Outlier Detection Using Three-Way Neighborhood Characteristic Regions and Corresponding Fusion Measurement

Xianyong Zhang, Zhong Yuan, and Duoqian Miao

Abstract—Outliers carry significant information to reflect an anomaly mechanism, so outlier detection facilitates relevant data mining. In terms of outlier detection, the classical approaches from distances apply to numerical data rather than nominal data, while the recent methods on basic rough sets deal with nominal data rather than numerical data. Aiming at wide outlier detection on numerical, nominal, and hybrid data, this paper investigates three-way neighborhood characteristic regions and corresponding fusion measurement to advance outlier detection. First, neighborhood rough sets are deepened via three-way decision, so they derive three-way neighborhood structures on model boundaries, inner regions, and characteristic regions. Second, the three-way neighborhood characteristic regions motivate the information fusion and weight measurement regarding all features, and thus, a multiple neighborhood outlier factor emerges to establish a new method of outlier detection; furthermore, a relevant outlier detection algorithm (called 3WNCROD) is designed to comprehensively process numerical, nominal, and mixed data. Finally, the 3WNCROD algorithm is experimentally validated, and it generally outperforms 13 contrast algorithms to perform better for outlier detection.

Index Terms—Outlier detection, neighborhood rough sets, three-way decision, uncertainty measurement, data mining.

1 Introduction

Data mining underlies knowledge discovery and emerging applications. In contrast to most data mining tasks, outlier detection finds rare data whose behavior is exceptional when compared with other mass data. As defined by Hawkins, an outlier is an observation that deviates so much from other observations to arouse suspicions that it is generated by a different mechanism [1]. Therefore, outliers usually adhere to a new perspective or a specific mechanism, and they can become more appealing than normal instances in data mining. Recently, outlier detection has been extensively studied [2]–[13]. Its applications include intrusion detection, image processing, medical treatment, and public security.

Regarding outlier detection, the traditional distance methods depend on object measures, and thus, they mainly apply to numerical data rather than categorical data. For this issue, some methods based on rough sets have been introduced to handle categorical data [14]–[18]. However, the classical rough set-based detection methods consider only the equivalent relation and classification, so they directly apply to categorical/nominal data (rather than numerical data). If these rough set methods are utilized for numerical data, then discretization is needed; however, the related processing usually causes a time increase and information loss. In the real world, numerical and categorical data are usually accompanied by their combination type, and the latter contains heterogeneous data. Hybrid data-driven research on outlier detection is required and challenging; however, its potential complexity causes only a small number of reports [19]–[21].

Neighborhood rough sets (NRSs) extend and improve classical rough sets, and they adopt the neighborhood and covering to effectively apply to multiple data types. They have already built a powerful platform for feature selection, pattern classification, and uncertainty reasoning [22]–[25]. NRSs and their robustness have also been introduced into outlier detection, especially for numerical and mixed data [19], [21], [26], [27]. Although NRS-based outlier detection methods have undergone some gradual development, they are worth deep exploration and efficient enhancement.

Outlier detection usually resorts to outlier factors; the corresponding detection measurement depends on the inherent characteristic structures of the related concepts. To promote outlier factors, we investigate structuring features via three-way decision (3WD). In particular, 3WD addresses structure cognition and partition processing, and it advocates three-way structuring and actions to effectively complete the trisecting-acting-outcome [28]. 3WD has become an important methodology for uncertainty measurement and data processing, thus encouraging popular studies [29]–[32]. Considering the underlying correlations of structural characterization and uncertainty measurement, 3WD technology for structurization and characterization is worth introducing into outlier detection to pursue development and achievement; however, there are few related reports.

Aiming at multiple data types (especially mixed data),

- *X. Y. Zhang is with School of Mathematical Sciences and Visual Computing and Virtual Reality Key Laboratory of Sichuan Province, Sichuan Normal University, Chengdu 610066, China (E-mail: xianyongzh@sina.com.cn). Z. Yuan is with College of Computer Science, Sichuan University, Chengdu 610065, China (E-mail: yuanzhong@scu.edu.cn). D. Q. Miao is with Department of Computer Science and Technology and Key Laboratory of Embedded System and Service Computing, Tongji University, Shanghai 201804, China (E-mail: dqmiao@tongji.edu.cn).*

Manuscript received X X, XX; revised X X, XX. (Corresponding author: Zhong Yuan)

this paper utilizes NRSs to establish a novel approach for outlier detection based on 3WD. The relevant processes and contents are presented as follows.

1) Three-way boundaries and three-way inner regions are proposed in NRSs.
2) Three-way neighborhood characteristic regions are constructed to motivate the multiple neighborhood outlier factor (MNOF). The latter generates a detection method, simply called 3WNCROD (three-way neighborhood characteristic region-based outlier detection). A relevant example and corresponding algorithm are provided.
3) The new detection approach is experimentally compared to 13 existing methods via 10 real outlier datasets (with 30 outlier data subsets). The superiority of 3WNCROD is eventually validated.

3WNCROD benefits from both the neighborhood extension and region structuring, so it applies to categorical, numerical, and hybrid data to acquire improvements. Categorical and numerical data have qualitative and quantitative natures, respectively, while hybrid data have a mixed feature; fortunately, all data associations of qualitative, quantitative, and heterogeneous mining can be uniformly realized via neighborhood construction and granulation. In 3WNCROD processing, a general measure is utilized to characterize the object distance regarding categorical, numerical, or mixed data, and robust neighborhood granulation further facilitates outlier detection. Overall, 3WNCROD depends on systematic three-way neighborhood characteristic regions; it adopts a regional integration measure: MNOF, which applies to diverse data. Therefore, 3WNCROD is effective for the three data detection cases, and its relevant complexity is related to distance measurement.

This study has the research novelty of neighborhood rough computation, three-way structuring measurement, and outlier detection construction; thus, it makes two contributions. In terms of theory, multiple three-way region structures of NRSs are deeply constructed to accelerate the feature extraction and outlier measurement. In terms of applications, a new outlier detection method is effectively proposed for categorical, numerical, and hybrid data detection to achieve better performances.

The remainder of this paper is organized as follows. Section 2 introduces the related work. Section 3 reviews NRSs. Section 4 studies three-way boundaries and three-way inner regions of NRSs. Section 5 establishes outlier detection based on three-way neighborhood characteristic regions and also provides the relevant example and algorithm. Section 6 conducts data experiments and comparison analyses. Finally, Section 7 presents the conclusion.

2 Related Work

Recently, outlier detection has attracted much attention from scholars. Over time, a large number of outlier detection algorithms have emerged in multiple research areas. In this section, only some outlier detection work relevant to this paper is reviewed. More details can be found in some good surveys, such as overviews [33]–[35] and their references.

Outlier detection first appeared in the statistics field and then entered the data mining field. It has four traditional methods: the statistical method [36], proximity approach [37]–[39], clustering method [40], and neural network-based method [41]–[43]. The statistical method assures that normal data objects are generated by a statistical model, and thus, abnormal points that never obey the model become outliers. This approach applies to data with known distributions and simple attributes. To improve the statistical method, the proximity approach adopts two basic strategies: distance-based and density-based detection. Moreover, the clustering method utilizes different clustering methods to exhibit distinctive effects, while the neural network-based method adheres to the advancement to generate satisfactory performances. Overall, most of these methods implement outlier detection via deterministic strategies, and corresponding treatments for uncertain information can be further developed.

Rough set theory is useful for data mining with imprecise, inconsistent, and incomplete information. This uncertainty methodology has been successfully utilized in outlier detection. For example, Jiang et al. [14] presented a detection method by adopting rough membership functions; Chen et al. [15] proposed a granular computing-based detection approach by relying on roughness granulation. Shaari et al. [16] studied a new detection method by proposing the nonreduction of rough sets. Jiang et al. [17] designed a detection algorithm by combining rough boundary-based and usual distance-based methods. Albanese et al. [44] used a new rough set approach to extend outlier detection to spatiotemporal data. Jiang et al. [18] implemented outlier detection based on rough approximation accuracy. By survey, these outlier detection studies all embrace the rough set theory, but they mainly resort to the equivalent relation and classified granulation in the classical case. Hence, these relevant detection treatments directly apply to only nominal data.

As extension models, NRSs have a robust description and broad applicability to support outlier detection, especially for numerical and mixed data mining. Chen et al. [19] proposed a neighborhood-based outlier detection algorithm for numerical data. Yuan et al. [21] investigated hybrid data-driven outlier detection based on neighborhood information entropy. Goh et al. [26] used NRSs to detect prototype outliers. Wang and Li [27] designed a detection method based on a weighted neighborhood information network for mixed-value datasets.

In this paper, 3WNCROD promotes NRS-based outlier detection and pursues good learning performances. This new method adheres to the uncertainty theory and 3WD structuring to differ from usual detection methods on deterministic methodologies and NRSs (or rough sets). Thus, 3WNCROD offers better semantic interpretations of uncertainty measurement and structuring decision, such as when comparing neural network-based methods. Regarding the technical contribution, 3WNCROD combines NRSs and 3WD to contain advanced structures and measures, so relevant mechanisms bring both the good theoretical interpretability of uncertainty information processing and the practical applicability of mixed data learning.

3 NEIGHBORHOOD ROUGH SETS

Rough set theory provides an effective method for data mining and knowledge discovery. Its NRS model is recalled by [45], [46]. As a preparation, the main notations of this paper are shown in Table 1.

TABLE 1: Main notations of this paper

Symbol	Meaning
NRSs	Neighborhood rough sets
3WD	Three-way decision
3WNCROD	The newly proposed method of outlier detection
B	Condition attribute subset
$HEOM$	Heterogeneous Euclidean-overlap metric
$n_B(x)$, nr_B	Neighborhood of x, neighborhood relation
X	Target concept
POS, NEG, BND	Positive and negative regions, boundary
NIB, NOB	Neighborhood inner and outer boundaries
NEB, NPB	Neighborhood exceptional, principal boundaries
NOM	Neighborhood overlap metric
NDF	Neighborhood deviation factor
$MNOF$	Multiple neighborhood outlier factor
$\omega^X_{nr_{\{c_j\}}}(x)$	Integration weight for outlier factor
ROC_AUC	Receiver operating characteristic_Area under curve
AP	Average precision
τ_F, τ_{χ^2}	Friedman's test item
CD_α	Nemenyi's test item

Data granulation offers basic knowledge blocks; thus, a concept can be represented by its double approximations or three-way classified regions. The granular membership description for a concept establishes a cognitive mechanism for outlier detection because outliers closely adhere to a sort of statistical belongingness for an observed concept. By virtue of rough set theory, there are some outlier detection discussions [14]–[21], [44].

Classical rough sets strictly adopt equivalence classes and their classification, while NRSs flexibly consider neighborhoods and their coverage. NRSs make distance measurement more applicable and are fully utilized for our detection construction.

An information system implies $IS = (U, A, V, f)$. $U = \{x_1, x_2, \cdots, x_n\}$ is the finite universe with object x, A is the finite attribute set with attribute a, $V = \bigcup_{a\in A} V_a$ is the union of domain V_a of attribute a, and $f : U \times A \to V$ denotes an information function with $f(x, a) \in V_a, \forall x \in U$, $\forall a \in A$. The information system can be specialized into a decision table via $A = C \cup D$ and $C \cap D = \emptyset$; here, $C = \{c_1, c_2, \cdots, c_m\} = \{c_j |\ j = 1, \cdots, m\}$ and D denote the conditional and decisional attribute sets, respectively.

There are multiple distance types, such as the Minkowski distance, heterogeneous Euclidean-overlap metric (HEOM), value difference metric (VDM), and heterogeneous value difference metric (HVDM) [47]. For a distance function $d : U \times U \to \mathbf{R}^+ \cup \{0\}$, its measurement is usually constructed by attribute subset $B = \{c_{j_1}, \cdots, c_{j_k}\} = \{c_{j_h} |\ h = 1, \cdots, k\}$ $(1 \le k \le m)$, and thus, it is represented by d_B. The distance and its threshold induce the neighborhood system, and the neighborhood radius $\varepsilon \ge 0$ is next utilized. The neighborhood of object x on subset B is $n_B(x) = \{y \in U |\ d_B(x, y) \le \varepsilon\}$, and there are two matching notions, i.e., neighborhood relation $nr_B = \{(x, y) \in U \times U |\ d_B(x, y) \le \varepsilon\}$ and neighborhood covering $\{n_B(x) |\ x \in U\}$. Let $NR_C = \{nr_B |\ B \subseteq C\}$ denote all neighborhood relations on U. Thus, $NIS = (U, NR_C, V, f)$ constitutes a neighborhood information system for applications, and it degenerates into the classical information system if $\varepsilon = 0$. Moreover, a target concept $\emptyset \ne X \subseteq U$, which can be a decision class for pattern recognition, is given to produce structural regions.

