Outlier Detection Based on Fuzzy Rough Granules in Mixed Attribute Data

Zhong Yuan, Hongmei Chen, *Member, IEEE*, Tianrui Li, *Senior Member, IEEE*, Binbin Sang, and Shu Wang

***Abstract*—Outlier detection is one of the most important research directions in data mining. However, most of the current research focuses on outlier detection for categorical or numerical attribute data. There are few studies on the outlier detection of mixed attribute data. In this article, we introduce fuzzy rough sets (FRSs) to deal with the problem of outlier detection in mixed attribute data. Since the outlier detection model of the classical rough set is only applicable to the categorical attribute data, we use FRS to generalize the outlier detection model and construct a generalized outlier detection model based on fuzzy rough granules. First, the granule outlier degree (GOD) is defined to characterize the outlier degree of fuzzy rough granules by employing the fuzzy approximation accuracy. Then, the outlier factor based on fuzzy rough granules is constructed by integrating the GOD and the corresponding weights to characterize the outlier degree of objects. Furthermore, the corresponding fuzzy rough granules-based outlier detection (FRGOD) algorithm is designed. The effectiveness of the FRGOD algorithm is evaluated through experiments on 16 real-world datasets. The experimental results show that the algorithm is more flexible for detecting outliers and is suitable for numerical, categorical, and mixed attribute data.**

***Index Terms*—Approximation accuracy, fuzzy rough sets (FRSs), granular computing (GrC), mixed attribute, outlier detection.**

I. Introduction

OUTLIER detection is one of the most important research directions in the field of data mining. Its purpose is to find out objects whose behavior varies from the expected object. In general, outliers can be divided into three categories: 1) global outliers; 2) contextual (or conditional) outliers; and 3) collective outliers [1]. It plays an important role in many applications, such as intrusion detection, credit card fraud judgement, and medical diagnosis [2]. Lately, outlier detection has been applied to wireless sensor networks localization, time-series sequences, and process monitoring [3]–[5].

Recently, a growing number of researchers have attached importance to outlier detection and proposed many outlier detection methods. According to the specific theory or technical route adopted by outlier detection methods, the outlier detection methods can be roughly divided into: 1) statistical method [6]; 2) cluster-based method [7]; 3) depth-based method [8]; 4) distance-based method [9]; and 5) density-based method [10].

Statistical methods are mostly for single attribute, but many of the data usually involve multiple attributes, which makes them unsuitable for multidimensional datasets. The depth-based method is more effective for data in 2-D or 3-D space, and the detection efficiency of higher dimensional data with mixed attribute is lower.

To avoid problems with statistical and depth-based methods, Knorr and Ng proposed a distance-based method [9], which uses the distance between any two objects as a measure of the outlier degree. An object away from most of the rest is regarded as an outlier. Distance-based methods are widely used due to its ease of operation. However, for high-dimensional data, it is difficult to solve the sparse problem. For the reason that it uses two parameters, it is very sensitive to the choice of parameter.

The existing outlier detection work treats an outlier as a binary property. However, in real life, it makes more meaningful to assign an outlier degree to each object. Therefore, the density-based method was first proposed by Breunig *et al.* in 2000 [10], where its basic idea is that the density around the outliers is quite different from that around their nearest neighbors. Based on this, each object is assigned a local outlier factor (LOF) to indicate its outlier degree. The larger the LOF value of an object, the more likely it is to be an outlier. The density-based method still has the problem of sensitivity to parameter selection. What is more, most distance-based and density-based methods are designed by using the Euclidean distance. Therefore, in practical applications, distance-based and density-based methods may not be the best way to detect outliers for categorical (nominal) or mixed (hybrid or heterogenous) attribute data.

In recent years, new theories and paradigms have been used in data mining. As an important tool for knowledge acquisition, granular computing (GrC) is a new theory for simulating

Manuscript received January 1, 2020; revised May 16, 2020, September 5, 2020, and November 18, 2020; accepted February 8, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 61976182, Grant 61572406, and Grant 62076171; and in part by the Key Techniques of Integrated Operation and Maintenance for Urban Rail Train Dispatching Control System Based on Artificial Intelligence under Grant 2019YFH0097. This article was recommended by Associate Editor Z. Xu. *(Corresponding author: Hongmei Chen.)*

The authors are with the School of Information Science and Technology, Institute of Artificial Intelligence, Southwest Jiaotong University, Chengdu 611756, China, and also with the National Engineering Laboratory of Integrated Transportation Big Data Application Technology, Southwest Jiaotong University, Chengdu 611756, China (e-mail: yuanzhong2799@foxmail.com; hmchen@swjtu.edu.cn; trli@swjtu.edu.cn; sangbinbin@my.swjtu.edu.cn; swang@swjtu.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCYB.2021.3058780.

Digital Object Identifier 10.1109/TCYB.2021.3058780

the natural model of human thinking when solving large-scale complex problems [11], [12]. It provides a conceptual framework for studying many problems in data mining and pattern recognition. GrC models can be roughly divided into two categories: one is to deal with uncertainty as the main goal; and the other is to multigranule computing as the goal.

Rough set theory (RST) is one of implementations of GrC, which emphasizes the importance of the concept of granularity to some extent. Hence, it has become a popular mathematical framework of GrC, which is applied to many fields, such as feature selection, pattern recognition, and data mining [13]–[17]. In order to make up for the shortcomings of distance-based and density-based methods that cannot effectively process categorical attribute data, outlier detection algorithms based on rough sets or GrC are proposed. For example, Nguyen proposed a method for detecting and evaluating outliers using a multilevel approximate reasoning scheme in RST [18]. Chen *et al.* [19] introduced a GrC-based outlier detection method. Xue and Liu [20] put forward a semisupervised outlier detection method based on rough sets. Albanese *et al.* [21] used a new rough set approach to extend outlier detection to spatiotemporal data. In addition, Jiang *et al.* [22] built a method for detecting outliers by using attribute sequence in RST. Jiang *et al.* [23] proposed an outlier detection method based on information entropy and RST. On the basis of RST, Jiang and Chen [24] introduced a new outlier detection method by using classical approximation accuracy. Maciá-Pérez *et al.* [25] extended the mathematical framework of the basic theory of outlier detection based on RST, and designed an efficient algorithm for detecting outliers in a large amount of information. Lately, under the framework of RST, an outlier detection algorithm based on classical approximation accuracy and entropy is constructed in [26].

The above methods have proved the validity of the RST method for outlier detection. Nevertheless, it is worth noting that these methods establish mathematical models based on equivalence relations and equivalence classes, and their detection models are only applicable to categorical attributes rather than numerical (numeric) attribute data. These detection models require discretization in processing numerical attribute data, thereby increasing the time of data processing and accompanying significant information loss, which ultimately affects the accuracy of detection. To this end, Chen *et al.* [27] proposed a neighborhood outlier detection algorithm. Taking the neighborhood rough set as the basic framework, Yuan *et al.* [28], [30] and Yuan and Feng [29] proposed some outlier detection methods that are suitable for mixed attributes.

Because classical RST is only applicable for dealing with categorical or nominal data, Dubois and Prade proposed the fuzzy rough set (FRS) model [31]. Since FRSs use fuzzy binary relations to characterize the similarity between objects, it can directly process numerical or continuous attribute data without discretization, thus retaining more information on numerical or continuous values. Inspired by this idea, a series of extensions on FRSs [32]–[34] was introduced. For example, in order to explore more complex relationshinch per second and structures between objects, Yeung *et al.* [34] established a generalized FRS model based on t-norm and t-conorm. Moreover, how to effectively apply the FRS model to data mining has also attracted widespread attention. FRSs have been successfully applied to numerical or mixed feature selection, classification, clustering, etc., [35]–[40]. However, the research on the outlier detection of mixed attributes under the FRS model has not been reported.

Based on the above discussions, a new definition of global outlier based on fuzzy rough granules is proposed. First, we give the definition of fuzzy approximation accuracy. Second, based on the fuzzy approximation accuracy, the granule outlier degree (GOD) is defined to characterize the outlier degree of fuzzy rough granules. Third, the fuzzy rough granules-based outlier factor (FRGOF) is constructed by integrating GODs and corresponding weights to characterize the outlier degree of objects. Finally, the specific fuzzy rough granules-based outlier detection (FRGOD) algorithm is designed. Specifically, our main contributions in this article are as follows.

1) A new FRS model is constructed by defining a mixed fuzzy similarity relation.
2) The fuzzy approximation accuracy is defined to construct GOD and outlier factor.
3) A generalized outlier detection model based on fuzzy rough granules is proposed.
4) The experimental results show that the proposed model has better validity and adaptability for numerical, categorical, and mixed attribute data.

The remainder of this article is organized as follows. In the next section, we introduce preliminary knowledge on FRS. In Section III, we construct a generalized outlier detection model, analyze the specific outlier detection model, and propose the corresponding outlier detection algorithm. The results of our experiments are reported in Section IV. Finally, Section V summarizes this article.

II. Preliminaries

This section will review some of the definitions and symbols used in the subsequent sections [34], [41], [42].

A. Granular Computing

GrC is a new theory for simulating the natural model of human thinking when solving large-scale complex problems, which provides a conceptual framework for studying many problems in data mining and pattern recognition. GrC models can be roughly divided into two categories: one is to deal with uncertainty as the main goal, such as word computing models based on the fuzzy set theory; and the other is to multigranule computing as the goal, such as the quotient space theory, RST, cloud model, etc. In addition to the above-mentioned main GrC models, with the in-depth research of GrC, some extended GrC fuzzy has been put forward one after another, such as fuzzy RST, neighborhood RST, and cloud rough set model [11], [12].

B. Fuzzy Rough Set

FRS is an important model of GrC, which is a powerful tool to deal with uncertainty information of numerical or mixed

data. In an FRS model, the information is stored in a table, where each row represents an object (an instance). The data table is also called the fuzzy information system (FIS), and its formal definition is as follows.

An FIS is a quadruple FIS $= (U, A, V, f)$, where U is a nonempty finite set of objects, called the universe of discourse (universe); A is a nonempty finite set of attributes; V is the attribute domain, that is, $V = \bigcup_{a\in A} V_a$, where V_a is the domain of the attribute a; and $f : U \times A \to V$ is an information function that satisfies for $\forall a \in A$ and $x \in U$, with $f(x, a) \in V_a$. When $A = C \cup D$ and $C \cap D = \emptyset$, the FIS is called the fuzzy decision system (FDS), where C denotes the condition attribute and D denotes decision attributes.

Definition 1: Let $U = \{x_1, x_2, \ldots, x_n\}$, if $\mathcal{A}$ is a map of U to [0, 1] , which is $\mathcal{A} : U \to [0, 1]$, then $\mathcal{A}$ is called a fuzzy set on U.

For $\forall x_i \in U$, $\mathcal{A}(x_i)$ is called the membership function of $\mathcal{A}$, or the membership of x_i for $\mathcal{A}$. The entire fuzzy sets on U are recorded as $\mathcal{F}(U)$. Obviously, $P(U) \subseteq \mathcal{F}(U)$, where $P(U)$ is the power set of U. The fuzzy set can be denoted as $\mathcal{A} = (\mathcal{A}(x_1), \mathcal{A}(x_2), \ldots, \mathcal{A}(x_n))$ or $\mathcal{A} = \sum_{i=1}^{n} \mathcal{A}(x_i)/x_i$.

Let $\mathcal{A}, \mathcal{B} \in \mathcal{F}(U)$. For $\forall x \in U$, some operations of fuzzy sets are defined as follows.

1) $\mathcal{A} = \mathcal{B} \Leftrightarrow \mathcal{A}(x) = \mathcal{B}(x)$, $\mathcal{A} \subseteq \mathcal{B} \Leftrightarrow \mathcal{A}(x) \le \mathcal{B}(x)$.
2) $(\mathcal{A} \cup \mathcal{B})(x) = \max\{\mathcal{A}(x), \mathcal{B}(x)\} = \mathcal{A}(x) \vee \mathcal{B}(x)$,$(\mathcal{A} \cap \mathcal{B})(x) = \min\{\mathcal{A}(x), \mathcal{B}(x)\} = \mathcal{A}(x) \wedge \mathcal{B}(x)$.
3) $\mathcal{A}^c(x) = 1 - \mathcal{A}(x)$.

Definition 2: Let $U = \{u_1, u_2, \ldots, u_m\}$ and $V = \{v_1, v_2, \ldots, v_n\}$, the fuzzy relation $\mathcal{R}$ from U to V is defined as a fuzzy set $\mathcal{R} : U \times V \to [0, 1]$.

For $\forall(u, v) \in U \times V$, the membership degree $\mathcal{R}(u, v)$ indicates the degree to which u has a relation $\mathcal{R}$ with v. In particular, the fuzzy relation from U to U is called the fuzzy relation on U.

A fuzzy relation $\mathcal{R}$ from U to V can be represented by a fuzzy relation matrix, that is, $M_{\mathcal{R}} = (r_{ij})_{m\times n}$, where $r_{ij} = \mathcal{R}(u_i, v_j)$, each row vector $(r_{i1}, r_{i2}, \ldots, r_{in})$ represents a fuzzy set. The set of all fuzzy relations from U to V is denoted as $\mathcal{F}(U \times V)$.

Definition 3: Let $\mathcal{R}$ be a fuzzy relation on U, that is, $\mathcal{R} \in \mathcal{F}(U \times U)$. For $\forall x, y, z \in U$, we have:

1) $\mathcal{R}$ is reflexive $\Leftrightarrow \mathcal{R}(x, x) = 1$;
2) $\mathcal{R}$ is symmetric $\Leftrightarrow \mathcal{R}(x, y) = \mathcal{R}(y, x)$;
3) $\mathcal{R}$ is transitive $\Leftrightarrow \mathcal{R}(x, z) \ge \vee_{y\in U}(\mathcal{R}(x, y) \wedge \mathcal{R}(y, z))$.

