Applied Soft Computing Journal 165 (2024) 112070

Contents lists available at ScienceDirect

Applied Soft Computing

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

Consistency-guided semi-supervised outlier detection in heterogeneous data using fuzzy rough sets

Baiyang Chen [a], Zhong Yuan [a,*], Dezhong Peng [a,d], Xiaoliang Chen [b], Hongmei Chen [c]

[a] College of Computer Science, Sichuan University, Chengdu, 610065, China
[b] Department of Computer Science and Operations Research, University of Montreal, Montreal, QC H3C3J7, Canada
[c] School of Computing and Artificial Intelligence, Southwest Jiaotong University, Chengdu, 611756, China
[d] Sichuan Newstrong UHD Video Technology Co., Ltd., Chengdu, 610095, China

ARTICLE INFO

Keywords:
Semi-supervised outlier detection
Fuzzy Rough Sets
Heterogeneous data
Label-informed fuzzy similarity relation
Classification consistency

ABSTRACT

Outlier detection aims to find objects that behave differently from the majority of the data. Semi-supervised detection methods can utilize the supervision of partial labels, thus reducing false positive rates. However, most of the current semi-supervised methods focus on numerical data and neglect the heterogeneity of data information. In this paper, we propose a consistency-guided outlier detection algorithm (COD) for heterogeneous data with the fuzzy rough set theory in a semi-supervised manner. First, a few labeled outliers are leveraged to construct label-informed fuzzy similarity relations. Next, the consistency of the fuzzy decision system is introduced to evaluate attributes' contributions to knowledge classification. Subsequently, we define the outlier factor based on the fuzzy similarity class and predict outliers by integrating the classification consistency and the outlier factor. The proposed algorithm is extensively evaluated on 20 freshly proposed datasets. Experimental results demonstrate that COD is better than or comparable with the leading outlier detectors.

1. Introduction

Outlier detection (OD), also known as anomaly detection or novelty detection, is a process that identifies patterns or data points in a dataset that deviate from what is expected or normal. OD allows for the identification of unusual or unexpected behavior that may have important implications for the system being studied. Therefore, it has numerous applications in various domains such as fraud detection [1], software defect prediction [2], industry control [3], medical anomaly detection [4], etc.

Since outliers are often rare events and labeled data are frequently insufficient, many OD methods are designed to be unsupervised [5–15]. However, in some cases, it may be possible to obtain a limited amount of labeled outliers to guide the detection process. Hence, a number of semi-supervised detection algorithms [16–22] have been proposed in the last decades. However, most of them are mainly designed to deal with numerical data and neglect the heterogeneity of information. Real-world applications generally include heterogeneous attributes (referred to as mixed attribute data or simply mixed data) where the attributes of objects take various types of values [23]. For instance, in fraud detection, the data may contain the gender and age of a customer, as well as the date and amount of a transaction. In this case, the attribute gender is nominal (categorical), the age is integer-valued, the transaction date is time-valued, and the transaction amount is real-valued (numerical). Detection of outliers in such a scenario may be more practical. However, many detectors inherently regard mixed data as numerical, such that a handful of convenient measures can be utilized to extract features or data structures. Unfortunately, this may bring in extra properties that do not exist. Taking the attribute gender as an example, one may assign a number 0 to the male, and 1 to the female. Then we would obtain that the female is greater than the male, which does not make any sense and is likely to harm the performance of detection algorithms.

The fuzzy rough set (FRS) theory [24], which unifies rough sets and fuzzy sets, is a popular mathematical model for processing data with uncertainty. It has been successfully applied to various domains including feature selection [23,25] and outlier detection [26,27] in mixed data over the last years. In FRS, the concepts of lower and upper approximations from classic Rough Sets are adapted to work within the framework of fuzzy logic, leading to fuzzy lower and upper approximations. This integration offers three advantages: (1) It provides a flexible framework for modeling data with uncertainty and

* Corresponding author.
E-mail addresses: chenby@stu.scu.edu.cn (B. Chen), yuanzhong@scu.edu.cn (Z. Yuan), pengdz@scu.edu.cn (D. Peng), chexiaol@iro.umontreal.ca (X. Chen), hmchen@swjtu.edu.cn (H. Chen).

https://doi.org/10.1016/j.asoc.2024.112070
Received 1 May 2023; Received in revised form 19 July 2024; Accepted 30 July 2024
Available online 8 August 2024

imprecision by representing the boundaries of a concept with fuzzy membership functions. This allows for partial membership where elements can belong to a set to varying degrees, thus more precisely reflecting real-world scenarios. (2) It facilitates the direct processing of various data types, including numerical values, categorical variables, symbols, and more, without the necessity of data type transformation, thereby preserving the data's intrinsic diversity for subsequent analyses. (3) It enhances reasoning and decision-making in ambiguous situations, enabling the classification of objects even when their attribute values do not precisely match the criteria of a specific class, offering a robust approach to dealing with data ambiguity. Therefore, FRS holds significant potential for identifying outliers in heterogeneous data with uncertainty and imprecision.

The central idea of this paper can be described as follows. Since the goal of outlier detection is to find the minority objects whose behavior is abnormal in data, any object less similar to the others has a higher probability of being an outlier. From the perspective of FRS, the similarity is reflected by the fuzzy similarity class, which can describe how similar an instance is to others. Therefore, this paper employs the fuzzy similarity class to characterize outliers and uses the unlabeled data to construct the outlier factor for each object based on fuzzy similarity classes. Then, with a few labeled data, we formulate a fuzzy decision system, where the unlabeled data constitute the conditional attributes (i.e., features) and the class labels form the decision attribute. In this context, given a set of conditional attributes, we can take some measure (e.g., dependency) from the FRS theory to evaluate its classification consistency with the decision. If an attribute set has greater consistency, then it is of high quality for producing a more separable classification. Finally, we combine the outlier factor and the classification consistency to predict outliers in the dataset.

With the above ideas, we propose a consistency-guided outlier detection algorithm (COD) based on FRS in a semi-supervised manner. The contributions of this paper include:

- This paper introduces a novel label-informed fuzzy similarity relation for the representation of heterogeneous data. It enables a downstream task-guided approach to determine the optimal fuzzy radius for mixed data from a wide range of applications.
- We propose to characterize outliers with the fuzzy similarity class and use the classification consistency to guide the scoring of outliers, which improves the accuracy and efficiency of outlier detection with very limited labeled data.
- To our best knowledge, we are the first to propose a novel FRS-based outlier detection model with a semi-supervised approach in mixed data. It may have the potential to advance the application of the FRS theory in real-world scenarios.
- Extensive experiments on various types of public datasets demonstrate that the proposed algorithm is better than or comparable with the state-of-the-art methods.

2. Related works

Depending on the availability of labeled data on which the algorithm relies, outlier detection methods are broadly classified into unsupervised [28], semi-supervised [29] and supervised detection algorithms [30]. This paper focuses on semi-supervised outlier detection (SSOD) methods for tabular data that is organized in rows and columns of a table. In particular, we mainly investigate the rank-based, representation learning-based, active learning-based, and reinforcement learning-based detection approaches. It is notable that some methods may belong to more than one group or may combine elements from different groups.