Definition 1 ([45], [46]). The lower and upper approximations of X on nr_B are defined by

$$\begin{cases} \underline{nr}_B(X) = \{x \in U |\ n_B(x) \subseteq X\}, \\ \overline{nr}_B(X) = \{x \in U |\ n_B(x) \cap X \ne \emptyset\}. \end{cases}$$

The positive region, negative region, and boundary of X on nr_B are defined by

$$\begin{cases} POS_{nr_B}(X) = \{x \in U |\ n_B(x) \subseteq X\}, \\ NEG_{nr_B}(X) = \{x \in U |\ n_B(x) \cap X = \emptyset\}, \\ BND_{nr_B}(X) = \{x \in U |\ n_B(x) \cap X, \neg X \ne \emptyset\}. \end{cases} \tag{1}$$

Dual approximations and three-way regions are basic notions with clear semantics that can be mutually derived to exhibit equivalency. The two types of model notions are illustrated in detail in Example 1. The former notions focus on the lower and upper bounds to bidirectionally approximate the central concept. The latter regions highlight the classified structure around the target pattern. Clearly, three-way regions are related to 3WD to promote the structuring and certainty/uncertainty. They constitute a universe partition, as shown in Fig. 1 with degenerate granules. They represent positive/negative certainty and uncertainty for concept cognition. In later studies, three-way regions and their further partition are utilized for outlier detection, and detection learning usually needs a rational division strategy.

4 THREE-WAY BOUNDARIES AND THREE-WAY INNER REGIONS OF NEIGHBORHOOD ROUGH SETS

The boundary $BND_{nr_B}(X)$ collects such objects that cannot be precisely classified into concept X (or complement $\neg X$) via granulation covering $\{n_B(x) |\ x \in U\}$. The boundary contains the margin objects and uncertainty information, so it is essentially linked to outlier detection. In fact, outlier detection aims to determine a small number of objects with unexpected behaviors or abnormal properties, so it needs in-depth descriptions, especially from the boundary. In this section, the boundary is divided into a two-partition first and a tripartition subsequently; thus, the constructional three-way boundaries further induce three-way inner regions, which finally underlie our latter outlier detection.

Definition 2. The neighborhood inner and outer boundaries of X on nr_B are defined by

$$\begin{cases} NIB_{nr_B}(X) = \{x \in X |\ n_B(x) \cap X, \neg X \ne \emptyset\}, \\ NOB_{nr_B}(X) = \{x \in \neg X |\ n_B(x) \cap X, \neg X \ne \emptyset\}. \end{cases} \tag{2}$$

Proposition 1. $NIB_{nr_B}(X), NOB_{nr_B}(X)$ have properties:

1) $NIB_{nr_B}(X) = BND_{nr_B}(X) \cap X \subseteq X$,
   $NOB_{nr_B}(X) = BND_{nr_B}(X) \cap \neg X \subseteq \neg X$;
2) $NIB_{nr_B}(X) = X - POS_{nr_B}(X)$,
   $NOB_{nr_B}(X) = \neg X - NEG_{nr_B}(X)$;

3) $NIB_{nr_B}(X) = \{x \in X |\, n_B(x) \nsubseteq X\}$,
$NOB_{nr_B}(X) = \{x \in \neg X |\, n_B(x) \nsubseteq \neg X\}$.

$BND_{nr_B}(X)$ includes all elements whose neighborhoods exhibit nonempty intersections with X and $\neg X$, and it induces a two-way partition around concept X. Specifically, $NIB_{nr_B}(X)$ and $NOB_{nr_B}(X)$ classify boundary $BND_{nr_B}(X)$; they are inside and outside X, respectively. The two-way boundaries have similar boundary connotations but different concept extensions, and they contain clear semantics and offer more exact uncertainty descriptions. They have the basic properties in Proposition 1, as shown in Fig. 1. Their relevant cases are described in Example 1. The two boundaries obey the symmetry; the inner $NIB_{nr_B}(X)$ adheres to outlier detection of target X and, thus, is the focus.

Proposition 2. $NIB_{nr_B}(X)$ has the infimum $NIB_{nr_C}(X)$, where $\forall B \subseteq C$. That is, there exists the monotonicity:

$$\forall B \subseteq C \Rightarrow NIB_{nr_B}(X) \supseteq NIB_{nr_C}(X).$$

Proposition 3. $NIB_{nr_B}(X)$ has a two-way classification on $NIB_{nr_C}(X)$, i.e.,

$$\begin{cases} NIB^{\circ}(X) = NIB_{nr_C}(X) \subseteq X, \\ NIB^{*}_{nr_B}(X) = NIB_{nr_B}(X) - NIB_{nr_C}(X) \subseteq X. \end{cases}$$

Regarding attribute subsets, constant $NIB_{nr_C}(X)$ is included in an arbitrary neighborhood inner boundary to support variational $NIB_{nr_B}(X)$. By virtue of $NIB_{nr_C}(X)$, the notion $NIB_{nr_B}(X)$ further produces a two-way partition. $NIB^{\circ}(X) = NIB_{nr_C}(X)$ corresponds to the infimum of inner boundaries to express a fixed feature of concept X, while $NIB^{*}_{nr_B}(X)$ collects the remaining part. Propositions 1–3 can be directly or further proven, and their correctness is shown in Fig. 1.

$NIB_{nr_C}(X)$ exists in m high-dimensional space, so a modified feature in low-dimensional space is worth mining to make a similar but efficient division, especially when facing complex data. Since the entire attribute set C carries m single attributes with constant properties, the inner boundaries of all single attributes and their integration can be considered. Single attributes induce a family of neighborhood inner boundaries: $NIB_{nr_{\{c_j\}}}(X)$ $(j = 1, \cdots, m)$. These inner boundaries act as constants and concern only one dimension, so their system information is worth extracting. Next, their intersection integration is adopted.

Definition 3. The neighborhood exceptional boundary of X is defined as $NEB(X) = \bigcap_{j=1}^{m} NIB_{nr_{\{c_j\}}}(X)$.

$NEB(X)$ utilizes the basic constant $NIB_{nr_{\{c_j\}}}(X)$ and logical intersection, implying an inclusion relationship: $NEB(X) \subseteq NIB_{nr_{\{c_j\}}}(X) \subseteq X$ $(\forall j = 1, \cdots, m)$. Therefore, $NEB(X)$ becomes an inherent feature in concept X. Specifically, the neighborhood exceptional boundary acts as the common region of all neighborhood inner boundaries on single attributes, so it represents the abnormality and becomes the core. This new notion is worth utilizing for outlier detection. Next, $NEB(X)$ first generates a kind of new boundary to divide the neighborhood inner boundary, mainly aiming at arbitrary subset B.

Definition 4. The neighborhood principal boundary of X on nr_B refers to $NPB_{nr_B}(X) = NIB_{nr_B}(X) - NEB(X)$.

The neighborhood principal boundary acts as the difference between neighborhood inner and exceptional boundaries, so it has the logical difference semantics of the two. $NIB_{nr_B}(X)$ and $NEB(X)$ do not necessarily have an inclusion relationship, and the noninclusion $NIB_{nr_B}(X) \not\supseteq NEB(X)$ can exist. Thus, the neighborhood exceptional and principal boundaries help the neighborhood inner boundary to form a two-way division, i.e., $NIB_{nr_B}(X)$ is classified into $NEB(X) \cap NIB_{nr_B}(X)$ and $NPB_{nr_B}(X)$. In particular, if $B = \{c_j\}$, then $NEB(X) \subseteq NIB_{nr_{\{c_j\}}}(X)$ denotes that $NEB(X)$ and $NPB_{nr_{\{c_j\}}}(X)$ completely classify $NPB_{nr_{\{c_j\}}}(X)$. The degeneration will come into play in later outlier factors and detection experiments, which concern only single attributes and relevant integrations.

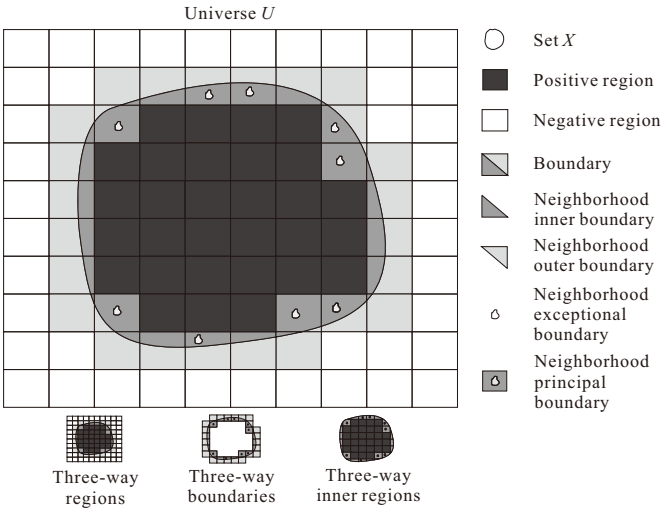


Fig. 1: Structural schematic diagram of three-way regions, three-way boundaries, three-way inner regions, as well as relevant parts

Thus far, there are three types of boundaries. The boundary $BND_{nr_B}(X)$ is first divided into the neighborhood inner and outer boundaries: $NIB_{nr_B}(X), NOB_{nr_B}(X)$; furthermore, the former $NIB_{nr_B}(X)$ is divided into the exceptional boundary $NEB(X)$(i.e., $NEB(X) \cap NIB_{nr_B}(X)$) and the principal boundary $NPB_{nr_B}(X)$. Accordingly, three-way boundaries $NEB(X)$, $NPB_{nr_B}(X)$, and $NOB_{nr_B}(X)$ emerge, and they imply uncertainty semantics, concept belongingness, and boundary division. Except for the boundaries, we now consider the inner regions inside the concept. Concept X is first divided into the positive region $POS_{nr_B}(X)$ and neighborhood inner boundary $NIB_{nr_B}(X)$, while the latter is further related to the neighborhood exceptional and principal boundaries: $NEB(X), NPB_{nr_B}(X)$. Accordingly, three-way inner regions $POS_{nr_B}(X)$, $NEB(X)$, $NPB_{nr_B}(X)$ emerge, and their semantics and structure can also be clarified by Fig. 1. These features in X become fundamental for our next construction of outlier factors and detection algorithms.

5 Outlier Detection Based on Three-Way Neighborhood Characteristic Regions

Three-way inner regions are formed above based on NRSs, and here, they are utilized to produce three-way neighbor-

hood characteristic regions. Furthermore, three-way neighborhood characteristic regions gain their fusion measurement to motivate outlier detection. Next, the relevant method, example, and algorithm are stated in three subsections.

5.1 Basic method of outlier detection

Herein, we propose a basic detection method based on three-way inner regions, where a restriction $x \in X$ is needed. First, we provide general discussions, including the three-way neighborhood characteristic regions, deviation/outlier factors, and detection discrimination. Then, we concretize the distance, weight, and parameter for further implementation.

Definition 5. Given distance measure $dist_C(x, y)$ and its thresholds d_1, d_2, d_3. Three-way neighborhood characteristic regions of x on X are defined by

$$\begin{cases} NEB^X(x) = \{y \in NEB(X) |\ dist_C(x, y) \leq d_1\}, \\ NPB^X_{nr_B}(x) = \{y \in NPB_{nr_B}(X) |\ dist_C(x, y) \geq d_2\}, \\ POS^X_{nr_B}(x) = \{y \in POS_{nr_B}(X) |\ dist_C(x, y) \geq d_3\}. \end{cases} \tag{3}$$

Eq. (3) concerns an interaction framework of distances and regions. $dist_C(x, y)$ represents the distance degree between objects x and y in concept X, and d_1, d_2, and d_3 limit the outlier features for $NEB(X)$, $NPB_{nr_B}(X)$, and $POS_{nr_B}(X)$, respectively. Thus, three-way neighborhood characteristic regions can be explained by three-way inner regions and underlying distance degrees. 1) Objects in $NEB(X)$ have the greatest possibility of being outliers, so $NEB(X)$ is chosen as a positive direction to estimate the outlier belongingness of object x. By measuring $dist_C$, the greater number of object y in $NEB(X)$, where y has a shorter distance from object x (i.e., $dist_C(x, y) \leq d_1$), will lead to a greater possibility of x for outliers. 2) Objects in $POS_{nr_B}(X)$ have the least possibility of being outliers, so $POS_{nr_B}(X)$ is chosen as a negative direction to value the outlier belongingness of object x. The greater number of object y in $POS_{nr_B}(X)$, where y has a greater distance from object x (i.e., $dist_C(x, y) \geq d_3$), will cause a greater outlier possibility for x. 3) Similarly, objects in $NPB_{nr_B}(X)$ have a moderate possibility of being outliers, and the negative direction can be chosen because of outlier sparseness. The greater number of object y in $NPB_{nr_B}(X)$, where y has a longer distance from object x (i.e., $dist_C(x, y) \geq d_2$), will induce a greater outlier possibility for x. In short, when determining outliers in X, three-way notions $NEB^X(x)$, $NPB^X_{nr_B}(x)$, $POS^X_{nr_B}(x)$ adopt both the distinctive qualitative attitudes for three-way inner regions and different parametric discriminations for measure distances, so they become three-way neighborhood characteristic regions for x's deviation detection around X. The definition construction follows the 3WD mechanism [28], and it underlies the next measurement development based on regional cardinalities. For convenience, $d_2 < d_3$ is set when considering that objects in $NPB_{nr_B}(X)$ have a greater outlier possibility than objects in $POS_{nr_B}(X)$; moreover, $d_1 < d_2$ is similarly required due to the greatest outlier possibility of objects in $NEB^X(x)$; hence, in practice, $d_1 < d_2 < d_3$ can be adopted for better operability.