If $\mathcal{R}$ satisfies reflexive and symmetric, then $\mathcal{R}$ is said to be a fuzzy similarity relation on U; and if $\mathcal{R}$ satisfies reflexive, symmetric, and transitive, then $\mathcal{R}$ is said to be a fuzzy equivalence relation on U.

The union, intersection, and complement operations of the fuzzy set are extended to the general triangular norm (t-norm), triangular conorm (s-norm or t-residual norm), and the negator operator.

Definition 4: Let the function $\mathcal{T} : [0, 1] \times [0, 1] \to [0, 1]$, if $\forall a, b, c \in [0, 1]$, it satisfies the following conditions.

1) *Commutativity:* $\mathcal{T}(a, b) = \mathcal{T}(b, a)$.
2) *Associativity:* $\mathcal{T}(a, \mathcal{T}(b, c)) = \mathcal{T}(\mathcal{T}(a, b), c)$.
3) *Monotonicity:* $b \le c \Rightarrow \mathcal{T}(a, b) < \mathcal{T}(a, c)$.
4) *Boundary Condition:* $\mathcal{T}(a, 1) = a$.

Then, $\mathcal{T}$ is called the triangular norm on [0, 1].

TABLE I
SOME TRIANGULAR NORMS AND CONORMS

$\mathcal{T}(a, b)$	$\mathcal{S}(a, b)$
$\min\{a, b\}$	$\max\{a, b\}$
ab	$a + b - ab$
$\max\{a + b - 1, 0\}$	$\min\{a + b, 1\}$

If the binary operator $\mathcal{S}$ satisfies the above commutativity, associativity, monotonicity, and boundary condition $\mathcal{S}(a, 0) = a$, then the binary operator $\mathcal{S}$ is called the triangle conorm on [0, 1].

Definition 5: Negator operator $\mathcal{N}$ is a descending map on $[0, 1] \to [0, 1]$ that satisfies the boundary conditions $\mathcal{N}(0) = 1$ and $\mathcal{N}(1) = 0$. Usually, $\mathcal{N}(a) = 1 - a$ is called the standard negator operator. If $\mathcal{N}(\mathcal{N}(a)) = a$, then $\mathcal{N}$ is called involutive.

Table I shows some triangular norms and conorms commonly used in fuzzy reasoning.

Presently, the FRS refers to the concept first proposed by Dubois and Prade [31], which is defined as follows.

Definition 6: Given FIS $= (U, A, V, f)$, let $\mathcal{R}$ be a fuzzy equivalence relation on U. For $\forall \mathcal{X} \in \mathcal{F}(U)$, the lower approximation $\underline{\mathcal{R}}\mathcal{X}$ and upper approximation $\overline{\mathcal{R}}\mathcal{X}$ of $\mathcal{X}$ are a pair of fuzzy sets on U whose membership functions, respectively, are

$$\underline{\mathcal{R}}\mathcal{X}(x) = \inf_{y\in U} \max\{1 - \mathcal{R}(x, y), \mathcal{X}(y)\} \tag{1}$$

$$\overline{\mathcal{R}}\mathcal{X}(x) = \sup_{y\in U} \min\{\mathcal{R}(x, y), \mathcal{X}(y)\}. \tag{2}$$

In order to explore more complex relationships per second and structures between objects, Yeung *et al.* established a generalized FRS model based on $\mathcal{T}$ and $\mathcal{S}$ [34].

Definition 7: Let $\mathcal{R}$ be a fuzzy similarity relation on U. For $\forall \mathcal{X} \in \mathcal{F}(U)$, the lower approximation $\underline{\mathcal{R}}\mathcal{X}$ and upper approximation $\overline{\mathcal{R}}\mathcal{X}$ of $\mathcal{X}$ are a pair of fuzzy sets on U whose membership functions, respectively, are

$$\underline{\mathcal{R}}\mathcal{X}(x) = \inf_{y\in U} \mathcal{S}(\mathcal{N}(\mathcal{R}(x, y)), \mathcal{X}(y)) \tag{3}$$

$$\overline{\mathcal{R}}\mathcal{X}(x) = \sup_{y\in U} \mathcal{T}(\mathcal{R}(x, y), \mathcal{X}(y)). \tag{4}$$

We can discovery that different FRS models can be obtained when different $\mathcal{S}$ and $\mathcal{T}$ are used. In practice, we can choose different $\mathcal{S}$ and $\mathcal{T}$ as needed. Clearly, when $\mathcal{T}(x, y) = \min\{x, y\}$ and $\mathcal{S}(x, y) = \max\{x, y\}$, the FRS model defined is the same as that in Definition 6.

C. Some Definitions of Outlier

Based on Hawkins' definition of outlier [2], the formal definition of outlier is as follows.

Definition 8: Let $R = \{r(x_1), r(x_2), \ldots, r(x_n)\}$ and $S = \{g(x_1), g(x_2), \ldots, g(x_n)\}$ be the value set of objects x_i, respectively, generated by two mechanisms $r(\cdot)$ and $g(\cdot)$, and $M(\cdot)$ is a measure that can express the characteristics of outliers. For $\forall x_i \in U$, $g(x_i) \ne r(x_i)$ and $\forall x_j (j \ne i)$, if $M(x_i) \gg M(x_j)$, then x_i is called an outlier in U.

Furthermore, Knorr and Ng introduced a formal definition of the distance-based outlier [9].

Definition 9: Let $r(r \geq 0)$ be a given distance threshold, $\xi(0 \leq \xi \leq 1)$ be a given fraction threshold, and $\text{dist}(x, x')$ be a given distance metric. For $\forall x \in U$, if

$$\frac{\left|\left\{x' | \text{dist}(x, x') \leq r\right\}\right|}{|U|} \leq \xi \tag{5}$$

then x is called a distance-based outlier, where $|\cdot|$ denotes the cardinality of a classical set.

In view of an outlier is regarded as a binary property in the existing outlier detection work (such as the above distance-based method), Breunig *et al.* proposed a density-based method [10]. In this method, each data object is assigned a LOF to indicate its degree of outlier. The larger the LOF value of a data object, the more likely it is to be an outlier. Recently, Jiang and Chen [24] defined a GrC and rough sets-based outlier factor to detect outliers. Following their ideas, this article defines FRGOF to detect outliers.

III. Outlier Detection Based on Fuzzy Rough Granules in Mixed Attribute Data

In this section, we first introduce a generalized outlier detection model, and then analyze it based on a specific detection model. Finally, we design a specific algorithm.

A. Generalized Outlier Detection Model

Let $U = \{x_1, x_2, \ldots, x_n\}$ and $C = \{c_1, c_2, \ldots, c_m\}$. For $\forall B \subseteq C$, let $B = \{c_{k_1}, c_{k_2}, \ldots, c_{k_h}\}(h \in [1, m])$, a fuzzy similarity relation $\mathcal{R}_B$ with respect to B on U can be induced. $M_{\mathcal{R}_B} = (r_{ij}^B)_{n\times n}$ denotes the fuzzy similarity relation matrix on $\mathcal{R}_B$, where $r_{ij}^B = \mathcal{R}_B(x_i, x_j)$.

Generally, the membership degree $\mathcal{R}_B(x_i, x_j)$ is defined by the following methods [43]: 1) quantitative product method; 2) angle cosine method; 3) correlation coefficient method; 4) closeness method; 5) distance method; and 6) conjunction method. Among them, the conjunction method is used in most literatures [36], [39], [42], [44], which is calculated by $\mathcal{R}_B(x_i, x_j) = \wedge_{l=1}^{h} \mathcal{R}_{c_{k_l}}(x_i, x_j)$. In this article, the conjunction method is adopted.

A fuzzy similarity relation induces a general fuzzy partition.

Definition 10: Given $\text{FIS} = (U, C, V, f)$ and $\forall B \subset C$. The general fuzzy partition $U/\mathcal{R}_B$ of U induced by the fuzzy similarity relationship $\mathcal{R}_B$ is defined as

$$U/\mathcal{R}_B = \{[x_1]_{\mathcal{R}_B}, [x_2]_{\mathcal{R}_B}, \ldots, [x_n]_{\mathcal{R}_B}\} \tag{6}$$

where $[x_i]_{\mathcal{R}_B} = (r_{i1}^B/x_1) + (r_{i2}^B/x_2) + \cdots + (r_{in}^B/x_n) = (r_{i1}^B, r_{i2}^B, \ldots, r_{in}^B)$ is called a fuzzy rough granule generated by the fuzzy similarity relation $\mathcal{R}_B$.

Obviously, $[x_i]_{\mathcal{R}_B}$ is a fuzzy set on $\mathcal{R}_B$. We have $[x_i]_{\mathcal{R}_B}(x_j) = \mathcal{R}_B(x_i, x_j) = r_{ij}^B$. If $\mathcal{R}_B(x_i, x_j) = 1$, then it means that x_j certainly belongs to $[x_i]_{\mathcal{R}_B}$. If $\mathcal{R}_B(x_i, x_j) = 0$, then x_j definitely does not belong to $[x_i]_{\mathcal{R}_B}$. The cardinality of the fuzzy rough granule $[x_i]_{\mathcal{R}_B}$ is calculated by $|[x_i]_{\mathcal{R}_B}| = \sum_{j=1}^{n} \mathcal{R}_B(x_i, x_j)$. We can obtain $1 \leq |[x_i]_{\mathcal{R}_B}| \leq n$.

Definition 11: For $\forall P, Q \subset C$, where $U/\mathcal{R}_Q = \{[x_1]_{\mathcal{R}_Q}, [x_2]_{\mathcal{R}_Q}, \ldots, [x_n]_{\mathcal{R}_Q}\}$. For $\forall [x_i]_{\mathcal{R}_Q} \in U/\mathcal{R}_Q$, the lower approximation $\underline{\mathcal{R}_P}[x_i]_{\mathcal{R}_Q}$ and upper approximation $\overline{\mathcal{R}_P}[x_i]_{\mathcal{R}_Q}$ of $[x_i]_{\mathcal{R}_Q}$ with respect to $\mathcal{R}_P$ are a pair of fuzzy sets on U whose membership functions, respectively, are

$$\underline{\mathcal{R}_P}[x_i]_{\mathcal{R}_Q}(x) = \inf_{y\in U} \mathcal{S}\Big(\mathcal{N}(\mathcal{R}_P(x, y)), [x_i]_{\mathcal{R}_Q}(y)\Big) \tag{7}$$

$$\overline{\mathcal{R}_P}[x_i]_{\mathcal{R}_Q}(x) = \sup_{y\in U} \mathcal{T}\Big(\mathcal{R}_P(x, y), [x_i]_{\mathcal{R}_Q}(y)\Big). \tag{8}$$

As mentioned above, for $x_i \in U$ and a fuzzy similarity relation $\mathcal{R}_B$, we can obtain the fuzzy rough granule $[x_i]_{\mathcal{R}_B}$. Before calculating the outlier factor of x_i, we first calculate the outlier degree of $[x_i]_{\mathcal{R}_B}$, and then define the outlier factor of x_i by the outlier degree of $[x_i]_{\mathcal{R}_B}$. Generally speaking, the larger the outlier factor of an object, the more likely it is to be an outlier.

In order to calculate the outlier degree of a given fuzzy rough granule, the approximation accuracy of the FRS is defined. The approximation accuracy is an important concepts in RST [45], which is used to analyze the roughness of set through the perspective of the rough boundary. On the basis of this, the approximation accuracy based on FRSs is defined as follows.

Definition 12: For $\forall B \subseteq C$, where $|C-B| \geq 2$, let $U/\mathcal{R}_B = \{[x_1]_{\mathcal{R}_B}, [x_2]_{\mathcal{R}_B}, \ldots, [x_n]_{\mathcal{R}_B}\}$. For $\forall [x_i]_{\mathcal{R}_B} \in U/\mathcal{R}_B$ and $P \subseteq C - B$, the approximation accuracy of $[x_i]_{\mathcal{R}_B}$ with respect to $\mathcal{R}_P$ is defined as

$$\alpha_{\mathcal{R}_P}\big([x_i]_{\mathcal{R}_B}\big) = \frac{\left|\underline{\mathcal{R}_P}[x_i]_{\mathcal{R}_B}\right|}{\left|\overline{\mathcal{R}_P}[x_i]_{\mathcal{R}_B}\right|}. \tag{9}$$

It can be seen that $0 \leq \alpha_{\mathcal{R}_P}([x_i]_{\mathcal{R}_B}) \leq 1$.

Herein, we use the approximation accuracy to define the outlier degree of a fuzzy rough granule. For a given x_i, we calculate the approximation accuracy of $[x_i]_{\mathcal{R}_B}$ relative to a set of fuzzy similarity relations. If the approximation accuracy of these relations is always low, then we can consider x_i as an anomalous object, so the outlier degree of x_i will be high.

For $\forall x_i \in U$, most current outlier detection methods only give a binary property of x_i, that is, x_i is or is not an outlier [10], [24]. In many cases, it makes more sense to specify an outlier factor for x_i. In the following section, similar to the above idea, we introduce two concepts: 1) GOD and 2) FRGOF. GOD is used to quantify the outlier degree of a given fuzzy rough granule and FRGOF is used to indicate the outlier degree of a given object.