Rank-based SSOD methods typically leverage a limited number of labeled samples to train a ranking model that scores the outlier degree of all objects. For instance, Pang et al. (2019) [18] introduce an end-to-end deep framework (DevNet) to learn outlier scores, and incorporate a Gaussian prior and deviation loss to detect outliers. Inspired by DevNet, the model PReNet [19] ranks the outliers by ordinal regression without involving any assumptions about the probability distribution of outlier scores. Besides, Zhou et al. (2022) [22] propose a weakly supervised outlier detection model (FEAWAD), which attempts to utilize a deep autoencoder to fit the normal data. The resulting representations are then leveraged to facilitate outlier score learning. Lately, Stradiotti et al. (2024) [31] introduce the Semi-Supervised Isolation Forest (SSIF) that integrates labeled and unlabeled data within a probabilistic framework, enhancing performance by leveraging informative split distributions and computing anomaly scores based on labeled instances within tree leaves.

Representation learning-based detection models can be viewed as an indirect way of learning anomaly scores, and they usually improve unsupervised representation learning by partially available labeled data. The earlier solution (e.g., OE [32], XGBOD [17]) uses multiple unsupervised detection algorithms to learn useful representations and appends their output anomaly scores to the original features for training a supervised classifier. Recent deep learning-based SSOD methods employ end-to-end frameworks to extract anomaly-oriented representations for discovering outliers. One such model is DeepSAD by Ruff et al. (2020) [20]. It extends the unsupervised outlier detection algorithm DeepSVDD [33] so that it can make use of labeled data. Moreover, DeepSAD penalizes the inverse of the distances of outliers' representation such that they are projected to further away from the center of a hypersphere where normal data reside. Following DeepSAD, Huang et al. (2021) [21] leverage the mutual information between data and their representations as well as an entropy measure, and devise an encoder–decoder–encoder architecture to optimize a KL-divergence-based objective function for semi-supervised outlier detection. The other detection method REPEN [16] uses a ranking-based approach to learn feature representations of ultra-high dimension data and optimizes the detection method based on random distances. A small number of labeled outliers are leveraged to refine the sampling for the training triplets, and help the model learn anomaly-oriented representations. Recently, Zhu et al. (2024) [34] propose Anomaly Heterogeneity Learning (AHL) to simulate diverse heterogeneous anomaly distributions from limited anomaly examples, allowing for detecting both seen and unseen anomalies more effectively than traditional closed-set approaches. Another line of research based on representation learning involves the use of generative adversarial networks (GAN) to improve model training and/or data augmentation. Among them, Tian et al. (2022) [35] propose the anomaly-aware bidirectional GAN model (AA-BiGAN) based on the BiGAN [36] architecture. It uses labeled outliers to learn a probability distribution that is guaranteed to assign low-density values to the collected anomalies. Li et al. (2022) [37] integrates multiple GANs to realize reference distribution construction and data augmentation for detecting both discrete and grouped anomalies. Liu et al. (2024) [38] address challenges in semi-supervised anomaly detection by regularizing the latent representations learned by deep generative models using mutual information maximization, thereby improving separation between normal and abnormal samples.

In addition to the above-mentioned approaches, some researchers have adopted active learning and reinforcement learning for SSOD. Active learning-based methods focus on acquiring more informative labeled data with the aid of humans. In one of these works, Gornitz et al. [39] propose a semi-supervised anomaly detection (SSAD) algorithm that follows the unsupervised learning paradigm, and additionally devises an active learning strategy that simply chooses borderline points for labeling. Another important method that leverages active learning is active anomaly discovery (AAD) [40]. It greedily selects the most likely abnormal samples for labeling and maximizes the number of true outliers under a query budget. Reinforcement learning-based SSOD models usually consider the discovery of outliers as a sequential decision process, where human feedback is utilized to help identify more outliers. For instance, the model Meta-AAD [41] employs deep

reinforcement learning to optimize a meta-policy to select the most appropriate samples for manual labeling, and optimizes the number of outliers discovered throughout the querying process. Unlike Meta-AAD, Pang et al. [42] try to explore unlabeled data without human help. They design an anomaly-biased simulation environment to enable an RL-based model to discover known and unknown outliers and develop a deep Q-learning-based detection model DPLAN. Chen et al. (2024) [43] introduce the Deep Anomaly Detection and Search (DADS) that integrates reinforcement learning to model a Markov decision process for searching possible anomalies in unlabeled data, effectively leveraging both labeled and unlabeled data to enhance detection performance.

3. Preliminaries

In information processing systems, data is usually stored in a table (also called an information system, information table, etc.), where every row corresponds to an object (sample), and every column denotes an attribute (feature). This section reviews some fundamental concepts of FRS that help understand the subsequent sections of this paper.

Definition 1. An information system is a tuple (U, A), where $U = \{x_1, x_2, \dots, x_n\}$ is the set of objects, also referred to as the universe of discourse, and $A = \{a_1, a_2, \dots, a_m\}$ is the set of attributes that every object has.

Definition 2. Given an information system (U, A). If $\widetilde{X}$ is a map from U to $[0, 1]$, then $\widetilde{X}$ is a fuzzy set on U, i.e. $\widetilde{X} : U \to [0, 1]$.

$\forall x_i \in U$, $\widetilde{X}(x_i)$ is the membership of x_i to $\widetilde{X}$, or the membership function of $\widetilde{X}$. The fuzzy set is often denoted by $\widetilde{X} = \left(\widetilde{X}(x_1), \widetilde{X}(x_2), \dots, \widetilde{X}(x_n)\right)$, and the fuzzy cardinality of $\widetilde{X}$ is computed by $|\widetilde{X}| = \sum_i \widetilde{X}(x_i)$.

Definition 3. Let U be a set of objects, a fuzzy relation $\widetilde{R}$ on U is defined as a family of fuzzy sets $\widetilde{R} : U \times U \to [0, 1]$.

Some commonly employed operations of fuzzy relations are listed as follows.

(1) $\widetilde{R}_1 = \widetilde{R}_2 \Leftrightarrow \forall (x_i, x_j) \in U \times U, \widetilde{R}_1(x_i, x_j) = \widetilde{R}_2(x_i, x_j)$;
(2) $\widetilde{R}_1 \subseteq \widetilde{R}_2 \Leftrightarrow \forall (x_i, x_j) \in U \times U, \widetilde{R}_1(x_i, x_j) \leq \widetilde{R}_2(x_i, x_j)$;
(3) $(\widetilde{R}_1 \cup \widetilde{R}_2)(x_i, x_j) = \max\left\{\widetilde{R}_1(x_i, x_j), \widetilde{R}_2(x_i, x_j)\right\}$;
(4) $(\widetilde{R}_1 \cap \widetilde{R}_2)(x_i, x_j) = \min\left\{\widetilde{R}_1(x_i, x_j), \widetilde{R}_2(x_i, x_j)\right\}$.

$\forall (x_i, x_j) \in U \times U$, the membership $\widetilde{R}(x_i, x_j)$ expresses the degree to which x_i has a relation $\widetilde{R}$ with x_j. A fuzzy relation $\widetilde{R}$ on U is usually denoted by a fuzzy relation matrix $M(\widetilde{R}) = (r_{ij})_{n \times n}$, where $r_{ij} = \widetilde{R}(x_i, x_j)$. $\forall x_1, x_2, x_3 \in U$, if a fuzzy relation $\widetilde{R}$ meets: (1) reflexive: $\widetilde{R}(x_1, x_1) = 1$, (2) symmetric: $\widetilde{R}(x_1, x_2) = \widetilde{R}(x_2, x_1)$, then $\widetilde{R}$ is also called a fuzzy similarity relation.

In Fuzzy Rough Sets, the tuple $(U, \widetilde{R})$ is called a fuzzy approximation space [24], where the approximation of a fuzzy set can be defined.

