Available online at www.sciencedirect.com

ScienceDirect

Fuzzy Sets and Systems ••• (••••) •••–•••

FUZZY sets and systems

www.elsevier.com/locate/fss

Fuzzy information entropy-based adaptive approach for hybrid feature outlier detection

Zhong Yuan [a,b,c], Hongmei Chen [a,b,c,*], Tianrui Li [a,b,c], Jia Liu [a,b,c], Shu Wang [a,b,c]

[a] *School of Information Science and Technology, Southwest Jiaotong University, Chengdu 611756, China*
[b] *Institute of Artificial Intelligence, Southwest Jiaotong University, Chengdu 611756, China*
[c] *National Engineering Laboratory of Integrated Transportation Big Data Application Technology, Southwest Jiaotong University, Chengdu 611756, China*

Received 31 January 2020; received in revised form 24 August 2020; accepted 29 October 2020

Abstract

Fuzzy information entropy based on fuzzy relation in fuzzy rough set theory is an important metric of uncertainty. However, the research of fuzzy information entropy for hybrid feature outlier detection has not been reported. On this basis, this paper constructs a hybrid feature outlier detection method based on fuzzy information entropy by using fuzzy approximate space with fuzzy similarity relation. Firstly, the adaptive fuzzy radius of standard deviation and hybrid fuzzy similarity are employed to construct the fuzzy approximate space, and the relative fuzzy entropy is defined based on the fuzzy information entropy. Then, two kinds of metrics are constructed to describe the outlier degree of object. Finally, the fuzzy entropy-based outlier factor is integrated to implement outlier detection, and the relevant fuzzy information entropy-based outlier detection algorithm (FIEOD) is designed. The FIEOD algorithm is compared with the main outlier detection algorithms on public data. The experimental results reveal that the proposed method has better effectiveness and adaptability.

Keywords: Outlier detection; Fuzzy rough set theory; Fuzzy approximation space; Fuzzy information entropy; Hybrid (or Mixed) feature

1. Introduction

Outlier detection is an important branch of data mining. Its purpose is to find objects whose behavior is very different from normal objects. It plays an important role in many applications such as fraud detection, medical treatment, and intrusion detection [1]. Besides, outlier detection has been applied to the localization for wireless sensor networks,

* Corresponding author at: School of Information Science and Technology, Southwest Jiaotong University, Chengdu 611756, China.
E-mail addresses: yuanzhong2799@foxmail.com (Z. Yuan), hmchen@swjtu.edu.cn (H. Chen), trli@swjtu.edu.cn (T. Li), xiaoke92@foxmail.com (J. Liu), swang@swjtu.edu.cn (S. Wang).

https://doi.org/10.1016/j.fss.2020.10.017
0165-0114/

traffic outlier detection, and process monitoring [2–4]. Recently, more and more researchers have begun to pay attention to outlier detection and put forward many outlier detection methods. According to the assumptions about normal objects and outliers of various detection methods, outlier detection methods can be generally divided into three types of methods: (1) statistical methods [5], (2) proximity-based methods [6–8], and (3) clustering-based methods [9]. For proximity-based methods, there are three mainly types: depth-based methods [6], distance-based methods [7,10], and density-based methods [8]. To the best of our knowledge, most methods based on distance and density are designed by using Euclidean distance. Therefore, in applications, they may not be the proper way to detect outliers of nominal (categorical or symbolic) or hybrid (mixed or heterogeneous) feature data.

Pawlak rough set theory has become a popular mathematical framework, which has been applied to many fields such as feature selection, pattern recognition, and data mining, etc [11–15]. In order to make up for the shortcomings of distance-based and density-based methods, outlier detection methods based on rough set theory were also proposed [16–29]. Although these methods have proved the effectiveness of rough set theory in outlier detection. However, these methods build mathematical models based on equivalence relations, and their detection models are only applicable to nominal feature data. Numerical (numeric) feature data need to be discretized before using these detection models. It increases the time required for data processing and is accompanied by significant loss of information.

In order to solve the problem that rough set theory is only applicable to nominal feature data, Dubois and Prade proposed fuzzy rough set [30,31]. Fuzzy relation is used in fuzzy rough set to describe the similarity between objects, so it can directly process numerical or continuous feature data without discretization. Therefore, fuzzy rough set retains more information of numerical or continuous feature values. Information entropy proposed by Shannon is an effective mechanism to measure uncertainty [32], which has been introduced into fuzzy rough set to measure and express the uncertainty [33–36]. Yager studied the concept of information entropy based on fuzzy similarity relation for the first time, and then discussed the uncertainty measurement of fuzzy information [33]. Hernandez et al. further proposed the joint entropy and conditional entropy based on fuzzy similarity relation [34]. Based on fuzzy partition, Mesiar introduced a generalized fuzzy information entropy model [37]. Bertoluzza further discussed the uncertainty of the fuzzy partition [38]. Mi et al. put forward uncertainty measurement based on partition by using fuzzy rough theory [39]. Hu et al. redefined joint entropy and conditional entropy and applied fuzzy information entropy to numerical feature selection [35]. Dai et al. proposed a feature selection strategy based on normative fuzzy information weights based on fuzzy conditional mutual information [40]. So far, fuzzy rough set has been successfully applied to feature selection, clustering analysis and pattern recognition of numerical or mixed features [41–46]. However, outlier detection based on fuzzy information entropy in fuzzy rough set has not been proposed.

From the above discussion, we propose a Fuzzy Information Entropy-based Outlier Detection (FIEOD) method for hybrid feature data in this paper. First, the adaptive fuzzy radius of standard deviation and hybrid fuzzy similarity are utilized to construct the fuzzy approximate space, and the relative fuzzy entropy is defined according to the fuzzy information entropy. Then, two kinds of metrics (relative fuzzy cardinality and fuzzy relation outlier degree) are constructed to describe the outlier degree of objects. Finally, the fuzzy entropy-based outlier factor is defined for detecting outliers, and the corresponding FIEOD algorithm is designed. The proposed method extends the traditional distance-based and the rough set-based detection method, and thus is applicable to both nominal, numerical, and hybrid feature data. In the relevant experiments, the FIEOD algorithm is compared with the main outlier detection methods (including DIS, kNN, LOF, FindCBLOF, GrC, SEQ, IE, and ODGrCR algorithms). The experimental results show that the FIEOD algorithm has better effectiveness and adaptability.

The rest of the paper is organized as follows. The second section briefly introduces the classical outlier detection methods and rough set-based outlier detection methods. The next section gives a review of some preliminary knowledge about fuzzy approximation space and fuzzy information entropy. The fourth section proposes a fuzzy information entropy-based outlier detection method for hybrid feature data. The fifth section presents the experimental results and analyses. The conclusion is given in the last section.

2. Related work

This section mainly reviews some classical outlier detection methods and rough sets-based outlier detection methods.

2.1. Classical outlier detection methods

According to the assumptions made by the outlier detection methods on the outliers and the remaining data objects, the outlier detection methods can be roughly categorized as: statistical methods, proximity-based methods, and cluster-based methods.

The statistical method assumes that the normal objects in the data set are generated by a probability distribution model, and objects appearing in the low probability region of the model are regarded as outliers [5]. It has many advantages. One advantage is that it is statistically very efficient when the statistical assumptions made on the data satisfy the actual constraints. Statistical methods are mostly for individual features, which makes them unsuitable for multidimensional data. In addition, the statistical method needs to presuppose that the data conforms to a certain distribution, so it is not suitable for situations where the distribution is unknown.

The proximity-based approaches mainly include: depth-based methods [6], distance-based methods [7,10], and density-based methods [8]. The depth-based methods assign a depth value to each object, and map objects to the corresponding layer of two-dimensional space according to the assigned depth value, wherein objects on the shallow layer may be outliers [6]. To a certain extent, it compensates for the shortcomings of statistical methods. However, the depth-based method is more effective for data in 2D and 3D space, the detection efficiency of hybrid feature data with higher dimensionality is lower. To avoid such problems with statistical methods and depth-based methods, Knorr and Ng proposed a distance-based method [7], which used the distance between two objects as a measure of outlier degree. The object away from most of the rest is considered to be an outlier. Distance-based methods are widely used due to their ease of operation. However, the sparse problem in high-dimensional data is difficult to be solved. On account of its use of two global parameters, it is very sensitive to the choice of parameters. At the same time, it does not consider the change of local density. To overcome the problem of local density mining in the distance-based approaches, Breunig et al. first proposed a density-based method [8]. Its basic idea is that the density around the outliers is significantly different from the density around its 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 LOF method is currently the most common used method for mining local outliers. Although the density-based method solves the problem of mining local outliers, it is still sensitive to the choice of parameters.

The cluster-based methods divide the data into clusters. The data objects that do not belong to any cluster or the small clusters whose data volume is significantly smaller than other clusters are considered to be outliers [9]. Since the cluster-based methods use the cluster structure of the data to detect outliers, they are very robust unsupervised methods and are applicable to data of multiple feature types. A disadvantage of the cluster-based approaches is that their effectiveness is highly dependent on the clustering method used, while these methods may not be optimal for outlier detection. In addition, the computation consumption may be a bottleneck for large data sets.

In addition, most distance-based and density-based methods are designed by using Euclidean distance. However, it is unreasonable to use Euclidean distance to measure the similarity or dissimilarity of nominal features. Therefore, in practical applications, they may not be the optimal way to detect outliers of nominal or mixed feature data.

2.2. Rough sets based outlier detection methods

In recent years, to compensate for the shortcomings of distance-based and density-based methods, outlier detection algorithms based on rough sets have been proposed. For example, Jiang et al. proposed a boundary-based outlier detection method using rough set theory [16]. Nguyen introduced a method for detecting and evaluating outliers using multi-level approximate reasoning scheme [17]. Jiang et al. presented a method for detecting outliers by using rough membership functions [18]. Chen et al. put forward a Granular Computing (GrC)-based outlier detection method [19]. Shaari et al. studied a new method for detecting outliers using the non-reduction concept in rough set theory [20]. Xue and Liu constructed a semi-supervised outlier detection method on the basis of rough set [21]. [22] introduced a new definition of outlier detection based on rough sequence, and discussed the definition of distance-based outlier detection metric in rough set theory. Using the rough information entropy model, [47] constructed a novel definition of outliers based on information entropy. Yang et al. used the concept of rough set to construct an outlier reduction method based on outlier detection and analysis system [23]. In addition, [24] presented an outlier detection method based on boundary and distance. Albanese et al. used a new rough set approach to extend outlier detection to spatio-temporal data [25]. Starting from GrC and rough set theory, [26] built a new outlier detection method based on

Table 1
The comparison of outlier detection methods.

Method	Advantage	Disadvantage	Feature type
Statistic-based	Statistically very effective	Not suitable for multidimensional and unknown distribution data	Single case
Depth-based	Suitable for unknown distribution data, very effective for 2D or 3D data	Low detection efficiency for high-dimensional data	2D or 3D
Distance-based	Suitable for high-dimensional and unknown distribution data, easy to implement	The curse of dimensionality, sensitive to parameter selection, without considering local density changes	Numerical
Density-based	Suitable for mining local outliers	Sensitive to parameter selection	Numerical
Cluster-based	Robust unsupervised method	Highly dependent on the clustering method	Related to clustering algorithms
Classical rough set-based	Effectively deal with uncertain nominal data	Build a detection model through equivalence relations	Categorical
Neighborhood rough set-based	Inherit the advantages of rough set methods, suitable for multiple data types	Without considering fuzziness, complex neighborhood calculations	Numerical, categorical and Mixed

classical approximation accuracy. Maciá-Pérez et al. extended the mathematical framework of the basic theory of outlier detection based on rough set theory, and proposed an efficient method for detecting outliers in a large amount of information [27]. Jiang et al. further introduced a rough set outlier detection method based on partition entropy [28]. [29] constructed an outlier detection algorithm based on the classical approximation accuracy and entropy. However, in these methods, equivalence relation is used to produce disjoint objects. Their detection model is only applicable to nominal feature data.

To effectively process numerical or hybrid feature data, Chen et al. proposed a neighborhood-based outlier detection algorithm [48]. However, many parameters are involved in this detection model. Therefore, it is sensitive to parameter settings. To this end, [49] proposed an outlier detection algorithm based on the neighborhood value difference metric by using the standard deviation neighborhood radius and the preferred hybrid distance metric. Yuan et al. extended a hybrid outlier detection method based on neighborhood sequence [50]. Adopting the neighborhood information entropy as the basic framework, Yuan et al. proposed an outlier detection method for mixed features [51]. Aiming at some problems existing in the past methods, this paper proposes a hybrid feature outlier detection method based on fuzzy information entropy. Finally, the comparison of outlier detection methods are summarized in Table 1.

3. Preliminaries

Fuzzy information theory is a powerful tool for dealing with information uncertainty. This section will review some of the definitions and symbols for fuzzy information entropy used in subsequent sections of this study [35,52,53].