Definition 6. The neighborhood deviation factor of x on X is defined as

$$\begin{aligned} &NDF^X_{nr_B}(x) \\ =&\frac{|NEB^X(x)| + |NPB^X_{nr_B}(x)| + |POS^X_{nr_B}(x)|}{|X|}. \end{aligned} \tag{4}$$

Deviation factors act as a tool for outlier detection, and they are usually constructed by traditional distances. In contrast, the neighborhood deviation factor comes from the cardinality fusion of three-way neighborhood characteristic regions. Three-way neighborhood characteristic regions describe the outlier features of x in terms of X's interior, and thus, $NDF^X_{nr_B}(x)$ adopts the ratio between the regional cardinality sum and concept cardinality. $NDF^X_{nr_B}(x)$ represents the likelihood that object x becomes an outlier for concept X, and its greater value corresponds to the more maximal deviation. Hence, it becomes a basic measure to characterize and underlie the next outlier detection.

For deviation factors, $NDF^X_{nr_B}(x)$ refers to subset B to become high-dimensional, so concrete attribute c_j in one-dimensional space is worth utilizing for simplicity. By setting $B = \{c_j\}$ ($j \in \{1, \cdots, m\}$), multiple and specific neighborhood deviation factors emerge, i.e., $NDF^X_{nr_{\{c_j\}}}(x)$ ($j \in \{1, \cdots, m\}$), and they can be systematically integrated into a discrimination measure.

Definition 7. The multiple neighborhood outlier factor (MNOF) of x on X is defined as

$$MNOF^X(x) = \sum_{j=1}^{m} NDF^X_{nr_{\{c_j\}}}(x) \times \omega^X_{nr_{\{c_j\}}}(x), \tag{5}$$

where $\omega^X_{nr_{\{c_j\}}}(x) \in [0, 1]$ ($j \in \{1, \cdots, m\}$) are weights.

$MNOF^X(x)$ adopts the weighted summation to integrate characteristic constants $NDF^X_{nr_{\{c_j\}}}(x)$ ($j \in \{1, \cdots, m\}$), and weight $\omega^X_{nr_{\{c_j\}}}(x)$ can be empirically determined to appropriately extract $NDF^X_{nr_{\{c_j\}}}(x)$. $NDF^X(x)$ exhibits systematicness and stability, and it becomes our eventual measure for outlier detection. Based on $NDF^X(x)$, we provide a formal outlier definition for illustration, but our later algorithm design and experiment evaluation consider only factor values and their sorting.

Definition 8. If $MNOF^X(x) > \mu$, then x is viewed as an outlier on concept X. The set of all outliers is $OS(X)$.

Thus far, we established a new method of outlier detection, i.e., 3WNCROD (three-way neighborhood characteristic region-based outlier detection). The related development comprises five parts: 1) three-way inner regions $NEB(X)$, $NPB_{nr_B}(X)$, $POS_{nr_B}(X)$, 2) three-way neighborhood characteristic regions $NEB^X(x)$, $NPB^X_{nr_B}(x)$, $POS^X_{nr_B}(x)$, 3) neighborhood deviation factor $NDF^X_{nr_B}(x)$, 4) multiple neighborhood outlier factor $MNOF^X(x)$ (based on the single-attribute integration, where $B = \{c_j\}$ ($j = 1, \cdots, m$)), and 5) outlier detection $MNOF^X(x) > \mu$. 3WNCROD has powerful generalization because its relevant distance, weight, and parameter are general. The notion concretization includes four parts: 1) neighborhood distance d_B and its radius ε, 2) distance measure $dist_B$ and its thresholds d_1, d_2, d_3, 3) weights $\omega^X_{\{c_j\}}(x)$ ($1 \leq j \leq m$),

and 4) detection threshold μ. Next, these notions d_B, ε, $dist_B$, d_1, d_2, d_3, $\omega^X_{\{c_j\}}(x)$, and μ are realized or analyzed, and concrete parameter values will be given in later examples and experiments.

First, data preprocessing on $IS = (U, A, V, f)$ uses min-max normalization:

$$f(x_i, c_j) \leftarrow \frac{f(x_i, c_j) - \min_{\{c_j\}}}{\max_{\{c_j\}} - \min_{\{c_j\}}} \in [0, 1] \quad (1 \le i \le n, 1 \le j \le m). \tag{6}$$

Other normalization approaches can be consulted [48], [49].

The neighborhood distance resorts to the heterogeneous Euclidean-overlap metric (HEOM) [47]:

$$\forall x, y \in X, HEOM_B(x, y) = \sqrt{\sum_{h=1}^{k} d_{c_{j_h}}(x, y)^2}, \tag{7}$$

where $d_{c_{j_h}}(x, y) =$

$$\begin{cases} 1, & \text{if attribute values of } x \text{ and } y \text{ are unknown on attribute } c_{j_h}; \\ 0, & \text{if } c_{j_h} \text{ is a categorical attribute and } f(x, c_{j_h}) = f(y, c_{j_h}); \\ 1, & \text{if } c_{j_h} \text{ is a categorical attribute and } f(x, c_{j_h}) \neq f(y, c_{j_h}); \\ |f(x, c_{j_h}) - f(y, c_{j_h})|, & \text{if } c_{j_h} \text{ is a numerical attribute.} \end{cases}$$

The single case $B = \{c_j\}$ offers $HEOM_{\{c_j\}}(x, y) = d_{\{c_j\}}(x, y)$. HEOM can deal with not only numerical data but also hybrid or complex data, and it also considers unknown attribute values; hence, HEOM is effective for extensive outlier detection. Moreover, HEOM is a piecewise function that contains a core numerical part on the Euclidean distance, so other distances can be introduced to produce more heterogeneous metrics.

Neighborhood radii can be given by expert experience [19], but this method easily causes more subjectivity and sensitivity. By the statistical strategy [50], we determine an adaptive neighborhood threshold on attribute c_j:

$$\varepsilon_{c_j} = \begin{cases} 0, & \text{if } c_j \text{ is a categorical attribute,} \\ \frac{std(c_j)}{\lambda}, & \text{if } c_j \text{ is a numerical atttibute.} \end{cases} \tag{8}$$

In Eq. (8), $std(c_j)$ is the standard deviation of attribute values on numerical c_j. λ is a given parameter for radius adjustments. 1) The standard deviation represents the degree of data dispersion, and a smaller value of $std(c_j)$ implies that the data are closer to the average, so the statistics endow the neighborhood radii with objectivity and reasonability. 2) λ adjusts neighborhood sizes in terms of data granulation accuracy, and neighborhood radii with $\lambda < 1$, $\lambda = 1$, and $\lambda > 1$ are more than, equal to, or less than the standard deviation of attribute values, respectively.

Next, measure $dist_C$ (Definition 5) is considered. The traditional distances cannot process categorical data, while the classical rough set-based distances easily lose efficiency for numerical or mixed data. To solve this issue, we define the neighborhood overlap metric (NOM), which can generate an effective heterogeneous measure in NRSs.

Definition 9. The neighborhood overlap metric (NOM) of x and y in concept X is defined as

$$NOM(x, y) = |\{c \in C | (x, y) \notin nr_{\{c\}}\}|. \tag{9}$$

$NOM = (NOM(x, y))_{|X| \times |X|}$ $(\forall x, y \in X)$ is the related and symmetrical NOM matrix, and its upper-triangular form (with zero principal diagonal) is directly used.

Function NOM introduces and improves the basic measure OM (i.e., the overlap metric) into NRSs to deal with broader data. The NOM matrix saves all NOM information to underlie the distance measurement of $dist_C$.

Furthermore, the weight coefficient regarding attribute c_j (Definition 7) is realized. The weight determines outlier factors and detection results, but it tends to be an empirical function. Through theoretical analyses and practical experiments, the specific weight is constructed by

$$\omega^X_{nr_{\{c_j\}}}(x) = \frac{1}{|C|}\left(1 - \sqrt{\frac{|n_{\{c_j\}}(x) \cap X|}{|X|}}\right). \tag{10}$$

This weight setting abides by the basic idea that outlier detection always considers the minority group in datasets. Objects in the minority group are more likely to become outliers, so they should have higher weights. In Eq. (10), if the objects in both x's neighborhood and concept X are few, then x has a small percentage in X to correspond to a minority group and a high weight.

Finally, the detection parameter μ (Definition 8) and its case are clarified. In terms of outlier factors of 3WNCROD, μ denotes the cut threshold, and thus, it can be considered by expert experience or the actual situation. Fortunately, parameter μ is not needed in later experiments, and the ordering of outlier factors is sufficient for detection estimation.

In summary, 3WNCROD benefits from three-way region structuring and single-attribute information integration, while its universality and effectiveness are also related to the above concretization. Furthermore, there are only two groups of required parameters, i.e., λ and d_1, d_2, d_3. In contrast, d_1, d_2, d_3 can be considered for settings, and their stationarity implies complexity reduction to support main effect analyses of fundamental neighborhood parameter λ. By Definition 5 and its explanation, $d_1 < d_2 < d_3$ is rational; meanwhile, d_1, d_2, d_3 adhere to the distance $dist_C(x, y)$, so their values can be connected with attribute number $|C|$. Based on procedural simulations and experimental observations, we adopt

$$d_1 = |C|/3 < d_2 = |C|/2 < d_3 = 0.9|C|. \tag{11}$$

This result is related to the relevant approach and empirical assignment in [51]. As shown by experiments, d_1, d_2, d_3 may impact the outlier factor and detection result to different degrees, but their settings in Eq. (11) can induce satisfying performances for 3WNCROD improvements.

5.2 Illustrative example of outlier detection

Example 1. 3WNCROD and its basic notions are illustrated by an example. An information system $IS = (U, A, V, f)$ with hybrid data is provided on the left of Table 2.

In Table 2, the 5th column concerns categorical data; the 6th and 7th columns embody numerical data from

TABLE 4: Multiple regions on single attributes

c_j	$POS_{nr_{\{c_j\}}}(X)$	$NEG_{nr_{\{c_j\}}}(X)$	$BND_{nr_{\{c_j\}}}(X)$	$NOB_{nr_{\{c_j\}}}(X)$	$NIB_{nr_{\{c_j\}}}(X)$	$NEB(X)$	$NPB_{nr_{\{c_j\}}}(X)$
c_1	$\{x_6\}$	$\emptyset$	$\{x_1, \cdots, x_5\}$	$\{x_3, x_4\}$	$\{x_1, x_2, x_5\}$	$\{x_1\}$	$\{x_2, x_5\}$
c_2	$\{x_2, x_5, x_6\}$	$\emptyset$	$\{x_1, x_3, x_4\}$	$\{x_3, x_4\}$	$\{x_1\}$	$\{x_1\}$	$\emptyset$
c_3	$\{x_6\}$	$\emptyset$	$\{x_1, \cdots, x_5\}$	$\{x_3, x_4\}$	$\{x_1, x_2, x_5\}$	$\{x_1\}$	$\{x_2, x_5\}$

TABLE 2: Initial and standard information systems

U	c_1	c_2	c_3	c_1	c_2	c_3
x_1	D	4	0.7	D	1/3	4/5
x_2	B	7	0.4	B	2/3	1/5
x_3	D	1	0.6	D	0	3/5
x_4	B	2	0.3	B	1/9	0
x_5	B	8	0.5	B	7/9	2/5
x_6	C	10	0.8	C	1	1

TABLE 3: Neighborhoods on all single attributes

U	c_1	c_2	c_3
x_1	$\{x_1, x_3\}$	$\{x_1, x_2, x_3, x_4\}$	$\{x_1, x_3, x_6\}$
x_2	$\{x_2, x_4, x_5\}$	$\{x_1, x_2, x_5, x_6\}$	$\{x_2, x_4, x_5\}$
x_3	$\{x_1, x_3\}$	$\{x_1, x_3, x_4\}$	$\{x_1, x_3, x_5\}$
x_4	$\{x_2, x_4, x_5\}$	$\{x_1, x_3, x_4\}$	$\{x_2, x_4\}$
x_5	$\{x_2, x_4, x_5\}$	$\{x_2, x_5, x_6\}$	$\{x_2, x_3, x_5\}$
x_6	$\{x_6\}$	$\{x_2, x_5, x_6\}$	$\{x_1, x_6\}$

standardized Eq. (6), and they produce standard deviations $std(c_2) \approx 0.3610, std(c_3) \approx 0.3416$. Consider Eq. (8) and let $\lambda = 1$, and then we obtain neighborhood radii $\varepsilon_{c_1} = 0, \varepsilon_{c_2} \approx 0.3610, \varepsilon_{c_3} \approx 0.3416$. By HEOM (Eq. (7)), the neighborhood granulation is established, and neighborhoods of single attributes c_1, c_2, c_3 are given in Table 3.