Definition 4. Let $(U, \widetilde{R})$ be a fuzzy approximation space, and $\widetilde{X}$ is a fuzzy set on U, the lower approximation of $\widetilde{X}$ is a fuzzy set. Its membership function is defined as

$$\underline{\widetilde{R}}\widetilde{X}(x_i) = \inf_{x_j \in U} \max\left\{1 - \widetilde{R}(x_i, x_j), \widetilde{X}(x_j)\right\}. \tag{1}$$

The fuzzy lower approximations $\underline{\widetilde{R}}\widetilde{X}$ of $\widetilde{X}$ are used to express the degrees of elements certainly belonging to $\widetilde{X}$. Specifically, when the fuzzy set to be approximated is crisp, Eq. (1) can be rewritten as

$$\underline{\widetilde{R}}X(x_i) = \inf_{x_j \notin X} \{1 - \widetilde{R}(x_i, x_j)\}. \tag{2}$$

Given a fuzzy approximation space $(U, \widetilde{R})$, the fuzzy similarity relation $\widetilde{R}$ can induce a generalized fuzzy partition of U, i.e., a set of fuzzy similarity classes which are constructed by collecting a group of fuzzy targets.

Definition 5 ([44]). The generalized fuzzy partition of U induced by a fuzzy similarity relation $\widetilde{R}$ is defined as

$$U/\widetilde{R} = \left\{[x_i]_{\widetilde{R}}\right\}_{x_i \in U}, \tag{3}$$

where $[x_i]_{\widetilde{R}} = \left(r_{i1}^{\widetilde{R}}, r_{i2}^{\widetilde{R}}, \dots, r_{in}^{\widetilde{R}}\right)$ is a fuzzy similarity class containing x_i.

Each fuzzy similarity class $[x_i]_{\widetilde{R}}$ is a fuzzy set, which clearly describes how similar the object x_i is to all objects in U. The family of $U/\widetilde{R}$ forms a fuzzy concept system on U [45]. Obviously, $[x_i]_{\widetilde{R}}(x_j) = \widetilde{R}(x_i, x_j) = r_{ij}^{\widetilde{R}}$. If $\widetilde{R}(x_i, x_j) = 1$, then it suggests that x_j certainly belongs to $[x_i]_{\widetilde{R}}$; If $\widetilde{R}(x_i, x_j) = 0$, then x_j definitely does not belong to $[x_i]_{\widetilde{R}}$. The fuzzy cardinality of $[x_i]_{\widetilde{R}}$ is calculated by

$$\left|[x_i]_{\widetilde{R}}\right| = \sum_{j=1}^{n} \widetilde{R}(x_i, x_j). \tag{4}$$

We can easily obtain $1 \leq \left|[x_i]_{\widetilde{R}}\right| \leq n$. The cardinality of $[x_i]_{\widetilde{R}}$ reflects the overall similarity of the object x_i to others in U based on the knowledge $\widetilde{R}$.

In the application of Fuzzy Rough Sets, the information system (U, A) is expressed as a fuzzy information system. If $A = C \cup d$ and $C \cap d = \emptyset$, the fuzzy information system is also referred to as a fuzzy decision system defined as

Definition 6. A fuzzy decision system is a tuple $(U, C \cup d)$, where $U = \{x_1, x_2, \dots, x_n\}$ is the set of objects, C represents the conditional attributes and d denotes the decision attribute.

Example 1. Let $U = \{x_1, x_2, x_3\}$ be the set of objects, $A = \{a_1, a_2\}$ is the set of attributes, and the fuzzy similarity relation matrices on U induced from the attribute a_1 and a_2 respectively are

$$M(\widetilde{a}_1) = \begin{pmatrix} 1 & 0 & 0.9 \\ 0 & 1 & 0 \\ 0.9 & 0 & 1 \end{pmatrix}, M(\widetilde{a}_2) = \begin{pmatrix} 1 & 0.8 & 0 \\ 0.8 & 1 & 0.9 \\ 0 & 0.9 & 1 \end{pmatrix}.$$

Then, the fuzzy similarity class $[x_1]_{\widetilde{a}_1}$ generated by $\widetilde{a}_1$ containing x_1 is $(1, 0, 0.9)$, and $\left|[x_1]_{\widetilde{a}_1}\right| = 1.9$. Similarly, $[x_2]_{\widetilde{a}_1} = (0, 1, 0)$, $\left|[x_2]_{\widetilde{a}_1}\right| = 1$, $[x_3]_{\widetilde{a}_1} = (0.9, 0, 1)$, $\left|[x_3]_{\widetilde{a}_1}\right| = 1.9$.

Let $\widetilde{X} = (0.2, 0.5, 0.8)$ be a fuzzy set on U, then the fuzzy lower approximation $\underline{\widetilde{a}_1}\widetilde{X}$ of $\widetilde{X}$ with regard to $\widetilde{a}_1$ can be computed using Eq. (1) as

$$\max\left\{1 - \widetilde{a}_1(x_1, x_j), \widetilde{X}(x_j)\right\}\Big|_{x_j = x_1} = \max\{0, 0.2\} = 0.2,$$

$$\max\left\{1 - \widetilde{a}_1(x_1, x_j), \widetilde{X}(x_j)\right\}\Big|_{x_j = x_2} = \max\{1, 0.5\} = 1,$$

$$\max\left\{1 - \widetilde{a}_1(x_1, x_j), \widetilde{X}(x_j)\right\}\Big|_{x_j = x_3} = \max\{0.1, 0.8\} = 0.8.$$

Then we have $\underline{\widetilde{a}_1}\widetilde{X}(x_1) = \inf_{x_j \in U} \max\left\{1 - \widetilde{a}_1(x_1, x_j), \widetilde{X}(x_j)\right\} = 0.2$. Similarly, we can obtain $\underline{\widetilde{a}_1}\widetilde{X}(x_2) = 0.5$, $\underline{\widetilde{a}_1}\widetilde{X}(x_3) = 0.2$. Therefore, $\underline{\widetilde{a}_1}\widetilde{X} = (0.2, 0.5, 0.2)$.

Having illustrated the basic calculations in FRS through the preceding example, we will now proceed to an in-depth exploration of our methods.

4. Methodology

This section presents the consistency-guided outlier detection (COD) method in detail. First, we leverage a few labeled outliers to construct label-informed fuzzy similarity relations, which enable a flexible and adaptable representation of heterogeneous data. Next, the consistency of the fuzzy decision system is introduced to assess attributes' contributions to knowledge classification. Subsequently, we define the outlier factor based on the fuzzy similarity class and predict outlier scores by integrating the classification consistency and the outlier factor. Finally, a threshold is determined for binary outlier predictions. The overall framework of COD is illustrated in Fig. 1.

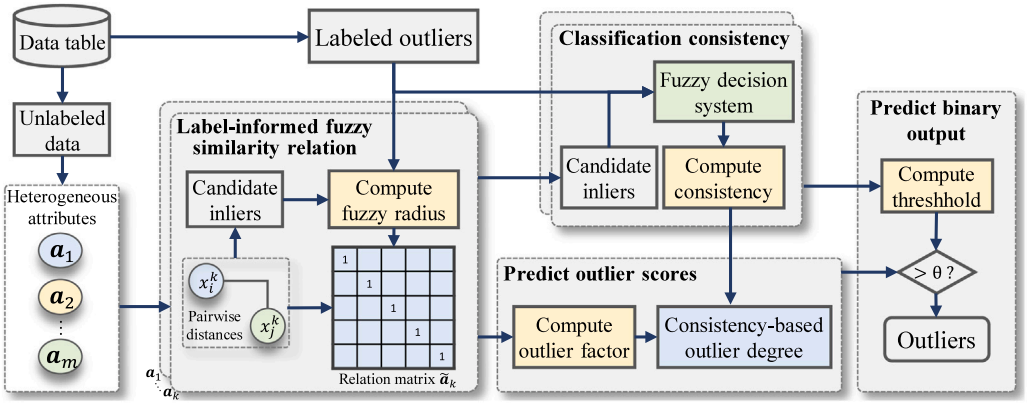

Fig. 1. Overall framework of COD.