In fuzzy information theory, the data table is also called Fuzzy Information System (FIS), and its formal definition is as follows. A FIS is a two-tuple $FIS = \langle U, A \rangle$, where: U is a non-empty finite object set called a universe; A is a non-empty finite feature set. For $\forall a \in A$ and $x \in U$, $a(x)$ denotes the value of x under the feature a. When $A = C \cup D$ and $C \cap D = \emptyset$, the data table is called the fuzzy decision system (FDS), where C indicates the conditional features and D indicates the decision features.

Definition 1. Let $U = \{x_1, x_2, \ldots, x_n\}$, if $\widetilde{A}$ is a mapping from U to $[0, 1]$, i.e.,

$$\widetilde{A} : U \to [0, 1]. \tag{1}$$

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

Similar to classical sets, fuzzy sets also have operations such as inclusion, equality, and complement. Let $\widetilde{A}, \widetilde{B} \in \widetilde{F}(U)$, some operations are defined as follows.

1) $\widetilde{A} = \widetilde{B} \Leftrightarrow \widetilde{A}(x) = \widetilde{B}(x)$;
2) $\widetilde{A} \subseteq \widetilde{B} \Leftrightarrow \widetilde{A}(x) \leq \widetilde{B}(x)$;
3) $(\widetilde{A} \cup \widetilde{B})(x) = \max\{\widetilde{A}(x), \widetilde{B}(x)\} = \widetilde{A}(x) \vee \widetilde{B}(x)$;
4) $(\widetilde{A} \cap \widetilde{B})(x) = \min\{\widetilde{A}(x), \widetilde{B}(x)\} = \widetilde{A}(x) \wedge \widetilde{B}(x)$;
5) $\widetilde{A}^c(x) = 1 - \widetilde{A}(x)$.

Definition 2. Let $U = \{u_1, u_2, \cdots, u_m\}$, $V = \{v_1, v_2, \cdots, v_n\}$, the fuzzy relation $\widetilde{R}$ from U to V refers to a fuzzy set on $U \times V$, i.e.,

$$\widetilde{R} : U \times V \to [0, 1]. \tag{2}$$

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

A fuzzy relation $\widetilde{R}$ from U to V can be represented by a fuzzy matrix, i.e., $M_{\widetilde{R}} = (r_{ij})_{m\times n}$, where $r_{ij} = \widetilde{R}(u_i, v_j)$, each row vector $(r_{i1}, r_{ij}, \ldots, r_{in})$ denotes a fuzzy set. The all fuzzy relations from U to V, denoted as $\widetilde{F}(U \times V)$. Obviously, fuzzy relations are a special kind of fuzzy sets, so there are also operations such as union, intersection, and complement.

Definition 3. Let $\widetilde{R}$ be the fuzzy relation on U, i.e., $\widetilde{R} \subseteq \widetilde{F}(U \times U)$. For $\forall x, y, x \in U$, there may be three properties as follows.

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

If $\widetilde{R}$ satisfies (1) and (2), then $\widetilde{R}$ is called a fuzzy similarity relation on U; if $\widetilde{R}$ satisfies (1)-(3), then $\widetilde{R}$ is called a fuzzy equivalence relation on U.

Definition 4. Let $B \subseteq C = \{c_1, c_2, \ldots, c_m\}$, $\widetilde{R}_B$ be a fuzzy similarity relation with respect to B on U, then the fuzzy relation matrix $M_{\widetilde{R}_B}$ is expressed as follows.

$$M_{\widetilde{R}_B} = \begin{pmatrix} r_{11}^B & r_{12}^B & \cdots & r_{1n}^B \\ r_{21}^B & r_{22}^B & \cdots & r_{2n}^B \\ \vdots & \vdots & \ddots & \vdots \\ r_{n1}^B & r_{n2}^B & \cdots & r_{nn}^B \end{pmatrix}. \tag{3}$$

Definition 5. The generalized fuzzy partition $U/\widetilde{R}_B$ generated by $\widetilde{R}_B$ is defined as

$$U/\widetilde{R}_B = \{[x_1]_{\widetilde{R}_B}, [x_2]_{\widetilde{R}_B}, \cdots, [x_n]_{\widetilde{R}_B}\}, \tag{4}$$

where $[x_i]_{\widetilde{R}_B} = \frac{r_{i1}^B}{x_1} + \frac{r_{i2}^B}{x_2} + \cdots + \frac{r_{in}^B}{x_n} = (r_{i1}^B, r_{i2}^B, \ldots, r_{in}^B)$ is a fuzzy information granule containing x_i induced by $\widetilde{R}_B$. In the absence of confusion, we also use B instead of $\widetilde{R}_B$.

Let $B = \{c_{j_1}, c_{j_2}, \cdots, c_{j_h}\}$ $(1 \leq h \leq m)$. Obviously, $[x_i]_{\widetilde{R}_B}$ is a fuzzy set on $\widetilde{R}_B$. We have $[x_i]_{\widetilde{R}_B}(x_j) = \widetilde{R}_B(x_i, x_j) = r_{ij}^B$. If $\widetilde{R}_B(x_i, x_j) = 1$, this means that x_j definitely belongs to $[x_i]_{\widetilde{R}_B}$; if $\widetilde{R}_B(x_i, x_j) = 0$, then x_j definitely does not belong to $[x_i]_{\widetilde{R}_B}$. The membership degree $\widetilde{R}_B(x_i, x_j)$ can be determined by the following methods [54]: (1) Quantitative product method, (2) Angle cosine method, (3) Correlation coefficient method and (4) Conjunction (intersection) method. Among them, the conjunction method is used in most literature [36,53,55]. In this paper, the conjunction method is used, and its calculation is as follows.

$$\widetilde{R}_B(x_i, x_j) = \bigwedge_{l=1}^{h} \widetilde{R}_{c_{j_l}}(x_i, x_j). \tag{5}$$

Definition 6. The cardinality of the fuzzy set $[x_i]_{\widetilde{R}_B}$ is defined as

$$|[x_i]_B| = |[x_i]_{\widetilde{R}_B}| = \sum_{j=1}^{n} r_{ij}^B. \tag{6}$$

Obviously, we get $0 \leq \left|[x_i]_{\widetilde{R}_B}\right| \leq n$.

Information entropy is an abstract concept in mathematics that represents the probability of occurrence on a particular information, and can represent the uncertainty of information in a FIS. The more certain a FIS is, the lower its information entropy will be; conversely, the higher the information entropy will be. Information entropy is introduced into the fuzzy rough set for representing uncertainty measurement, resulting in different forms [33–36]. Hu et al. studied fuzzy information entropy and applied it to numerical feature selection [35].

Definition 7 *(Fuzzy entropy).* [35] The fuzzy entropy $FE(B)$ of $\widetilde{R}_B$ is defined as

$$FE(B) = FE(\widetilde{R}_B) = -\frac{1}{|U|} \sum_{i=1}^{n} \log_2 \frac{|[x_i]_{\widetilde{R}_B}|}{|U|}. \tag{7}$$

If $\widetilde{R}_B$ degenerates into a crisp equivalence relation, the above fuzzy information entropy is the same as the classical information entropy. It lays a theoretical foundation for the construction of the hybrid feature outlier detection method.

Proposition 1. *[35] Let $B_1 \subseteq B_2 \subseteq C$, we have*

$$FE(B_1) \leq FE(B_2). \tag{8}$$

Proposition 1 reflects that the fuzzy information entropy changes monotonously with the feature. As the feature increases, the fuzzy information entropy becomes larger and the uncertainty increases. Conversely, the fuzzy information entropy becomes smaller and the uncertainty decreases accordingly. It can be seen that the increase or decrease of features causes the fuzzy information entropy and uncertainty to change. Therefore, fuzzy information entropy can be used to construct outlier detection methods.

4. Outlier detection

Based on FIS and fuzzy information entropy, this section defines fuzzy relative entropy to establish an outlier detection method gradually for mixed feature data, mainly involving the detection method, a corresponding algorithm, and an example.

4.1. Method

This section discusses the detection method based on fuzzy information entropy, which mainly involves normalization preprocessing, fuzzy similarity selection, standard deviation setting of fuzzy similarity radius, construction of metric based on fuzzy information entropy and outlier detection.

In the data processing of FISs, the magnitude and dimension are usually different in various data. In order to obtain accurate data processing results, the original numerical feature data is normalized firstly. There are many commonly used normalization methods, such as min-max, fractional, and decimal-scale normalization methods. In this paper, the min-max normalization method is used for preprocessing so that all numerical feature ranges are in range of [0, 1]. For $\forall c_k \in C = \{c_1, c_2, \ldots, c_m\}$, the relevant formula is [51]

$$f(c_k(x_i)) = \frac{c_k(x_i) - \min_{c_k}}{\max_{c_k} - \min_{c_k}}, \tag{9}$$

where $\max_{c_k}$ and $\min_{c_k}$ are the maximum and minimum values of $c_k(x_i)$ in U, respectively.

In order to effectively deal with numerical, nominal or hybrid feature data, a hybrid fuzzy similarity is constructed as follows.

Definition 8 *(Hybrid fuzzy similarity).* The hybrid fuzzy similarity $r_{ij}^{c_k}$ between x_i and x_j relative to c_k is calculated as

$$r_{ij}^{c_k} = r_{ij}^{\{c_k\}} = \begin{cases} 1, & \text{if } c_k(x_i) = c_k(x_j) \text{ and } c_k \text{ is nominal;} \\ 0, & \text{if } c_k(x_i) \neq c_k(x_j) \text{ and } c_k \text{ is nominal;} \\ 1 - |c_k(x_i) - c_k(x_j)|, & \text{if } |c_k(x_i) - c_k(x_j)| \leq \varepsilon_{c_k} \text{ and } c_k \text{ is numerical;} \\ 0, & \text{if } |c_k(x_i) - c_k(x_j)| > \varepsilon_{c_k} \text{ and } c_k \text{ is numerical.} \end{cases} \tag{10}$$

It can be seen from the above definition that the fuzzy similarity $r_{ij}^{c_k}$ is calculated in different ways for nominal and numerical features. Therefore, this processing method can handle nominal, numerical and mixed features. Where ε_{c_k} is an adaptive fuzzy radius, which is computed as follows [51].

$$\varepsilon_{c_k} = \varepsilon_{\{c_k\}} = \frac{std(c_k)}{\delta}, \tag{11}$$

where $std(c_k)$ is the standard deviation of the feature values of c_k, and the parameter δ is used to adjust the adaptive fuzzy radius. Obviously, $r_{ij}^{c_k} = r_{ji}^{c_k}$ and $r_{ii}^{c_k} = 0$, $0 \leq r_{ij}^{c_k} \leq 1$, $M_{\widetilde{R}_{c_k}}$ is a fuzzy similarity relation matrix.

The standard deviation $std(c_k)$ is used to measure the degree of dispersion of a given data. A large $std(c_k)$ means that differences among most of the feature values and their averages are large; while $std(c_k)$ is small, indicating that the feature values are close to their averages. If $\delta < 1$, then the fuzzy radius is more than $std(c_k)$; If $\delta = 1$, then it is equal to $std(c_k)$; If $\delta > 1$, then the fuzzy radius should be less than $std(c_k)$. In summary, $std(c_k)$ is an important factor for adjusting fuzzy radius, so more scientific statistics and objectivity are added. It lays a foundation for the follow-up adaptive outlier detection methods.

Through the above normalization, hybrid fuzzy similarity, and adaptive fuzzy radius, the corresponding FIS is constructed. Next, a fuzzy entropy-based outlier factor is constructed for outlier detection.

Fuzzy information entropy provides an overall uncertainty characterization of fuzzy knowledge. A new concept is constructed below: relative fuzzy entropy, which measures the uncertainty of each object and lays a foundation for outlier detection in FISs.

Definition 9 *(Relative fuzzy entropy).* Let $\forall x \in U$ and $\{U - \{x\}\}/\widetilde{R}_B = \{[x_{i_1}]_B, [x_{i_2}]_B, \cdots, [x_{i_{n-1}}]_B\}$, then relative fuzzy entropy $RFE_B(x)$ of x on $\widetilde{R}_B$ is defined as

$$RFE_B(x) = \begin{cases} 1 - \frac{FE_x(B)}{FE(B)}, & FE_x(B) < FE(B); \\ 0, & \text{otherwise,} \end{cases} \tag{12}$$

where $FE_x(B) = -\sum_{r=1}^{n-1} \frac{1}{|U-\{x\}|} \log_2 \frac{|[x_{i_r}]_B|}{|U-\{x\}|}$ denotes the fuzzy entropy after removing x. Obviously, we have $0 \leq RFE_B(x) \leq 1$.

In the above definition, if the fuzzy entropy $FE_x(B)$ decreases greatly after removing x, then the uncertainty of x on $\widetilde{R}_B$ may be considered high; if the fuzzy entropy $FE_x(B)$ changes little or even increases after removing x, then the uncertainty of x on $\widetilde{R}_B$ may be considered low. Specifically speaking, the higher $RFE_B(x)$ of x, the higher uncertainty of x and the more likely x is an outlier. Therefore, based on $FE_x(B)$ and its semantics, the $RFE_B(x)$ can be used to indicate the abnormality of an object.

To show whether x belongs to majority classes, the relative fuzzy cardinality is given below.

