN}),$$
$$\text{FPR} = \text{FP}/(\text{FP} + \text{TN}).$$

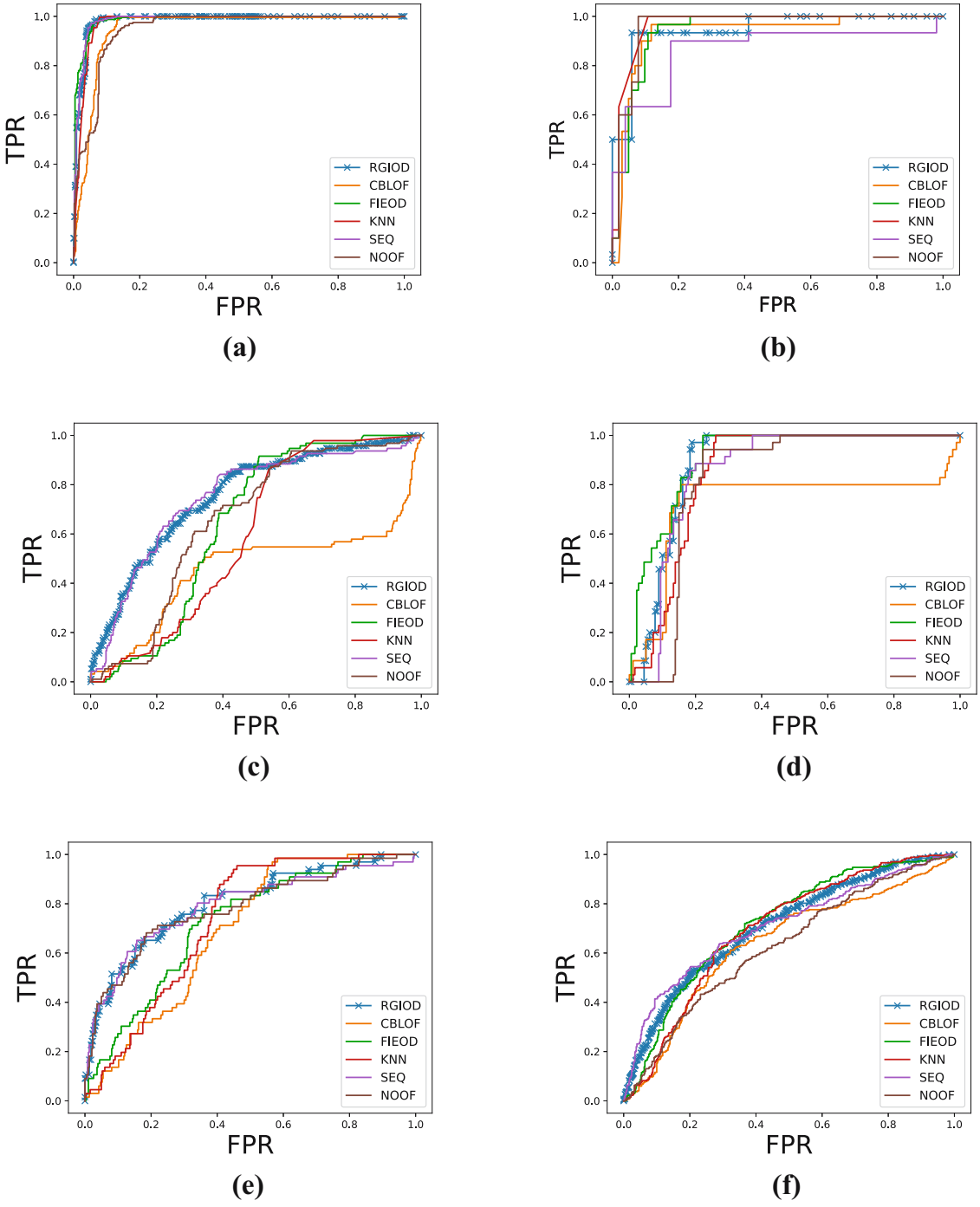

Figure 7: ROC results. (These image were created by the authors.) (a) Breast Cancer Wisconsin (Original), (b) Hayes, (c) Mammographic Masses, (d) Newthyroid, (e) Credit Approval, and (f) Pima.