4.1. Label-informed fuzzy similarity relation

This part begins with data preprocessing, and then constructs the fuzzy similarity relation for each attribute with the guidance of labeled data, which is termed the label-informed fuzzy similarity relation.

As real-world datasets are usually heterogeneous, this paper recognizes two broad types of data, i.e., the numerical data including real-valued and integer-valued numbers, and the nominal data including categorical and symbolic data. Since the magnitude and dimension of raw numerical values are diverse, the numerical data are first normalized into the interval of $[0, 1]$ by the min–max normalization during data preprocessing.

In order to jointly represent both numerical and nominal values, we adopt a hybrid fuzzy membership function [44] to construct the fuzzy similarity relation between each pair of objects in the datasets. Let f_i^k be the value of attribute a_k for object x_i, and $d_{ij}^k = \left|f_i^k - f_j^k\right|$ is the distance between x_i and x_j on the attribute a_k, the fuzzy similarity relation $\widetilde{a}_k(x_i, x_j)$ between x_i and x_j induced by the attribute a_k is calculated by

$$\widetilde{a}_k(x_i, x_j) = \begin{cases} 1, & \text{if } f_i^k = f_j^k, a_k \text{ is nominal;} \\ 0, & \text{if } f_i^k \neq f_j^k, a_k \text{ is nominal;} \\ 1 - d_{ij}^k, & \text{if } d_{ij}^k \leq \lambda_k, a_k \text{ is numerical;} \\ 0, & \text{if } d_{ij}^k > \lambda_k, a_k \text{ is numerical;} \end{cases} \tag{5}$$

where $\lambda_k \in [0, 1]$ is the fuzzy radius, which is usually determined empirically [26] or by some heuristic rules (e.g., coverage and specificity [46]) in unsupervised scenarios. However, in the label-informed setting, we are able to design a downstream task-guided approach to determine their values with a few labeled objects. The optimal λ_k is derived through the maximization of the following equation.

$$\lambda_k = \arg\max_{\lambda} \frac{\sum_{x_i \in U^-} \left|[x_i]_{\widetilde{a}_k}^{\lambda}\right|}{|U^-| \cdot |U|} - \frac{\sum_{x_j \in U^+} \left|[x_j]_{\widetilde{a}_k}^{\lambda}\right|}{|U^+| \cdot |U|}, \tag{6}$$

where U^- and U^+ denote the subset of the negative instances and the positive instances in U, respectively. The cardinality of the fuzzy similarity classes $[x_i]$ reflects the overall similarity of a negative object x_i to all others in U based on the knowledge $\widetilde{a}_k$, and likewise, the cardinality of $[x_j]$ reflects that of a positive object x_j. The objective is to find an optimal value of λ by balancing the average cardinality of labeled outliers and normal points. Given the assumption that outliers are less similar to others than the majority of data, the objects in U^- exhibit higher similarity to each other. Consequently, the cardinality $\|[x_i]_{\widetilde{a}_k}^{\lambda}\|$ should be greater. Conversely, outliers in U^+ tend to be less similar to the majority of U, resulting in smaller cardinality $\|[x_j]_{\widetilde{a}_k}^{\lambda}\|$.

The label-informed fuzzy similarity relation allows for modeling heterogeneous data unitedly and efficiently. However, in case normal objects in U^- are not explicitly identified, it is necessary to designate certain instances as candidate negative instances. Previous works [16, 19,22] simply treat unlabeled data as inliers under the assumption that outliers are often rare events. However, this approach may not work well in cases where the outlier contamination level is high (i.e., the proportion of outliers in the unlabeled data), potentially introducing noise. This paper introduces a more practical solution for selecting candidate inliers.

Let d_{ij}^k be the pairwise distance between x_i and x_j with respect to the attribute a_k. Then the average distance of x_i to others in U is $D_i^k = \frac{\sum_{x_j \in U} d_{ij}^k}{|U|}$. We first sort the objects in ascending order based on their average distances as $D_{i_1}^k \leq D_{i_2}^k \leq \cdots \leq D_{i_{|U|}}^k$. Then, the set of candidate inliers is defined as

$$U^- = \left\{ x_{i_k} \mid k = 1, 2, \ldots, N_- \right\}. \tag{7}$$

This selection process aims to identify the top N_- objects with the greatest average distance, reflecting their greater similarity to others, and designates them as candidate inliers.

Given an attribute subset $B \subseteq C$, the fuzzy similarity relation $\widetilde{B}$ on U can be derived from Eq. (5) using conjunction operation [47] as

$$\widetilde{B}(x_i, x_j) = \min_{a_k \in B} \widetilde{a}_k(x_i, x_j). \tag{8}$$

Proposition 1. *For any attribute set $B, P \subseteq C$, if $B \subseteq P$, then $\widetilde{P} \subseteq \widetilde{B}$.*

Proposition 1 expresses the inclusion relation between fuzzy relations induced by attributes: the more attributes adopted, the smaller (fine-grained) the fuzzy relation. Notably, given any x_i and x_j in U, if the fuzzy relation $\widetilde{P}$ is sufficiently fine-grained, the degree to which x_i has a relation $\widetilde{P}$ with x_j approaches 0. In other words, every object in U will be distinct from each other under the knowledge of $\widetilde{P}$.

4.2. Classification consistency

In a fuzzy decision system $(U, C \cup d)$, both the condition attributes in C and the decision attribute d can induce a fuzzy partition, which inherently indicates a knowledge classification. To evaluate attributes' contributions to knowledge classification, we introduce classification consistency based on the notion of fuzzy dependency in the following.

Definition 7 ([48]). The fuzzy dependency $\gamma_C(d)$ of the decision attribute d on the conditional attributes C is

$$\gamma_C(d) = \frac{\left|\bigcup_{\widetilde{X} \in U/\widetilde{d}} \underline{\widetilde{C}}\widetilde{X}\right|}{|U|}. \tag{9}$$

It is obvious that $0 < \gamma_C(d) \leq 1$, and $\gamma_C(d) = 1$ means that the decision d can be approximated accurately by the attributes C. In the

application of outlier detection, all objects in U are partitioned into two imbalanced crisp classes by the decision attribute d as

$$U/d = \left\{[x^-]_d, [x^+]_d\right\}, \tag{10}$$

where $[x^-]_d$ and $[x^+]_d$ represent the two crisp equivalence classes i.e., the negative class and the positive class, respectively. In this context, given an conditional attribute subset $B \subseteq C$, the fuzzy dependency between B and d is computed by

$$\gamma_B(d) = \frac{|POS_B(d)|}{|U|} = \frac{\left|\underline{\widetilde{B}}[x^-]_d \cup \underline{\widetilde{B}}[x^+]_d\right|}{|U|}. \tag{11}$$

To address the imbalance problem in the task of outlier detection, we define the consistency based on the fuzzy dependency through balancing the weights between the two classes.