Next, concept $X = \{x_1, x_2, x_5, x_6\}$ is given to illustrate three-way regions and outlier detection. Multiple regions of single attributes are given in Table 4, and they offer the following partitions. 1) Three-way regions $POS_{nr_{\{c_j\}}}(X)$, $NEG_{nr_{\{c_j\}}}(X)$, $BND_{nr_{\{c_j\}}}(X)$ divide universe U. 2) For boundaries, internal $NIB_{nr_{\{c_j\}}}(X)$ and external $NOB_{nr_{\{c_j\}}}(X)$ divide the entire $BND_{nr_{\{c_j\}}}(X)$; furthermore, exceptional $NEB(X)$ and principal $NPB_{nr_{\{c_j\}}}(X)$ divide the inner $NIB_{nr_{\{c_j\}}}(X)$. Hence, $NEB(X)$, $NPB_{nr_{\{c_j\}}}(X)$, and $NOB_{nr_{\{c_j\}}}(X)$ constitute three-way classified boundaries. 3) $POS_{nr_{\{c_j\}}}(X)$, $NEB(X)$, $NPB_{nr_{\{c_j\}}}(X)$ constitute three-way inner regions to divide concept X.

The NOM matrix is computed to yield

$$\begin{pmatrix} 0 & 2 & 3 & 2 \\ & 0 & 0 & 2 \\ & & 0 & 2 \\ & & & 0 \end{pmatrix}.$$

For Eq. (11), we have $d_1 = 1$, $d_2 = 1.5$, and $d_3 = 2.7$, so three-way neighborhood characteristic regions (Definition 5) are obtained. The measure MNOF (Definition 7) with weights (Eq. (10)) can be calculated. Let MNOF threshold $\mu = 0.13$, so final outliers (Definition 8) are gained. The processing of object x_1 is offered as a case. 1) For $\{c_1\}$, three-way neighborhood characteristic regions exhibit $(NEB^X(x_1), NPB^X_{nr_{\{c_1\}}}(x_1), POS^X_{nr_{\{c_1\}}}(x_1)) = (\{x_1\}, \{x_2, x_5\}, \emptyset)$, and their cardinalities induce $NDF^X_{nr_{\{c_1\}}}(x_1) = \frac{|\{x_1\}|+|\{x_2,x_5\}|+|\emptyset|}{|X|} = \frac{3}{4}$. Furthermore, we gain $NDF^X_{nr_{\{c_2\}}}(x_1) = \frac{2}{4}$ and $NDF^X_{nr_{\{c_3\}}}(x_1) = \frac{3}{4}$, so the factor vector is $(\frac{3}{4}, \frac{2}{4}, \frac{3}{4})$. 2) The weight vector is

$$(\omega^X_{nr_{\{c_1\}}}(x_1), \omega^X_{nr_{\{c_2\}}}(x_1), \omega^X_{nr_{\{c_3\}}}(x_1)) = \frac{1}{3}\left(1 - \sqrt{\tfrac{1}{4}}, 1 - \sqrt{\tfrac{2}{4}}, 1 - \sqrt{\tfrac{2}{4}}\right) = (\tfrac{1}{6}, \tfrac{\sqrt{2}-1}{3\sqrt{2}}, \tfrac{\sqrt{2}-1}{3\sqrt{2}}).$$

The detection factor concerns $MNOF^X(x_1) = (\frac{3}{4}, \frac{2}{4}, \frac{3}{4}) \cdot (\frac{1}{6}, \frac{\sqrt{2}-1}{3\sqrt{2}}, \frac{\sqrt{2}-1}{3\sqrt{2}}) \approx 0.2470 > 0.13 = \mu$, so x_1 is an outlier of X. Similarly, the remaining objects $x_2, x_5, x_6 \in X$ are accompanied by $MNOF^X(x_2) \approx 0 \approx MNOF^X(x_5)$, $MNOF^X(x_6) \approx 0.1321$, so we obtain $x_6 \in OS(X) = \{x_1, x_6\}$. Outlier factors offer an X-object order: $x_1 \succeq x_6 \succeq x_2 \succeq x_5$.

Algorithm 1: One-attribute-based three-way inner regions calculation (1A3WIRC)

Input: Information system $IS = (U, C, V, f)$ (with $|U| = n$ and $|C| = m$), concept X, and threshold λ.
Output: Three-way inner regions $POS_{nr_{\{c_j\}}}(X)$, $NEB(X)$, $NPB_{nr_{\{c_j\}}}(X)$ $(j = 1, \cdots, m)$.

1 $NEB(X) \leftarrow X$;
2 **for** $j \leftarrow 1$ *to* m **do**
3 Determine the covering $\{n_{\{c_j\}}(x) | \ x \in U\}$;
4 $POS_{nr_{\{c_j\}}}(X) \leftarrow \emptyset$;
5 **for** $i \leftarrow 1$ *to* n **do**
6 **if** $n^{\varepsilon}_{\{c_j\}}(x_i) \subseteq X$ **then**
7 $POS_{nr_{\{c_j\}}}(X) \leftarrow POS_{nr_{\{c_j\}}}(X) \cup \{x_i\}$;
8 **end**
9 continue;
10 **end**
11 $NIB_{nr_{\{c_j\}}}(X) \leftarrow X - POS_{nr_{\{c_j\}}}(X)$;
12 $NEB(X) \leftarrow NEB(X) \cap NIB_{nr_{\{c_j\}}}(X)$;
13 **end**
14 **for** $j \leftarrow 1$ *to* m **do**
15 $NPB_{nr_{\{c_j\}}}(X) \leftarrow NIB_{nr_{\{c_j\}}}(X) - NEB(X)$;
16 **end**
17 Return $POS_{nr_{\{c_j\}}}(X)$, $NEB(X)$, $NPB_{nr_{\{c_j\}}}(X)$ $(j = 1, \cdots, m)$.

5.3 Corresponding algorithm of outlier detection

Algorithm 1 calculates the family of three-way inner regions based on single attributes, and three "for" loops are considered. In the outside loop, Step 3 provides the initial covering. Steps 5-10 use an inside loop to offer the positive region, and Step 7 implements the cycle collection. Step 11 uses the basic formula to yield the inner boundary. Step 12 makes the cycle intersection to present the exceptional boundary. After the two loops, Steps 1–13 obtain the positive regions $POS_{nr_{\{c_j\}}}(X)$ $(j = 1, \cdots, m)$ and integrated

Algorithm 2: Three-way neighborhood characteristic regions-based outlier detection (3WNCROD)

Input: Information system $IS = (U, C, V, f)$ (with $|U| = n$ and $|C| = m$), concept X, parameter λ.
Output: Multiple neighborhood outlier factors $MNOF^X(x)$ $(x \in X)$.

1 Obtain three-way inner regions $POS_{nr_{\{c_j\}}}(X)$, $NEB(X)$, $NPB_{nr_{\{c_j\}}}(X)$ $(j = 1, \cdots, m)$ by Algorithm 1;
2 **for** $x \in X$ **do**
3 $MNOF^X(x) \leftarrow 0$;
4 **for** $j \leftarrow 1$ *to* m **do**
5 $|NEB^X(x)| \leftarrow 0$, $|NPB^X_{nr_{\{c_j\}}}(x)| \leftarrow 0$, $|POS^X_{nr_{\{c_j\}}}(x)| \leftarrow 0$;
6 **for** $y \in X$ **do**
7 Calculate $NOM(x, y)$;
8 **if** $NOM(x, y) \leq d_1$ *and* $y \in NEB(X)$ **then**
9 $|NEB^X(x)| \leftarrow |NEB^X(x)| + 1$;
10 **end**
11 **if** $NOM(x, y) \geq d_2$ *and* $y \in NPB_{nr_{\{c_j\}}}(X)$ **then**
12 $|NPB^X_{nr_{\{c_j\}}}(x)| \leftarrow |NPB^X_{nr_{\{c_j\}}}(x)| + 1$;
13 **end**
14 **if** $NOM(x, y) \geq d_3$ *and* $y \in POS_{nr_{\{c_j\}}}(X)$ **then**
15 $|POS^X_{nr_{\{c_j\}}}(x)| \leftarrow |POS^X_{nr_{\{c_j\}}}(x)| + 1$;
16 **end**
17 **end**
18 By referring to Eq. (4), $NDF^X_{nr_{\{c_j\}}}(x) = \frac{|NEB^X(x)| + |NPB^X_{nr_{\{c_j\}}}(x)| + |POS^X_{nr_{\{c_j\}}}(x)|}{|X|}$;
19 Assign weight $\omega^X_{nr_{\{c_j\}}}(x) \leftarrow 1 - \sqrt{\frac{|n_{\{c_j\}}(x) \cap X|}{|X|}}$;
20 According to Eq. (5), $MNOF^X(x) \leftarrow MNOF^X(x) + \omega^X_{nr_{\{c_j\}}}(x) \times NDF^X_{nr_{\{c_j\}}}(x)$;
21 **end**
22 **end**
23 **Return** $MNOF^X(x)$ $(x \in X)$.

feature $NEB(X)$. The last loop calculates all neighborhood principle boundaries $NPB_{nr_{\{c_j\}}}(X)$ $(j = 1, \cdots, m)$, and they are finally returned in Step 17.

Algorithm 2 extracts the outlier factor MONF by three "for" loops. 1) Step 1 invokes Algorithm 1 and offers three-inner regions of all single attributes. 2) Steps 6–17 concern the core loop, and they fix x and c_j to count characteristic regional cardinalities on $y \in X$. Specifically, Step 7 calculates distance $NOM(x, y)$, and Steps 8-16 use three conditional judgments to extract cardinalities of three-way neighborhood characteristic regions. 3) The middle loop continuously uses Steps 18 and 19 to calculate factor $NDF^X_{nr_{\{c_j\}}}(x)$ and weight $\omega^X_{nr_{\{c_j\}}}(x)$. Thus, Step 20 circularly computes detection factor $MNOF^X(x)$ (Eq. (5)). 4) The outside loop (embracing Steps 2 and 22) mainly moves object x in X. 5) Finally, Step 23 returns outlier factors of all objects in X.

Now focus on the algorithmic complexity analysis. For Algorithm 1, Step 3 uses the single-attribute neighborhood covering (SANC) algorithm proposed by [21], so it has the time complexity $O(n \log n)$. In addition, the frequency of Steps 2-13 is m, the frequency of Steps 5-10 is n, and the frequency of Steps 14-16 is m, so the total frequency becomes $m \times (n \times \log n + n) + m$. Therefore, the time complexity of Algorithm 1 is $O(mn \log n)$. For Algorithm 2, Step 1 invokes Algorithm 1 to concern the time complexity $O(mn \log n)$, and its total frequency is $m \times n \times \log n + m \times |X|^2$. Accordingly, the time complexity of Algorithm 2 eventually yields $O(mn \log n + m|X|^2)$.

By single attributes, Algorithms 1 and 2 focus on the attribute family and integration, respectively. Their symbols of POS, NEB, NPB correspond to three-way inner regions and three-way characteristic regions, respectively, and the latter regions related to Eq. (3) become pivotal for detection construction. Algorithm 2 invokes Algorithm 1 for three-way integration, and it realizes the 3WNCROD strategy, thus becoming effective and feasible. A diagram of the algorithmic framework is depicted in Fig. 2 to helpfully capture the relevant idea and flow.