Table 14: AUC results

Data sets	AUC value (rank)					
	ODIRG	**CBLOF**	**FIEOD**	**KNN**	**NOOF**	**SEQ**
Mammographic Masses	**0.7567 (1)**	0.4564 (6)	0.644 (4)	0.5977 (5)	0.7512 (2)	0.6575 (3)
Breast Cancer Wisconsin (Original)	0.9827 (3)	0.9477 (6)	**0.9856 (1)**	0.9755 (4)	0.9842 (2)	0.9483 (5)
Hayes	0.9471 (3)	0.9314 (5)	0.9425 (4)	**0.9668 (1)**	0.8634 (6)	0.9614 (2)
Breast Cancer Wisconsin (Diagnose)	**0.8875 (1)**	0.7914 (5)	0.8474 (3)	0.7848 (6)	0.8837 (2)	0.8045 (4)
Newthyriod	0.8849 (2)	0.7262 (6)	**0.9122 (1)**	0.8497 (4)	0.8554 (3)	0.823 (5)
Credit Approval	**0.8007 (1)**	0.6897 (6)	0.7213 (5)	0.7314 (4)	0.7904 (2)	0.7814 (3)
Pima	0.7086 (2)	0.6367 (5)	**0.7171 (1)**	0.6967 (4)	0.7085 (3)	0.6259 (6)
Average rank	**1.9**	5.6	2.7	4.0	2.9	4.0

The FPR represents the rate at which actual inliers are incorrectly identified as outliers and can be interpreted as a "false alarm rate." The TPR is numerically equivalent to the recall (R) and can be understood as the "detection rate." The goal of outlier detection is to maximize the TPR (or recall) while minimizing the FPR. Consequently, an algorithm whose ROC curve is closer to the upper left corner of the first quadrant exhibits better performance. AUC represents the ability of a model to correctly distinguish between positive and negative classes. If the AUC equals 1, it indicates a perfect classifier, meaning the model can make perfect predictions regardless of the decision threshold. In reality, a perfect classifier is rare. When the AUC is between 0.5 and 1, the classifier performs better than random guessing. An AUC of 0.5 corresponds to the performance of a random classifier. If the AUC is less than 0.5, the classifier performs worse than random guessing.

Take the Breast Cancer Wisconsin (Diagnose) dataset for example, it can be seen from Figure 6 that the curve of the ODIRG algorithm is closest to the upper left corner of the first quadrant. Meanwhile, the AUC is the largest compared with other algorithms, reaching 0.8875.

ROC results for the six datasets are indicated in Figure 7. ROC results show that the ODIRG algorithm implements well and operates smoothly on most datasets. Clear and robust conclusions can also be drawn from the AUC results. As shown in Table 14, numbers in bold are best. All in all, the ODIRG algorithm performs best.

6 Conclusions

Outlying observation identification finds wide-ranging deployments in expert and intelligent systems. Nevertheless, traditional detection methods often struggle to effectively handle an IRVIS. In this article, an outlier identification method built upon RST and GrC is raised. This article presents the ODIRG algorithm, which circumvents high computational costs by eliminating the need for data filling or deletion within an IRVIS. The proposed method outperforms five other detection methods across most datasets in the experiments conducted on seven datasets from UCI and KEEL. In addition, while the performance of the other five detection methods varies with different datasets, the proposed method exhibits a degree of robustness. Moving forward, our future endeavors will concentrate on minimizing the temporal complexity of this algorithm, aiming to enhance its compatibility with big data. In addition, we plan to explore the extension of this algorithm to address outlier detection in incomplete hybrid data.

Acknowledgements: The authors would like to thank the editors and the anonymous reviewers for their valuable comments and suggestions, which have helped immensely in improving the quality of the article.

Funding information: This work was supported by Doctoral Research of Guangdong University of Science and Technology (GKY-2024BSQDK-11), and Science Foundation in Guangdong University of Science and Technology.

Author contributions: Zhaowen Li designed the research and drafted the article, and Hongxuan He conducted the experiments, analyzed the data, and Wang Pei conducted the revisions and proofreading.

Conflict of interest: The authors state no conflict of interest.

Data availability statement: The datasets analyzed during the current study are available from the corresponding author upon reasonable request.