Definition 8. Given an attribute subset $B \subseteq C$, the consistency $\xi_B(d)$ of the decision attribute d on the conditional attributes B is defined as

$$\xi_B(d) = \frac{\left|\underline{\widetilde{B}}[x^-]_d\right|}{|U^-|} + \frac{\left|\underline{\widetilde{B}}[x^+]_d\right|}{|U^+|}. \tag{12}$$

Proposition 2. *$\forall B, P \subseteq C$, if $B \subseteq P$, then $\xi_B(d) \leq \xi_P(d)$.*

Proof. Given $B \subseteq P$, according to Proposition 1, we have $\widetilde{B} \supseteq \widetilde{P}$. Therefore, $\forall x_i, x_j \in U$, we can obtain $1 - \widetilde{B}(x_i, x_j) \leq 1 - \widetilde{P}(x_i, x_j)$. Let $[x^-]_d$ and $[x^+]_d$ be the negative class and the positive class of U, by Definition 4 and Eq. (2), we have $\underline{\widetilde{B}}[x^-]_d \subseteq \underline{\widetilde{P}}[x^-]_d$, and $\underline{\widetilde{B}}[x^+]_d \subseteq \underline{\widetilde{P}}[x^+]_d$. Further, $\frac{\left|\underline{\widetilde{B}}[x^-]_d\right|}{|U^-|} \leq \frac{\left|\underline{\widetilde{P}}[x^-]_d\right|}{|U^-|}$, and $\frac{\left|\underline{\widetilde{B}}[x^+]_d\right|}{|U^+|} \leq \frac{\left|\underline{\widetilde{P}}[x^+]_d\right|}{|U^+|}$. Therefore, $\xi_B(d) \leq \xi_P(d)$. □

The above proposition suggests that the more conditional attributes employed, the higher the degree of classification consistency of the fuzzy decision system will be.

4.3. Consistency-guided outlier degree

In a fuzzy decision system, a fuzzy similarity relation $\widetilde{B}$ can induce a fuzzy partition of U, i.e., a set of fuzzy similarity classes. Each fuzzy similarity class $[x_i]_{\widetilde{B}}$ is a fuzzy set, which clearly describes how similar the object x_i is to all objects in U. As the aim of outlier detection is to find the minority objects whose behavior is abnormal, any object which is relatively less similar to the other objects, has a higher probability of being an outlier. Therefore, we define the outlier factor based on the fuzzy similarity class.

Definition 9. Let $(U, C \cup d)$ be a fuzzy decision system, $\forall B \in C$, the outlier factor generated by the attribute subset B is defined as

$$OF_B(x_i) = 1 - \frac{1}{|U|}\left|[x_i]_{\widetilde{B}}\right|. \tag{13}$$

In the above definition, choosing an appropriate attribute set B is crucial for constructing an outlier factor. To achieve this, it is essential to have a quality metric for candidate attributes, and one such measure is the classification consistency defined in the previous section. If an attribute subset has greater classification consistency, then it is of higher quality for producing a more separable partition. Ideally, all the 2^m subsets of conditional attributes should be considered for constructing outlier factors. But this procedure is exhaustive and may be too costly and practically prohibitive even for a medium-sized m. Moreover, as reflected by Proposition 1, a fine-grained fuzzy relation tends to treat each object as a distinct category, which does not help to distinguish which objects are anomalous. Therefore, following the previous works [26,49,50], this paper also adopts a straightforward solution in that each singleton attribute is taken to construct a similarity class.

On this basis, we can predict the outlier score for each instance by integrating the OFs of every attribute and its corresponding consistency in the form of a weighted summation.

Table 1
A data table for patients.

U	a_1	a_2	a_3	d	a_1	a_2	a_3	d
x_1	♂	38	62.5	Negative	♂	0	0.503	Negative
x_2	♂	47	72.3	Negative	♂	0.692	1	Negative
x_3	♀	51	52.6	Positive	♀	1	0	Positive
x_4	♀	44	65.6	Negative	♀	0.462	0.66	Negative

Definition 10. Let $(U, C \cup d)$ be a fuzzy decision system, $\forall x_i \in U$, the consistency-guided outlier degree COD of x_i is defined as

$$\mathrm{COD}(x_i) = \frac{1}{|C|}\sum_{a_k \in C} OF_{a_k}(x_i) \cdot \xi_{a_k}(d). \tag{14}$$

To illustrate the computation process of the aforementioned method more clearly, a concrete example is provided in the following.

Example 2. Let $(U, C \cup d)$ be a data table for patients with heterogeneous attributes (as shown in the left of Table 1), and $U = \{x_1, x_2, x_3, x_4\}$ denotes the persons, $C = \{a_1, a_2, a_3\}$ represents the set of conditional attributes for each patient, where a_1, a_2 and a_3 represent the *gender*, the *age* and the *weight*, respectively. d is the decision attribute indicating whether the person has been diagnosed with a disease.

We first transform the numerical attributes a_2 and a_3 into the same magnitude by min–max normalization, the results are shown in the right of Table 1. Then we construct the label-informed fuzzy similarity relations in the following.

From the decision attribute d, we have the set of positive instances $U^+ = \{x_3\}$ and the set of negative instances $U^- = \{x_1, x_2, x_4\}$. For the categorical attribute a_1, we can easily obtain the fuzzy similarity relation matrix by Eq. (5) as

$$M(\widetilde{a}_1) = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix}.$$

For the numerical attributes, we first calculate the fuzzy radius by Eq. (6) as $\lambda_2 \approx 0.539$ for a_2 and $\lambda_3 \approx 0.498$ for a_3. Then, we have the fuzzy similarity relation matrix by Eq. (5) as

$$M(\widetilde{a}_2) = \begin{pmatrix} 1 & 0 & 0 & 0.539 \\ 0 & 1 & 0.692 & 0.769 \\ 0 & 0.692 & 1 & 0.462 \\ 0.539 & 0.769 & 0.462 & 1 \end{pmatrix},$$

$$M(\widetilde{a}_3) = \begin{pmatrix} 1 & 0.503 & 0 & 0.843 \\ 0.503 & 1 & 0 & 0.660 \\ 0 & 0 & 1 & 0 \\ 0.843 & 0.660 & 0 & 1 \end{pmatrix}.$$

Next, we compute the classification consistency for each attribute. From Table 1, we have the positive and negative class induced by d as $[x^+]_d = (0, 0, 1, 0)$, $[x^-]_d = (1, 1, 0, 1)$. Then we obtain the following low approximations (refer to Example 1):

$$\underline{\widetilde{a}_1}[x^+]_d = (0, 0, 0, 0), \underline{\widetilde{a}_1}[x^-]_d = (1, 1, 0, 0),$$

$$\underline{\widetilde{a}_2}[x^+]_d = (0, 0, 0.308, 0), \underline{\widetilde{a}_2}[x^-]_d = (1, 0.308, 0, 0.539),$$

$$\underline{\widetilde{a}_3}[x^+]_d = (0, 0, 1, 0), \underline{\widetilde{a}_3}[x^-]_d = (1, 1, 0, 1).$$

By Definition 8, we have $\xi_{a_1}(d) = \frac{2}{3} + \frac{0}{1} \approx 0.667, \xi_{a_2}(d) = \frac{1.846}{3} + \frac{0.308}{1} \approx 0.923, \xi_{a_3}(d) = \frac{3}{3} + \frac{1}{1} = 2$.

In the following, we calculate the outlier factor for each instance. By Definition 9, we have

$$OF_{a_1}(x_1) = 1 - \frac{1}{|U|}|[x_1]_{\widetilde{a_1}}| = 1 - \frac{1}{4} \times 2 = 0.5,$$

$$OF_{a_2}(x_1) = 1 - \frac{1}{4} \times 1.539 \approx 0.615,$$

$$OF_{a_3}(x_1) = 1 - \frac{1}{4} \times 2.346 \approx 0.414.$$

Similarly, we can compute the outlier factor for the other instances.