6 Outlier Detection Data Experiments

6.1 Experimental settings

The effectiveness, superiority, and adaptability of 3WNCROD are next verified by data experiments and comparison analyses. For this purpose, 3WNCROD is compared with 13 existing methods of outlier detection, and we utilize 13 algorithmic abbreviations: DIS, RMF, GrC, ITB, BD, ODGrCR, VOS, POD, WNINOD, MIX, WFRDA, AE, VAE, which respectively imply the distance (DIS) algorithm [52], rough membership function (RMF)-based algorithm [14], granular computing (GrC)-based algorithm [15], information theory-based (ITB) algorithm [5], boundary and distance (BD)-based algorithm [51], outlier detection based on granular computing and rough set (ODGrCR) [18], virtual outlier score (VOS) [53], practical outlier detection (POD) [54], weighted neighborhood information network-based outlier detection (WNINOD) [27], joint learning framework for outlier detection in MIXed-type (MIX) data [41], weighted fuzzy-rough density-based anomaly (WFRDA) [55], autoencoder (AE) [42], and variational autoencoder (VAE) [43]. Among the 14 algorithms, RMF, GrC, ITB, BD, and ODGrCR are suitable for nominal data; DIS, VOS, AE, and VAE are suitable for numerical data; and POD, WNINOD, MIX, WFRDA, and 3WNCROD are suitable for mixed data.

Experimental datasets are downloaded from relevant websites of outlier detection [1] [2]. Specifically, there are 10 public datasets recorded in Table 5, and they have been extensively utilized for outlier detection [53], [55]–[57], where the outlier determination follows the random downsampling method [58]. Among them, five datasets are numerical, two datasets are nominal, and the remaining three are mixed. Furthermore, they are imported into information systems $IS_A, IS_C, IS_G, IS_H, IS_L, IS_M, IS_{Ma}, IS_{Mu}, IS_T$, and IS_W. Some data subsets can be selected as detection units from each outlier dataset, and this strategy facilitates

1. http://odds.cs.stonybrook.edu
2. https://github.com/BElloney/Outlier-detection

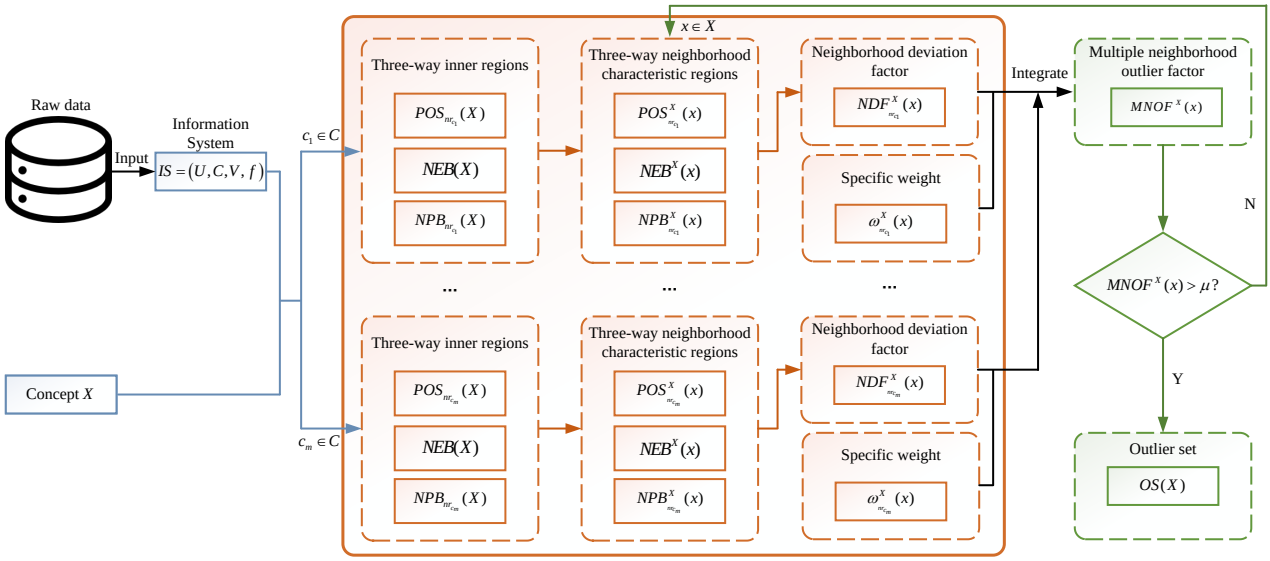


Fig. 2: Framework diagram of proposed algorithms

TABLE 5: Basic information of experimental data subsets

No.	Original dataset	Data subset	Selection criteria	Number of condition attributes		Number of objects	Number of outliers
				Numeric	Nominal		
1	Annthyroid	A1	$A1=\{x\in U_A \mid f_A(x,c_3)\in[0,0.025]\}$	6	0	6355	513
2		A2	$A2=\{x\in U_A \mid f_A(x,c_4)\in[0,0.12]\}$			5094	485
3		A3	$A3=\{x\in U_A \mid f_A(x,c_5)\in[0,0.1]\}$			4643	303
4	CreditA_plus_42_variant1	C1	$C1=\{x\in U_C \mid f_C(x,c_{10})=' f'\}$	6	9	307	10
5		C2	$C2=\{x\in U_C \mid f_C(x,c_{12})=' f'\}$			229	16
6		C3	$C3=\{x\in U_C \mid f_C(x,c_{13})=' g'\}$			375	37
7	German_1_14_variant1	G1	$G_1=\{x\in U_G \mid f_G(x,c_{18})=1\}$	7	13	600	9
8		G2	$G_2=\{x\in U_G \mid f_G(x,c_{20})=' A201'\}$			680	13
9		G3	$G_3=\{x\in U_G \mid f_G(x,c_{14})=' A143'\}$			594	4
10	Heart_2_16_variant1	H1	$H_1=\{x\in U_H \mid f_H(x,c_2)=2\}$	6	7	91	8
11		H2	$H_2=\{x\in U_H \mid f_H(x,c_6)=1\}$			137	10
12		H3	$H_3=\{x\in U_H \mid f_H(x,c_9)=1\}$			131	4
13	Lymphography	L1	$L_1=\{x\in U_L \mid f_L(x,c_3)=' no'\}$	0	18	122	4
14		L2	$L_2=\{x\in U_L \mid f_L(x,c_2)=' no'\}$			66	4
15		L3	$L_3=\{x\in U_L \mid f_L(x,c_{13})=' yes'\vee f_L(x,c_{18})=' no'\}$			85	4
16	Mammography	MA1	$MA1=\{x\in U_{MA} \mid f_{MA}(x,c_4)\in[-0.9,2]\}$	6	0	10951	204
17		MA2	$MA2=\{x\in U_{MA} \mid f_{MA}(x,c_5)\in[-0.4,1]\}$			10098	76
18		MA3	$MA3=\{x\in U_{MA} \mid f_{MA}(x,c_6)\in[-1,1]\}$			8263	54
19	Mushroom_p_221_variant1	M1	$M1=\{x\in U_M \mid f_M(x,c_4)=2\}$	0	22	1649	193
20		M2	$M2=\{x\in U_M \mid f_M(x,c_7)=1\}$			3218	210
21		M3	$M3=\{x\in U_M \mid f_M(x,c_7)=2\}$			1211	11
22	Musk	MU1	$MU1=\{x\in U_{MU} \mid f_{MU}(x,c_1)\in[28,60]\}$	166	0	2963	78
23		MU2	$MU2=\{x\in U_{MU} \mid f_{MU}(x,c_4)\in[-200,0]\}$			2348	85
24		MU3	$MU3=\{x\in U_{MU} \mid f_{MU}(x,c_9)\in[-165,20]\}$			2260	97
25	Thyroid	T1	$T1=\{x\in U_T \mid f_T(x,c_2)\in[0,0.5]\}$	6	0	3766	87
26		T2	$T2=\{x\in U_T \mid f_T(x,c_6)\in[0,0.3]\}$			3676	93
27		T3	$T3=\{x\in U_T \mid f_T(x,c_5)\in[0,0.5]\}$			3530	83
28	Wdbc_M_39_variant1	W1	$W1=\{x\in U_W \mid f_W(x,c_4)=2\}$	9	0	42	5
29		W2	$W2=\{x\in U_W \mid f_W(x,c_3)=6\vee f_W(x,c_5)=2\}$			363	7
30		W3	$W3=\{x\in U_W \mid f_W(x,c_1)=5\vee f_W(x,c_7)=3\}$			194	12

experiments and comparisons [18]. Although outlier determination may be a practical problem, our data subsets consider the outlier settings and use cases in the detection field [53], [55]–[58], and their relevant information is shown in Table 5.

The algorithm settings and dataset treatments refer to corresponding references. For example, the parameter values of GrC and BD come from the initial settings in [15] and [51], respectively, and the min-max normalization in Eq. (6) is uniformly used. Here, some cases are generally explained. The 3WNCROD parameter λ changes in [0.1, 2] with step length 0.1. The WFRDA [55] parameter similarly changes in [0.1, 2] with length 0.1. The WNINOD [27] parameter varies in [1, 10] with length 1. For DIS [52], the Euclidean distance is used, and all different nominal attribute values are replaced with different integer values. Rough set-based methods RMF, GrC, BD, ODGrCR and information-theoretical ITB require data discretization, and

TABLE 6: ROC_AUC values of comparison experiments

Data subset	DIS	RMF	GrC	ITB	BD	ODGrCR	VOS	POD	WNINOD	MIX	WFRDA	AE	VAE	3WNCROD
A1	0.600	0.471	0.661	0.659	0.315	0.658	0.729	0.672	0.682	0.646	0.674	0.680	0.679	**0.732**
A2	0.602	0.358	0.655	0.669	0.297	0.660	**0.747**	0.711	0.696	0.699	0.704	0.682	0.680	0.692
A3	0.545	0.480	0.636	0.619	0.423	0.618	0.744	0.625	0.633	0.612	0.618	0.716	0.693	**0.775**
C1	0.901	0.982	0.864	0.980	0.975	0.987	0.897	0.591	0.970	0.932	0.978	0.973	0.975	**0.990**
C2	0.986	0.999	0.995	0.999	0.999	0.999	0.987	0.910	0.988	0.824	0.995	0.979	0.981	**0.999**
C3	0.977	0.996	0.990	0.995	0.992	0.996	0.978	0.771	0.971	0.934	0.990	0.983	0.983	**0.997**
G1	0.960	0.980	0.980	0.968	0.031	0.980	0.970	0.537	0.982	0.899	0.984	0.957	0.963	**0.987**
G2	0.950	0.981	0.975	0.969	0.979	0.979	0.964	0.500	0.981	0.885	0.983	0.950	0.954	**0.986**
G3	0.947	0.992	0.986	0.978	0.994	0.992	0.981	0.475	0.992	0.977	0.993	0.947	0.953	**0.996**
H1	0.956	0.949	0.953	0.967	0.955	0.971	0.953	0.811	0.976	0.938	0.985	0.973	**0.991**	0.989
H2	0.979	0.992	0.982	0.991	0.989	0.991	0.968	0.847	0.996	0.972	0.997	0.985	0.988	**0.998**
H3	0.967	0.972	0.953	0.963	0.970	0.970	**0.990**	0.766	0.982	0.951	0.986	0.984	**0.990**	0.988
L1	**1.000**	0.998	0.998	0.994	0.998	**1.000**	0.985	0.636	0.998	0.996	**1.000**	0.994	0.987	**1.000**
L2	**1.000**	0.992	0.992	0.988	**1.000**	**1.000**	0.984	0.619	**1.000**	0.996	**1.000**	0.996	0.992	**1.000**
L3	0.997	0.994	**1.000**	**1.000**	**1.000**	**1.000**	0.985	0.745	0.997	**1.000**	**1.000**	**1.000**	0.997	**1.000**
MA1	0.847	0.833	0.817	0.816	0.845	0.813	0.803	0.769	0.853	0.812	0.836	0.875	0.874	**0.893**
MA2	0.745	0.681	0.667	0.668	0.651	0.673	0.672	0.664	0.742	0.720	0.738	0.748	0.752	**0.775**
MA3	0.744	0.735	0.663	0.752	0.694	0.749	**0.840**	0.645	0.770	0.729	0.763	0.790	0.783	0.788
M1	0.822	0.937	0.987	0.967	**0.999**	0.985	0.403	0.798	0.976	0.983	0.991	0.863	0.864	0.985
M2	0.918	0.976	0.962	0.977	0.919	**0.983**	0.421	0.672	0.969	0.980	0.973	0.916	0.918	0.939
M3	**1.000**	0.993	0.992	**1.000**	0.975	1.000	**1.000**	0.980	0.996	**1.000**	0.999	0.992	0.993	0.978
MU1	0.883	0.318	0.771	0.897	0.999	0.774	0.697	0.897	0.986	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
MU2	0.856	0.343	0.649	0.764	0.984	0.656	0.674	0.918	0.918	0.785	0.999	0.999	0.999	**1.000**
MU3	0.791	0.332	0.683	0.714	0.886	0.666	0.650	0.917	0.854	0.753	0.996	0.998	0.998	**1.000**
TH1	0.853	0.838	0.854	0.824	0.697	0.819	0.975	0.976	0.939	0.934	0.954	0.963	0.963	**0.985**
TH2	0.869	0.830	0.852	0.838	0.836	0.836	0.953	**0.993**	0.952	0.950	0.963	0.969	0.968	0.985
TH3	0.878	0.847	0.852	0.836	0.827	0.839	0.961	0.977	0.958	0.961	0.970	0.968	0.967	**0.992**
W1	**1.000**	0.984	0.984	0.995	0.995	0.995	0.914	0.749	**1.000**	**1.000**	**1.000**	0.995	0.995	**1.000**
W2	**1.000**	0.999	0.999	**1.000**	0.998	1.000	**1.000**	0.560	**1.000**	**1.000**	**1.000**	0.998	0.998	**1.000**
W3	0.989	0.993	0.993	0.995	0.993	0.992	0.987	0.658	0.995	0.992	**0.997**	0.991	0.991	0.995
Average	0.885	0.826	0.878	0.893	0.841	0.886	0.860	0.746	0.925	0.895	0.936	0.929	0.929	**0.948**