Definition 13: Let $P = C - B = \{p_{k_1}, p_{k_2}, \ldots, p_{k_q}\}$. For $\forall x_i \in U$, the $\text{GOD}([x_i]_{\mathcal{R}_B})$ of $[x_i]_{\mathcal{R}_B}$ is defined as

$$\text{GOD}\big([x_i]_{\mathcal{R}_B}\big) = 1 - \frac{|[x_i]_{\mathcal{R}_B}| \cdot \Big(\alpha_{\mathcal{R}_P}([x_i]_{\mathcal{R}_B}) + \sum_{r=1}^{q} \Big(\alpha_{\mathcal{R}_{P-\{p_{k_r}\}}}([x_i]_{\mathcal{R}_B}) + 1\Big)/2\Big)}{|U| \cdot (q+1)}. \tag{10}$$

In Definition 13, we use $\text{GOD}([x_i]_{\mathcal{R}_B})$ to measure the outlier degree of the fuzzy rough granule $[x_i]_{\mathcal{R}_B}$. Since $\alpha_{\mathcal{R}_P}([x_i]_{\mathcal{R}_B})$ can be used to measure the uncertainty of $[x_i]_{\mathcal{R}_B}$, we calculate $\text{GOD}([x_i]_{\mathcal{R}_B})$ by using it. Given a set of fuzzy similarity relations on U, if the approximation accuracies for these relations are always low, then we think that x_i is not behaving properly and $\text{GOD}([x_i]_{\mathcal{R}_B})$ will be high.

In order to calculate the approximation accuracies for various fuzzy similarity relations, we should not attempt to check all subsets of C because each subset of C determines a fuzzy similarity relations on U, so we will obtain $2^{|C|}$ fuzzy similarity relations. Calculating all approximation accuracies on these relations is impractical because the time complexity will be exponential with respect to $|C|$. Therefore, we only calculate the approximation accuracies with respect to $\mathcal{R}_P$ and $\mathcal{R}_{P-\{c_k\}}$.

Definition 14: For $\forall x_i \in U$, the FRGOF(x_i) of x_i can be defined as

$$\mathrm{FRGOF}(x_i) = \frac{\sum_{k=1}^{m} \left(\mathrm{GOD}([x_i]_{\mathcal{R}_{c_k}}) \cdot W_{c_k}(x_i)\right)}{|C|} \tag{11}$$

where $W_{c_k} : U \to [0, 1]$ is a weight function that satisfies $W_{c_k}(x) = 1 - \sqrt[3]{|[x_i]_{\mathcal{R}_{c_k}}|/|U|}$.

Each $B \subseteq C$ can determine a $\mathcal{R}_B$, then we can obtain a GOD on it, so there are $2^{|C|}$ GODs. It is also impractical to calculate the outlier factors of all GODs because time complexity is exponential with respect to $|C|$. Therefore, in Definition 14, we only use $|C|$ GODs, which is determined by $c_k \in C$ to calculate FRGOF(x_i).

The weight function $W_{c_k}(x_i)$ in Definition 14 conforms to the opinion that outlier detection always involves a small amount of objects in a dataset, and these objects are more likely to be outliers. As can be seen from Definition 14, the higher the weight of x_i, the larger FRGOF(x_i) of x_i, so a minority objects should have higher weight than most objects. For $\forall x_i \in U$, if the cardinality of $[x_i]_{\mathcal{R}_{c_k}}$ is less than that of the other fuzzy rough granule, then we treat x_i as a minority of objects in U and give x_i a higher weight.

Equations (10) and (11) origin from Hawkins' view of outliers, that is, for $\forall x_i \in U$, if x_i has some anomalous characteristics relative to other objects in U, then we can treat x_i as an outlier in U. In (10) and (11), the uncertainty is regarded as an anomalous characteristic. If $\mathrm{GOD}([x_i]_{\mathcal{R}_{c_k}})$ always stays very high, then we might consider that x_i is not performing properly, and the outlier degree of x_i will be quite high. Hence, in (11), the outlier degree of x_i is proportional to $\mathrm{GOD}([x_i]_{\mathcal{R}_{c_k}})$.

Definition 15: Let μ be a given threshold, for $\forall x \in U$, if FRGOF(x) $> \mu$, then we call x as an outlier based on fuzzy rough granules.

Until now, we have established a generalized outlier detection model called FRGOD. The route to construct the model consists of five parts: 1) FRS model; 2) approximation accuracy; 3) GOD; 4) FRGOF; and 5) outlier detection. The FRGOD model has strong generalization ability, because different $\mathcal{T}$ and $\mathcal{S}$ can obtain different FRS models, so that different FRGOD models can be obtained. Subsequently, a specific FRGOD method will be analyzed.

The description of an object generally differs in magnitude and dimension. In order to obtain accurate data processing results and avoid the impact of different data, we must first normalize the original numerical attribute values [30], where this article uses min–max normalization. Its calculation formula is as follows:

$$F(f(x_i, c_k)) = \frac{f(x_i, c_k) - \min_{c_k}}{\max_{c_k} - \min_{c_k}} \tag{12}$$

where $\max_{c_k}$ and $\min_{c_k}$ are the maximum and minimum values of the attribute $c_k \in C$, respectively. After normalization, the attribute values of these attributes are in the range [0, 1].

To efficiently process nominal, numeric, and mixed attribute data, the mixed fuzzy similarity degree $r_{ij}^{c_k}$ between the object x_i and x_j with respect to the attribute c_k is calculated as

$$r_{ij}^{c_k} = \begin{cases} 1, f(x_i, c_k) = f\left(x_j, c_k\right) \text{ and} \\ c_k \text{ is categorical} \\ 0, f(x_i, c_k) \neq f\left(x_j, c_k\right) \text{ and} \\ c_k \text{ is categorical} \\ 1 - \left|f(x_i, c_k) - f\left(x_j, c_k\right)\right| \\ \left|f(x_i, c_k) - f\left(x_j, c_k\right)\right| \leq \varepsilon_k \text{ and} \\ c_k \text{ is numerical} \\ 0, |f(x_i, c_k) - f\left(x_j, c_k\right)| > \varepsilon_k \text{ and} \\ c_k \text{ is numerical} \end{cases} \tag{13}$$

where ε_{c_k} is the radius of adjustable fuzzy similarity degree, which is calculated as $\varepsilon_{c_k} = (\mathrm{std}(c_k)/\lambda)$, where $\mathrm{std}(c_k)$ is the standard deviation of the attribute values of c_k, and the default parameter λ is used to adjust the fuzzy similarity degree radius. Obviously, there are $r_{ij}^{c_k} = r_{ij}^{c_k}$, $r_{ii}^{c_k} = 1$, and $0 \leq r_{ij}^{c_k} \leq 1$, so $M_{\mathcal{R}_{c_k}}$ is a fuzzy similarity relation matrix.

We set $\mathcal{T}(x, y) = \min\{x, y\}$ and $\mathcal{S}(x, y) = \max\{x, y\}$. When the relation between objects is a crisp equivalence relation and the subset of objects to be approximated is a fuzzy set, then the model will degenerate into a rough fuzzy set model; if the subset of objects to be approximated is crisp and the relations between objects is a fuzzy similarity relation, then the model is an FRS model; in addition, when both are crisp equivalence relations, it is an RS model. These characteristics will make the FRGOD model suitable for outlier detection in mixed attribute data.

B. Specific Algorithm

In this section, we design a specific FRGOD algorithm and analyze its complexity.

Algorithm 1 obtains OS by four "for" loops. The first for loop calculates the fuzzy relation matrix of a single attribute. The second to fourth nested for loops calculate the GODs and the corresponding weights, and further calculates the outlier factor based on fuzzy rough granules. As a result, the number of loops for steps 2–4 is m, and the number of loops in step 3 is $n \times n$. Likewise, the number of loops for steps 5–22 is n, the number of loops for steps 6–17 is m, and the number of loops for steps 11–14 is $(m-1)$. Thus, the total number of loops for Algorithm 1 is $m \times n \times n + n \times m \times (m-1)$. Therefore, in the worst case, the time complexity of Algorithm 1 is $O(mn(m+n))$, and the space complexity is $O(mn)$. In order to explain the process of Algorithm 1 more clearly, its flowchart is depicted in Fig. 1.

IV. Experiments and Comparative Analyses

In this section, in order to evaluate the effectiveness of the FRGOD algorithm, 16 datasets (including numeric, categorical, and mixed attributes) are selected from the UCI machine learning repository for experiments [46].

Algorithm 1: FRGOD Algorithm

Input: $FIS = (U, C, V, f)$, μ, λ.
Output: Outlier Set (OS).

$OS \leftarrow \emptyset$;
for $k \leftarrow 1$ *to* m **do**
  Calculate $M_{\mathcal{R}_{c_k}}$;
end
for $i \leftarrow 1$ *to* n **do**
  for $k \leftarrow 1$ *to* m **do**
    Calculate $M_{\mathcal{R}_{C-\{c_k\}}}$;
    //Let $B = \{c_k\}$.
    Calculate $\alpha_{C-\{c_k\}}([x_i]_{\mathcal{R}_{c_k}})$;
    //Let $P = C - B = \{p_{k_1}, p_{k_2}, \ldots, p_{k_{m-1}}\}$.
    for $r \leftarrow 1$ *to* $m-1$ **do**
      Calculate $M_{\mathcal{R}_{C-\{c_k,p_{k_r}\}}}$;
      Calculate $\alpha_{C-\{c_k,p_{k_r}\}}([x_i]_{\mathcal{R}_{c_k}})$;
    end
    Calculate $GOD([x_i]_{\mathcal{R}_{c_k}})$;
    Calculate $W_{c_k}(x_i)$;
  end
  Calculate $FRGOF(x_i)$;
  if $FRGOF(x_i) > \mu$ **then**
    $OS \leftarrow OS \cup \{x_i\}$;
  end
end
return OS.

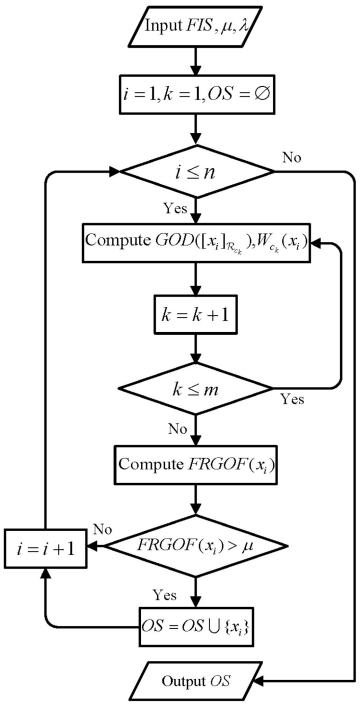


Fig. 1. Flowchart of FRGOD algorithm.

On 16 datasets, we compared the performance of the FRGOD algorithm with the distance-based method (DIS) [9], k-nearest neighbor-based method (kNN) [47], density-based method (LOF) [10], cluster-based method (find cluster-based LOF, FindCBLOF) [7], GrC-based method (GrC) [19], sequence and RS-based method (SEQ) [22], information entropy and RS-based method (IE) [23], and GrC and RST-based method (outlier detection based on GrC and rough set, ODGrCR) [24]. Among them, DIS, kNN, and LOF algorithms are relatively simple and only applicable to numerical attribute data. The FindCBLOF algorithm is well related to clustering methods, whereas it is only applicable to nominal attribute data. GrC, SEQ, IE, and ODGrCR algorithms are outlier detection algorithms using rough set as the framework. They are only applicable to categorical attribute data, while for numerical attribute data, discretization is required as one of preprocessing steps. Finally, time and space complexities of different outlier detection algorithms are listed in Table II, where $e = \sum_{c \in C} e_c$, e_c is the range of nominal attribute values in domain V_c. M denotes the maximum value of e_c.

TABLE II
COMPLEXITIES OF DIFFERENT ALGORITHMS

Algorithms	The worst time complexity	The space complexity
DIS	$O(mn^2)$	$O(mn)$
kNN	$O(mn^2)$	$O(mn)$
LOF	$O(mn^2)$	$O(mn)$
FindCBLOF	$O(mn)$	$O(m+n)$
GrC	$O(mn^2)$	$O(mn^2)$
SEQ	$O(m^2 n \log n)$	$O(mn)$
IE	$O(m^2 n \log n)$	$O(m(m+n))$
ODGrCR	$O(m^2(mn+e))$	$O(mn+M)$
FRGOD	$O(mn(m+n))$	$O(mn+n)$

A. Experimental Dataset

Most of public datasets are used for the evaluation of classification and clustering methods. For the evaluation of outlier detection, there are very few existing datasets. Accordingly, this article uses the downsampling method proposed in the document [48] to obtain some datasets for evaluating outlier detection methods. The method randomly downsamples a particular class to produce outliers while preserving all objects of the remaining classes to form a dataset for evaluating outlier detection methods. In addition, for the missing values of data set, this article uses the maximum probability value method to complete the missing values, that is, the value of attribute with the highest frequency on other objects is used to fill the missing attribute values. The details of data preprocessing and description are shown in Table III.