Finally, the outlier degree of each instance is obtained by integrating the OFs of all attributes and their corresponding consistencies as

$$\begin{aligned} COD(x_1) &= \frac{1}{|C|}\left(OF_{a_1}(x_1)\cdot\xi_{a_1}(d) + OF_{a_2}(x_1)\cdot\xi_{a_2}(d) + OF_{a_3}(x_1)\cdot\xi_{a_3}(d)\right) \\ &= \frac{1}{3}(0.5\times 0.667 + 0.615\times 0.923 + 0.414\times 2) \approx 0.576. \end{aligned}$$

Similarly, we have $COD(x_2) \approx 0.536$, $COD(x_3) \approx 0.753$, $COD(x_4) \approx 0.455$.

4.4. Binary outlier classification

After associating each object with an outlier score, a binary output is generated to indicate whether an input instance is an outlier or not. This is achieved by determining a threshold θ as follows.

Definition 11. Let θ be a real-valued threshold, $\forall x_i \in U$, if the outlier score $COD(x_i) > \theta$, then x_i is regarded as a consistency-guided outlier.

In the label-informed scenario, where labeled outliers are available, the optimal threshold value is obtained adaptively. A straightforward approach is to utilize the smallest outlier degree of the labeled outliers. Let U^+ be the set of labeled outliers, and the optimal threshold θ^* can be determined as the minimum outlier degree of the labeled outliers:

$$\theta^* = \min_{x_i \in U^+} COD(x_i) \tag{15}$$

The whole procedure of the proposed outlier detection algorithm COD is illustrated in Algorithm 1. COD begins by finding an optimal fuzzy radius for each attribute. It then constructs label-informed fuzzy similarity relations and calculates the classification consistency for each attribute. Subsequently, COD computes the outlier factor and the outlier degree for each input instance. Finally, a threshold value is computed, which is used to predict a binary output designating outliers. Since COD's primary operation involves computing the distance between each pair of instances for each attribute, the worst time complexity for COD is $O(mn^2)$, where m represents the number of attributes and n represents the number of instances in the dataset.

Algorithm 1: Consistency-guided outlier detection

Input: A set of n objects U with m conditional attributes C, a subset of labeled outliers $U^+ \subseteq U$.
Output: Outlier set O

```
O ← ∅;
for Each a_k ∈ C do
    Compute fuzzy radius λ_k by Eq. (6);
    Construct fuzzy relation matrix M(a_k) using Eq. (5);
    Compute the consistency ξ_{a_k}(d) by Definition 8;
end
for Each x_i ∈ U do
    for Each a_k ∈ C do
        Compute outlier factor OF_{a_k}(x_i) using Eq. (13);
    end
    Compute outlier degree COD(x_i) using Eq. (14);
end
Compute the threshold θ by Eq. (15);
for Each x_i ∈ U do
    if COD(x_i) > θ then
        O ← O ∪ {x_i};
    end
end
return O.
```

5. Experiments

This section conducts comparison experiments with some state-of-the-art methods to evaluate our proposed algorithm on 20 public datasets, which range over various fields including images, medics, biologics, etc. All the datasets and source codes are publicly available online.[1]

5.1. Datasets

The experimental datasets include 4 categorical, 5 mixed, and 11 numerical datasets. The number of samples in a dataset ranges from 226 to 11 183, and the ratio of anomalies varies between 0.8% and 35.9%. The details of the datasets are provided in Table 2.

[1] https://github.com/ChenBaiyang/COD.

5.2. Experiment settings

Following previous works [29,30,51], we adapt the unsupervised detection algorithms for predicting the new coming data, i.e., a fixed number (e.g., 5) of labeled outliers (unlabeled data are taken as negative instances) are used as validation set for tuning their hyper-parameters. For SSOD, we use various levels of supervision data for training and the rest for testing. Specifically, the number of labeled outliers ranges from 5 to 30. For COD, the number of candidate negative instances N_- is set to 100. Each experiment is repeated 10 times independently and the average is reported.

5.3. Comparison methods and settings

We select the following 10 baseline methods to evaluate the comparison performances of the detection algorithms, including 5 unsupervised and 5 semi-supervised ones. Most of them, except for WFRDA, are implemented by the AD library PyOD [30,51] and WSAD [29].

Unsupervised detection methods:

- IForest (2008) [52]: An ensemble model which isolates instances and constructs outlier scores using binary trees. The number of base estimators is tuned among {5, 10, 50, 100, 500}.
- SOD (2009) [53]: A subspace learning-based model that recognizes outliers by constructing subspaces of high-dimensional data. The parameter number of neighbors is found among {2, 3, 5, 10, 20, 50}.
- DeepSVDD (2018) [33]: A deep learning-based model that learns embeddings of objects by minimizing the volume of a data hypersphere. The hyper-parameter number of epochs is selected in {20, 50, 100, 200}.
- ECOD (2022) [49]: A probabilistic model that predicts the empirical cumulative distribution of each attribute with the assumption that anomalies lie at the tails of the distribution. This model is parameter-free.
- WFRDA (2023) [26]: A FRS-based algorithm that uses fuzzy-rough density and fuzzy entropy to describe the outlier degree of data. The parameter fuzzy radius is chosen from 0.1 to 2.0 with a stepsize of 0.1.

Table 2
The statistics of the datasets used in the experiments.

No.	Datasets	# Samples	# Attributes	# Outlier	% Outlier	Category	Data type
1	Annealing	798	38	42	5.3%	Physical	Mixed
2	Annthyroid	7200	6	534	7.4%	Healthcare	Numerical
3	Arrhythmia	452	279	66	14.6%	Medical	Mixed
4	Audiology	226	69	57	25.2%	Healthcare	Categorical
5	Breast	286	9	85	29.7%	Medical	Categorical
6	Breastw	683	9	239	35.0%	Healthcare	Numerical
7	Cardio	1831	21	176	9.6%	Healthcare	Numerical
8	CreditA	425	15	42	9.9%	Commercial	Mixed
9	Ionosphere	351	32	126	35.9%	Oryctognosy	Numerical
10	Mammography	11 183	6	260	2.3%	Healthcare	Numerical
11	Mushroom1	4429	22	221	5.0%	Botany	Categorical
12	Mushroom2	4781	22	573	12.0%	Botany	Categorical
13	Musk	3062	166	97	3.2%	Chemistry	Numerical
14	Optdigits	5216	64	150	2.9%	Image	Numerical
15	PageBlocks	5393	10	510	9.5%	Document	Numerical
16	Sick	3613	29	72	2.0%	Medical	Mixed
17	Thyroid	9172	28	74	0.8%	Medical	Mixed
18	Waveform	3443	21	100	2.9%	Physical	Numerical
19	Wilt	4819	5	257	5.3%	Botany	Numerical
20	Yeast	1484	8	507	34.2%	Biology	Numerical

Semi-supervised approaches:

- REPEN (2018) [16]: A deep learning-based model that uses a ranking approach to learn feature representations of ultra-high dimensional data.
- DevNet (2019) [18]: A rank-based method that introduces an end-to-end deep framework and incorporates a Gaussian prior and deviation loss to detect outliers. The confidence margin parameter a is set to 5.
- DeepSAD (2020) [20]: A deep learning-based model that penalizes the inverse of the distances of outliers such that they are projected away from normal data. The balancing parameter η in the objective function is fixed to 1.
- FEAWAD (2022) [22]: A rank-based method that utilizes a deep autoencoder to fit the normal data.
- PReNet (2023) [19]: A rank-based method that scores the outliers by ordinal regression without involving any assumptions about outlier scores.