TABLE 7: AP values of comparison experiments

Data subset	DIS	RMF	GrC	ITB	BD	ODGrCR	VOS	POD	WNINOD	MIX	WFRDA	AE	VAE	3WNCROD
A1	0.150	0.076	0.170	0.248	0.054	0.230	0.284	0.201	0.252	0.193	0.272	0.275	0.274	**0.437**
A2	0.190	0.068	0.192	0.278	0.062	0.243	0.313	0.274	0.330	0.309	0.334	0.343	0.341	**0.427**
A3	0.075	0.058	0.106	0.130	0.052	0.095	0.209	0.122	0.121	0.096	0.117	0.193	0.167	**0.278**
C1	0.608	0.825	0.597	0.790	0.705	**0.834**	0.329	0.424	0.654	0.668	0.689	0.597	0.610	0.800
C2	0.837	0.988	0.914	0.992	0.990	0.984	0.886	0.714	0.879	0.212	0.942	0.721	0.742	**0.993**
C3	0.826	0.973	0.889	0.971	0.958	0.969	0.867	0.549	0.809	0.453	0.928	0.804	0.808	**0.978**
G1	0.398	0.374	0.388	0.309	0.009	0.372	0.459	0.031	0.485	0.315	0.402	0.335	0.389	**0.561**
G2	0.279	0.396	0.359	0.337	0.397	0.371	0.362	0.028	0.479	0.372	0.452	0.278	0.288	**0.543**
G3	0.148	0.399	0.222	0.243	0.408	0.346	0.352	0.011	0.417	0.360	0.406	0.148	0.152	**0.501**
H1	0.811	0.774	0.692	0.854	0.762	0.846	0.780	0.238	0.903	0.667	0.900	0.777	0.896	**0.931**
H2	0.837	0.925	0.887	0.932	0.902	0.919	0.819	0.266	0.967	0.843	0.962	0.763	0.791	**0.973**
H3	0.681	0.679	0.358	0.541	0.592	0.540	**0.861**	0.105	0.771	0.641	0.788	0.678	0.793	0.817
L1	**1.000**	0.950	0.950	0.893	0.950	**1.000**	0.787	0.067	0.950	0.888	**1.000**	0.893	0.799	**1.000**
L2	**1.000**	0.917	0.917	0.893	**1.000**	**1.000**	0.817	0.137	**1.000**	0.950	**1.000**	0.950	0.888	**1.000**
L3	0.950	0.917	**1.000**	**1.000**	**1.000**	**1.000**	0.759	0.238	0.950	**1.000**	**1.000**	**1.000**	0.950	**1.000**
MA1	0.073	0.098	0.065	0.094	0.087	0.067	0.165	0.058	0.077	0.052	0.075	0.141	0.141	**0.355**
MA2	0.027	0.042	0.012	0.028	0.015	0.030	**0.076**	0.021	0.027	0.015	0.065	0.025	0.025	0.073
MA3	0.024	0.022	0.011	0.149	0.048	0.144	0.181	0.012	0.065	0.024	**0.213**	0.145	0.149	0.187
M1	0.420	0.684	0.918	0.819	**0.988**	0.905	0.156	0.278	0.875	0.853	0.933	0.464	0.465	0.920
M2	0.429	0.915	0.893	0.844	0.874	**0.920**	0.228	0.086	0.907	0.871	0.898	0.416	0.417	0.898
M3	**1.000**	0.796	0.848	**1.000**	0.750	0.960	**1.000**	0.129	0.827	**1.000**	0.896	0.791	0.797	0.753
MU1	0.130	0.018	0.074	0.113	0.976	0.052	**1.000**	0.667	0.653	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
MU2	0.117	0.025	0.049	0.067	0.923	0.047	**1.000**	0.668	0.227	0.073	0.977	0.982	0.982	**1.000**
MU3	0.088	0.030	0.065	0.067	0.236	0.059	**1.000**	0.708	0.131	0.075	0.943	0.951	0.948	**1.000**
TH1	0.091	0.174	0.159	0.363	0.320	0.136	0.600	0.341	0.241	0.161	0.296	0.402	0.397	**0.716**
TH2	0.196	0.092	0.115	0.426	0.168	0.158	0.328	**0.807**	0.375	0.334	0.424	0.506	0.499	0.761
TH3	0.168	0.119	0.108	0.425	0.147	0.134	0.339	0.361	0.371	0.331	0.412	0.413	0.409	**0.852**
W1	**1.000**	0.925	0.925	0.967	0.967	0.967	0.563	0.527	**1.000**	**1.000**	**1.000**	0.967	0.967	**1.000**
W2	**1.000**	0.957	0.938	**1.000**	0.924	0.982	**1.000**	0.081	**1.000**	**1.000**	**1.000**	0.874	0.874	**1.000**
W3	0.864	0.884	0.889	0.941	0.884	0.877	0.832	0.369	0.933	0.910	**0.961**	0.881	0.881	0.921
Average	0.481	0.503	0.490	0.557	0.572	0.540	0.578	0.284	0.589	0.522	0.676	0.590	0.595	**0.756**

the discretization intervals concern number 3. The Fuzzy C-Means (FCM) discretization method [56] is used for datasets containing numeric attributes. Note that MIX, AE, and VAE imply neural network-based methods.

These detection algorithms finally output outlier measures, such as outlier factors. A greater measurement value implies a higher outlier possibility, and thus, the value ordering of data samples is usually utilized. Based on algorithmic results, we need scientific indices for performance estimation and algorithm comparison. Since ROC_AUC has

TABLE 8: Running times of comparison experiments (unit: seconds)

Data subset	DIS	RMF	GrC	ITB	BD	ODGrCR	VOS	POD	WNINOD	MIX	WFRDA	AE	VAE	3WNCROD
A1	0.104	2988.470	48.847	1.039	12.845	27.792	267.953	0.956	1409.919	0.646	77.312	19.283	42.090	15.748
A2	0.075	2398.664	25.973	0.704	14.333	17.231	1047.051	0.814	720.456	0.699	37.998	38.429	35.648	19.689
A3	0.049	2186.457	20.660	0.580	12.938	13.722	873.156	0.622	547.306	0.612	28.184	15.041	32.573	18.799
C1	0.001	2.076	0.064	0.008	0.094	0.366	2.970	0.015	0.050	0.932	0.020	3.727	5.035	0.219
C2	0.001	1.539	0.042	0.008	0.086	0.267	1.641	0.010	0.020	0.824	0.010	3.522	4.483	0.084
C3	0.001	2.520	0.102	0.012	0.112	0.937	5.987	0.017	0.079	0.934	0.032	3.979	4.911	0.467
G1	0.001	12.201	0.342	0.034	0.325	3.066	7.416	0.001	0.453	0.899	0.120	4.644	6.140	0.831
G2	0.002	14.004	0.435	0.040	0.355	3.481	23.579	0.029	0.651	0.885	0.162	7.540	6.186	0.398
G3	0.002	12.124	0.339	0.037	0.311	2.913	13.842	0.001	0.457	0.977	0.118	4.909	6.018	0.393
H1	0.001	0.074	0.010	0.002	0.012	0.096	0.067	0.001	0.031	0.938	0.002	3.343	3.998	0.009
H2	0.000	0.109	0.014	0.002	0.015	0.124	0.185	0.001	0.537	0.972	0.004	4.474	4.061	0.012
H3	0.000	0.119	0.014	0.002	0.015	0.121	0.196	0.001	0.018	0.951	0.003	4.857	4.428	0.011
L1	0.000	0.135	0.022	0.003	0.018	0.235	0.168	0.001	0.366	0.996	0.003	3.392	4.125	0.040
L2	0.000	0.066	0.015	0.001	0.012	0.180	0.050	0.001	0.003	0.996	0.001	5.428	3.841	0.011
L3	0.001	0.089	0.017	0.002	0.020	0.204	0.382	0.001	0.004	1.000	0.002	5.801	3.929	0.013
MA1	0.292	28525.475	219.932	3.265	78.064	111.549	4763.525	1.891	7356.592	0.812	873.601	36.074	39.793	45.042
MA2	0.215	30285.588	179.618	2.668	78.844	90.787	4224.372	1.689	5147.417	0.720	765.326	28.005	31.748	61.293
MA3	0.136	26352.883	121.187	1.724	42.370	55.840	886.872	1.008	2890.884	0.729	354.858	26.219	31.378	53.314
M1	0.006	1716.268	4.000	0.172	20.982	94.616	35.113	0.003	13.762	0.983	2.416	7.966	9.868	18.533
M2	0.036	3421.398	25.108	0.937	37.800	335.407	127.146	0.014	333.273	0.980	29.838	12.877	15.721	26.294
M3	0.005	1268.725	1.761	0.089	14.901	34.214	13.557	0.002	5.774	1.000	0.855	6.753	8.640	17.812
MU1	0.059	3152.342	165.241	16.646	169.528	28772.884	138.263	0.329	2015.849	1.000	59.462	18.350	21.225	110.730
MU2	0.042	2457.401	82.942	10.097	119.508	27402.444	82.141	0.138	912.274	0.785	32.760	15.413	17.340	93.271
MU3	0.037	2411.021	75.341	9.660	115.532	24754.673	94.666	0.142	800.720	0.753	31.636	14.717	17.668	101.388
TH1	0.030	1252.493	13.672	1.184	6.367	31.484	152.053	0.354	153.390	0.934	27.168	13.223	15.865	3.940
TH2	0.033	1117.679	12.146	0.396	3.723	27.587	152.539	0.353	140.150	0.950	24.787	13.175	15.992	6.714
TH3	0.041	719.030	11.093	0.383	7.743	23.409	128.135	0.332	125.913	0.961	16.954	13.077	15.182	6.941
W1	0.000	0.268	0.036	0.001	0.079	0.131	0.012	0.003	0.001	1.000	0.001	3.108	3.682	0.064
W2	0.001	2.219	0.131	0.004	0.121	0.774	2.827	0.024	0.437	1.000	0.026	4.334	5.088	0.131
W3	0.001	1.175	0.054	0.002	0.077	0.264	0.558	0.009	0.009	0.992	0.005	3.522	4.206	0.061
Average	0.039	3676.754	33.639	1.657	24.571	2726.893	434.881	0.292	752.560	0.895	78.789	11.506	14.029	20.075

already become a very effective and popular evaluation indicator for outlier detection [52], [55], [57]–[59], it is used to evaluate algorithm performances here; moreover, we add average precision (AP) [58] as an auxiliary index to more fully complete the algorithmic comparison and assessment. Greater values of ROC_AUC and AP imply better detection performances for outlier algorithms, and the main case of ROC_AUC will be emphatically analyzed.