References

[1] Hawkins DM. Identification of outliers. London: Chapman and Hall. 1980. doi: 10.1007/978-94-015-3994-4.
[2] Wu R, Keogh EJ. Current time series anomaly detection benchmarks are flawed and are creating the illusion of progress. IEEE Trans Knowl Data Eng. 2023;35(3):2421–9. doi: 10.1109/TKDE.2021.3112126.
[3] Bolton RJ, Hand DJ, Provost F, Breiman L, Bolton RJ, Hand DJ. Statistical fraud detection: a review comment comment rejoinder. Stat Sci. 2002;17(3):235–55. http://www.jstor.org/stable/3182782.
[4] Paola AD, Gaglio S, Re GL, Milazzo F, Ortolani M. Adaptive distributed outlier detection for WSNs. IEEE Trans Cybernet. 2015;45(5):902–13. doi: 10.1109/TCYB.2014.2338611.
[5] Rasheed F, Alhajj R. A framework for periodic outlier pattern detection in time-series sequences. IEEE Trans Cybernet. 2014;44(5):569–82. doi: 10.1109/TSMCC.2013.2261984.
[6] Wang B, Mao Z. Outlier detection based on a dynamic ensemble model: Applied to process monitoring. Inform Fusion. 2019;51:244–58. doi: 10.1016/j.inffus.2019.02.006.
[7] Markou M, Singh S. Novelty detection: a review part 1, statistical approaches. Signal Process 2003;83(12):2481–97. doi: 10.1016/j.sigpro.2003.07.018.
[8] Piepel GF. Robust regression and outlier detection. Technometrics. 2005;31(2):260–1. doi: 10.1080/00401706.1989.10488524.
[9] He Z, Xu X, Deng S. Discovering cluster-based local outliers. Pattern Recognit Let. 2003;24(9–10):1641–50. doi: 10.1016/S0167-8655(03)00003-5.
[10] Jayakumar G, Thomas BJ. A new procedure of clustering based on multivariate outlier detection. J Data Sci. 2013;11(1):69–84. doi: 10.6339/JDS.2013.11(1).1091.
[11] Johnson T, Kwok I, Ng RT. Fast computation of 2-dimensional depth contours. in: Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining. 1998. p. 224–8. doi: 10.5555/3000292.3000332.
[12] Knorr EM, Ng RT, Tucakov V. Distance-based outliers: algorithms and applications. VLDB J. 2000;8(3):237–53. doi: 10.1007/s007780050006.
[13] Tao Y, Xiao X, Zhou S. Mining distance-based outliers from large databases in any metric space. In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '06). New York, NY, USA: Association for Computing Machinery; 2006. p. 394–403. doi: 10.1145/1150402.1150447.
[14] Breunig MM, Kriegel HP, Ng RT, Sander J. Lof: identifying density-based local outliers. in: Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data. 2000. p. 93–104. doi: 10.1145/342009.335388.
[15] Yao Y. Granular computing: Past, present, and future. 2008 IEEE International Conference on Granular Computing. 2008. p. 80–5. doi: 10.1109/GRC.2008.4664800.
[16] Yao JT, Vasilakos AV, Pedrycz W. Granular computing: Perspectives and challenges. IEEE Trans Cybernet. 2013;43(6):1977–89. doi: 10.1109/TSMCC.2012.2236648.
[17] Zadeh LA. Fuzzy sets and information granularity. Fuzzy sets, fuzzy logic, and fuzzy systems: selected papers by Lotfi A Zadeh. 1996. p. 433–48. doi: 10.1142/9789814261302_0022.
[18] Pawlak Z. Rough sets. Int J Comput Inform Sci. 1982;11:341–56. doi: 10.1007/BF01001956.
[19] Yao YY. Granular computing for data mining. In: Dasarathy BV, (ed). Proceedings of SPIE Conference on Data Mining, Intrusion Detection, Information Assurance, and Data Networks Security. 2006. p. 1–12. doi: 10.1117/12.669023.
[20] Miao DQ, Wang GY, Liu Q, Lin TY, Yao YY. Granular computing: past, present and future prospect. Beijing: Science Press; 2007. doi: 10.1109/GRC.2008.4664800.
[21] Jiang F, Zhao HB, Du JW, Xue Y, Peng YJ. Outlier detection based on approximation accuracy entropy. Int J Machine Learn Cybernet. 2018;10(9):2483–99. doi: 10.1007/s13042-018-0884-8.
[22] Chen YM, Miao DQ, Zhang HY. Neighborhood outlier detection. Expert Syst Appl. 2010;37(12):8745–49. doi: 10.1016/j.eswa.2010.06.040.
[23] Jiang F, Chen YM. Outlier detection based on granular computing and rough set theory. Appl Intel. 2015;42(2):303–22. doi: 10.1007/s10489-014-0591-4.