Definition 10 *(Relative fuzzy cardinality).* For $\forall x \in U$, the relative fuzzy cardinality $RFC_B(x)$ of x on $\widetilde{R}_B$ is defined by:

$$RFC_B(x) = |[x]_B| - \frac{\sum_{r=1}^{n-1} |[x_{i_r}]_B|}{n-1}, \tag{13}$$

From the above definition, it is easy to prove that $RFC_B(x)$ can be positive, zero or negative. If $RFC_B(x) \leq 0$, then x is considered to belong to majority classes; If $RFC_B(x) < 0$, then x is considered to belong to minority classes.

Next, we construct two kinds of sequences: feature sequences and feature subset sequences. By calculating the fuzzy entropy of each feature in C, all the features in C can be sorted according to the fuzzy information entropy to obtain a feature sequence.

Definition 11 *(Feature sequence).* Feature sequence FS is constructed as

$$FS = \langle c'_1, c'_2, \cdots, c'_m \rangle, \tag{14}$$

where $FE(c'_k) \geq FE(c'_{k+1})$, $1 \leq k < m$.

Further starting with the feature set c'_1, each time the feature with the largest fuzzy information entropy is added until a set containing the entire feature is obtained. Through this way, we can determine a feature subset sequence.

Definition 12 *(Feature subset sequence).* Feature subset sequence FSS is constructed as

$$FSS = \langle C_1, C_2, \cdots, C_m \rangle, \tag{15}$$

where $C_k \subseteq C$, $C_1 = \{c'_1\}$, $C_m = C$ and $C_{k+1} = C_k \bigcup \{c'_{k+1}\}$, $1 \leq k < m$.

In outlier detection, it is often necessary to construct an outlier factor to measure the outlier degree of an object. Jiang et al. provides a classical entropy outlier factor [47], but it only applies to nominal feature data rather than numerical feature data because it only considers equivalence relations and equivalence classes. Here, the classical entropy outlier factor is mainly extended to the fuzzy information system. Specifically, the relative fuzzy entropy and fuzzy cardinality are used to construct the fuzzy information entropy-based outlier factor. To this end, the outlier degree is firstly established to characterize the outlier degree of each object.

Definition 13 *(Outlier degree).* Let $\forall x \in U$, the fuzzy relation-based outlier degree $FOD_B(x)$ of x is defined as

$$FOD_B(x) = \begin{cases} RFE_B(x) \times \left(\frac{|U| - RFC_B(x)}{2|U|}\right), & RFC_B(x) > 0; \\ RFE_B(x) \times \sqrt{\frac{|U| + abs(RFC_B(x))}{2|U|}}, & RFC_B(x) \leq 0, \end{cases} \tag{16}$$

where $abs(\cdot)$ means the absolute value of $\cdot$.

The above definition expresses the idea that outliers always pay attention to the minority classes of objects in U, and objects belonging to the minority classes are more likely to be outliers. If $RFC_B(x) \leq 0$, that is, x belongs to the minority class, then x is more likely to become an outlier. Therefore, we use $\sqrt{\frac{|U| + abs(RFC_B(x))}{2|U|}}$ to make the value of $FOD_B(x)$ larger. On the contrary, if $RFC_B(x) > 0$, we use $\frac{|U| - RFC_B(x)}{2|U|}$ to make the value of $FOD_B(x)$ smaller.

The $FOD_B(x)$ is used to measure the outlier degree of x under $\widetilde{R}_B$. It is used for the feature subset B. However, since C has $2^{|C|}$ subsets, it is impractical to calculate fuzzy relation-based outlier degrees for all subsets. Therefore, this paper considers the metric based on FS and FSS, for which the fuzzy relation-based outlier degree needs to be specialized as $FOD_{c_k}(x)$ and $FOD_{C_k}(x)$. In other words, we need to calculate $FOD_{c_k}(x)$ and $FOD_{C_k}(x)$ to characterize the outlier degree of x. Then the fuzzy information entropy-based outlier factor is constructed by integrating the outlier degrees of FS and FSS.

Definition 14 *(Outlier factor).* Let $\forall x \in U$, the fuzzy entropy-based outlier factor $FEOF(x)$ of x is computed by

$$FEOF(x) = 1 - \frac{\sum_{k=1}^{m} \left(1 - FOD_{c'_k}(x)\right) W_{c'_k}(x) + \sum_{k=1}^{m} \left(1 - FOD_{C_k}(x)\right) W_{C_k}(x)}{2|C|}, \tag{17}$$

where, for $\forall B \subseteq C$, weight function $W_B : U \to (0, 1], s.t., W_B(x) = \sqrt{\frac{|[x]_B|}{|U|}}$.

In accordance with the above definition, $FEOF(x)$ is inversely proportional to its weight function W_B, that is, the smaller the $W_B(x)$ of x is, the more likely x is an outlier. Given a fuzzy similarity relation, if $|[x]_B|$ is always small, then we consider x belongs to a minority class in U. Correspondingly, we will give x a smaller $W_B(x)$, making x more likely to be an outlier. Therefore, the weight function W_B embodies the idea that outliers often belong to the objects in the minority class, and the objects in the minority class are more likely to be outliers than the objects in the majority class.

Definition 15 *(Outlier detection).* Given a judgment threshold μ. $\forall x \in U$, if $FEOF(x) > \mu$, x is called an outlier based on fuzzy information entropy in U.

4.2. Algorithm

Section 4.1 proposes an outlier detection method based on fuzzy information entropy, including stepwise the construct of metric and final outlier decision. This section mainly proposes related algorithms, namely, FIEOD algorithm.

Algorithm 1 FIEOD algorithm.

Input: $FIS = (U, C)$, μ, δ
Output: Outlier Set (OS)
1: $OS \leftarrow \emptyset$
2: **for** $k \leftarrow 1$ to m **do**
3: Compute $M_{\widetilde{R}_{c_k}}$
4: Compute $FE(\{c_k\})$
5: **end for**
6: Determine $FS = \{c'_1, c'_2, \cdots, c'_m\}$
7: Build $FSS = \{C_1, C_2, \cdots, C_m\}$
8: **for** $k \leftarrow 1$ to m **do**
9: Compute $FE(C_k)$
10: **end for**
11: **for** $i = 1$ to n **do**
12: **for** $k \leftarrow 1$ to m **do**
13: Compute $RFE_{\{c'_k\}}(x_i)$ and $RFE_{C_k}(x_i)$
14: Compute $RFC_{\{c'_k\}}(x_i)$ and $RFC_{C_k}(x_i)$
15: Compute $FOD_{\{c'_k\}}(x_i)$ and $FOD_{C_k}(x_i)$
16: Compute $W_{\{c'_k\}}(x_i)$ and $W_{C_k}(x_i)$
17: **end for**
18: Compute $FEOF(x_i)$
19: **if** $NEOF(x_i) > \mu$ **then**
20: $O \leftarrow O \cup \{x_i\}$
21: **end if**
22: **end for**
23: **return** OS

The frequency of steps 2–4 in the Algorithm 1 is m, the frequency of step 3 is $n \times n$, and the frequency of steps 8–10 is m. Besides, the frequency of steps 11–22 is n, and the frequency of steps 12–17 is m. Thus, the total frequency of Algorithm 1 is $m \times n \times n + m + n \times m$. Therefore, in the worst case, the time complexity of the Algorithm 1 is $O(mn^2)$.

Table 2
Initial and normalized FISs.

Initial				Normalized			
U	c_1	c_2	c_3	U	c_1	c_2	c_3
x_1	D	4	0.7	x_1	D	$\frac{1}{3}$	$\frac{4}{5}$
x_2	B	7	0.4	x_2	B	$\frac{2}{3}$	$\frac{1}{5}$
x_3	D	1	0.6	x_3	D	0	$\frac{3}{5}$
x_4	B	2	0.3	x_4	B	$\frac{1}{9}$	0
x_5	B	8	0.5	x_5	B	$\frac{7}{9}$	$\frac{2}{5}$
x_6	C	10	0.8	x_6	C	1	1

4.3. An example

This section employs an example to illustrate the proposed outlier detection method.

Example 1. Let $FIS = (U, C)$ be a fuzzy information system listed in the left column of Table 2, where $U = \{x_1, x_2, x_3, x_4, x_5, x_6\}$, $C = \{c_1, c_2, c_3\}$. The original numerical feature data is normalized by the min-max normalization method (Eq. (9)), and the data obtained after normalization is shown on the right column of Table 2.

This fuzzy information system involves hybrid features. As is shown in Table 2, the second column is symbolic feature data, and the third column and the fourth column are both numeric feature data. The corresponding data standard deviation are $std(c_2) \approx 0.3610$ and $std(c_3) \approx 0.3416$, respectively. Take $\delta = 1$, and the fuzzy radii on features c_2 and c_3 are obtained by Eq. (11): $\varepsilon_{c_2} \approx 0.3610$ and $\varepsilon_{c_3} \approx 0.3416$.

After calculation, the fuzzy relation matrix of each feature in C is as follows.

$$M_{\widetilde{R}_{c_1}} = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix},$$

$$M_{\widetilde{R}_{c_2}} = \begin{pmatrix} 1 & 0.6667 & 0.6667 & 0.7778 & 0 & 0 \\ 0.6667 & 1 & 0 & 0 & 0.8889 & 0.6667 \\ 0.6667 & 0 & 1 & 0.8889 & 0 & \\ 0.7778 & 0 & 0.8889 & 1 & 0 & 0 \\ 0 & 0.8889 & 0 & 0 & 1 & 0.7778 \\ 0 & 0.6667 & 0 & 0 & 0.7778 & 1 \end{pmatrix},$$

$$M_{\widetilde{R}_{c_3}} = \begin{pmatrix} 1 & 0 & 0.8000 & 0 & 0 & 0.8000 \\ 0 & 1 & 0 & 0.8000 & 0.8000 & 0 \\ 0.8000 & 0 & 1 & 0 & 0.8000 & 0 \\ 0 & 0.8000 & 0 & 1 & 0 & 0 \\ 0 & 0.8000 & 0.8000 & 0 & 1 & 0 \\ 0.8000 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

The related metrics for all feature sequences and feature subset sequences are calculated below, and the outliers are finally detected. The process involves the following seven steps.

1) The fuzzy information entropy given in Definition 7 of each feature is as follows.

$FE(\{c_1\}) = -\frac{1}{|U|}\sum_{i=1}^{n} \log_2 \frac{|[x_i]_{\widetilde{R}_B}|}{|U|} = -(\frac{1}{6}\log_2\frac{2}{6} + \frac{1}{6}\log_2\frac{3}{6} + \frac{1}{6}\log_2\frac{2}{6} + \frac{1}{6}\log_2\frac{3}{6} + \frac{1}{6}\log_2\frac{3}{6} + \frac{1}{6}\log_2\frac{1}{6}) \approx 1.4591,$

Similarly, there are $FE(\{c_2\}) \approx 1.1185$, $FE(\{c_3\}) \approx 1.3833$.

2) The fuzzy information entropy after removing an object is calculated as follows.
$FE_{x_1}(\{c_1\}) = FE_{x_3}(\{c_1\}) = -(\frac{1}{5}\log_2\frac{3}{5} + \frac{1}{5}\log_2\frac{1}{5} + \frac{1}{5}\log_2\frac{3}{5} + \frac{1}{5}\log_2\frac{3}{5} + \frac{1}{5}\log_2\frac{1}{5}) \approx 1.3710$,
$FE_{x_2}(\{c_1\}) = FE_{x_4}(\{c_1\}) = FE_{x_5}(\{c_1\}) \approx 1.5219$, $FE_{x_6}(\{c_1\}) \approx 0.9710$;
$FE_{x_1}(\{c_2\}) \approx 1.1433$, $FE_{x_2}(\{c_2\}) \approx 1.1783$, $FE_{x_3}(\{c_2\}) \approx 1.0195$,
$FE_{x_4}(\{c_2\}) \approx 1.0515$, $FE_{x_5}(\{c_2\}) \approx 1.0477$, $FE_{x_6}(\{c_2\}) \approx 0.9865$;
$FE_{x_1}(\{c_3\}) = FE_{x_2}(\{c_3\}) \approx 1.4313$, $FE_{x_3}(\{c_3\}) = FE_{x_5}(\{c_3\}) \approx 1.3678$, $FE_{x_4}(\{c_3\}) = FE_{x_6}(\{c_3\}) \approx 1.1556$.

Further, the relative fuzzy entropy can be obtained as follows according to Definition 9
$RFE_{\{c_1\}}(x_1) = RFE_{\{c_1\}}(x_3) \approx 0.0604$, $RFE_{\{c_1\}}(x_2) = RFE_{\{c_1\}}(x_4) = RFE_{\{c_1\}}(x_5) = 0$, $RFE_{\{c_1\}}(x_6) \approx 0.3346$;
$RFE_{\{c_2\}}(x_1) = RFE_{\{c_2\}}(x_2) = 0$, $RFE_{\{c_2\}}(x_3) \approx 0.0885$, $RFE_{\{c_2\}}(x_4) \approx 0.0599$,
$RFE_{\{c_2\}}(x_5) \approx 0.6223$, $RFE_{\{c_2\}}(x_6) \approx 0.1180$;
$RFE_{\{c_3\}}(x_1) = RFE_{\{c_3\}}(x_2) = 0$, $RFE_{\{c_3\}}(x_3) = RFE_{\{c_3\}}(x_5) \approx 0.0112$, $RFE_{\{c_3\}}(x_4) = RFE_{\{c_3\}}(x_6) \approx 0.1646$.