5.4. Evaluation metrics

We assess the detection methods by two popular metrics: AUC-ROC (Area Under Curve-Receiver Operating Characteristic) and AUC-PR (Area Under Curve-Precision Recall). They are both computed based on the outlier scores of objects. A larger AUC-ROC or AUC-PR value indicates better detection performance. It is worth noting that AUC-ROC measures the overall ability of a model to discriminate between positive (outliers) and negative (inliers) instances and is less sensitive to class imbalance. AUC-PR, on the other hand, focuses on the performance of the model in capturing true outliers while minimizing false positives, and is more informative when dealing with imbalanced datasets with rare outliers. The paired Wilcoxon signed-rank test [19] is employed to assess the statistical significance of COD in comparison to its rival methods.

5.5. Experimental results and analysis

5.5.1. Comparative performance of detectors in real-world datasets

This part examines the detection performances in real-world datasets where there are a lot of unlabeled data and a few labeled outliers are available. Therefore, only 5 labeled outliers (other supervision levels are further examined in the following part) are taken to train the detectors, and all the unlabeled data are taken as the test set. The labeled outlier ratio of the experimental datasets ranges from 0.87% to 11.9%. To ensure a fair comparison with unsupervised methods, we use the same training set with semi-supervised models to tune hyper-parameters, and the same testing set to evaluate the unsupervised algorithms.

The overall experimental results on 20 datasets are listed in Tables 3 and 4 (@ 5 labeled outliers). One can observe that COD achieves the best AUC-ROC on 8 datasets, including Annealing, Annthyroid, Arrhythmia, Audiology, Mushroom2, Musk, Sick and Thyroid. The improvements of COD over 10 comparison methods are statistically significant at the 95% confidence level. In the case of AUC-PR, COD demonstrates superior performance on 7 datasets, namely Annthyroid, Arrhythmia, Audiology, Breast, Mushroom1, Mushroom2, and Musk. Notably, COD's enhancements over 7 out of 10 comparison methods (with the exceptions of DevNet, FEAWAD, and PReNet) are statistically significant at the 99% confidence level. The results confirm the effectiveness of COD in detecting outliers with very limited labeled data from a wide range of applications.

In addition, COD does not perform as well as other semi-supervised methods on Waveform or Wilt. For instance, DevNet beats COD on Waveform by 14.7% on AUC-ROC, PReNet achieves 59.2% higher AUC-ROC than COD on Wilt. Similar results can be found in terms of AUC-PR. The reason may be that COD's assumption of similarity-based outliers fails in this case where there is an unknown outlier type.

5.5.2. Detection performances in handling heterogeneous data

As real-world datasets often include heterogeneous attributes that take different types of values, this part investigates the model performances in the scenarios with the categorical data types. Table 5 summarizes the average results of detectors on three types of data, i.e., mixed, categorical and numerical data. COD achieves the best average scores with both AUC-ROC and AUC-PR across 5 mixed datasets and 4 categorical datasets, and also performs the best on average of numerical datasets in AUC-ROC. For example, COD performs better than PReNet by 11.6% on mixed datasets in AUC-ROC. This is due to the COD's utilization of FRS to directly process nominal attributes without introducing extra data assumptions, which is pretty effective in outlier detection in mixed attribute data. However, we also observed that COD's average AUC-PR score on numerical datasets is relatively lower than that of PReNet (8.6%) or DevNet (5.8%). This may be due to the different characteristics or distributions of outliers in some cases (e.g., Waveform and Wilt).

5.5.3. Model performances in various levels of label supervision

Since it is hard to obtain labeled outliers in most outlier detection tasks, this part aims to study the detection performances of the comparison methods with respect to various levels of label supervision, and

Table 3
AUC-ROC performances of comparison methods on 20 public datasets (@ 5 labeled outliers). The highest score is marked in bold.

Datasets	IForest	SOD	DeepSVDD	ECOD	WFRDA	DeepSAD	REPEN	DevNet	FEAWAD	PReNet	COD
Annealing	0.802	0.581	0.519	0.795	0.729	0.584	0.751	0.851	0.832	0.847	**0.857**
Annthyroid	0.828	0.792	0.753	0.789	0.637	0.760	0.693	0.935	0.794	0.931	**0.986**
Arrhythmia	0.798	0.731	0.619	0.807	0.755	0.612	0.759	0.596	0.634	0.626	**0.811**
Audiology	0.774	0.563	0.552	0.837	0.834	0.592	0.633	0.619	0.616	0.639	**0.877**
Breast	**0.673**	0.627	0.581	0.659	0.657	0.570	0.649	0.507	0.528	0.493	0.649
Breastw	0.979	0.947	0.924	0.991	**0.992**	0.838	0.987	0.764	0.827	0.704	0.989
Cardio	0.927	0.634	0.522	**0.935**	0.914	0.733	0.889	0.775	0.795	0.776	0.871
CreditA	0.978	0.844	0.865	**0.991**	0.975	0.739	0.948	0.805	0.709	0.813	0.963
Ionosphere	0.843	**0.900**	0.770	0.728	0.784	0.818	0.846	0.637	0.617	0.597	0.785
Mammography	0.866	0.798	0.615	**0.906**	0.839	0.870	0.872	0.896	0.857	0.903	0.855
Mushroom1	0.931	**0.983**	0.542	0.949	0.971	0.942	0.929	0.902	0.910	0.886	0.964
Mushroom2	0.879	0.883	0.628	0.866	0.882	0.904	0.888	0.892	0.828	0.890	**0.976**
Musk	0.999	0.738	0.669	0.959	0.999	0.989	0.999	0.859	0.929	0.973	**1.000**
Optdigits	0.739	0.492	0.667	0.606	0.942	0.869	0.609	0.988	0.994	**0.999**	0.989
PageBlocks	0.904	0.658	0.708	**0.914**	0.868	0.888	0.902	0.800	0.730	0.776	0.889
Sick	0.801	0.698	0.535	0.844	0.837	0.859	0.748	0.837	0.859	0.809	**0.902**
Thyroid	0.662	0.588	0.534	0.579	0.516	0.663	0.648	0.730	0.723	0.735	**0.740**
Waveform	0.684	0.634	0.628	0.601	0.699	0.733	0.668	**0.859**	0.821	0.854	0.749
Wilt	0.489	0.586	0.460	0.394	0.331	0.696	0.341	0.909	0.867	**0.928**	0.583
Yeast	0.430	0.449	0.491	0.445	0.395	0.474	0.383	**0.602**	0.573	0.600	0.446
Average	0.799	0.706	0.629	0.780	0.778	0.757	0.757	0.788	0.772	0.789	**0.844**
p-value	0.012	0.001	0.001	0.01	0.001	0.001	0.002	0.02	0.009	0.027	–

Table 4
AUC-PR performances of comparison methods on 20 public datasets (@ 5 labeled outliers). The highest score is marked in bold.