From Table III, we can see that there are two datasets containing only categorical attributes, nine datasets contain only numerical attributes, and the other five datasets contain both categorical attributes and numerical attributes, that is, mixed attributes. In addition, the number of objects is between 94 and 5171, and the number of attributes is between 8 and 60. To ensure repeatability of all experiments, the relevant datasets can be downloaded from github homepage.[1]

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

TABLE III
DESCRIPTION OF DATA AND THE DETAILS OF DATA PREPROCESSING

No	Data set	Abbreviation	Preprocessing	Conditional attribute		Oulier ($\|OS^\circ\|$)	Normal
				numerical	categorical		
1	Lymphography	Lymp	Classes "1" and "4" are considered as outliers [7]	0	8	6	142
2	Mushroom	Mush	Downsampling the class "+" to 221 objects	0	22	221	4208
3	Diabetes	Diab	Downsampling the class "tested_positive" to 26 objects	8	0	26	500
4	Ionosphere	Iono	Downsampling the class "b" to 24 objects	34	0	24	225
5	Iris	Iris	Downsampling the class "Iris-virginica" to 11 objects	4	0	11	100
6	Pima	Pima	Downsampling the class "TURE" to 55 objects	9	0	55	500
7	Sonar	Sonar	Downsampling the class "M" to 10 objects	60	0	10	97
8	Wisconsin diagnostic breast cancer	Wdbc	Downsampling the class "M" to 39 objects	31	0	39	357
9	Page blocks	Page	Downsampling the class "Non-text" to 258 objects	10	0	258	4913
10	Wisconsin breast cancer	Wbc	202 "malignant" (outlier) and 14 "benign" objects were removed [7]	9	0	39	444
11	Yeast	Yeast	Classes "ERL" (outlier), "CYT", "NUC", and "MIT" are selected [24]	8	0	5	1136
12	Credit approval	Cred	Downsampling the class "+" to 42 objects	6	9	42	383
13	German	Germ	Downsampling the class "2" to 14 objects	7	13	14	700
14	Heart disease	Heart	Downsampling the class "2" to 16 objects	6	7	16	150
15	Hepatitis	Hepa	Downsampling the class "2" to 9 objects	6	13	9	85
16	Horse	Horse	Downsampling the class "1" to 12 objects	8	19	12	244

B. Experimental Setup

In the experiments, we repeat the kNN and LOF algorithms for 16 datasets and calculate the optimal values of their respective parameters k and MinPts in the range $[1, n/4]$ with the step size of 1. The two parameters α and β required by the FindCBLOF algorithm are set to 90% and 5, respectively, [7]. For its similarity threshold s, the optimal values of s are calculated in the range $[1, 10]$ with step size of 1. For the GrC algorithm, the overlap distance metric is used to calculate the distance between any two objects [19], whose parameter $d = |C|/w$. We calculate integer optimal values of w in the range $[1, 10]$. In addition, for the SEQ, IE, and ODGrCR algorithms, all attribute values in the Lymp and Wbc datasets are considered to be categorical type [7]. For the remaining datasets with numerical attributes, we use Weka's discretization method based on equal width (EW) and equal frequency (EF) to convert all numerical attribute values into discrete forms with three interval numbers, and finally adopt the most excellent discretization method. For the DIS, kNN, and LOF algorithms, the Euclidean distance metric is used. All different nominal attribute values are replaced with different integer values, and then all attribute values are normalized to [0, 1] interval by using the min–max normalization (12). For the FRGOD algorithm, we calculate the optimal values for parameter λ in the range $[0.1, 5]$ with the step size of 0.1. Furthermore, outliers are treated as binary classifications in traditional distance-based methods. Obviously, this is unreasonable. Therefore, we use the strategy given in document [22], [23] to define the distance outlier factor $\mathrm{DOF}(x_i) = \sum_{j=1}^{n} \mathrm{dist}(x_i, x_j)$, which is used to indicate the outlier degree of each object. Finally, the optimal parameter settings and discretization methods for different datasets are described in Table IV.

The experiments in this section are carried out on a computer with the Intel core i7-8700 processor platform, 3.20-GHz frequency, 16-GB memory. The operating system is Windows 10. The experiments are performed in MATLAB R2015b.

In this article, precision (P), recall (R), and receiver operating characteristic (ROC) curves are used to evaluate the effectiveness of the proposed method [1]. The specific steps are as follows.

In the outlier detection, most of the outlier detection methods finally output the outlier factor value of each object in U, and the larger the outlier factor of an object, the more likely the object is a outlier. These objects can be sorted by their outlier factors in descending order and numbered from 1. Given a sequence number t, objects with a sequence number greater than or equal to t are considered as outliers. If t is too small, it will cause the method to miss the true outliers. Conversely, judging too many objects as outliers can lead to too many false positives. This tradeoff can usually be measured by P and R. Let $\mathrm{OS}(t)$ denote the set of outliers detected by a given t, that is, $\mathrm{OS}(t)$ is a function of t. OS° represents the set of true outlier in the dataset. $P(t)$ and $R(t)$ are, respectively, calculated by

$$P(t) = \frac{|\mathrm{OS}(t) \cap OS^\circ|}{|\mathrm{OS}(t)|} \times 100\% \tag{14}$$

$$R(t) = \frac{|\mathrm{OS}(t) \cap OS^\circ|}{|OS^\circ|} \times 100\% \tag{15}$$

where $P(t)$ denotes the proportion of true outliers detected under a given t. $R(t)$ represents the proportion of true outliers detected under a given t in the total number of true outliers.

The maximum possible value of $P(t)$ and $R(t)$ is 100%, and the minimum possible value is 0. Given the value of t, the larger the value of $P(t)$ and $R(t)$, the better outlier detection results. Obviously, when $P(t)$ and $R(t)$ are given, the smaller the value of t, the better the detection effect. In addition, it can be proved that $P(t)$ and $R(t)$ are equal when $t = |\mathrm{OS}^\circ|$.

The ROC curve is a curve with the false positive rate (FPR) as the abscissa and the true positive rate (TPR) as the ordinate. $\mathrm{FPR}(t)$ and $\mathrm{TPR}(t)$ are computed, respectively, as

$$\mathrm{FPR}(t) = \frac{|\mathrm{OS}(t) - OS^\circ|}{|U - OS^\circ|} \times 100\% \tag{16}$$

$$\mathrm{TPR}(t) = R(t) = \frac{|\mathrm{OS}(t) \cap OS^\circ|}{|OS^\circ|} \times 100\%. \tag{17}$$

The ROC curve is used to compare the performance of different outlier detection algorithms. If the ROC curve of a detection algorithm is as close as possible to the upper left corner of the first quadrant, and the larger the area under the curve, the better its performance.

TABLE IV
OPTIMAL PARAMETER SETTING AND DISCRETIZATION METHOD FOR DIFFERENT DATASETS

Data set	k(kNN)	MinPts(LOF)	s(FindCBLOF)	w(GrC)	λ(FRGOD)	Discretization method
Lymp	37	36	3 [7]	2 [19]	-	-
Mush	212	2	9	10	-	-
Diab	7	76	7	3	4.5	EW
Iono	2	2	4	5	1.5	EW
Iris	13	2	1	3	2.1	EW
Pima	136	109	6	5	0.9	EW
Sonar	13	19	2	2	1.2	EW
Wdbc	74	41	2	2	0.7	EW
Page	35	1281	2	7	0.9	EW
Wbc	39	15	3 [7]	2 [19]	0.4	-
Yeast	243	8	5	3	0.1	EW
Cred	6	106	10	3	1.3	EW
Germ	31	135	4	4	0.7	EW
Heart	31	30	5	4	1.8	EW
Hepa	16	8	8	4	0.6	EW
Horse	29	11	4	3	1.5	EF

TABLE V
COMPARISON OF EXPERIMENTAL RESULTS ON CATEGORICAL ATTRIBUTE DATASET

Data set	t	DIS		kNN		LOF		FindCBLOF		GrC		SEQ		IE		ODGrCR		FRGOD	
		$P(t)$	$R(t)$	$P(t)$	$R(t)$	$P(t)$	$R(t)$	$P(t)$	$R(t)$	$P(t)$	$R(t)$	$P(t)$	$R(t)$	$P(t)$	$R(t)$	$P(t)$	$R(t)$	$P(t)$	$R(t)$
Lymp	3	66.67	33.33	100.00	50.00	66.67	33.33	66.67	33.33	100.00	50.00	100.00	50.00	100.00	50.00	100.00	50.00	100.00	50.00
	4	75.00	50.00	100.00	66.67	75.00	50.00	50.00	33.33	100.00	66.67	100.00	66.67	100.00	66.67	100.00	66.67	100.00	66.67
	5	80.00	66.67	80.00	66.67	60.00	50.00	60.00	50.00	80.00	66.67	80.00	66.67	80.00	66.67	80.00	66.67	80.00	66.67
	6	66.67	66.67	66.67	66.67	66.67	66.67	50.00	50.00	83.33	83.33	66.67	66.67	83.33	83.33	83.33	83.33	83.33	83.33
	7	71.43	83.33	71.43	83.33	71.43	83.33	57.14	66.67	85.71	100.00	71.43	83.33	71.43	83.33	71.43	83.33	71.43	83.33
	9	55.56	83.33	66.67	100.00	55.56	83.33	44.44	66.67	66.67	100.00	55.56	83.33	66.67	100.00	66.67	100.00	66.67	100.00
	10	50.00	83.33	60.00	100.00	60.00	100.00	40.00	66.67	60.00	100.00	50.00	83.33	60.00	100.00	60.00	100.00	60.00	100.00
	11	54.55	100.00	54.55	100.00	54.55	100.00	36.36	66.67	54.55	100.00	45.45	83.33	54.55	100.00	54.55	100.00	54.55	100.00
	12	50.00	100.00	50.00	100.00	50.00	100.00	33.33	66.67	50.00	100.00	20.00	100.00	50.00	100.00	50.00	100.00	50.00	100.00
	23	26.09	100.00	26.09	100.00	26.09	100.00	21.74	83.33	26.09	100.00	26.09	100.00	26.09	100.00	26.09	100.00	26.09	100.00
	30	20.00	100.00	20.00	100.00	20.00	100.00	20.00	100.00	20.00	100.00	20.00	100.00	20.00	100.00	20.00	100.00	20.00	100.00
	Average	56.00	78.79	63.22	84.85	55.09	78.79	43.61	62.12	**66.03**	**87.88**	60.47	80.30	64.73	86.36	64.73	86.36	64.73	86.36
Mush	100	44.00	19.91	47.00	21.27	44.00	19.91	89.00	40.27	99.00	44.80	99.00	44.80	100.00	45.25	100.00	45.25	100.00	45.25
	221	19.91	19.91	22.14	22.17	20.81	20.81	87.33	87.33	87.33	87.33	87.33	87.33	85.52	85.52	87.33	87.33	87.33	87.33
	1011	19.39	88.69	15.53	71.04	20.38	93.21	21.46	98.19	19.39	88.69	19.68	90.05	21.46	98.19	21.86	100.00	21.86	100.00
	1185	17.05	91.40	18.65	100.00	18.06	96.83	18.57	99.55	16.79	90.05	16.79	90.05	18.31	98.19	18.65	100.00	18.65	100.00
	1231	16.49	91.86	17.95	100.00	17.38	96.83	17.95	100.00	16.17	90.05	16.17	90.05	17.63	98.19	17.95	100.00	17.95	100.00
	1440	14.31	93.21	15.35	100.00	14.86	96.83	15.53	100.00	13.82	90.05	13.82	90.05	5.35	100.00	15.35	100.00	15.35	100.00
	2143	9.89	95.93	10.31	100.00	10.31	100.00	10.31	100.00	10.03	97.29	10.22	99.10	10.31	100.00	10.31	100.00	10.31	100.00
	2171	9.77	95.93	10.18	100.00	10.18	100.00	10.18	100.00	10.00	98.19	10.18	100.00	10.18	100.00	10.18	100.00	10.18	100.00
	2337	9.16	96.83	9.46	100.00	9.46	100.00	9.46	100.00	9.46	100.00	9.46	100.00	9.46	100.00	9.46	100.00	9.46	100.00
	2700	8.19	100.00	8.19	100.00	8.19	100.00	8.19	100.00	8.19	100.00	8.19	100.00	8.19	100.00	8.19	100.00	8.19	100.00
	Average	16.81	79.37	17.48	81.45	17.36	82.44	28.78	92.53	29.02	88.64	29.08	89.14	29.64	92.53	**29.93**	**93.26**	**29.93**	**93.26**

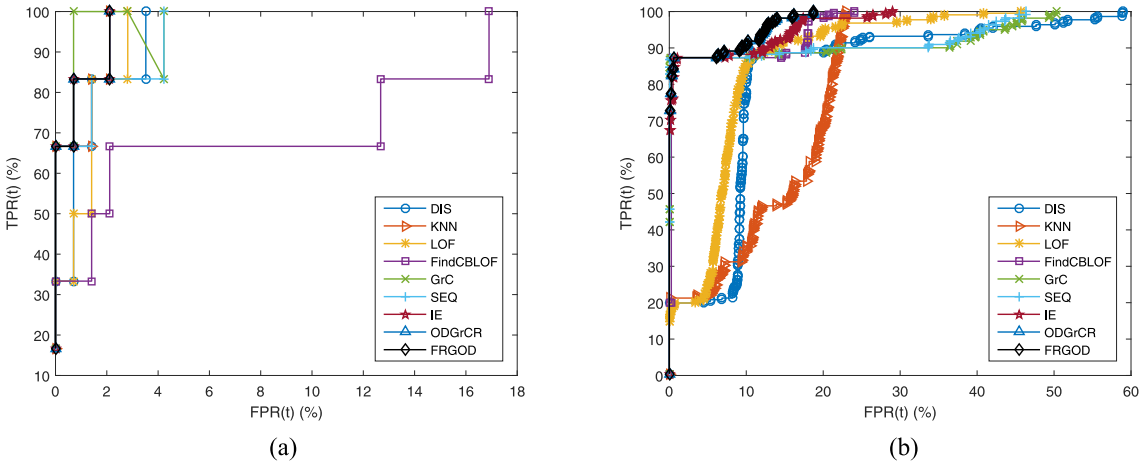


Fig. 2. ROC curve of categorical attribute dataset. (a) Lymp. (b) Mush.

C. Experimental Results Analyses

The comparison of experimental results is listed in Tables V–VII, and the ROC curves are depicted in Figs. 2–4.

First, we test the effectiveness of the FRGOD algorithm on the categorical data. Two datasets are used in the experiments. Table V and Fig. 2 give the comparison experimental results and ROC curves, respectively. Among them, the comparative experimental results show that the values of $P(t)$ and $R(t)$ of nine algorithms varying with t.