3) According to Definition 10, the relative fuzzy cardinality is calculated as follows.
$RFC([x_1]_{\{c_1\}}) = RFC([x_3]_{\{c_1\}}) = -0.2000$, $RFC([x_2]_{\{c_1\}}) = RFC([x_4]_{\{c_1\}}) = RFC([x_5]_{\{c_1\}}) = 1.2000$,
$RFC([x_6]_{\{c_1\}}) = -1.6000$;
$RFC([x_1]_{\{c_2\}}) \approx 0.8222$, $RFC([x_2]_{\{c_2\}}) \approx 0.9778$, $RFC([x_3]_{\{c_2\}}) \approx 0.0444$, $RFC([x_4]_{\{c_2\}}) = RFC([x_5]_{\{c_2\}}) \approx 0.2000$, $RFC([x_6]_{\{c_2\}}) \approx -0.1111$;
$RFC([x_1]_{\{c_3\}}) = RFC([x_2]_{\{c_3\}}) = RFC([x_3]_{\{c_3\}}) = RFC([x_5]_{\{c_3\}}) \approx 0.6400$, $RFC([x_4]_{\{c_3\}}) = RFC([x_6]_{\{c_3\}}) \approx -0.4800$.

4) In the light of Definitions 11 and 12, we can construct the following two sequences
$FS = \langle c_1', c_2', c_3'\rangle = \langle c_1, c_3, c_2\rangle$, $FSS = \langle C_1, C_2, C_3\rangle = \langle\{c_1\}, \{c_1, c_3\}, \{c_1, c_2, c_3\}\rangle$;
Similarly, we can get
$RFE_{C_1}(x_1) = RFE_{C_1}(x_3) \approx 0.0604$, $RFE_{C_1}(x_2) = RFE_{C_1}(x_4) = RFE_{C_1}(x_5) = 0$, $RFE_{C_1}(x_6) \approx 0.3346$;
$RFE_{C_2}(x_1) = RFE_{C_2}(x_3) \approx 0.0463$, $RFE_{C_2}(x_2) = 0$, $RFE_{C_2}(x_4) = RFE_{C_2}(x_5) \approx 0.0818$, $RFE_{C_2}(x_6) \approx 0.2358$;
$RFE_{C_3}(x_1) = RFE_{C_3}(x_3) \approx 0.0359$, $RFE_{C_3}(x_2) = RFE_{C_3}(x_5) \approx 0.0143$, $RFE_{C_3}(x_4) = RFE_{C_3}(x_6) \approx 0.1793$.
$RFC([x_1]_{C_1}) = RFC([x_3]_{C_1}) = -0.2000$, $RFC([x_2]_{C_1}) = RFC([x_4]_{C_1}) = RFC([x_5]_{C_1}) = 1.2000$, $RFC([x_6]_{C_1}) = -1.6000$;
$RFC([x_1]_{C_2}) = RFC([x_3]_{C_2}) = RFC([x_4]_{C_2}) = RFC([x_5]_{C_2}) = 0.1600$, $RFC([x_2]_{C_2}) = 1.2800$, $RFC([x_6]_{C_2}) = -0.9600$;
$RFC([x_1]_{C_3}) = RFC([x_3]_{C_3}) \approx 0.3467$, $RFC([x_2]_{C_3}) = RFC([x_5]_{C_3}) \approx 0.5333$, $RFC([x_4]_{C_3}) = RFC([x_6]_{C_3}) \approx -0.5867$.

5) According to Definition 13, the outlier degree of fuzzy relation is calculated as follows.
$FOD_{\{c_1\}}(x_1) = FOD_{\{c_1\}}(x_3) \approx 0.0434$, $FOD_{\{c_1\}}(x_2) = FOD_{\{c_1\}}(x_4) = FOD_{\{c_1\}}(x_5) = 0$, $FOD_{\{c_1\}}(x_6) \approx 0.2663$,
$FOD_{\{c_2\}}(x_1) = FOD_{\{c_2\}}(x_2) = 0$, $FOD_{\{c_2\}}(x_3) = 0.0439$, $FOD_{\{c_2\}}(x_4) \approx 0.0289$, $FOD_{\{c_2\}}(x_5) \approx 0.0301$,
$FOD_{\{c_2\}}(x_6) \approx 0.0842$, $FOD_{\{c_3\}}(x_1) = FOD_{\{c_3\}}(x_2) = 0$, $FOD_{\{c_3\}}(x_3) = FOD_{\{c_3\}}(x_5) \approx 0.0050$, $FOD_{\{c_3\}}(x_4) = FOD_{\{c_3\}}(x_6) \approx 0.1209$;
$FOD_{C_1}(x_1) = FOD_{C_1}(x_3) \approx 0.0434$, $FOD_{C_1}(x_2) = FOD_{C_1}(x_4) = FOD_{C_1}(x_5) = 0$, $FOD_{C_1}(x_6) \approx 0.2663$,
$FOD_{C_2}(x_1) = FOD_{C_2}(x_3) \approx 0.0225$, $FOD_{C_2}(x_3) = 0$, $FOD_{C_2}(x_4) = FOD_{C_2}(x_5) \approx 0.0398$, $FOD_{C_2}(x_6) \approx 0.1796$, $FOD_{C_3}(x_1) = FOD_{C_3}(x_3) \approx 0.0169$, $FOD_{C_3}(x_2) = FOD_{C_3}(x_5) \approx 0.0065$, $FOD_{C_3}(x_4) = FOD_{C_3}(x_6) \approx 0.1328$.

6) By Definition 14, the fuzzy entropy outlier factor of x is calculated as follows.
$FEOF(x_1) = 1 - \frac{(1-0.0434)\times\sqrt{\frac{2}{6}}+(1-0)\times\sqrt{\frac{3.1111}{6}}+(1-0)\times\sqrt{\frac{2.6000}{6}}}{2\times 3} - \frac{(1-0.0434)\times\sqrt{\frac{2}{6}}+(1-0.0225)\times\sqrt{\frac{1.8}{6}}+(1-0.0169)\times\sqrt{\frac{1.6667}{6}}}{2\times 3} \approx 0.4106$.

Analogously, $FEOF(x_2) \approx 0.3320$, $FEOF(x_3) \approx 0.4272$, $FEOF(x_4) \approx 0.4295$, $FEOF(x_5) \approx 0.3690$, $FEOF(x_6) \approx 0.6077$.

7) Let $\mu = 0.60$, outliers determined by Definition 15 are as follows.
$FEOF(x_1), FEOF(x_2), FEOF(x_3), FEOF(x_4), FEOF(x_5) < \mu$, $FEOF(x_6) > \mu$. It can be seen that only the fuzzy entropy-based outlier factor of x_6 is greater than μ, so $OS = \{x_6\}$.

Table 3
The description of data sets.

ID	Data set	Abbreviation	Preprocessing	Conditional feature		Outlier ($\lvert OS^{\circ}\rvert$)	Normal
				Numerical	Nominal		
1	Credit approval	Cred	Downsampling class "+" to 42 objects	6	9	42	383
2	German	Germ	Downsampling class "2" to 14 objects	7	13	14	700
3	Heart disease	Heart	Downsampling class "2" to 16 objects	6	7	16	150
4	Hepatitis	Hepa	Downsampling class "2" to 9 objects	6	13	9	85
5	Horse	Horse	Downsampling class "1" to 12 objects	8	19	12	244
6	Diabetes	Diab	Downsampling the class "tested_positive" to 26 objects	8	0	26	500
7	Ionosphere	Iono	Downsampling class "b" to 24 objects	34	0	24	225
8	Iris	Iris	Downsampling the class "Iris-virginica" to 11 objects	4	0	11	100
9	Pima	Pima	Downsampling class "TRUE" to 55 objects	9	0	55	500
10	Sonar	Sonar	Downsampling class "M" to 10 objects	60	0	10	97
11	Wisconsin diagnostic breast cancer	Wdbc	Downsampling class "M" to 39 objects	31	0	39	357
12	Page blocks	Page	Downsampling the class "Non-text" to 258 objects	10	0	258	4913
13	Wisconsin breast cancer	Wbc	202 "malignant" (outliers) and 14 "benign" objects were removed	9	0	39	444
14	Yeast	Yeast	Classes "ERL" (outliers), "CYT", "NUC", and "MIT" are selected	8	0	5	1136
15	Lymphography	Lymp	Classes "1" and "4" are treated as outliers	0	8	6	142
16	Mushroom	Mush	Downsampling class "+" to 221 objects	0	22	221	4208

5. Experiments

5.1. Data sets and experimental settings

To evaluate the effectiveness of the FIEOD algorithm, 16 data sets (including numerical, nominal, and mixed features) are selected from UCI for experiments [56]. On 16 data sets, we compare the performance of the FIEOD algorithm with DIStance-based (DIS) [10], k-Nearest Neighbor (kNN) [57], density-based (Local Outlier Factor, LOF) [8], Find Cluster-Based Local Outlier Factor (FindCBLOF) [9], Granular Computing-based (GrC) [19], SEQuence-based (SEQ) [22], Information Entropy-based (IE) [47], and Outlier Detection based on Granular Computing and Rough set (ODGrCR) [26] algorithms. DIS, kNN, and LOF algorithms are relatively simple and only applicable to numerical feature data. The FindCBLOF algorithm is closely related to clustering methods, which is only applied to nominal feature data. Besides, GrC, SEQ, IE, and ODGrCR algorithms are outlier detection algorithms based on rough sets. They are only applicable to nominal feature data, while discretization preprocessing is required for numerical feature data.

Most of the selected 16 data sets are used for the evaluation of classification and clustering methods. However, for the evaluation of outlier detection, there are few existing data sets. Accordingly, this paper utilizes the downsampling method proposed in [58] to obtain valid data sets for outlier detection evaluation. A particular class is randomly downsampled to produce outliers, while all objects of the remaining classes are preserved to form a data set for outlier detection. In addition, the maximum probability value method is used to fill the missing value in data set, that is, the missing feature values are filled by the highest frequency value on the other objects. The data preprocessing and description of data sets are demonstrated in Table 3.

In Table 3, two data sets contain only nominal features, nine data sets contain only numerical features, and five data sets contain both nominal features and numerical features, i.e., hybrid feature. In addition, the number of objects

Table 4
Optimal parameter setting and discretization method for different data sets.

Data set	k(kNN)	$MinPts$(LOF)	s(FindCBLOF)	w(GrC)	δ(FIEOD)	Discretization method
Cred	6	106	10	3	1.0	EW
Germ	31	135	4	4	0.6	EW
Heart	31	30	5	4	3.9	EW
Hepa	16	8	8	4	0.5	EW
Horse	29	11	4	3	1.4	EF
Diab	7	76	7	3	2.8	EW
Iono	2	2	4	5	0.8	EW
Iris	13	2	1	3	0.1	EW
Pima	136	109	6	5	0.8	EW
Sonar	13	19	2	2	1.0	EW
Wdbc	74	41	2	2	0.6	EW
Page	35	1281	2	7	0.9	EW
Wbc	39	15	3 [9]	2 [19]	0.6	–
Yeast	243	8	5	3	0.1	EW
Lymp	37	36	3 [9]	2 [19]	–	–
Mush	212	2	9	10	–	–

is between 94 and 5171, and the number of features is between 8 and 60. To ensure repeatability of the experiments, the relevant data sets can be downloaded from the github homepage.[1]

During the experiment, we repeat the kNN and LOF algorithms for the 16 data sets and calculate the optimal values of their respective parameters k and $MinPts$ in the range of $[1, n/4]$ with step size 1. The two parameters α and β required by the FindCBLOF algorithm are set to 90% and 5, respectively, and for the similarity threshold parameter s, its optimal value is calculated in the range of [1,10] with steps size 1 [9]. For the GrC algorithm, the overlap distance metric is conducted to calculate the distance between any two objects [19], whose parameter $d = |C|/w$, the w's integer optimal value is calculated in the range [1,10]. In addition, for SEQ, IE, and ODGrCR algorithms, all features values in Lymp and Wbc data sets are considered to be the nominal type [9]. For the rest of the data sets with numerical features, we use the discretization method of Equal Width (EW) and Equal Frequency (EF) in Weka to convert all numerical feature values into discrete feature with 3 interval numbers [59]. Finally, the optimal discretization method is adopted. For DIS, kNN, and LOF algorithms, the Euclidean distance metric is used. What's more, all the different nominal feature values are replaced with different integer values. The feature value is normalized to the interval [0,1] by using the min-max normalization (Eq. (9). For the FIEOD algorithm, we calculate the optimal parameter for δ in [0.1,5] with step size 0.1. Furthermore, in traditional DIS methods, it is unreasonable that outliers are treated as binary classification. Therefore, this article uses strategy in documents [22,47] to define the distance-based outlier factor $DOF(x_i) = \sum_{j=1}^{n} dist(x_i, x_j)$ for each object. In the end, the optimal parameter settings and discretization methods for different data sets are illustrated in Table 4.