6.2 Experimental results

In terms of ROC_AUC, the experimental results of 14 algorithms on 10 datasets (with 30 outlier data subsets) are shown in Table 6, where the bold labels optimal values on data subsets. Table 6 fully reflects the detection performance and comparison superiority of 3WNCROD. Clearly, 3WNCROD achieves better ROC_AUC values in most cases. For example, on data subset C1, 3WNCROD reaches the optimal value of 0.990, while the other 13 algorithms exhibit lower values of 0.901, 0.982, 0.864, 0.980, 0.975, 0.987, 0.897, 0.591, 0.970, 0.932, 0.978, 0.973, and 0.975. From the frequency statistics, 3WNCROD achieves the best detection result on 22 data subsets, whose proportion is $22/30 \approx 73.33\%$; in contrast, the other 13 algorithms achieve subset numbers of maximum effects: 5, 0, 1, 3, 3, 5, 3, 1, 3, 5, 7, 2, and 3, corresponding to lower percentages of 16.67%, 0.00%, 3.33%, 10.00%, 10.00%, 16.67%, 10.00%, 3.33%, 10.00%, 16.67%, 23.33%, 6.67%, and 10.00%, respectively. Final arithmetic averages of ROC_AUC can better reveal and validate contrast performances. The 14 algorithms correspond to 0.885, 0.826, 0.878, 0.893, 0.841, 0.886, 0.860, 0.756, 0.925, 0.895, 0.936, 0.929, 0.929, and 0.948. Hence, 3WNCROD achieves the optimal value of 0.948, and this value is greater than the suboptimal 0.936 reached by the recent WFRDA [55]. If concerning only mixed data subsets (C1–C3, G1–G3, H1–H3), actual algorithms for mixed features may not function well even in contrast to some other algorithms, but 3WNCROD never follows this case. Moreover, for AP, the algorithmic results are recorded in Table 7, and 3WNCROD still obtains the optimal performance based on similar analyses. For example, 3WNCROD obtains the optimal average 0.756 to surpass the suboptimal mean of 0.676 reached by WFRDA.

The running times of all comparison experiments are reported in Table 8. By Table 8, 3WNCROD has a moderate running time, and its temporal cost is also less than that of WFRDA. Thus, it is feasible and effective in practical applications. For 3WNCROD, its experimental time accords with and validates its theoretical complexity, i.e., $O(mn \log n + m|X|^2)$ of Algorithm 2.

Since these experiments concern three data types, 3WNCROD can effectively process categorical, numerical, and mixed attribute data. It acquires comparatively optimal performances for outlier detection. Note that the effectiveness and superiority of 3WNCROD are derived by concrete observations and general analyses from Tables 6, 7, 8. In the later induction of statistical analysis on ROC_AUC, 3 approaches, 3WNCROD, WFRDA, and WNINOD, are located in the first echelon, while 3WNCROD significantly outperforms the other 11 contrast algorithms. Therefore, 3WNCROD truly improves multiple existing algorithms, and its advancement benefits from the neighborhood rough computation, three-way structuring measurement, and robust hybrid processing.

6.3 Statistical test analysis

Friedman's test [60] and Nemenyi's post hoc test [61] are adopted here to evaluate the statistical significance, mainly in terms of the index ROC_AUC.

First, the ROC_AUC values of each algorithm on all datasets are sorted from low to high, and the sequence

numbers are $(1, 2, \cdots)$. If the ROC_AUC values of the two algorithms are the same, the ordinal values are equally divided. Then, Friedman's test is used to determine whether these algorithms have the same performance. Assume we compare M algorithms on N datasets, and Friedman's test is calculated by τ_F, τ_{χ^2} [60], where τ_F obeys the F distribution with $(M-1)$ and $(M-1)(N-1)$ degrees of freedom. If the null hypothesis "all algorithms have the same performance" is rejected, then the algorithmic performances are significantly different. At this time, Nemenyi's post hoc test is needed to further distinguish these algorithms. Thus, the critical difference (CD) of the average ordinal value is calculated by $CD_\alpha = q_\alpha\sqrt{\frac{M(M+1)}{6N}}$ [61], where q_α is the critical value of Tukey's distribution. Note that Nemenyi's test figure intuitively represents the significant differences between two algorithms or among multiple algorithms.

Our experimental cases involve $M = 14, N = 30$, and τ_F follows 13 and 377 degrees of freedom. According to Friedman's test, when $\alpha = 0.05$, the value 10.0433 of τ_F is greater than the critical value 1.7462. Therefore, the null hypothesis "all algorithms have the same performance" is rejected, and the detection performances of all outlier algorithms are significantly different.

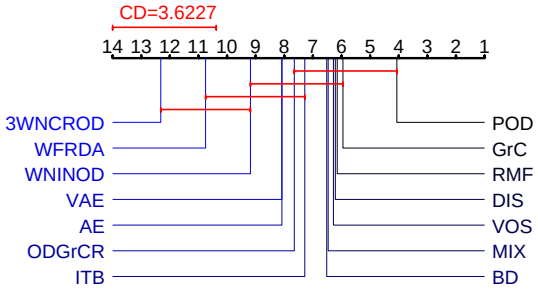


Fig. 3: Nemenyi's test figure on ROC_AUC

For significance level $\alpha = 0.05$, the corresponding critical distance $CD_{0.05} = 3.6227$ is obtained, and Nemenyi's test figure on ROC_AUC is shown in Fig. 3. Thus, 3WNCROD is significantly different from 11 algorithms: DIS, RMF, GrC, ITB, BD, ODGrCR, VOS, POD, MIX, AE, and VAE; meanwhile, there is no consistent evidences to indicate 3WNCROD's significant difference from WFRDA and WNINOD, but the three algorithms actually exhibit an optimal ordering: 3WNCROD$\succeq$WFRDA$\succeq$WNINOD. From the statistical perspective, 3WNCROD has the main development and specific advantage for outlier detection.

6.4 Parametric sensitivity analysis

Finally, 3WNCROD is the focus of the parameter analyses. In our experiments, only λ is variable, and it determines neighborhood radii to impact detection effects. λ-based change curves of ROC_AUC are depicted in Fig. 4, and its eight subfigures correspond to eight datasets (each one contains three distinctive data subsets). The remaining datasets Lymphography and Mushroom (with L1–L3 and M1–M3, respectively) are categorical to never consider λ. Fig. 4 can uncover the change association between parameter λ and ROC_AUC result, and the corresponding sensitivity.

By Fig. 4, three subset-based ROC_AUC lines of each dataset/subfigure exhibit generally similar trends when λ increases, so they have roughly coincident data distributions. For the dataset system, eight subfigures can be classified into three categories. 1) In subfigures (b) (c) (d) (g), ROC_RUC first rapidly increases and then tends to a stable maximum. This monotonic phenomenon facilitates the optimization selection of λ because only the segment point is basically needed. 2) In subfigures (a) (f) (h), ROC_AUC first rapidly increases, then suddenly decreases, and finally increases (may reach stable maximums). Although complex fluctuations emerge, the result optimization tends to the large parametric value (such as $\lambda = 2$). 3) In subfigure (e), ROC_AUC lines become convex, and the vertex reaches the optimal. In summary, the optimal ROC_RUC may be reached by great λ values, while the latter are multiple in most datasets; moreover, the optimal may be acquired by the peak value, such as in Mammography's MA1-MA3 (subfigure (e)). By the above analyses, most data subsets are generally sensitive to the parameter change in λ, and thus, λ is effective for 3WNCROD. λ's optimization requires in-depth discussions to further promote 3WNCROD.

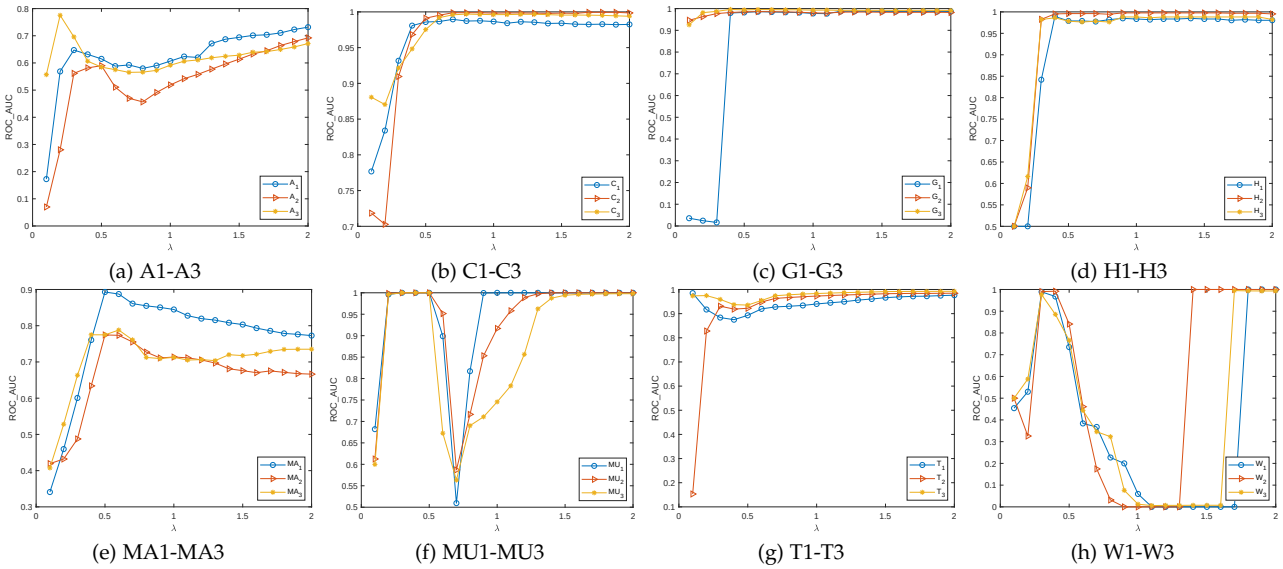
7 Conclusion

In this paper, an outlier detection strategy – 3WNCROD – is established via several three-way structures and measures, mainly based on NRSs and 3WD. First, three-way inner regions have specific semantics to divide detection set X, and they resort to distance measure $dist_C(x, y)$ to induce three-way neighborhood characteristic regions. Then, characteristic cardinalities of single attributes are utilized, neighborhood deviation factors $NDF^X_{nr_{\{c_j\}}}(x)$ $(j = 1, \cdots, m)$ are calculated, and they are integrated into the outlier degree $MNOF^X(x)$ via adjustable weight coefficients. Furthermore, MNOF is sorted in decreasing order, and instances with greater values tend to be outliers. Finally, 3WNCROD is realized by Algorithm 2 and is validated by examples and experiments. As a result, 3WNCROD effectively applies to outlier detection of categorical, numerical, and hybrid data. Its improvement leads to better performances. 3WNCROD also has good generalization on measures and parameters. However, 3WNCROD only considers the integration of neighborhood deviation factors of all single attributes, and this treatment may cause some measurement limitations. The 3WNCROD processing time is feasible, but its relevant optimization is also a new issue.

In the future, 3WNCROD needs in-depth comparisons with other outlier detection methods, and its scalability is worth further verification by using larger datasets. Regarding outlier detection, multiple cases in existing studies, such as unsupervised learning in [2], imperfect data labels in [3], data streams in [4], large-scale categorical data in [5], and temporal data in [6], are worth extensively considering to further develop 3WNCROD. Moreover, 3WNCROD and its underlying neighborhood structure and measure can be combined with other uncertainty methodologies (such as fuzzy sets and soft sets) so that relevant studies come into play in data mining and knowledge discovery.

Acknowledgments

The authors thank the reviewers and editors for their valuable suggestions, which substantially improve this paper.



(a) A1-A3 (b) C1-C3 (c) G1-G3 (d) H1-H3

(e) MA1-MA3 (f) MU1-MU3 (g) T1-T3 (h) W1-W3

Fig. 4: Variation curve of ROC_AUC with parameter λ regarding algorithm 3WNCROD

The work was supported by National Key Research and Development Program of China (2017YFC0821300), National Natural Science Foundation of China (61976158, 61673285), Natural Science Foundation of Sichuan Province of China (2022NSFSC0929), Sichuan Science and Technology Program of China (2023YFQ0020, 2022ZYD0001).

REFERENCES

[1] D. M. Hawkins, *Identification of outliers*. Springer, 1980.

[2] Y. Z. Liu, Z. Li, C. Zhou, Y. C. Jiang, J. S. Sun, M. Wang, and X. N. He, "Generative adversarial active learning for unsupervised outlier detection," *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, no. 8, pp. 1517–1528, 2019.

[3] B. Liu, Y. S. Xiao, S. Y. Philip, Z. F. Hao, and L. B. Cao, "An efficient approach for outlier detection with imperfect data labels," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 7, pp. 1602–1616, 2013.

[4] M. Salehi, C. Leckie, J. C. Bezdek, T. Vaithianathan, and X. Y. Zhang, "Fast memory efficient local outlier detection in data streams," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 12, pp. 3246–3260, 2016.

[5] S. Wu and S. R. Wang, "Information-theoretic outlier detection for large-scale categorical data," *IEEE Transactions on Knowledge and data Engineering*, vol. 25, no. 3, pp. 589–602, 2011.