[24] Nguyen TT. Outlier detection: An approximate reasoning approach. Rough Sets and Intelligent Systems Paradigms: International Conference. 2007. p. 495–504. doi: 10.1007/978-3-540-73451-2_52.
[25] Dubois D, Prade H. Rough fuzzy sets and fuzzy rough sets. Int J General Syst. 1990;17:191–209. doi: 10.1080/03081079008935107.
[26] Yuan Z, Chen HM, Li TR, Liu J, Wang S. Fuzzy information entropy-based adaptive approach for hybrid feature outlier detection. Fuzzy Sets Syst. 2021;421:1–28. doi: 10.1016/j.fss.2020.10.017.
[27] Shaari F, Bakar AA, Hamdan AR. Outlier detection based on rough sets theory. Intel Data Anal. 2009;13(2):191–206. doi: 10.3233/IDA-2009-0363.
[28] Jiang F, Sui YF, Cao CG. Some issues about outlier detection in rough set theory. Expert Syst Appl. 2009;36:4680–7. doi: 10.1016/j.eswa.2008.06.019.
[29] Albanese A, Pal SK, Petrosino A. Rough sets, kernel set, and spatiotemporal outlier detection. IEEE Trans Knowl Data Eng. 2014;26(1):194–207. doi: 10.1109/TKDE.2012.234.
[30] Grzymala-Busse JW, Hu M. A comparison of several approaches to missing attribute values in data mining. In: International Conference on Rough Sets and Current Trends in Comput. Berlin, Heidelberg: Springer; 2001. p. 378–85. doi: 10.1007/3-540-45554-X_46.
[31] Yuan Z, Chen H, Li T. Fuzzy information entropy-based adaptive approach for hybrid feature outlier detection. Fuzzy Sets Syst. 2021;421:1–28. doi: 10.1016/j.fss.2020.10.017.
[32] Ramaswamy S, Rastogi R, Shim K. Efficient algorithms for mining outliers from large datasets. in: Proceedings of the 2000 ACM Sigmod International Conference on Management of Data. 2000. p. 427–38. doi: 10.1145/342009.335437.
[33] Chen YM, Miao DQ, Wang RZ. Outlier detection based on granular computing. Rough Sets and Current Trends in Computing: 6th International Conference. 2008. p. 283–92. doi: 10.1007/978-3-540-88425-5_29.
[34] Kryszkiewicz M. Rough set approach to incomplete information systems. Inform Sci. 1998;112:39–49. doi: 10.1016/S0020-0255(98)10019-1.
[35] Wang P, He JL, Li ZW. Attribute reduction for hybrid data based on fuzzy rough iterative computation model. Inform Sci. 2023;632:555–75. doi: 10.1016/j.ins.2023.03.027.
[36] Li ZW, Zhang QL, Wang P, Song Y, Wen CF. Uncertainty measurement for a gene space based on class-consistent technology: an application in gene selection. Appl Intel. 2023;53:5416–36. doi: 10.1016/j.asoc.2023.110645.
[37] Aggarwal CC, Yu PS. Outlier detection for high dimensional data. in: Proceedings of the 2001 ACM Sigmod international conference on Management of data. 2001. p. 37–46. doi: 10.1145/375663.375668.
[38] Campos GO, Zimek A, Sander J, Campello RJ, Micenkova B, Schubert E, et al. On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study. Data Mining Knowl Discovery. 2016;30(4):891–927. doi: 10.1007/s10618-015-0444-8.
[39] Hawkins S, He H, Williams GJ, Baxter RA. Outlier detection using replicator neural networks, CiteSeer. International Conference on Data Warehousing and Knowledge Discovery. 2002. p. 170–80. doi: 10.1007/3-540-46145-0_17.