The comparative experiments are conducted on a computer with the Intel (R) core (TM) i7-8700 processor platform; 3.20 GHz frequency; 16 G memory. The operating system is Windows 10. The experimental results are performed in Matlab R2015b.

5.2. Evaluation metrics

In this paper, Precision (P), Recall (R), and Receiver Operating Characteristic (ROC) curves are used to evaluate the effectiveness of the proposed method [60]. The specific steps are as follows. In the outlier detection, most of the detection methods ultimately output the outlier factor of each object in U, and the larger the outlier factor of an object, the more likely it is the outlier. These objects can be arranged in descending order according to their outlier factor values. Given an order number t, objects with a sequence number greater than or equal to t are treated as outliers. If the given t is too small, it will cause the method to miss the true outliers. Conversely, too many objects are judged to

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

be outliers, which leads to too excessive false positives. This trade-off can usually be measured by P and R. For $\forall t$, $OS(t)$ is a function of t. It denotes the outlier set detected by the given t. OS° represents the true outlier set in the data set, and the $P(t)$ is calculated as

$$P(t) = \frac{|OS(t) \cap OS^{\circ}|}{|OS(t)|} \times 100\%. \tag{18}$$

The $P(t)$ indicates the ratio of outliers detected at a given t to true outliers. The $R(t)$ is calculated as follows.

$$R(t) = \frac{|OS(t) \cap OS^{\circ}|}{|OS^{\circ}|} \times 100\%. \tag{19}$$

The $R(t)$ represents the ratio of true outliers detected at a given t to all true outliers.

The range of $P(t)$ and $R(t)$ is $[0, 100\%]$. Under a given t, the larger the values of $P(t)$ and $R(t)$, the better the outlier detection results. In addition, when $P(t)$ and $R(t)$ are given, the smaller the t, the better the detection effect. What's more, it can be shown that $P(t)$ and $R(t)$ are equal when $t = |OS^{\circ}|$.

The ROC is a curve with the False Positive Rate (FPR) as the abscissa and the True Positive Rate (TPR) as the ordinate. The $FPR(t)$ is calculated by

$$FPR(t) = \frac{|OS(t) - OS^{\circ}|}{|U - OS^{\circ}|} \times 100\%. \tag{20}$$

The calculation method of $TPR(t)$ is as follows.

$$TPR(t) = R(t) = \frac{|OS(t) \cap OS^{\circ}|}{|OS^{\circ}|} \times 100\%. \tag{21}$$

The ROC curve can be used to compare the performance of different outlier detection models. If the curve of a detection method is closer to the upper left corner of the first quadrant, and the larger the area under the curve, the better the performance.

5.3. Experimental results and analyses

5.3.1. Comparison by P(t) and R(t)

Tables 5-7 show the experimental results for $P(t)$ and $R(t)$ on hybrid, numerical, and nominal feature data sets, respectively. They illustrate the results of the $P(t)$ and $R(t)$ change with t.

From Table 5, it can be seen that the FIEOD algorithm achieves superior performance on the mixed feature data sets. The analyses are mainly carried out from the following aspects.

1) Given $t = |OS^{\circ}|$, the FIEOD algorithm has a larger $P(t)$. For example, for the Cred data set, the FIEOD algorithm's $P(t)$ is 90.48%. However, for DIS, kNN, LOF, FindCBLOF, GrC, SEQ, IE, and ODGrCR algorithms, their $P(t)$ are 73.81%, 35.71%, 30.95%, 47.62%, 76.19%, 66.67%, 71.43%, and 73.81%, respectively. The $P(t)$ of the FIEOD algorithm is larger than that of other algorithms. On Germ, Hepa, and Horse data sets, the FIEOD algorithm's $P(t)$ is greater than or equal to that of other algorithms. For the Heart data set, the $P(t)$ of the FIEOD algorithm is slightly smaller than that of the GrC algorithm, but greater than or equal to other algorithms.
2) When $R(t)$ reaches 100.00% for the first time, the FIEOD algorithm has a smaller t. For example, the t of the FIEOD algorithm on the Cred data set is 72. For DIS, kNN, LOF, FindCBLOF, GrC, SEQ, IE, and ODGrCR algorithms, their t are 277, 166, 325, 154, 130, 137, 111, and 154, respectively. The FIEOD algorithm's t is significantly smaller than other algorithms. It indicates that the FIEOD algorithm can detect all 42 outliers in the Cred data set at $t = 72$, but other algorithms can only detect less than 42 outliers. For Germ, Heart, and Horse data sets, the FIEOD algorithm's t is also smaller than other algorithms. The t of the FIEOD algorithm is 11 on the Heap data set. It is equal to t of LOF, but less than that of all other algorithms.
3) For the average of $P(t)$ and $R(t)$, the FIEOD algorithm achieves maximum values on the Cred, Germ, Hepa, and Horse data sets. For example, the average $P(t)$ and $R(t)$ of the FIEOD algorithm on the Cred data set are 43.03% and 94.05%, respectively, which is significantly larger than other algorithms. However, for the Heart data set, the average $P(t)$ and $R(t)$ of the FIEOD algorithm are larger than that of all algorithms except the FindCBLOF and GrC algorithms.

Table 5
The comparison of experimental results on hybrid feature data sets.

Data set	t	DIS		kNN		LOF		FindCBLOF		GrC		SEQ		IE		ODGrCR		FIEOD	
		$P(t)$	$R(t)$	$P(t)$	$R(t)$	$P(t)$	$R(t)$	$P(t)$	$R(t)$	$P(t)$	$R(t)$	$P(t)$	$R(t)$	$P(t)$	$R(t)$	$P(t)$	$R(t)$	$P(t)$	$R(t)$
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
	72	52.78	90.48	33.33	57.14	25.00	42.86	47.22	80.95	54.17	92.86	48.61	83.33	50.00	85.71	54.17	92.86	58.33	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	39.12	88.57	27.05	75.71	23.37	61.67	31.25	81.90	39.57	90.71	37.33	87.86	37.91	89.52	39.58	90.00	**43.03**	**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	42.86	21.43
	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	35.71	35.71
	36	13.89	35.71	22.22	57.14	13.89	35.71	19.44	50.00	19.44	50.00	25.00	64.29	22.22	57.14	27.78	71.43	38.89	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	15.42	64.29	21.07	73.02	15.24	62.70	21.51	78.57	21.51	78.57	18.12	66.67	20.82	77.78	22.44	80.95	**23.67**	**84.13**
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	87.50	43.75
	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	81.25	81.25
	19	68.42	81.25	68.42	81.25	68.42	81.25	78.95	93.75	78.95	93.75	63.16	75.00	78.95	93.75	78.95	93.75	84.21	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
	43	37.21	100.00	37.21	100.00	37.21	100.00	37.21	100.00	37.21	100.00	37.21	100.00	37.21	100.00	37.21	100.00	37.21	100.00
	Average	54.11	82.64	56.65	88.19	52.60	82.64	59.58	91.67	**60.27**	**92.36**	53.68	85.42	57.15	89.58	57.86	90.97	58.45	91.67

(*continued on next page*)

Table 5 (*continued*)

Data set	t	DIS		kNN		LOF		FindCBLOF		GrC		SEQ		IE		ODGrCR		FIEOD	
		$P(t)$	$R(t)$	$P(t)$	$R(t)$	$P(t)$	$R(t)$	$P(t)$	$R(t)$	$P(t)$	$R(t)$	$P(t)$	$R(t)$	$P(t)$	$R(t)$	$P(t)$	$R(t)$	$P(t)$	$R(t)$
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
	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	71.69	87.30	68.51	84.13	72.59	88.89	73.44	87.30	71.86	85.71	63.51	77.78	72.96	87.30	74.15	88.89	**77.03**	**92.06**
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
	16	37.50	50.00	31.25	41.67	12.50	16.67	62.50	83.33	62.50	83.33	56.25	75.00	56.25	75.00	62.50	83.33	75.00	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	28.85	59.26	26.76	59.26	13.26	35.19	46.98	85.19	48.04	87.04	46.28	84.26	47.90	87.04	44.58	82.41	**49.06**	**89.81**

Table 6
The comparison of experimental results on numerical feature data sets.

Data set	t	DIS		kNN		LOF		FindCBLOF		GrC		SEQ		IE		ODGrCR		FIEOD	
		$P(t)$	$R(t)$	$P(t)$	$R(t)$	$P(t)$	$R(t)$	$P(t)$	$R(t)$	$P(t)$	$R(t)$	$P(t)$	$R(t)$	$P(t)$	$R(t)$	$P(t)$	$R(t)$	$P(t)$	$R(t)$
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	68.75	42.31
	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
	58	34.48	76.92	34.48	76.92	36.21	80.77	32.76	73.08	34.48	76.92	32.76	73.08	34.48	76.92	36.21	80.77	44.83	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	30.25	83.85	29.87	85.00	31.67	85.77	26.35	77.69	30.64	84.23	29.31	79.62	30.82	83.08	27.85	81.54	**35.71**	**91.15**
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	95.83	95.83
	25	92.00	95.83	96.00	100.00	88.00	91.67	72.00	75.00	72.00	75.00	72.00	75.00	68.00	70.83	80.00	83.33	96.00	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	68.99	90.74	**70.36**	**92.13**	68.18	89.81	60.33	81.94	60.33	81.94	60.12	81.48	57.57	80.09	65.26	87.50	69.89	91.67
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
	127	40.94	94.55	40.94	94.55	40.94	94.55	35.43	81.82	38.58	89.09	22.05	50.91	33.86	78.18	38.58	89.09	43.31	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	35.00	86.73	35.69	87.09	35.40	86.91	35.20	85.27	35.35	84.91	25.16	58.55	34.43	84.00	34.85	84.00	**38.87**	**89.82**

(*continued on next page*)

Table 6 (*continued*)

Data set	t	DIS		kNN		LOF		FindCBLOF		GrC		SEQ		IE		ODGrCR		FIEOD	
		$P(t)$	$R(t)$	$P(t)$	$R(t)$	$P(t)$	$R(t)$	$P(t)$	$R(t)$	$P(t)$	$R(t)$	$P(t)$	$R(t)$	$P(t)$	$R(t)$	$P(t)$	$R(t)$	$P(t)$	$R(t)$
Sonar	5	80.00	40.00	80.00	40.00	100.00	50.00	80.00	40.00	100.00	50.00	100.00	50.00	100.00	50.00	80.00	40.00	80.00	40.00
	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	75.00	90.00	83.33	100.00	66.67	80.00	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	76.92	100.00	69.23	90.00	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	71.43	100.00	64.29	90.00	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	**71.38**	**90.00**	69.92	**90.00**	68.48	86.25	68.00	85.00	65.79	82.50	67.02	86.25	69.92	**90.00**
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	92.31	92.31
	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	88.10	94.87
	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	88.37	97.44
	44	86.36	97.44	86.36	97.44	88.64	100.00	86.36	97.44	88.64	100.00	59.09	66.67	86.36	97.44	86.36	97.44	88.64	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	83.53	91.17	82.69	90.31	85.69	93.16	84.71	92.02	**86.30**	**93.45**	58.44	65.81	84.71	92.02	84.95	92.31	84.90	92.31
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	51.94	25.97
	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	44.19	44.19
	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	43.57	70.93
	879	26.39	89.92	27.08	92.25	25.82	87.98	15.47	52.71	15.47	52.71	13.20	44.96	26.28	89.53	24.69	84.11	29.35	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	26.97	82.34	**29.76**	**84.45**	26.64	81.70	23.74	72.18	23.04	71.53	19.44	53.45	25.02	76.49	25.57	76.10	26.03	82.34
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
	47	80.85	97.44	78.72	94.87	80.85	97.44	72.34	87.18	76.60	92.31	74.47	89.74	74.47	89.74	78.72	94.87	82.98	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	77.73	91.67	77.71	91.67	78.30	92.31	71.67	85.58	73.64	88.14	74.32	88.46	74.08	88.14	75.40	89.74	**78.57**	**92.63**

(*continued on next page*)

Table 6 (*continued*)

Data set	t	DIS		kNN		LOF		FindCBLOF		GrC		SEQ		IE		ODGrCR		FIEOD	
		$P(t)$	$R(t)$	$P(t)$	$R(t)$	$P(t)$	$R(t)$	$P(t)$	$R(t)$	$P(t)$	$R(t)$	$P(t)$	$R(t)$	$P(t)$	$R(t)$	$P(t)$	$R(t)$	$P(t)$	$R(t)$
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	66.67	40.00
	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
	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	57.14	80.00
	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	17.26	40.00	**62.46**	**95.00**	**62.46**	**95.00**	40.67	77.50	47.45	80.00	25.75	52.50	26.74	47.50	37.72	72.50	54.01	87.50

Table 7
The comparison of experimental results on nominal feature data sets.