TABLE VI (a)
COMPARISON OF EXPERIMENTAL RESULTS ON NUMERICAL ATTRIBUTE DATASETS

Data set	t	DIS		kNN		LOF		FindCBLOF		GrC		SEQ		IE		ODGrCR		FRGOD	
		$P(t)$	$R(t)$	$P(t)$	$R(t)$	$P(t)$	$R(t)$	$P(t)$	$R(t)$	$P(t)$	$R(t)$	$P(t)$	$R(t)$	$P(t)$	$R(t)$	$P(t)$	$R(t)$	$P(t)$	$R(t)$
Diab	16	56.25	34.62	43.75	26.92	62.50	38.46	37.50	23.08	56.25	34.62	62.50	38.46	62.50	38.46	37.50	23.08	62.50	38.46
	26	42.31	42.31	46.15	46.15	46.15	46.15	34.62	34.62	46.15	46.15	38.46	38.46	46.15	46.15	38.46	38.46	69.23	69.23
	39	33.33	50.00	41.03	61.54	35.90	53.85	33.33	50.00	41.03	61.54	30.77	46.15	41.03	61.54	43.59	65.38	66.67	100.00
	81	29.63	92.31	32.10	100.00	29.63	92.31	27.16	84.62	29.63	92.31	28.40	88.46	28.40	88.46	29.63	92.31	32.10	100.00
	83	28.92	92.31	31.33	100.00	31.33	100.00	26.51	84.62	30.12	96.15	27.71	88.46	28.92	92.31	28.92	92.31	31.33	100.00
	91	28.57	100.00	28.57	100.00	28.57	100.00	26.37	92.31	27.47	96.15	26.37	92.31	26.37	92.31	26.37	92.31	28.57	100.00
	106	24.53	100.00	24.53	100.00	24.53	100.00	22.64	92.31	24.53	100.00	22.64	92.31	23.58	96.15	23.58	96.15	24.53	100.00
	108	24.07	100.00	24.07	100.00	24.07	100.00	22.22	92.31	24.07	100.00	22.22	92.31	24.07	100.00	24.07	100.00	24.07	100.00
	117	22.22	100.00	22.22	100.00	22.22	100.00	22.22	100.00	22.22	100.00	20.51	92.31	22.22	100.00	22.22	100.00	22.22	100.00
	226	11.50	100.00	11.50	100.00	11.50	100.00	11.50	100.00	11.50	100.00	11.50	100.00	11.50	100.00	11.50	100.00	11.50	100.00
	Average	30.13	81.15	30.53	83.46	31.64	83.08	26.41	75.38	31.30	82.69	29.11	76.92	31.47	81.54	28.59	80.00	**37.27**	**90.77**
Iono	13	100.00	54.17	100.00	54.17	100.00	54.17	92.31	50.00	92.31	50.00	92.31	50.00	76.92	41.67	92.31	50.00	100.00	54.17
	18	100.00	75.00	100.00	75.00	100.00	75.00	83.33	62.50	83.33	62.50	83.33	62.50	83.33	62.50	94.44	70.83	100.00	75.00
	24	91.67	91.67	100.00	100.00	91.67	91.67	75.00	75.00	75.00	75.00	75.00	75.00	70.83	70.83	83.33	83.33	91.67	91.67
	26	88.46	95.83	92.31	100.00	84.62	91.67	73.08	79.17	73.08	79.17	69.23	75.00	69.23	75.00	80.77	87.50	92.31	100.00
	31	77.42	100.00	77.42	100.00	74.19	95.83	64.52	83.33	64.52	83.33	64.52	83.33	61.29	79.17	77.42	100.00	77.42	100.00
	48	50.00	100.00	50.00	100.00	50.00	100.00	47.92	95.83	47.92	95.83	47.92	95.83	47.92	95.83	50.00	100.00	50.00	100.00
	52	46.15	100.00	46.15	100.00	46.15	100.00	44.23	95.83	44.23	95.83	44.23	95.83	46.15	100.00	46.15	100.00	46.15	100.00
	54	44.44	100.00	44.44	100.00	44.44	100.00	44.44	100.00	44.44	100.00	42.59	95.83	44.44	100.00	44.44	100.00	44.44	100.00
	125	19.20	100.00	19.20	100.00	19.20	100.00	19.20	100.00	19.20	100.00	19.20	100.00	19.20	100.00	19.20	100.00	19.20	100.00
	Average	68.59	90.74	**69.95**	**92.13**	67.81	89.81	60.45	82.41	60.45	82.41	59.81	81.48	57.70	80.56	65.34	87.96	69.02	91.20
Iris	6	100.00	54.55	83.33	45.45	100.00	54.55	100.00	54.55	100.00	54.55	100.00	54.55	33.33	18.18	100.00	54.55	100.00	54.55
	11	100.00	100.00	72.73	72.73	100.00	100.00	90.91	90.91	81.82	81.82	81.82	81.82	63.64	63.64	81.82	81.82	100.00	100.00
	12	91.67	100.00	75.00	81.82	91.67	100.00	91.67	100.00	83.33	90.91	83.33	90.91	66.67	72.73	83.33	90.91	91.67	100.00
	13	84.62	100.00	76.92	90.91	84.62	100.00	84.62	100.00	84.62	100.00	84.62	100.00	69.23	81.82	84.62	100.00	84.62	100.00
	14	78.57	100.00	78.57	100.00	78.57	100.00	78.57	100.00	78.57	100.00	78.57	100.00	71.43	90.91	78.57	100.00	78.57	100.00
	15	73.33	100.00	73.33	100.00	73.33	100.00	73.33	100.00	73.33	100.00	73.33	100.00	73.33	100.00	73.33	100.00	73.33	100.00
	Average	**88.03**	**92.42**	76.65	81.82	**88.03**	**92.42**	86.52	90.91	83.61	87.88	83.61	87.88	62.94	71.21	83.61	87.88	**88.03**	**92.42**
Pima	25	52.00	23.64	56.00	25.45	56.00	25.45	56.00	25.45	64.00	29.09	56.00	25.45	60.00	27.27	60.00	27.27	72.00	32.73
	55	49.09	49.09	52.73	52.73	49.09	49.09	56.36	56.36	49.09	49.09	40.00	40.00	47.27	47.27	50.91	50.91	65.45	65.45
	86	47.67	74.55	52.33	81.82	46.51	72.73	43.02	67.27	47.67	74.55	31.40	49.09	39.53	61.82	51.16	80.00	63.95	100.00
	145	37.93	100.00	37.24	98.18	37.93	100.00	36.55	96.36	35.86	94.55	22.76	60.00	33.79	89.09	35.86	94.55	37.93	100.00
	165	33.33	100.00	33.33	100.00	33.33	100.00	32.12	96.36	31.52	94.55	20.61	61.82	32.73	98.18	31.52	94.55	33.33	100.00
	175	31.43	100.00	31.43	100.00	31.43	100.00	30.29	96.36	29.71	94.55	19.43	61.82	31.43	100.00	29.71	94.55	31.43	100.00
	180	30.56	100.00	30.56	100.00	30.56	100.00	30.56	100.00	30.00	98.18	18.89	61.82	30.56	100.00	28.89	94.55	30.56	100.00
	181	30.39	100.00	30.39	100.00	30.39	100.00	30.39	100.00	30.39	100.00	18.78	61.82	30.39	100.00	28.73	94.55	30.39	100.00
	187	29.41	100.00	29.41	100.00	29.41	100.00	29.41	100.00	29.41	100.00	18.18	61.82	29.41	100.00	29.41	100.00	29.41	100.00
	369	14.91	100.00	14.91	100.00	14.91	100.00	14.91	100.00	14.91	100.00	14.91	100.00	14.91	100.00	14.91	100.00	14.91	100.00
	Average	35.67	84.73	36.83	85.82	35.96	84.73	35.96	83.82	36.26	83.45	26.09	58.36	35.00	82.36	36.11	83.09	**40.94**	**89.82**
Sonar	5	80.00	40.00	80.00	40.00	80.00	40.00	100.00	50.00	100.00	50.00	100.00	50.00	100.00	50.00	80.00	40.00	100.00	50.00
	10	80.00	80.00	70.00	70.00	80.00	80.00	80.00	80.00	80.00	80.00	80.00	80.00	80.00	80.00	80.00	80.00	80.00	80.00
	12	66.67	80.00	66.67	80.00	83.33	100.00	66.67	80.00	75.00	90.00	75.00	90.00	66.67	80.00	75.00	90.00	83.33	100.00
	13	69.23	90.00	69.23	90.00	76.92	100.00	69.23	90.00	76.92	100.00	69.23	90.00	61.54	80.00	69.23	90.00	76.92	100.00
	14	64.29	90.00	71.43	100.00	71.43	100.00	64.29	90.00	71.43	100.00	64.29	90.00	57.14	80.00	64.29	90.00	71.43	100.00
	15	66.67	100.00	66.67	100.00	66.67	100.00	66.67	100.00	66.67	100.00	60.00	90.00	60.00	90.00	66.67	100.00	66.67	100.00
	18	55.56	100.00	55.56	100.00	55.56	100.00	55.56	100.00	55.56	100.00	50.00	90.00	55.56	100.00	55.56	100.00	55.56	100.00
	22	45.45	100.00	45.45	100.00	45.45	100.00	45.45	100.00	45.45	100.00	45.45	100.00	45.45	100.00	45.45	100.00	45.45	100.00
	Average	65.98	85.00	65.63	85.00	69.92	90.00	68.48	86.25	71.38	90.00	68.00	85.00	65.79	82.50	67.02	86.25	**72.42**	**91.25**

It can be shown from Table V that the FRGOD algorithm achieves superior performance on the categorical attribute dataset. For example, in the Lymp dataset, on the one hand, when $t = 6$, $P(t)$ of the FRGOD algorithm is 83.33%, while for DIS, kNN, LOF, FindCBLOF, GrC, SEQ, IE, and ODGrCR, their $P(t)$ are 66.67%, 66.67%, 66.67%, 50.00%, 83.33%, 66.67%, 83.33%, and 83.33%, respectively. $P(t)$ of the FRGOD algorithm is greater than or equal to other algorithms. On the other hand, when $R(t)$ reaches 100.00% for the first time, the FRGOD algorithm's t is slightly larger than that of the GrC algorithm, but less than or equal to other algorithms. What is more, in terms of the average of $R(t)$, the FRGOD algorithm is greater than or equal to all other algorithms except for the GrC algorithm.

For the Mush dataset, when $t = 221$, $P(t)$ of the FRGOD algorithm is 87.33%, which is also greater than or equal to other algorithms. When $R(t)$ reaches 100.00% for the first time, the FRGOD algorithm's t is 1011, but for DIS, kNN, LOF, FindCBLOF, GrC, SEQ, IE, and ODGrCR algorithm, their t are 2700, 1185, 2143, 1231, 2337, 2171, 1440, and 1011, respectively, and t of the FRGOD algorithm is less than or equal to other algorithms. This indicates that when $t = 1011$, FRGOD and ODGrCR algorithms can detect all 221 outliers, but other algorithms can only detect outliers less than 221. In addition, for the average of $R(t)$, the FRGOD algorithm is larger than all other algorithms except the ODGrCR algorithm.

From Fig. 2, we can also see the validity of the FRGOD algorithm more vividly. For the Lymp dataset, it can be observed from Fig. 2(a) that the FRGOD algorithm is closer to the upper left corner of first quadrant, and the area under the curve is the larger one. It only slightly weaker than that of the GrC algorithm. For the Mush dataset, it can be observed from Fig. 2(b) that the FRGOD algorithm is the closest to the upper left corner of the first quadrant and the area under the curve is the largest one.

From the above analyses, we can conclude that the FRGOD algorithm can be applied to nominal attribute datasets.

Next, we test the validity of the FRGOD algorithm for nine numerical attribute datasets. The comparative experimental results are given in Tables VI (a) and VI (b), and the ROC curves are shown in Fig. 3.

From Tables VI (a) and VI (b), we can see that the FRGOD algorithm performs well on most datasets. For example, for the Diab dataset, on the one hand, when $t = 26$, $P(t)$ of the FRGOD algorithm is 69.23%, but for DIS, kNN, LOF,