[6] M. Gupta, J. Gao, C. C. Aggarwal, and J. W. Han, "Outlier detection for temporal data: A survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 9, pp. 2250–2267, 2013.

[7] C. Yu, Q. G. Wang, D. Zhang, L. Wang, and J. S. Huang, "System identification in presence of outliers," *IEEE Transactions on Cybernetics*, vol. 46, no. 5, pp. 1202–1216, 2015.

[8] X. J. Li, J. C. Lv, and Z. Yi, "Outlier detection using structural scores in a high-dimensional space," *IEEE Transactions on Cybernetics*, vol. 50, no. 5, pp. 2302–2310, 2018.

[9] H. D. Zhao, H. F. Liu, Z. M. Ding, and Y. Fu, "Consensus regularized multi-view outlier detection," *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 236–248, 2017.

[10] H. W. Wang, H. B. Li, J. Fang, and H. P. Wang, "Robust gaussian kalman filter with outlier detection," *IEEE Signal Processing Letters*, vol. 25, no. 8, pp. 1236–1240, 2018.

[11] J. Mao, T. Wang, C. Jin, and A. Zhou, "Feature grouping-based outlier detection upon streaming trajectories," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 12, pp. 2696–2709, 2017.

[12] A. De Paola, S. Gaglio, G. L. Re, F. Milazzo, and M. Ortolani, "Adaptive distributed outlier detection for wsns," *IEEE Transactions on Cybernetics*, vol. 45, no. 5, pp. 902–913, 2014.

[13] R. Domingues, M. Filippone, P. Michiardi, and J. Zouaoui, "A comparative evaluation of outlier detection algorithms: Experiments and analyses," *Pattern Recognition*, vol. 74, pp. 406–421, 2018.

[14] F. Jiang, Y. F. Sui, and C. G. Cao, "A rough set approach to outlier detection," *International Journal of General Systems*, vol. 37, no. 5, pp. 519–536, 2008.

[15] Y. M. Chen, D. Q. Miao, and R. Z. Wang, "Outlier detection based on granular computing," in *International Conference on Rough Sets and Current Trends in Computing*, pp. 283–292, 2008.

[16] F. Shaari, A. A. Bakar, and A. R. Hamdan, "Outlier detection based on rough sets theory," *Intelligent Data Analysis*, vol. 13, no. 2, pp. 191–206, 2009.

[17] F. Jiang, Y. F. Sui, and C. G. Cao, "A hybrid approach to outlier detection based on boundary region," *Pattern Recognition Letters*, vol. 32, no. 14, pp. 1860–1870, 2011.

[18] F. Jiang and Y. M. Chen, "Outlier detection based on granular computing and rough set theory," *Applied Intelligence*, vol. 42, no. 2, pp. 303–322, 2015.

[19] Y. M. Chen, D. Q. Miao, and H. Y. Zhang, "Neighborhood outlier detection," *Expert Systems with Applications*, vol. 37, no. 12, pp. 8745–8749, 2010.

[20] X. J. Li and F. Rao, "Outlier detection using the information entropy of neighborhood rough sets," *Journal of Information & Computational Science*, vol. 9, no. 12, pp. 3339–3350, 2012.

[21] Z. Yuan, X. Y. Zhang, and S. Feng, "Hybrid data-driven outlier detection based on neighborhood information entropy and its developmental measures," *Expert Systems with Applications*, vol. 112, pp. 243–257, 2018.

[22] Y. M. Chen, Y. Xue, Y. Ma, and F. F. Xu, "Measures of uncertainty for neighborhood rough sets," *Knowledge-Based Systems*, vol. 120, pp. 226–235, 2017.

[23] X. Y. Zhang, H. Y. Gou, Z. Y. Lv, and D. Q. Miao, "Double-quantitative distance measurement and classification learning based on the tri-level granular structure of neighborhood system," *Knowledge-Based Systems*, vol. 217, p. 106799, 2021.

[24] H. M. Chen, T. R. Li, Y. Cai, C. Luo, and H. Fujita, "Parallel attribute reduction in dominance-based neighborhood rough set," *Information Sciences*, vol. 373, pp. 351–368, 2016.

[25] R. Jensen and Q. Shen, "Semantics-preserving dimensionality reduction: rough and fuzzy-rough-based approaches," *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 12, pp. 1457–1471, 2004.

[26] P. Y. Goh, S. C. Tan, W. P. Cheah, and C. P. Lim, "Adaptive rough radial basis function neural network with prototype outlier removal," *Information Sciences*, vol. 505, pp. 127–143, 2019.

[27] Y. Wang and Y. P. Li, "Outlier detection based on weighted neighbourhood information network for mixed-valued datasets," *Information Sciences*, vol. 564, pp. 396–415, 2021.

[28] Y. Y. Yao, "Three-way decisions and cognitive computing," *Cognitive Computation*, vol. 8, no. 4, pp. 543–554, 2016.
[29] Y. F. Li, L. B. Zhang, Y. Xu, Y. Y. Yao, R. Y. K. Lau, and Y. T. Wu, "Enhancing binary classification by modeling uncertain boundary in three-way decisions," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 7, pp. 1438–1451, 2017.
[30] X. Y. Zhang and Y. Y. Yao, "Tri-level attribute reduction in rough set theory," *Expert Systems with Applications*, vol. 190, p. 116187, 2022.
[31] D. C. Liang, W. Pedrycz, and D. Liu, "Determining three-way decisions with decision-theoretic rough sets using a relative value approach," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 47, no. 8, pp. 1785–1799, 2016.
[32] J. T. Yao and N. Azam, "Web-based medical decision support systems for three-way medical decision making with game-theoretic rough sets," *IEEE Transactions on Fuzzy Systems*, vol. 23, no. 1, pp. 3–15, 2015.
[33] G. S. Pang, C. H. Shen, L. B. Cao, and A. V. D. Hengel, "Deep learning for anomaly detection: A review," *ACM computing surveys (CSUR)*, vol. 54, no. 2, pp. 1–38, 2021.
[34] L. Ruff, J. R. Kauffmann, R. A. Vandermeulen, G. Montavon, W. Samek, M. Kloft, T. G. Dietterich, and K. R. Müller, "A unifying review of deep and shallow anomaly detection," *Proceedings of the IEEE*, vol. 109, no. 5, pp. 756–795, 2021.
[35] S. Q. Han, X. Y. Hu, H. L. Huang, M. Q. Jiang, and Y. Zhao, "Adbench: Anomaly detection benchmark," *Advances in Neural Information Processing Systems*, vol. 35, pp. 32142–32159, 2022.
[36] P. J. Rousseeuw and A. M. Leroy, *Robust regression and outlier detection*. John Wiley & Sons, 2005.
[37] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "Lof: identifying density-based local outliers," in *Proceedings of The 2000 ACM SIGMOD International Conference on Management of Data*, pp. 93–104, 2000.
[38] E. M. Knorr and R. T. Ng, "A unified notion of outliers: Properties and computation," in *KDD*, vol. 97, pp. 219–222, 1997.
[39] E. M. Knorr, R. T. Ng, and V. Tucakov, "Distance-based outliers: algorithms and applications," *The VLDB Journal*, vol. 8, no. 3, pp. 237–253, 2000.
[40] Z. Y. He, X. F. Xu, and S. C. Deng, "Discovering cluster-based local outliers," *Pattern Recognition Letters*, vol. 24, no. 9–10, pp. 1641–1650, 2003.
[41] H. Z. Xu, Y. J. Wang, Y. J. Wang, and Z. Y. Wu, "Mix: A joint learning framework for detecting both clustered and scattered outliers in mixed-type data," in *2019 IEEE International Conference on Data Mining (ICDM)*, pp. 1408–1413, IEEE, 2019.
[42] Y. Zhao, Z. Nasrullah, and Z. Li, "Pyod: A python toolbox for scalable outlier detection," *Journal of Machine Learning Research*, vol. 20, no. 96, pp. 1–7, 2019.
[43] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.
[44] A. Albanese, S. K. Pal, and A. Petrosino, "Rough sets, kernel set, and spatiotemporal outlier detection," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 1, pp. 194–207, 2014.
[45] Q. H. Hu, D. R. Yu, J. F. Liu, and C. X. Wu, "Neighborhood rough set based heterogeneous feature subset selection," *Information Sciences*, vol. 178, no. 18, pp. 3577–3594, 2008.
[46] Q. H. Hu, D. R. Yu, and Z. X. Xie, "Neighborhood classifiers," *Expert Systems with Applications*, vol. 34, no. 2, pp. 866–876, 2008.
[47] D. R. Wilson and T. R. Martinez, "Improved heterogeneous distance functions," *Journal of Artificial Intelligence Research*, vol. 6, pp. 1–34, 1997.
[48] R. L. Kennedy, Y. Lee, B. Van Roy, C. D. Reed, and R. P. Lippmann, *Solving data mining problems through pattern recognition*. Prentice Hall, 1997.
[49] S. M. Weiss and N. Indurkhya, *Predictive data mining: A practical guide*. Morgan Kaufmann, 1998.
[50] D. W. Zhang, P. Wang, J. Q. Qiu, and Y. Jiang, "An improved approach to feature selection," in *2010 International Conference on Machine Learning and Cybernetics*, vol. 1, pp. 488–493, IEEE, 2010.
[51] F. Jiang, Y. F. Sui, and C. G. Cao, "A hybrid approach to outlier detection based on boundary region," *Pattern Recognition Letters*, vol. 32, no. 14, pp. 1860–1870, 2011.
[52] E. M. Knorr and R. T. Ng, "Algorithms for mining distancebased outliers in large datasets," in *Proceedings of The International Conference on Very Large Data Bases*, pp. 392–403, Citeseer, 1998.
[53] C. Wang, Z. Liu, H. Gao, and Y. Fu, "Vos: A new outlier detection model using virtual graph," *Knowledge-Based Systems*, vol. 185, p. 104907, 2019.
[54] M. Bouguessa, "A practical outlier detection approach for mixed-attribute data," *Expert Systems with Applications*, vol. 42, no. 22, pp. 8637–8649, 2015.
[55] Z. Yuan, B. Y. Chen, J. Liu, H. M. Chen, D. Z. Peng, and P. L. Li, "Anomaly detection based on weighted fuzzy-rough density," *Applied Soft Computing*, vol. 134, p. 109995, 2023.
[56] D. R. Yu, Q. H. Hu, and W. Bao, "Combining rough set methodology and fuzzy clustering for knowledge discovery from quantitative data," *Proceedings of the CSEE*, vol. 24, no. 6, pp. 205–210, 2004.
[57] Z. Yuan, H. Chen, C. Luo, and D. Z. Peng, "Mfgad: Multi-fuzzy granules anomaly detection," *Information Fusion*, vol. 95, pp. 17–25, 2023.
[58] G. O. Campos, A. Zimek, J. Sander, R. J. Campello, B. Micenková, E. Schubert, I. Assent, and M. E. Houle, "On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study," *Data Mining and Knowledge Discovery*, vol. 30, no. 4, pp. 891–927, 2016.
[59] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (roc) curve.," *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.
[60] M. Friedman, "A comparison of alternative tests of significance for the problem of m rankings," *The Annals of Mathematical Statistics*, vol. 11, no. 1, pp. 86–92, 1940.
[61] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.

Xianyong Zhang received the BS, MS, and PhD degrees in mathematics from Sichuan Normal University, Chengdu, China, in 2001, 2004, and 2011, respectively. He completed two-and-a-half-year visiting research in University of Regina, Saskatchewan, Canada and two-year postdoctoral work in Tongji University, Shanghai, China. He is currently a professor and a doctoral tutor with School of Mathematical Sciences, Visual Computing and Virtual Reality Key Laboratory of Sichuan Province, Sichuan Normal University, Chengdu, China. His research interests include uncertainty analysis, intelligent computing, data mining, and machine learning.

Zhong Yuan received the MS degree in mathematics from Sichuan Normal University, Chengdu, China, in 2018. He received the PhD degree in computer science and technology from Southwest Jiaotong University, Chengdu, China, in 2022. He is currently a distinguished associate researcher with College of Computer Science, Sichuan University, Chengdu, China. His research interests include granular computing, uncertainty information processing, anomaly detection, and knowledge discovery.

Duoqian Miao received the BS and MS degrees in mathematics and statistics from Shanxi University, Taiyuan, China, in 1985 and 1991, respectively. He received the PhD degree in pattern recognition and intelligent system at Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 1997. He is currently a professor and a doctoral tutor with Department of Computer Science and Technology, Key Laboratory of Embedded System and Service Computing, Tongji University, Shanghai, China. His broad research interests include data analysis, granular computing, soft computing, machine learning, and pattern recognition.