Data set	t	DIS		kNN		LOF		FindCBLOF		GrC		SEQ		IE		ODGrCR		FIEOD	
		$P(t)$	$R(t)$	$P(t)$	$R(t)$	$P(t)$	$R(t)$	$P(t)$	$R(t)$	$P(t)$	$R(t)$	$P(t)$	$R(t)$	$P(t)$	$R(t)$	$P(t)$	$R(t)$	$P(t)$	$R(t)$
Lymp	3	100.00	50.00	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	100.00	66.67	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	75.00	50.00
	5	100.00	83.33	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	83.33	83.33	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	60.00	100.00	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	65.54	86.36	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	84.85
Mush	100	32.00	14.48	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	21.46	98.19	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	21.76	99.55
	1028	21.11	98.19	15.86	73.76	20.04	93.21	21.11	98.19	19.07	88.69	19.36	90.05	21.11	98.19	21.50	100.00	21.50	100.00
	1185	18.57	99.55	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	17.87	99.55	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
	1342	16.47	100.00	16.47	100.00	15.95	96.83	16.47	100.00	14.83	90.05	14.83	90.05	6.32	99.10	16.47	100.00	16.47	100.00
	1440	15.35	100.00	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	10.31	100.00	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	10.18	100.00	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	9.46	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	9.46	100.00
	Average	17.52	84.53	18.08	80.75	18.31	83.13	28.84	93.05	28.72	87.74	28.80	88.32	29.60	92.97	**29.91**	**93.87**	**29.91**	93.83

According to Table 6, we can do the following analyses.

1) Given $t = |OS^{\circ}|$, the FIEOD algorithm has a larger $P(t)$. For example, for the Diab data set, the FIEOD algorithm's P(t) is 69.23%, while for DIS, kNN, LOF, FindCBLOF, GrC, SEQ, IE, and ODGrCR algorithms, their $P(t)$ is 42.31%, 46.15%, 46.15%, 34.62%, 46.15%, 38.46%, 46.15%, and 38.46%, respectively. The $P(t)$ of the FIEOD algorithm is larger than other algorithms. However, for the Iono data set, the $P(t)$ of the FIEOD algorithm is slightly lower than that of the kNN algorithm, but larger than that of the other algorithms. The $P(t)$ of the FIEOD algorithm on the Page data set is 44.19%, slightly smaller than that of DIS, kNN, and LOF algorithms, but larger than FindCBLOF, GrC, SEQ, IE, and ODGrCR algorithms. Also for the Yeast data set, the FIEOD algorithm's $P(t)$ is just smaller than that of kNN and LOF algorithms, but larger than that of the others.
2) When $R(t)$ reaches 100.00% for the first time, the FIEOD algorithm has a smaller t. For example, for the Diab data set, the t of the FIEOD algorithm is 58, and 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 t of the FIEOD algorithm is smaller than that of other algorithms. It implies that FIEOD and ODGrCR algorithms can detect all 26 outliers at $t = 39$, but other algorithms cannot detect all outliers of the data set. For example, for the Iono data set, the FIEOD algorithm's t is 26, which is one greater than kNN's t, but equal to or less than that of other algorithms. On the Page data set, the t of the FIEOD algorithm is 879, which is smaller than that of other algorithms. It shows that the FIEOD algorithm is optimally effective. Also for the Yeast data set, the t of the FIEOD algorithm is only larger than that of kNN and LOF algorithms, but smaller than that of other algorithms.
3) In terms of the average of $R(t)$, the FRGOD algorithm achieves a larger value on most numerical feature data sets. For example, the average $R(t)$ of the FIEOD algorithm on the Diab, Iris, Pima, Sonar, and Wbc data sets is significantly greater than or equal to that of other algorithms. For Iono and Page data sets, the average $R(t)$ of the FIEOD algorithm is only smaller than the kNN algorithm, but larger than that of the other algorithms. The average $R(t)$ of the FIEOD algorithm on the Wdbc data set is only smaller than that of kNN and GrC algorithms. In addition, the average $R(t)$ of the FIEOD algorithm on the Yeast data set is just smaller than that of kNN and LOF algorithms.

From the above analysis, the FIEOD algorithm shows excellent results on most numerical data sets.

The same analyses can be done with Table 7, and we can find that the FIEOD algorithm is roughly superior to other algorithms on nominal feature data sets.

5.3.2. Comparison by ROC

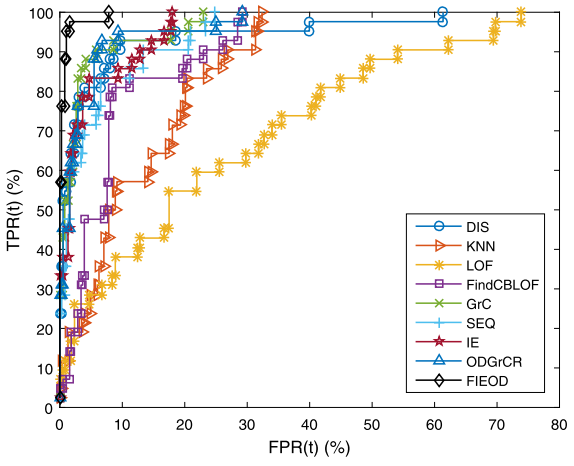
Figs. 1-3 show ROC curves on the hybrid, numerical, and nominal feature data sets, respectively.

From Fig. 1, we can also get the validity of the FIEOD algorithm more vividly. For example, for the Cred data set, it can be observed from Fig. 1(a) that the FIEOD algorithm is closest to the upper left corner of the first quadrant and the area under the curve is the largest. For the Germ data set, it can be observed from the Fig. 1(b) that the FIEOD algorithm is also closest to the upper left corner of the first quadrant and the area under the curve is the largest.

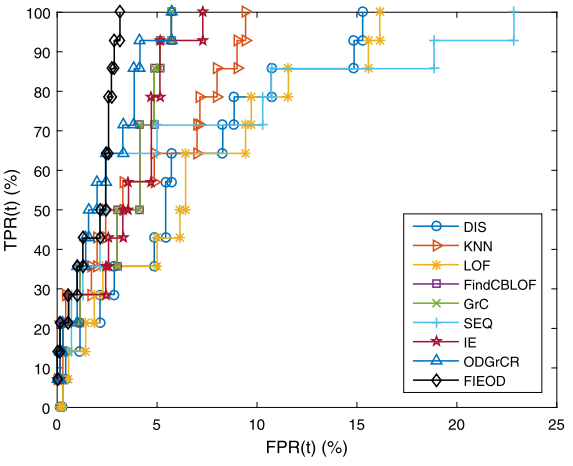
What's more, DIS, kNN, and LOF algorithms show a relatively worse effect on the mixed feature data sets. The reason is that because the nominal features are replaced with different integer values before the experiment, which may result in changes in the data structure. In addition, the discretization of numerical features in the mixed feature data set also significantly affects the precision of FindCBLOF, GrC, SEQ, IE, and ODGrCR algorithms. However, for the FIEOD algorithm, replacement and discretization do not need to be done, so that more real information of the data can be retained. Therefore, the FIEOD algorithm has relatively superior detection results.

In general, the FIEOD algorithm performs better on most hybrid data sets than other algorithms, so it can be effectively applied to hybrid feature data sets.

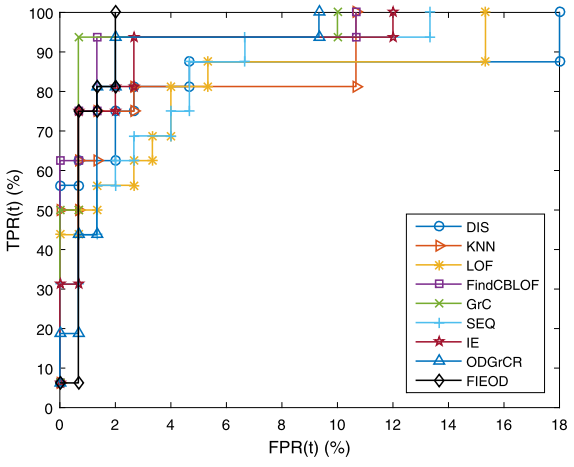
From Fig. 2, it can be seen that the FIEOD algorithm is the closest to the upper left corner of the first quadrant on Diab, Iris, and Pima data sets, and the area under the curve is the largest, which shows that the FIEOD algorithm is significantly better than other algorithms. On Iono, Page, Wbc, Wdbc, and Yeast data sets, the FIEOD algorithm is closer to the upper left corner of the first quadrant, and the area under the curve is larger. Its performance is comparable to or slightly weaker than other algorithms.



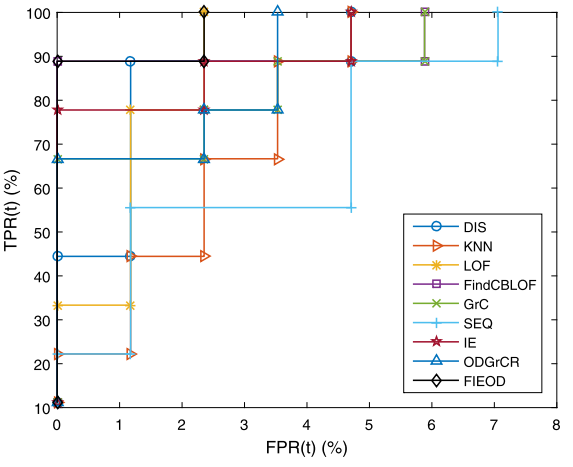
(a) Cred



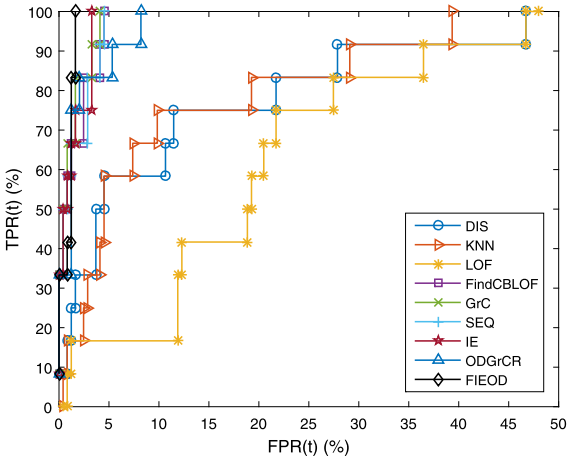
(b) Germ



(c) Heart

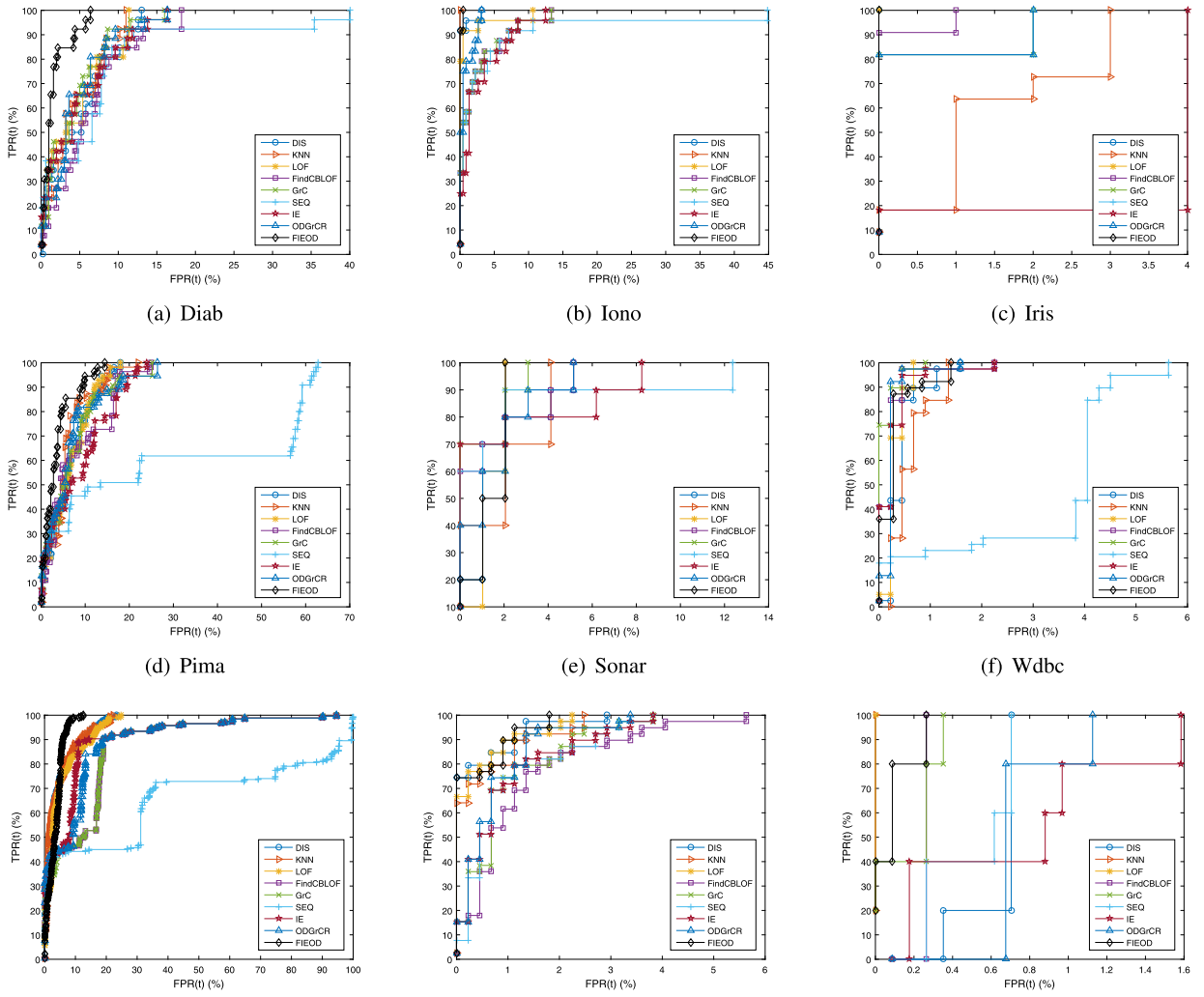


(d) Hepa



(e) Horse

Fig. 1. The ROC curves for mixed attribute data sets.