TABLE VI (b)
CONTINUED

Data set	t	DIS		kNN		LOF		FindCBLOF		GrC		SEQ		IE		ODGrCR		FRGOD	
		$P(t)$	$R(t)$	$P(t)$	$R(t)$	$P(t)$	$R(t)$	$P(t)$	$R(t)$	$P(t)$	$R(t)$	$P(t)$	$R(t)$	$P(t)$	$R(t)$	$P(t)$	$R(t)$	$P(t)$	$R(t)$
Wdbc	19	89.47	43.59	89.47	43.59	94.74	46.15	94.74	46.15	100.00	48.72	52.63	25.64	94.74	46.15	94.74	46.15	94.74	46.15
	39	89.74	89.74	84.62	84.62	92.31	92.31	94.87	94.87	94.87	94.87	53.85	53.85	94.87	94.87	94.87	94.87	94.87	94.87
	41	87.80	92.31	85.37	89.74	92.68	97.44	92.68	97.44	92.68	97.44	56.10	58.97	90.24	94.87	92.68	97.44	95.12	100.00
	42	88.10	94.87	85.71	92.31	92.86	100.00	90.48	97.44	90.48	97.44	57.14	61.54	90.48	97.44	90.48	97.44	92.86	100.00
	43	88.37	97.44	86.05	94.87	90.70	100.00	88.37	97.44	90.70	100.00	58.14	64.10	88.37	97.44	88.37	97.44	90.70	100.00
	45	84.44	97.44	86.67	100.00	86.67	100.00	84.44	97.44	86.67	100.00	60.00	69.23	84.44	97.44	84.44	97.44	86.67	100.00
	46	84.78	100.00	84.78	100.00	84.78	100.00	82.61	97.44	84.78	100.00	60.87	71.79	82.61	97.44	84.78	100.00	84.78	100.00
	49	79.59	100.00	79.59	100.00	79.59	100.00	79.59	100.00	79.59	100.00	63.27	79.49	79.59	100.00	79.59	100.00	79.59	100.00
	64	60.94	100.00	60.94	100.00	60.94	100.00	60.94	100.00	60.94	100.00	60.94	100.00	60.94	100.00	60.94	100.00	60.94	100.00
	Average	83.69	90.60	82.58	89.46	86.14	92.88	85.41	92.02	86.75	93.16	58.10	64.96	85.14	91.74	85.66	92.31	**86.70**	**93.45**
Page	129	55.04	27.52	71.32	35.66	54.26	27.13	66.67	33.33	65.89	32.95	58.91	29.46	65.89	32.95	73.64	36.82	45.74	22.87
	258	51.16	51.16	59.69	59.69	50.78	50.78	43.41	43.41	37.60	37.60	41.86	41.86	43.41	43.41	43.41	43.41	29.84	29.84
	420	45.00	73.26	44.52	72.48	44.05	71.71	27.38	44.57	27.62	44.96	26.67	43.41	28.81	46.90	27.62	44.96	33.33	54.26
	781	28.81	87.21	29.45	89.15	28.04	84.88	16.77	50.78	16.77	50.78	14.72	44.57	29.45	89.15	22.79	68.99	33.03	100.00
	1348	18.99	99.22	19.14	100.00	18.92	98.84	17.58	91.86	17.58	91.86	8.61	44.96	17.58	91.86	17.58	91.86	19.14	100.00
	1404	18.38	100.00	18.38	100.00	18.16	98.84	16.88	91.86	16.88	91.86	8.26	44.96	16.88	91.86	16.88	91.86	18.38	100.00
	1476	17.48	100.00	17.48	100.00	17.48	100.00	16.06	91.86	16.06	91.86	7.93	45.35	16.06	91.86	16.06	91.86	17.48	100.00
	4908	5.26	100.00	5.26	100.00	5.26	100.00	5.26	100.00	5.26	100.00	4.52	86.05	5.26	100.00	5.26	100.00	5.26	100.00
	5171	4.99	100.00	4.99	100.00	4.99	100.00	4.99	100.00	4.99	100.00	4.99	100.00	4.99	100.00	4.99	100.00	4.99	100.00
	Average	27.23	82.04	**30.02**	**84.11**	26.88	81.35	23.89	71.96	23.18	71.32	19.61	53.40	25.37	76.44	25.36	74.42	23.02	78.55
Wbc	20	100.00	51.28	100.00	51.28	100.00	51.28	85.00	43.59	85.00	43.59	90.00	46.15	90.00	46.15	90.00	46.15	100.00	51.28
	39	87.18	87.18	87.18	87.18	89.74	89.74	79.49	79.49	79.49	79.49	82.05	82.05	82.05	82.05	84.62	84.62	89.74	89.74
	48	79.17	97.44	77.08	94.87	79.17	97.44	72.92	89.74	77.08	94.87	75.00	92.31	75.00	92.31	77.08	94.87	81.25	100.00
	49	77.55	97.44	77.55	97.44	79.59	100.00	71.43	89.74	75.51	94.87	73.47	92.31	73.47	92.31	75.51	94.87	79.59	100.00
	50	76.00	97.44	78.00	100.00	78.00	100.00	70.00	89.74	74.00	94.87	74.00	94.87	74.00	94.87	74.00	94.87	78.00	100.00
	52	75.00	100.00	75.00	100.00	75.00	100.00	69.23	92.31	73.08	97.44	73.08	97.44	71.15	94.87	73.08	97.44	75.00	100.00
	54	72.22	100.00	72.22	100.00	72.22	100.00	68.52	94.87	70.37	97.44	70.37	97.44	70.37	97.44	72.22	100.00	72.22	100.00
	56	69.64	100.00	69.64	100.00	69.64	100.00	67.86	97.44	69.64	100.00	69.64	100.00	69.64	100.00	69.64	100.00	69.64	100.00
	64	60.94	100.00	60.94	100.00	60.94	100.00	60.94	100.00	60.94	100.00	60.94	100.00	60.94	100.00	60.94	100.00	60.94	100.00
	Average	77.52	92.31	77.51	92.31	78.26	93.16	71.71	86.32	73.90	89.17	74.28	89.17	74.07	88.89	75.23	90.31	**78.49**	**93.45**
Yeast	3	0.00	0.00	100.00	60.00	100.00	60.00	0.00	0.00	66.67	40.00	0.00	0.00	33.33	20.00	0.00	0.00	100.00	60.00
	5	20.00	20.00	100.00	100.00	100.00	100.00	40.00	40.00	40.00	40.00	40.00	40.00	40.00	40.00	40.00	40.00	80.00	80.00
	6	16.67	20.00	83.33	100.00	83.33	100.00	50.00	60.00	50.00	60.00	33.33	40.00	33.33	40.00	50.00	60.00	83.33	100.00
	7	14.29	20.00	71.43	100.00	71.43	100.00	57.14	80.00	57.14	80.00	28.57	40.00	28.57	40.00	57.14	80.00	71.43	100.00
	8	12.50	20.00	62.50	100.00	62.50	100.00	62.50	100.00	50.00	80.00	25.00	40.00	25.00	40.00	50.00	80.00	62.50	100.00
	9	11.11	20.00	55.56	100.00	55.56	100.00	55.56	100.00	55.56	100.00	22.22	40.00	22.22	40.00	44.44	80.00	55.56	100.00
	10	20.00	40.00	50.00	100.00	50.00	100.00	50.00	100.00	50.00	100.00	30.00	60.00	20.00	40.00	50.00	100.00	50.00	100.00
	13	38.46	100.00	38.46	100.00	38.46	100.00	38.46	100.00	38.46	100.00	38.46	100.00	23.08	60.00	38.46	100.00	38.46	100.00
	23	21.74	100.00	21.74	100.00	21.74	100.00	21.74	100.00	21.74	100.00	21.74	100.00	21.74	100.00	21.74	100.00	21.74	100.00
	Average	17.20	37.78	64.78	95.56	**64.78**	**95.56**	41.71	75.56	47.73	77.78	26.59	51.11	27.48	46.67	39.09	71.11	62.56	93.33

FindCBLOF, GrC, SEQ, IE, and ODGrCR algorithms, their $P(t)$ are 42.31%, 46.15%, 46.15%, 34.62%, 46.15%, 38.46%, 46.15%, and 38.46%, respectively. $P(t)$ of the FRGOD algorithm is larger than other algorithms. On the other hand, when $R(t)$ reaches 100.00% for the first time, t for the FRGOD algorithm is 39, while for DIS, kNN, LOF, FindCBLOF, GrC, SEQ, IE, and ODGrCR algorithms, their t are 91, 81, 83, 117, 106, 226, 108, and 108, respectively. The FRGOD algorithm's t is smaller than that of other algorithms. This shows that the FRGOD algorithm can detect all 26 outliers at $t = 39$. However, when detecting all 26 outliers, t of other algorithms must be greater than that of the FRGOD algorithm.

In addition, the FRGOD algorithm does not perform optimally in Iono, Page, and Yeast datasets. For example, for Iono dataset, when $t = 24$, $P(t)$ of the FRGOD algorithm is slightly lower than that of the kNN algorithm, but greater than the other algorithms. From the aspect of $R(t)$, when it first reaches at 100.00%, t of the FGROD algorithm is 26, which is greater than that of the kNN algorithm, but is less than that of the other algorithms. Besides, when $t = 258$, $P(t)$ of the FRGOD algorithm on Page dataset is 29.84%, and the FRGOD algorithm is the worst performer. But for $R(t)$, when it reaches 100.00% for the first time, t of the FRGOD algorithm is 781, which is smaller than that of other algorithms, indicating that its effect is optimal. Also, for the Yeast dataset, we can conclude that the FRGOD algorithm is only slightly worse than the LOF algorithm but better than other algorithms.

According to Fig. 3, it can be seen that the FRGOD algorithm is closer to the upper left corner of first quadrant on Diab, Iris, Pima, and Sonar datasets, and the area under the curve is the larger one. This shows that the FRGOD algorithm is obviously better than other algorithms. In Iono, Page, Wbc, Wdbc, and Yeast datasets, the FRGOD algorithm is close to the upper left corner of first quadrant, and its performance is equal or slightly weak when comparing with other algorithms.

From the average of $R(t)$, the FRGOD algorithm is larger than other algorithms on Diab, Pima, Sonar, Wdbc, and Wbc datasets. For the Iris dataset, it is greater than or equal to other algorithms. The average $R(t)$ of the FRGOD algorithm is only slightly smaller than that of the kNN algorithm on Iono dataset, but larger than that of all other algorithms. In Page dataset, the average $R(t)$ of the FRGOD algorithm is only smaller than that of DIS, kNN, and LOF algorithms. In terms of Yeast dataset, the FRGOD algorithm is only worse than kNN and LOF algorithms.

What is more, FindCBLOF, GrC, SEQ, IE, and ODGrCR algorithms show relatively poor effects in most datasets with numerical attributes. This is because that the numerical attributes need to be discretized before the experiments, which will obviously affect the precision of their experimental results.

In short, the FRGOD algorithm is also effective for numeric attribute datasets.

Finally, we analyze the validity of the FRGOD algorithm in five mixed attribute datasets. Table VII and Fig. 4, respectively, give the comparative experimental results and ROC

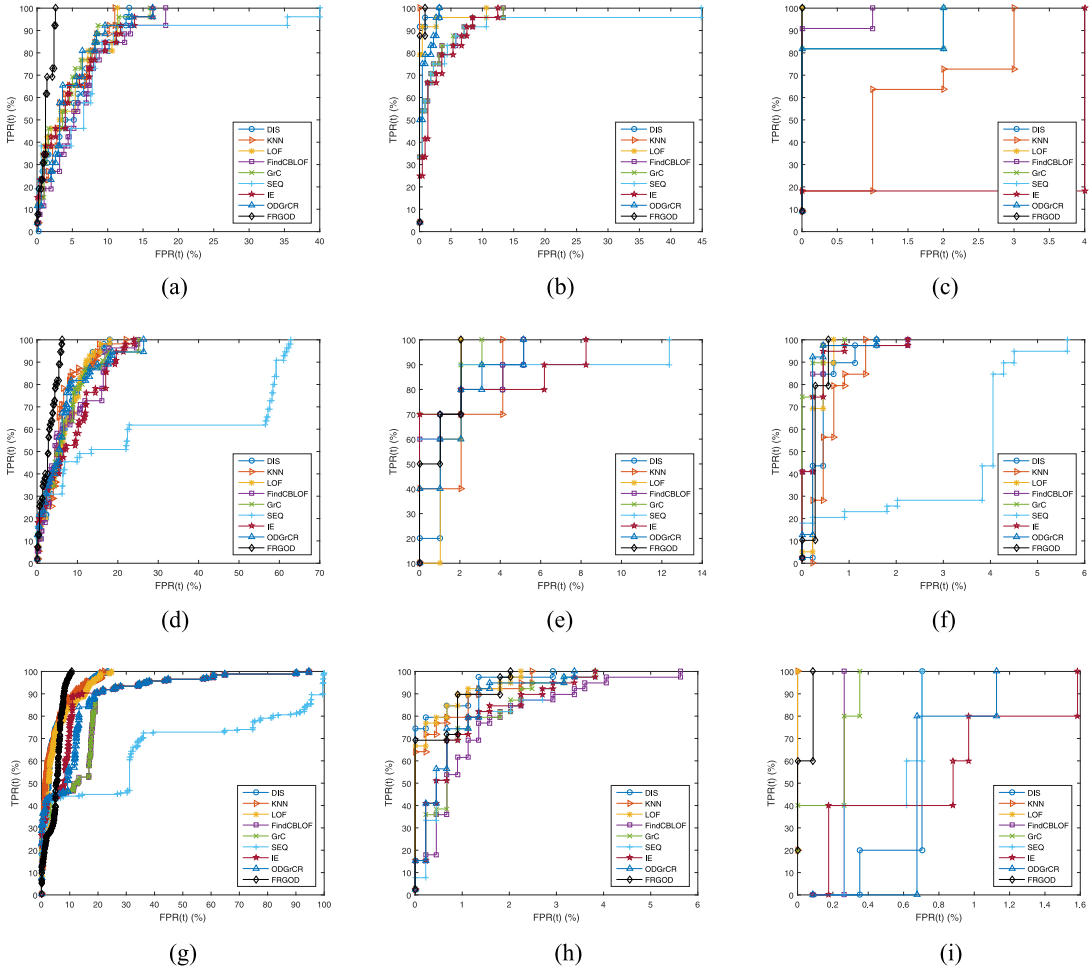


Fig. 3. ROC curves of numeric attribute datasets. (a) Diab. (b) Iono. (c) Iris. (d) Pima. (e) Sonar. (f) Wdbc. (g) Page. (h) Wbc. (i) Yeast.

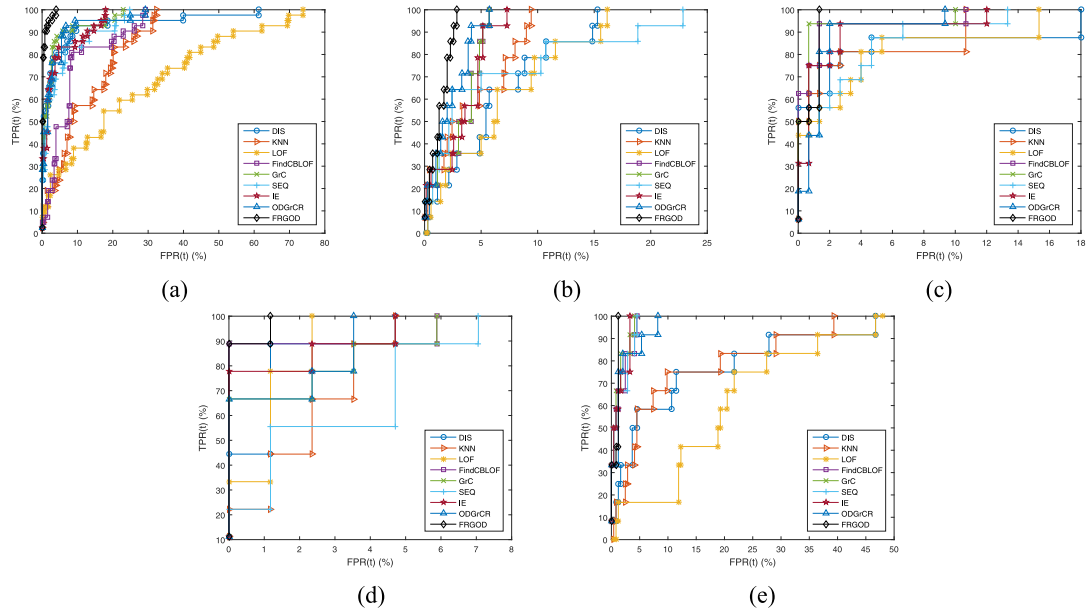


Fig. 4. ROC curves of mixed attribute datasets. (a) Cred. (b) German. (c) Heart. (d) Hepa. (e) Horse.

curves. In terms of the average of $R(t)$, the FRGOD algorithm is larger than or equal to that of all other algorithms in all mixed attribute datasets. The other same analysis can also be done. We can see that the FRGOD algorithm is significantly superior to other algorithms in mixed attribute datasets.

TABLE VII
COMPARISON OF EXPERIMENTAL RESULTS ON MIXED ATTRIBUTE DATASETS

Data set	t	DIS		kNN		LOF		FindCBLOF		GrC		SEQ		IE		ODGrCR		FRGOD	
		$P(t)$	$R(t)$	$P(t)$	$R(t)$	$P(t)$	$R(t)$	$P(t)$	$R(t)$	$P(t)$	$R(t)$	$P(t)$	$R(t)$	$P(t)$	$R(t)$	$P(t)$	$R(t)$	$P(t)$	$R(t)$
Cred	21	90.48	45.24	38.10	19.05	52.38	26.19	47.62	23.81	85.71	42.86	80.95	40.48	76.19	38.10	90.48	45.24	100.00	50.00
	42	73.81	73.81	35.71	35.71	30.95	30.95	47.62	47.62	76.19	76.19	66.67	66.67	71.43	71.43	73.81	73.81	90.48	90.48
	57	59.65	80.95	40.35	54.76	28.07	38.10	47.37	64.29	64.91	88.10	56.14	76.19	61.40	83.33	63.16	85.71	73.68	100.00
	111	36.04	95.24	30.63	80.95	22.52	59.52	32.43	85.71	36.04	95.24	34.23	90.48	37.84	100.00	36.04	95.24	37.84	100.00
	130	30.77	95.24	27.69	85.71	20.00	61.90	29.23	90.48	32.31	100.00	31.54	97.62	32.31	100.00	30.77	95.24	32.31	100.00
	137	29.20	95.24	27.01	88.10	18.98	61.90	27.74	90.48	30.66	100.00	30.66	100.00	30.66	100.00	29.93	97.62	30.66	100.00
	154	25.97	95.24	24.68	90.48	18.83	69.05	27.27	100.00	27.27	100.00	27.27	100.00	27.27	100.00	27.27	100.00	27.27	100.00
	166	24.10	95.24	25.30	100.00	18.07	71.43	25.30	100.00	25.30	100.00	25.30	100.00	25.30	100.00	25.30	100.00	25.30	100.00
	277	15.16	100.00	15.16	100.00	14.08	92.86	15.16	100.00	15.16	100.00	15.16	100.00	15.16	100.00	15.16	100.00	15.16	100.00
	325	12.92	100.00	12.92	100.00	12.92	100.00	12.92	100.00	12.92	100.00	12.92	100.00	12.92	100.00	12.92	100.00	12.92	100.00
	Average	39.81	87.62	27.76	75.48	23.68	61.19	31.27	80.24	40.65	90.24	38.08	87.14	39.05	89.29	40.48	89.29	**44.56**	**94.05**
Germ	7	28.57	14.29	57.14	28.57	28.57	14.29	42.86	21.43	42.86	21.43	28.57	14.29	42.86	21.43	42.86	21.43	57.14	28.57
	14	21.43	21.43	28.57	28.57	21.43	21.43	35.71	35.71	35.71	35.71	35.71	35.71	28.57	28.57	35.71	35.71	42.86	42.86
	34	14.71	35.71	23.53	57.14	14.71	35.71	20.59	50.00	20.59	50.00	26.47	64.29	23.53	57.14	29.41	71.43	41.18	100.00
	54	16.67	64.29	16.67	64.29	16.67	64.29	25.93	100.00	25.93	100.00	18.52	71.43	24.07	92.86	25.93	100.00	25.93	100.00
	65	13.85	64.29	16.92	78.57	13.85	64.29	21.54	100.00	21.54	100.00	15.38	71.43	21.54	100.00	21.54	100.00	21.54	100.00
	80	13.75	78.57	17.50	100.00	13.75	78.57	17.50	100.00	17.50	100.00	12.50	71.43	17.50	100.00	17.50	100.00	17.50	100.00
	121	11.57	100.00	11.57	100.00	9.92	85.71	11.57	100.00	11.57	100.00	9.92	85.71	11.57	100.00	11.57	100.00	11.57	100.00
	127	11.02	100.00	11.02	100.00	11.02	100.00	11.02	100.00	11.02	100.00	9.45	85.71	11.02	100.00	11.02	100.00	11.02	100.00
	174	8.05	100.00	8.05	100.00	8.05	100.00	8.05	100.00	8.05	100.00	8.05	100.00	8.05	100.00	8.05	100.00	8.05	100.00
	Average	15.51	64.29	21.22	73.02	15.33	62.70	21.64	78.57	21.64	78.57	18.29	66.67	20.97	77.78	22.62	80.95	**26.31**	**85.71**
Heart	8	100.00	50.00	100.00	50.00	87.50	43.75	100.00	50.00	100.00	50.00	87.50	43.75	87.50	43.75	87.50	43.75	100.00	50.00
	16	75.00	75.00	75.00	75.00	68.75	68.75	87.50	87.50	93.75	93.75	68.75	68.75	81.25	81.25	81.25	81.25	87.50	87.50
	18	72.22	81.25	72.22	81.25	66.67	75.00	83.33	93.75	83.33	93.75	66.67	75.00	77.78	87.50	83.33	93.75	88.89	100.00
	30	46.67	87.50	46.67	87.50	46.67	87.50	50.00	93.75	50.00	93.75	50.00	93.75	50.00	93.75	53.33	100.00	53.33	100.00
	32	43.75	87.50	50.00	100.00	43.75	87.50	50.00	100.00	50.00	100.00	46.88	93.75	46.88	93.75	50.00	100.00	50.00	100.00
	34	41.18	87.50	47.06	100.00	41.18	87.50	47.06	100.00	47.06	100.00	44.12	93.75	47.06	100.00	47.06	100.00	47.06	100.00
	36	38.89	87.50	44.44	100.00	38.89	87.50	44.44	100.00	44.44	100.00	44.44	100.00	44.44	100.00	44.44	100.00	44.44	100.00
	39	35.90	87.50	41.03	100.00	41.03	100.00	41.03	100.00	41.03	100.00	41.03	100.00	41.03	100.00	41.03	100.00	41.03	100.00
	Average	56.70	80.47	59.55	86.72	54.30	79.69	62.92	90.63	63.70	91.41	56.17	83.59	59.49	87.50	60.99	89.84	**64.03**	**92.19**
Hepa	5	80.00	44.44	80.00	44.44	80.00	44.44	100.00	55.56	100.00	55.56	80.00	44.44	100.00	55.56	100.00	55.56	100.00	55.56
	9	88.89	88.89	66.67	66.67	77.78	77.78	88.89	88.89	77.78	77.78	55.56	55.56	77.78	77.78	77.78	77.78	88.89	88.89
	10	80.00	88.89	70.00	77.78	80.00	88.89	80.00	88.89	70.00	77.78	60.00	66.67	80.00	88.89	70.00	77.78	90.00	100.00
	11	72.73	88.89	72.73	88.89	81.82	100.00	72.73	88.89	72.73	88.89	63.64	77.78	72.73	88.89	72.73	88.89	81.82	100.00
	12	66.67	88.89	66.67	88.89	75.00	100.00	66.67	88.89	66.67	88.89	66.67	88.89	66.67	88.89	75.00	100.00	75.00	100.00
	13	69.23	100.00	69.23	100.00	69.23	100.00	61.54	88.89	61.54	88.89	61.54	88.89	69.23	100.00	69.23	100.00	69.23	100.00
	14	64.29	100.00	64.29	100.00	64.29	100.00	64.29	100.00	64.29	100.00	57.14	88.89	64.29	100.00	64.29	100.00	64.29	100.00
	15	60.00	100.00	60.00	100.00	60.00	100.00	60.00	100.00	60.00	100.00	60.00	100.00	60.00	100.00	60.00	100.00	60.00	100.00
	Average	72.72	87.50	68.70	83.33	73.51	88.89	74.26	87.50	71.62	84.72	63.07	76.39	73.84	87.50	73.63	87.50	**78.65**	**93.06**
Horse	6	50.00	25.00	33.33	16.67	33.33	16.67	83.33	41.67	83.33	41.67	83.33	41.67	83.33	41.67	66.67	33.33	66.67	33.33
	12	33.33	33.33	33.33	33.33	16.67	16.67	66.67	66.67	66.67	66.67	66.67	66.67	66.67	66.67	75.00	75.00	75.00	75.00
	15	40.00	50.00	33.33	41.67	13.33	16.67	60.00	75.00	66.67	83.33	53.33	66.67	60.00	75.00	66.67	83.33	80.00	100.00
	20	35.00	58.33	35.00	58.33	10.00	16.67	50.00	83.33	55.00	91.67	50.00	83.33	60.00	100.00	50.00	83.33	60.00	100.00
	22	31.82	58.33	31.82	58.33	9.09	16.67	50.00	91.67	54.55	100.00	50.00	91.67	54.55	100.00	45.45	83.33	54.55	100.00
	23	30.43	58.33	30.43	58.33	8.70	16.67	52.17	100.00	52.17	100.00	52.17	100.00	52.17	100.00	43.48	83.33	52.17	100.00
	32	21.88	58.33	25.00	66.67	9.38	25.00	37.50	100.00	37.50	100.00	37.50	100.00	37.50	100.00	37.50	100.00	37.50	100.00
	108	10.19	91.67	11.11	100.00	10.19	91.67	11.11	100.00	11.11	100.00	11.11	100.00	11.11	100.00	11.11	100.00	11.11	100.00
	126	9.52	100.00	9.52	100.00	9.52	100.00	9.52	100.00	9.52	100.00	9.52	100.00	9.52	100.00	9.52	100.00	9.52	100.00
	Average	29.13	59.26	26.99	59.26	13.36	35.19	46.70	84.26	48.50	87.04	45.96	83.33	48.32	87.04	45.04	82.41	**49.61**	**89.81**

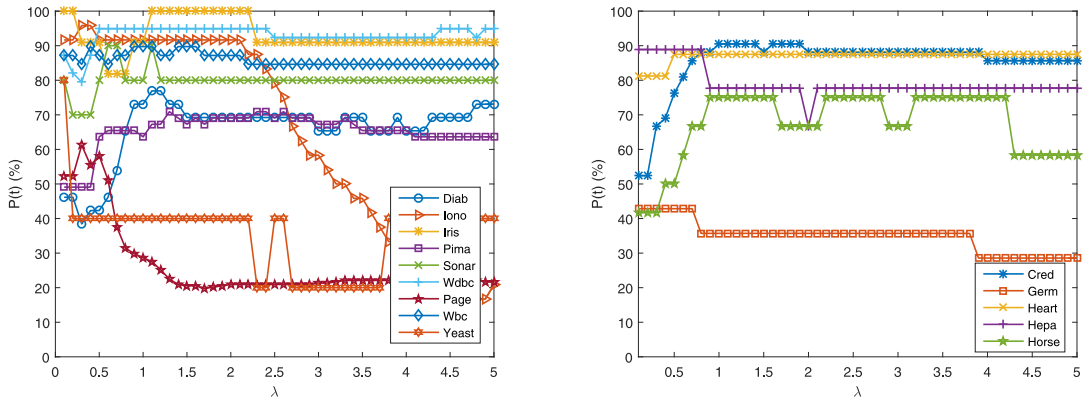


Fig. 5. Change of $P(t)$ with the variation of λ. (a) Numeric attribute datasets. (b) Mixed attribute datasets.

The results show that the FRGOD algorithm is also applicable to mixed attribute datasets.

D. Experimental Parameter Analyses

Threshold λ plays an important role for the FRGOD algorithm. It can be used as a parameter to control the granularity of data analysis. When t takes the number of true outliers, the change curve of $P(t)$ with λ is drawn in Fig. 5. Since $P(t)$ is equal to $R(t)$ when t takes the number of true outliers, $R(t)$ can obtain the same change curve.

For the numeric attribute datasets, their change curves of $P(t)$ with λ are depicted in Fig. 5(a). From Fig. 5(a), we can

see that with the increase of λ on Diab, Iris, Pima, Sonar, Wdbc, and Wbc datasets, the values of $P(t)$ increase first and then tend to balance. However, for Iono and Page datasets, $P(t)$ decreases as λ increases. On Yeast dataset, $P(t)$ increases with λ, first reaches the maximum at $\lambda = 0.1$, then suddenly decreases to a steady state, then starts to fluctuate and finally stabilizes.

For the mixed attribute datasets, their change curves of $P(t)$ with λ are drawn in Fig. 5(b). Through Fig. 5(b), we can see that with the change of λ on Cred, Heart, and Horse datasets, the values of $P(t)$ are also increased first and then balanced. However, on Germ and Hepa datasets, with the increase of λ, the values of $P(t)$ first maintain the maximum value and, finally, tend to balance. At the same time, we can also see that for different datasets, the optimal value can be obtained under multiple λ.