(a) Diab (b) Iono (c) Iris

(d) Pima (e) Sonar (f) Wdbc

(g) Page (h) Wbc (i) Yeast

Fig. 2. The ROC curves for numeric feature data sets.

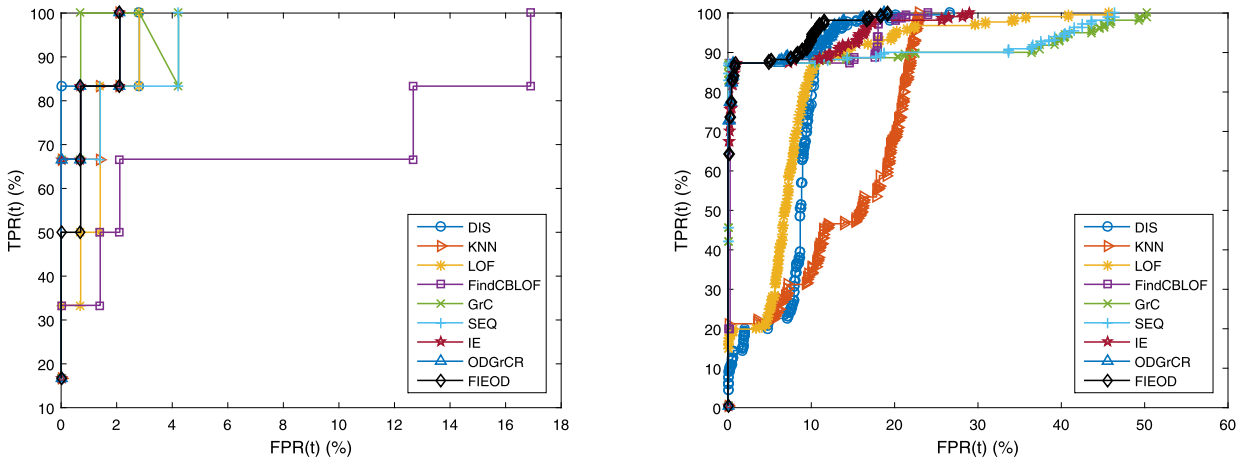
Likewise, the discretization of numerical features also significantly affects the accuracy of FindCBLOF, GrC, SEQ, IE, and ODGrCR algorithms. However, the FIEOD algorithm can retain more real information because it does not require discretization.

Therefore, the FIEOD algorithm is also applicable to numeric feature data sets.

Fig. 3 gives the ROC curves on the nominal feature data sets. The analysis results indicate that the FIEOD algorithm is also applicable to the nominal feature data set.

5.4. Parametric sensitivity analyses

The threshold δ plays an important role in the FIEOD algorithm. It can be used as a parameter to control the granularity of data analysis. The experimental results of the FIEOD algorithm will be affected by δ. To this end, the experiment of parameter changes is further carried out. When t takes a different value, the curves that $R(t)$ changes with δ on mixed feature data sets are depicted in Fig. 4. Taking the Dred and Germ data sets as examples, the corresponding analyses are as follows.



(a) Lymp (b) Mush

Fig. 3. The ROC curves for nominal feature data sets.

1) On the Cred data set, when $t = 19$, $R(t)$ first increases with δ. Then we get the maximum value at $\delta = 1$. Finally start to reduce until the end tends to balance.
2) For the Germ data set, when $t = 36$, $R(t)$ first increases with the parameter δ. Then it takes the maximum value at $\delta = 0.6$ and balances after three reductions. Finally, $R(t)$ grows to a steady state.

At the same time, we can also see that the maximum value can be obtained under multiple δ for different data sets. For example, on the Heart data set, $R(t)$ takes the maximum value at $t \in [3.9, 4.1]$. For the Horse data set, $R(t)$ gets the maximum value for $t \in [1.4, 3.1]$.

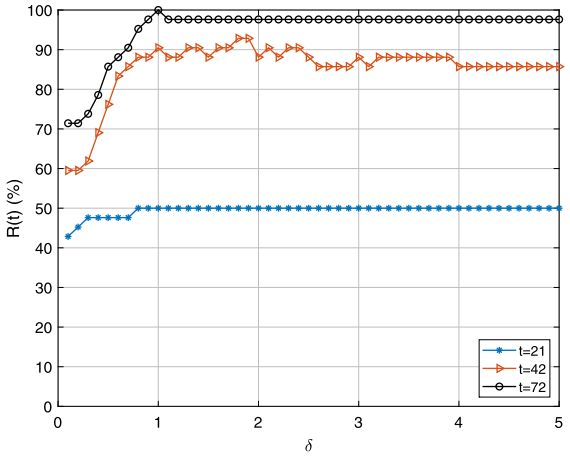
Fig. 5 shows that the precision $R(t)$ varies with the parameter δ when t is different for numerical feature data sets. Some corresponding analyses are as follows.

1) For the Diab, Iris, Poma, Sonar, Wdbc, and Wbc data sets, its $R(t)$ first increases with δ, and then roughly levels off.
2) In Fig. 5(b), $R(t)$ remains unchanged as δ increases. And then it starts to decrease.
3) On the Page data set, $R(t)$ first increases as δ increases. Then it starts to decrease until it flattens out.
4) Through Figs. 5(c) and 5(i), it can be observed that with the increase of δ, the $R(t)$ fluctuates greatly.

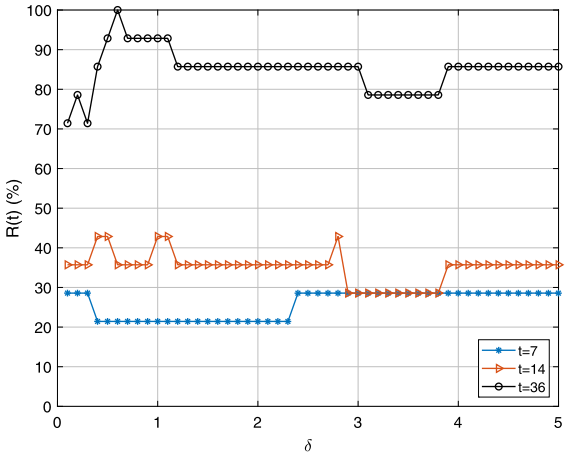
Obviously, we can also see that $R(t)$ can get the maximum value on multiple parameters. For example, on the Iono data set, $R(t)$ adopts the maximum value at $t \in [0.8, 1.2]$. When $t \in [1, 3.2]$ or $t \in [3.9, 4.1]$ or $t = 4.4$, the maximum value can be obtained for $R(t)$ on the Sonar data set.

6. Conclusion

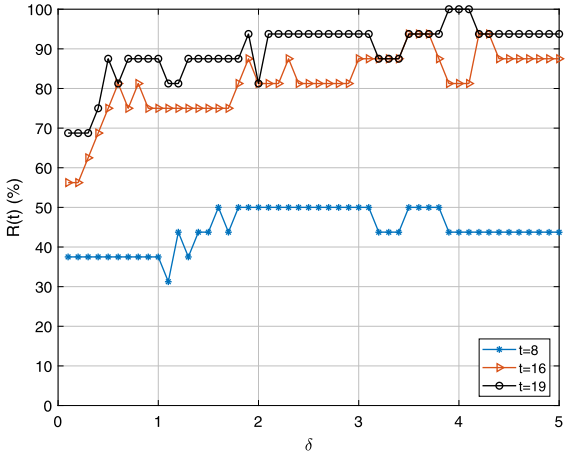
Based on the concept of fuzzy information entropy, this paper proposes a method of outlier detection based on fuzzy information entropy. This method is not only suitable for nominal feature data, but also for numerical and hybrid feature data. In terms of this method, we design the corresponding FIEOD algorithm, and carry out experiments to compare with some existing outlier detection algorithms on the UCI standard data sets. Experimental results demonstrate that our outlier detection is effective. The method of outlier detection based on fuzzy information entropy has not been studied before. The research of this paper expands the application scope of fuzzy information theory in data mining, and opens up a new application field for fuzzy information theory. With data constantly changing, we will consider the research of dynamic outlier detection in the future work.



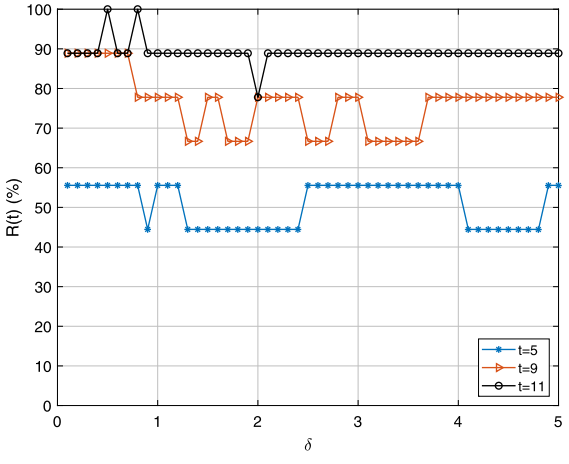
(a) Cred



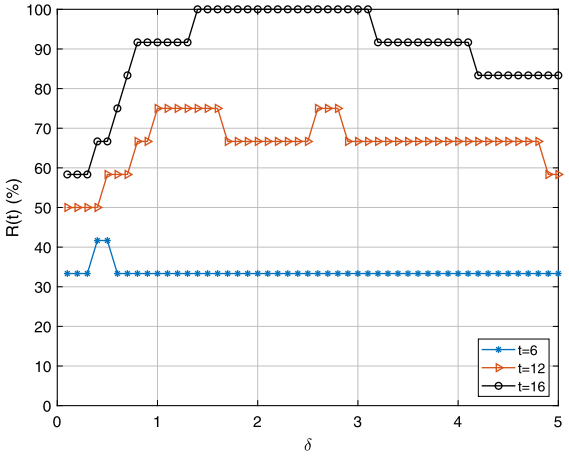
(b) Germ



(c) Heart

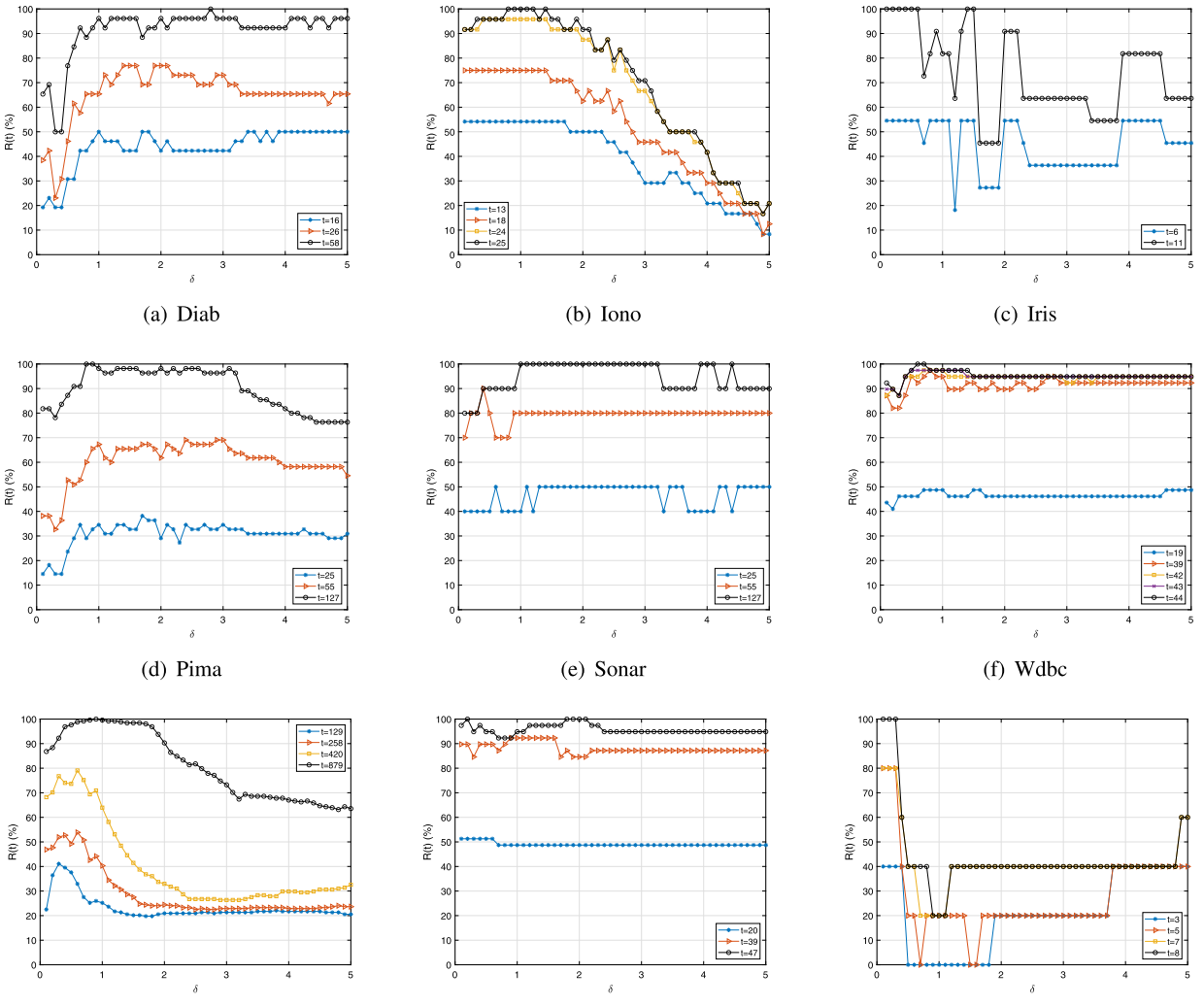


(d) Hepa



(e) Horse

Fig. 4. The change of $R(t)$ with the variation of δ on the mixed feature data sets.



(a) Diab (b) Iono (c) Iris

(d) Pima (e) Sonar (f) Wdbc

(g) Page (h) Wbc (i) Yeast

Fig. 5. The change of $R(t)$ with the variation of δ on the numerical feature data sets.

Declaration of competing interest

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

Acknowledgements

This work was supported by the National Natural Science Foundation of China (61976182, 61572406, and 61602327), the Key Techniques of Integrated Operation and Maintenance for Urban Rail Train Dispatching Control System based on Artificial Intelligence (2019YFH0097), and the Applied Basic Research Programs of Science and Technology Department of Sichuan Province (2019YJ0084).

References

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

[2] N. Saeed, T.Y. Al-Naffouri, M.-S. Alouini, Outlier detection and optimal anchor placement for 3-d underwater optical wireless sensor network localization, IEEE Trans. Commun. 67 (1) (2019) 611–622.
[3] Y.J. Chen, J.H. Pu, J.H. Du, Y. Wang, Z. Xiong, Spatial–temporal traffic outlier detection by coupling road level of service, IET Intell. Transp. Syst. 13 (6) (2019) 1016–1022.
[4] B. Wang, Z.Z. Mao, Outlier detection based on a dynamic ensemble model: applied to process monitoring, Inf. Fusion 51 (2019) 244–258.
[5] P.J. Rousseeuw, A.M. Leroy, Robust Regression and Outlier Detection, vol. 589, John Wiley & Sons, 2005.
[6] T. Johnson, I. Kwok, R. Ng, Fast computation of 2-dimensional depth contours, in: International Conference on Knowledge Discovery and Data Mining, 1998, pp. 224–228.
[7] E.M. Knox, R.T. Ng, Algorithms for mining distancebased outliers in large datasets, in: Proceedings of the International Conference on Very Large Data Bases, Citeseer, 1998, pp. 392–403.
[8] M.M. Breunig, H.P. Kriegel, R.T. Ng, J. Sander, Lof: identifying density-based local outliers, SIGMOD Rec. 29 (2) (2000) 93–104.
[9] Z.Y. He, X.F. Xu, S.C. Deng, Discovering cluster-based local outliers, Pattern Recognit. Lett. 24 (9–10) (2003) 1641–1650.
[10] E.M. Knorr, R.T. Ng, V. Tucakov, Distance-based outliers: algorithms and applications, VLDB J. 8 (3) (2000) 237–253.
[11] H.M. Chen, T.R. Li, Y. Cai, C. Luo, H. Fujita, Parallel attribute reduction in dominance-based neighborhood rough set, Inf. Sci. 373 (2016) 351–368.
[12] J.H. Dai, Q.H. Hu, H. Hu, D.B. Huang, Neighbor inconsistent pair selection for attribute reduction by rough set approach, IEEE Trans. Fuzzy Syst. 26 (2) (2018) 937–950.
[13] D.G. Chen, X.X. Zhang, X.Z. Wang, Y.J. Liu, Uncertainty learning of rough set-based prediction under a holistic framework, Inf. Sci. 463 (2018) 129–151.
[14] H.M. Chen, T.R. Li, X. Fan, C. Luo, Feature selection for imbalanced data based on neighborhood rough sets, Inf. Sci. 483 (2019) 1–20.
[15] B.B. Sang, H.M. Chen, T.R. Li, W.H. Xu, H. Yu, Incremental approaches for heterogeneous feature selection in dynamic ordered data, Inf. Sci. 541 (2020) 457–501.
[16] F. Jiang, Y.F. Sui, C.G. Cao, Outlier detection using rough set theory, in: International Workshop on Rough Sets, Fuzzy Sets, Data Mining, and Granular-Soft Computing, Springer, 2005, pp. 79–87.
[17] T.T. Nguyen, Outlier detection: an approximate reasoning approach, in: International Conference on Rough Sets and Intelligent Systems Paradigms, 2007, pp. 495–504.
[18] F. Jiang, Y.F. Sui, C.G. Cao, A rough set approach to outlier detection, Int. J. Gen. Syst. 37 (5) (2008) 519–536.
[19] Y.M. Chen, D.Q. Miao, R.Z. Wang, Outlier detection based on granular computing, in: International Conference on Rough Sets and Current Trends in Computing, 2008, pp. 283–292.
[20] F. Shaari, A.A. Bakar, A.R. Hamdan, Outlier detection based on rough sets theory, Intell. Data Anal. 13 (2) (2009) 191–206.
[21] Z.X. Xue, S.Y. Liu, Rough-based semi-supervised outlier detection, in: 2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery, vol. 1, IEEE, 2009, pp. 520–523.
[22] F. Jiang, Y.F. Sui, C.G. Cao, Some issues about outlier detection in rough set theory, Expert Syst. Appl. 36 (3) (2009) 4680–4687.
[23] P. Yang, Q.S. Zhu, Finding key attribute subset in dataset for outlier detection, Knowl.-Based Syst. 24 (2) (2011) 269–274.
[24] F. Jiang, Y.F. Sui, C.G. Cao, A hybrid approach to outlier detection based on boundary region, Pattern Recognit. Lett. 32 (14) (2011) 1860–1870.
[25] A. Albanese, S.K. Pal, A. Petrosino, Rough sets, kernel set, and spatiotemporal outlier detection, IEEE Trans. Knowl. Data Eng. 26 (1) (2014) 194–207.
[26] F. Jiang, Y.M. Chen, Outlier detection based on granular computing and rough set theory, Appl. Intell. 42 (2) (2015) 303–322.
[27] F. Maciá-Pérez, J.V. Berna-Martinez, A.F. Oliva, M.A.A. Ortega, Algorithm for the detection of outliers based on the theory of rough sets, Decis. Support Syst. 75 (2015) 63–75.
[28] F. Jiang, G.Z. Liu, J.W. Du, Y.F. Sui, Initialization of k-modes clustering using outlier detection techniques, Inf. Sci. 332 (2016) 167–183.
[29] F. Jiang, H. Zhao, J. Du, Y. Xue, Y. Peng, Outlier detection based on approximation accuracy entropy, Int. J. Mach. Learn. Cybern. (2018) 1–17.
[30] D. Dubois, H. Prade, Rough fuzzy sets and fuzzy rough sets, Int. J. Gen. Syst. 17 (2–3) (1990) 191–209.
[31] D. Dubois, H. Prade, Putting rough sets and fuzzy sets together, in: Intelligent Decision Support, Springer, 1992, pp. 203–232.
[32] C.E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. 27 (3) (1948) 379–423.
[33] R.R. Yager, Entropy measures under similarity relations, Int. J. Gen. Syst. 20 (4) (1992) 341–358.
[34] E. Hernández, J. Recasens, A reformulation of entropy in the presence of indistinguishability operators, Fuzzy Sets Syst. 128 (2) (2002) 185–196.
[35] Q.H. Hu, D.R. Yu, Z.X. Xie, J.F. Liu, Fuzzy probabilistic approximation spaces and their information measures, IEEE Trans. Fuzzy Syst. 14 (2) (2006) 191–201.
[36] C.Z. Wang, Y. Huang, M.W. Shao, D.G. Chen, Uncertainty measures for general fuzzy relations, Fuzzy Sets Syst. 360 (2019) 82–96.
[37] R. Mesiar, J. Rybárik, Entropy of fuzzy partitions: a general model, Fuzzy Sets Syst. 99 (1) (1998) 73–79.
[38] C. Bertoluzza, V. Doldi, G. Naval, Uncertainty measure on fuzzy partitions, Fuzzy Sets Syst. 142 (1) (2004) 105–116.
[39] J.S. Mi, Y. Leung, W.Z. Wu, An uncertainty measure in partition-based fuzzy rough sets, Int. J. Gen. Syst. 34 (1) (2005) 77–90.
[40] J.H. Dai, J.L. Chen, Feature selection via normative fuzzy information weight with application in biological data classification, Appl. Soft Comput. (2020) 106–299.
[41] Y.H. Qian, Q. Wang, H.H. Cheng, J.Y. Liang, C.Y. Dang, Fuzzy-rough feature selection accelerator, Fuzzy Sets Syst. 258 (2015) 61–78.
[42] Y.Y. Yang, D.G. Chen, H. Wang, E.C. Tsang, D.L. Zhang, Fuzzy rough set based incremental attribute reduction from dynamic data with sample arriving, Fuzzy Sets Syst. 312 (2017) 66–86.
[43] J.H. Dai, H. Hu, W.Z. Wu, Y.H. Qian, D.B. Huang, Maximal-discernibility-pair-based approach to attribute reduction in fuzzy rough sets, IEEE Trans. Fuzzy Syst. 26 (4) (2018) 2174–2187.

[44] J.T. Wang, Y.H. Qian, F.J. Li, J.Y. Liang, W.P. Ding, Fusing fuzzy monotonic decision trees, IEEE Trans. Fuzzy Syst. 28 (5) (2020) 887–900.
[45] J.K. Chen, J.S. Mi, Y.J. Lin, A graph approach for fuzzy-rough feature selection, Fuzzy Sets Syst. 391 (2020) 96–116.
[46] C.Z. Wang, Y. Wang, M.W. Shao, Y.H. Qian, D.G. Chen, Fuzzy rough attribute reduction for categorical data, IEEE Trans. Fuzzy Syst. 28 (5) (2020) 818–830.
[47] F. Jiang, Y.F. Sui, C.G. Cao, An information entropy-based approach to outlier detection in rough sets, Expert Syst. Appl. 37 (9) (2010) 6338–6344.
[48] Y.M. Chen, D.Q. Miao, H.Y. Zhang, Neighborhood outlier detection, Expert Syst. Appl. 37 (12) (2010) 8745–8749.
[49] Z. Yuan, S. Feng, Outlier detection algorithm based on neighborhood value difference metric, J. Comput. Appl. 38 (7) (2018) 81–85.
[50] Z. Yuan, X.Y. Zhang, S. Feng, Sequence-based mixed attribute outlier detection in neighborhood rough sets, J. Chin. Comput. Syst. 39 (6) (2018) 1317–1322.
[51] Z. Yuan, X.Y. Zhang, S. Feng, Hybrid data-driven outlier detection based on neighborhood information entropy and its developmental measures, Expert Syst. Appl. 112 (2018) 243–257.
[52] D.S. Yeung, D. Chen, E.C. Tsang, J.W. Lee, W. Xizhao, On the generalization of fuzzy rough sets, IEEE Trans. Fuzzy Syst. 13 (3) (2005) 343–361.
[53] Q.H. Hu, D.R. Yu, W. Pedrycz, D.G. Chen, Kernelized fuzzy rough sets and their applications, IEEE Trans. Knowl. Data Eng. 23 (11) (2010) 1649–1667.
[54] X.H. Zhang, D.W. Fei, J.H. Dai, Fuzzy Mathematics and Rough Set Theory, Tsinghua University Press, 2013.
[55] Q.H. Hu, S. An, X. Yu, D.R. Yu, Robust fuzzy rough classifiers, Fuzzy Sets Syst. 183 (1) (2011) 26–43.
[56] D. Dheeru, E. Karra Taniskidou, UCI Machine Learning Repository, http://archive.ics.uci.edu/ml, 2017.
[57] S. Ramaswamy, R. Rastogi, K. Shim, Efficient algorithms for mining outliers from large data sets, SIGMOD Rec. 29 (2) (2000) 427–438.
[58] G.O. Campos, A. Zimek, J. Sander, R.J. Campello, B. Micenková, E. Schubert, I. Assent, M.E. Houle, On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study, Data Min. Knowl. Discov. 30 (4) (2016) 891–927.
[59] F. Eibe, A.H. Mark, H.W. Ian, The WEKA Workbench, Morgan Kaufmann, 2016.
[60] C.C. Aggarwal, Outlier Analysis, Springer, 2016.