V. Conclusion

In many areas, outlier detection becomes increasingly important. Aiming at the problem that the outlier detection method based on classical RST cannot effectively deal with the numerical and mixed attribute datasets, an outlier detection method based on fuzzy rough granule by employing the concept of fuzzy approximation accuracy is proposed in this article, which is applicable not only to nominal attribute data but also to numerical attribute and mixed attribute data. It can also solve the problem that existing outlier detection methods cannot effectively deal with uncertainty data, and it does not need to discretize the data, which can reduce the data processing time and information loss. Aiming at the proposed method, we design a specific FRGOD algorithm, conduct experiments on UCI datasets, and compare with some existing outlier detection algorithms to verify the effectiveness of the FRGOD algorithm. The experimental results show that the proposed outlier detection method is effective. Presently, the research on outlier detection using FRS method has not been reported yet. The work of this article extends the application scope of FRS in data mining and other fields, and opens up a new application space for fuzzy RST. However, the time complexity and space complexity are relatively high due to the strategy adopted in the FRGOD algorithm. Therefore, the time and space complexity of the FRGOD algorithm need to be further optimized.

In future work, we will consider dividing a concept into three parts through three-way decision models [49], and then perform outlier detection in these three parts. What is more, the proposed model can be extended for multimodality data [50]. In addition, the conditional outlier detection is also a future research direction.

References

[1] C. C. Aggarwal, *Outlier Analysis*. Cham, Switzerland: Springer, 2016.

[2] D. M. Hawkins, *Identification of Outliers*. Dordrecht, The Netherlands: Springer, 1980.

[3] A. De Paola, S. Gaglio, G. L. Re, F. Milazzo, and M. Ortolani, "Adaptive distributed outlier detection for WSNs," *IEEE Trans. Cybern.*, vol. 45, no. 5, pp. 902–913, May 2015.

[4] F. Rasheed and R. Alhajj, "A framework for periodic outlier pattern detection in time-series sequences," *IEEE Trans. Cybern.*, vol. 44, no. 5, pp. 569–582, May 2014.

[5] B. Wang and Z. Mao, "Outlier detection based on a dynamic ensemble model: Applied to process monitoring," *Inf. Fusion*, vol. 51, pp. 244–258, Nov. 2019.

[6] P. J. Rousseeuw and A. M. Leroy, *Robust Regression and Outlier Detection*, vol. 589. Hoboken, NJ, USA: Wiley, 2005.

[7] Z. Y. He, X. F. Xu, and S. C. Deng, "Discovering cluster-based local outliers," *Pattern Recognit. Lett.*, vol. 24, nos. 9–10, pp. 1641–1650, 2003.

[8] T. Johnson, I. Kwok, and R. T. Ng, "Fast computation of 2-dimensional depth contours," in *Proc. Int. Conf. Knowl. Discov. Data Min. (KDD)*, 1999, pp. 224–228.

[9] E. M. Knorr and R. T. Ng, "Algorithms for mining distance-based outliers in large datasets," in *Proc. Int. Conf. Very Large Data Bases*, 1998, pp. 392–403.

[10] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying density-based local outliers," *ACM SIGMOD Rec.*, vol. 29, no. 2, pp. 93–104, 2000.

[11] Y. Y. Yao, "Granular computing: Past, present, and future," in *Rough Sets and Knowledge Technology*. Heidelberg, Germany: Springer, 2008, pp. 27–28.

[12] J. T. Yao, A. V. Vasilakos, and W. Pedrycz, "Granular computing: Perspectives and challenges," *IEEE Trans. Cybern.*, vol. 43, no. 6, pp. 1977–1989, Dec. 2013.

[13] H. M. Chen, T. R. Li, R. Da, J. H. Lin, and C. X. Hu, "A rough-set-based incremental approach for updating approximations under dynamic maintenance environments," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 2, pp. 274–284, Feb. 2013.

[14] J. H. Dai, W. T. Wang, and Q. Xu, "An uncertainty measure for incomplete decision tables and its applications," *IEEE Trans. Cybern.*, vol. 43, no. 4, pp. 1277–1289, Aug. 2013.

[15] Q. H. Zhang, S. H. Yang, and G. Y. Wang, "Measuring uncertainty of probabilistic rough set model from its three regions," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 47, no. 12, pp. 3299–3309, Dec. 2017.

[16] J. H. Dai, Q. H. Hu, J. H. Zhang, H. Hu, and N. G. Zheng, "Attribute selection for partially labeled categorical data by rough set approach," *IEEE Trans. Cybern.*, vol. 47, no. 9, pp. 2460–2471, Sep. 2017.

[17] C. Z. Wang, Y. Huang, M. W. Shao, Q. H. Hu, and D. G. Chen, "Feature selection based on neighborhood self-information," *IEEE Trans. Cybern.*, vol. 50, no. 9, pp. 4031–4042, Sep. 2020.

[18] T. T. Nguyen, "Outlier detection: An approximate reasoning approach," in *Proc. Int. Conf. Rough Sets Intell. Syst. Paradigms*, 2007, pp. 495–504.

[19] Y. M. Chen, D. Q. Miao, and R. Z. Wang, "Outlier detection based on granular computing," in *Proc. Int. Conf. Rough Sets Current Trends Comput.*, 2008, pp. 283–292.

[20] Z. X. Xue and S. Y. Liu, "Rough-based semi-supervised outlier detection," in *Proc. 6th Int. Conf. Fuzzy Syst. Knowl. Discov.*, vol. 1. Tianjin, China, 2009, pp. 520–523.

[21] A. Albanese, S. K. Pal, and A. Petrosino, "Rough sets, kernel set, and spatiotemporal outlier detection," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 1, pp. 194–207, Jan. 2014.

[22] F. Jiang, Y. F. Sui, and C. G. Cao, "Some issues about outlier detection in rough set theory," *Expert Syst. Appl.*, vol. 36, no. 3, pp. 4680–4687, 2009.

[23] F. Jiang, Y. F. Sui, and C. G. Cao, "An information entropy-based approach to outlier detection in rough sets," *Expert Syst. Appl.*, vol. 37, no. 9, pp. 6338–6344, 2010.

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

[25] F. Maciá-Pérez, J. V. Berna-Martinez, A. F. Oliva, and M. A. A. Ortega, "Algorithm for the detection of outliers based on the theory of rough sets," *Decis. Support Syst.*, vol. 75, pp. 63–75, Jul. 2015.

[26] F. Jiang, H. B. Zhao, J. W. Du, Y. Xue, and Y. J. Peng, "Outlier detection based on approximation accuracy entropy," *Int. J. Mach. Learn. Cybern.*, vol. 10, no. 9, pp. 2483–2499, 2019.

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

[28] Z. Yuan, X. Y. Zhang, and S. Feng, "Sequence-based mixed attribute outlier detection in neighborhood rough sets," *J. Chin. Comput. Syst.*, vol. 39, no. 6, pp. 1317–1322, 2018.

[29] Z. Yuan and S. Feng, "Outlier detection algorithm based on neighborhood value difference metric," *J. Comput. Appl.*, vol. 38, no. 7, pp. 81–85, 2018.

[30] Z. Yuan, X. Y. Zhang, and S. Feng, "Hybrid data-driven outlier detection based on neighborhood information entropy and its developmental measures," *Expert Syst. Appl.*, vol. 112, pp. 243–257, Dec. 2018.
[31] D. Dubois and H. Prade, "Rough fuzzy sets and fuzzy rough sets," *Int. J. Gen. Syst.*, vol. 17, nos. 2–3, pp. 191–209, 1990.
[32] W.-Z. Wu, J.-S. Mi, and W.-X. Zhang, "Generalized fuzzy rough sets," *Inf. Sci.*, vol. 151, pp. 263–282, May 2003.
[33] J.-S. Mi and W.-X. Zhang, "An axiomatic characterization of a fuzzy generalization of rough sets," *Inf. Sci.*, vol. 160, nos. 1–4, pp. 235–249, 2004.
[34] D. S. Yeung, D. Chen, E. C. C. Tsang, J. W. T. Lee, and W. Xizhao, "On the generalization of fuzzy rough sets," *IEEE Trans. Fuzzy Syst.*, vol. 13, no. 3, pp. 343–361, Jun. 2005.
[35] P. Maji and P. Garai, "Fuzzy–rough simultaneous attribute selection and feature extraction algorithm," *IEEE Trans. Cybern.*, vol. 43, no. 4, pp. 1166–1177, Aug. 2013.
[36] Q. H. Hu, S. An, X. Yu, and D. R. Yu, "Robust fuzzy rough classifiers," *Fuzzy Sets Syst.*, vol. 183, no. 1, pp. 26–43, 2011.
[37] D. G. Chen and Y. Y. Yang, "Attribute reduction for heterogeneous data based on the combination of classical and fuzzy rough set models," *IEEE Trans. Fuzzy Syst.*, vol. 22, no. 5, pp. 1325–1334, Oct. 2014.
[38] S. Y. Zhao, H. Chen, C. P. Li, X. Y. Du, and H. Sun, "A novel approach to building a robust fuzzy rough classifier," *IEEE Trans. Fuzzy Syst.*, vol. 23, no. 4, pp. 769–786, Aug. 2015.
[39] C. Z. Wang *et al.*, "A fitting model for feature selection with fuzzy rough sets," *IEEE Trans. Fuzzy Syst.*, vol. 25, no. 4, pp. 741–753, Aug. 2017.
[40] S. An, Q. H. Hu, W. Pedrycz, P. F. Zhu, and E. C. C. Tsang, "Data-distribution-aware fuzzy rough set model and its application to robust classification," *IEEE Trans. Cybern.*, vol. 46, no. 12, pp. 3073–3085, Dec. 2016.
[41] J. S. Mi, Y. Leung, H.-Y. Zhao, and T. Feng, "Generalized fuzzy rough sets determined by a triangular norm," *Inf. Sci.*, vol. 178, no. 16, pp. 3203–3213, 2008.
[42] Q. H. Hu, D. R. Yu, W. Pedrycz, and D. G. Chen, "Kernelized fuzzy rough sets and their applications," *IEEE Trans. Knowl. Data Eng.*, vol. 23, no. 11, pp. 1649–1667, Nov. 2011.
[43] X. H. Zhang, D. W. Fei, and J. H. Dai, *Fuzzy Mathematics and Rough Set Theory*, Tsinghua Univ. Press, Beijing, 2013.
[44] C. Z. Wang, Y. Huang, M. W. Shao, and D. G. Chen, "Uncertainty measures for general fuzzy relations," *Fuzzy Sets Syst.*, vol. 360, pp. 82–96, Apr. 2019.
[45] Z. Pawlak, J. Grzymala-Busse, R. Slowinski, and W. Ziarko, "Rough sets," *Int. J. Comput. Inf. Sci.*, vol. 11, no. 5, pp. 341–356, 1982.
[46] D. Dheeru and E. Karra Taniskidou, "UCI machine learning repository," University of California, School of Information and Computer Sciences, Irvine, CA, U.S., 2017.
[47] S. Ramaswamy, R. Rastogi, and K. Shim, "Efficient algorithms for mining outliers from large data sets," *ACM SIGMOD Rec.*, vol. 29, no. 2, pp. 427–438, 2000.
[48] G. O. Campos *et al.*, "On the evaluation of unsupervised outlier detection: Measures, datasets, and an empirical study," *Data Min. Knowl. Discov.*, vol. 30, no. 4, pp. 891–927, 2016.
[49] Y. Y. Yao, "An outline of a theory of three-way decisions," in *Proc. Int. Conf. Rough Sets Current Trends Comput.*, 2012, pp. 1–17.
[50] Q. H. Hu, L. J. Zhang, Y. C. Zhou, and W. Pedrycz, "Large-scale multimodality attribute reduction with multi-kernel fuzzy rough sets," *IEEE Trans. Fuzzy Syst.*, vol. 26, no. 1, pp. 226–238, Feb. 2018.

Zhong Yuan received the B.S. degree in mathematics from the Sichuan University for Nationalities, Kangding, China, in 2015, and the M.S. degree in mathematics from Sichuan Normal University, Chengdu, China, in 2018. He is currently pursuing the Ph.D. degree with Southwest Jiaotong University, Chengdu.

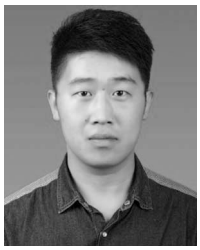
His research interests include rough sets, granular computing, and data mining.

Hongmei Chen (Member, IEEE) received the M.S. degree from the University of Electronic Science and Technology of China, Chengdu, China, in 2000, and the Ph.D. degree from Southwest Jiaotong University, Chengdu, in 2013.

She is currently a Professor with the School of Information Science Technology, Southwest Jiaotong University. Her research interests include the areas of data mining, pattern recognition, fuzzy sets, and rough sets.

Tianrui Li (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees from Southwest Jiaotong University, Chengdu, China, in 1992, 1995, and 2002, respectively.

He was a Postdoctoral Researcher with the Belgian Nuclear Research Centre (SCK.CEN), Auderghem, Belgium, from 2005 to 2006, a Visiting Professor with Hasselt University, Hasselt, Belgium, in 2008, and the University of Technology, Sydney, NSW, Australia, in 2009. He is currently a Professor and the Director of the Key Laboratory of Cloud Computing and Intelligent Technology, Southwest Jiaotong University. He has authored or coauthored more than 300 research papers in refereed journals and conferences. His research interests include big data, cloud computing, data mining, granular computing, and rough sets.



Binbin Sang received the B.S. degree in mathematics from Yangtze Normal University, Chongqing, China, in 2015, and the M.S. degree in mathematics from Chongqing University of Technology, Chongqing, in 2018. He is currently pursuing the Ph.D. degree with Southwest Jiaotong University, Chengdu, China.

His research interests include rough sets, granular computing, and data mining.

Shu Wang received the B.S. degree in computer science and technology, the M.S. degree in communication and information systems, and the Ph.D. degree in computer science and technology from Southwest Jiaotong University, Chengdu, China, in 2000, 2007, and 2020, respectively.

He is currently an Engineer with the School of Information Science Technology, Southwest Jiaotong University. His research interests include rough sets, granular computing, and data mining.