Datasets	IForest	SOD	DeepSVDD	ECOD	WFRDA	DeepSAD	REPEN	DevNet	FEAWAD	PReNet	COD
Annealing	0.211	0.076	0.091	0.199	0.110	0.085	0.115	0.494	0.482	**0.506**	0.269
Annthyroid	0.325	0.233	0.209	0.269	0.169	0.253	0.198	0.616	0.370	0.592	**0.784**
Arrhythmia	0.436	0.314	0.233	0.448	0.373	0.227	0.373	0.251	0.301	0.279	**0.502**
Audiology	0.572	0.335	0.292	0.649	0.706	0.341	0.400	0.542	0.473	0.515	**0.816**
Breast	0.446	0.385	0.378	0.467	0.462	0.355	0.418	0.320	0.338	0.318	**0.469**
Breastw	0.945	0.861	0.805	**0.984**	0.979	0.777	0.967	0.791	0.839	0.735	0.969
Cardio	0.553	0.221	0.160	0.562	0.516	0.310	0.474	0.543	**0.621**	0.577	0.595
CreditA	0.815	0.459	0.479	**0.916**	0.855	0.331	0.643	0.621	0.501	0.650	0.845
Ionosphere	0.780	**0.896**	0.589	0.638	0.655	0.764	0.777	0.611	0.585	0.599	0.669
Mammography	0.215	0.116	0.056	0.432	0.094	0.314	0.177	0.507	0.275	**0.515**	0.258
Mushroom1	0.432	0.675	0.179	0.483	0.899	0.675	0.411	0.835	0.852	0.830	**0.905**
Mushroom2	0.382	0.610	0.398	0.365	0.673	0.610	0.401	0.835	0.778	0.828	**0.929**
Musk	0.976	0.123	0.307	0.504	0.981	0.893	0.972	0.802	0.902	0.945	**1.000**
Optdigits	0.056	0.026	0.053	0.033	0.366	0.206	0.036	0.968	0.842	**0.985**	0.843
PageBlocks	0.493	0.298	0.339	0.517	0.373	0.555	0.538	**0.580**	0.522	0.568	0.452
Sick	0.056	0.064	0.043	0.063	0.059	0.107	0.051	0.315	**0.355**	0.302	0.267
Thyroid	0.015	0.013	0.011	0.009	0.008	0.018	0.011	0.051	0.046	**0.067**	0.049
Waveform	0.051	0.054	0.070	0.038	0.047	**0.235**	0.053	0.191	0.185	0.210	0.063
Wilt	0.047	0.065	0.045	0.041	0.036	0.106	0.036	0.339	**0.442**	0.393	0.061
Yeast	0.319	0.307	0.328	0.331	0.324	0.319	0.287	**0.431**	0.409	0.427	0.335
Average	0.406	0.307	0.253	0.397	0.434	0.374	0.367	0.532	0.506	0.542	**0.554**
p-value	0.001	0.001	0.001	0.006	0.001	0.001	0.001	0.298	0.139	0.378	–

Table 5
Average performance of detectors w.r.t. data type (@ 5 labeled outliers).

Metric	Data type	IForest	SOD	DeepSVDD	ECOD	WFRDA	DeepSAD	REPEN	DevNet	FEAWAD	PReNet	COD
AUC-ROC	Mixed	0.808	0.689	0.614	0.803	0.762	0.691	0.771	0.764	0.751	0.766	**0.855**
	Categorical	0.814	0.764	0.576	0.828	0.836	0.752	0.775	0.730	0.720	0.727	**0.866**
	Numerical	0.790	0.693	0.655	0.752	0.764	0.788	0.745	0.820	0.800	0.822	**0.831**
AUC-PR	Mixed	0.307	0.185	0.172	0.327	0.281	0.154	0.239	0.346	0.337	0.361	**0.386**
	Categorical	0.458	0.501	0.312	0.491	0.685	0.495	0.407	0.633	0.610	0.623	**0.780**
	Numerical	0.433	0.291	0.269	0.395	0.413	0.430	0.411	0.580	0.545	**0.595**	0.548

investigate the improvements of the semi-supervised methods gained from partial labels compared with the best-unsupervised algorithm. The number of labeled outliers for training is set to vary from 5 to 30, while other unlabeled data are treated as normal instances during training. This setting complies with the previous works [16,18,19] which is viewed as training the detectors with noise. However, we do not deliberately manipulate the value of outlier contamination, since it is more practical in real-world scenarios with various levels of outlier contamination.

Fig. 2 depicts the AUC-ROC performances of the comparison methods w.r.t. multiple levels of supervision. These semi-supervised algorithms tend to improve with the increase of labeled outliers, as more

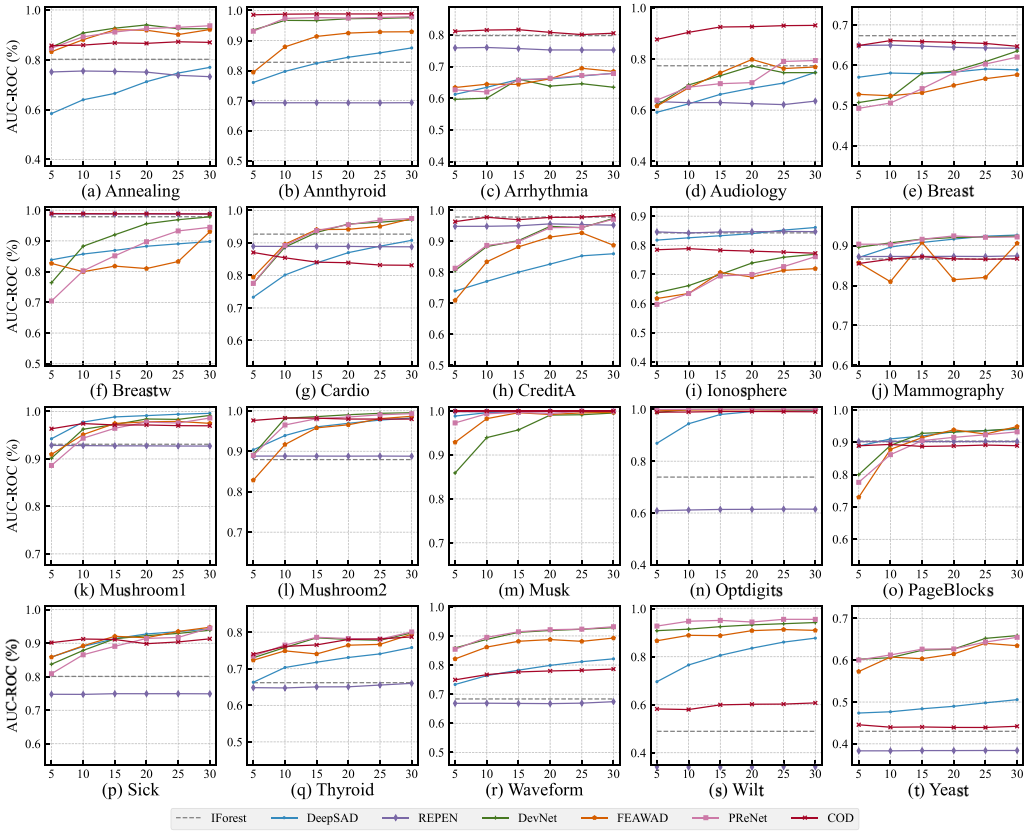

Fig. 2. AUC-ROC scores w.r.t. multiple levels of supervision. The best-unsupervised method is IForest which achieves the highest average score.

labeled data provide more information for training. However, the AUC-ROC of some comparison deep learning-based models falls with more labeled data in some datasets, e.g., FEAWAD and DevNet on Audiology, COD on Cardio. This may be caused by the different distribution of outliers, and they may degrade detection performance when the added data have anomalous behavior that conflicts with other data. However, we also noticed that COD does not improve significantly with an increasing level of supervision in some cases, e.g., COD's improvements on Waveform, Wilt and Yeast are much lower than that of PReNet or FEAWAD. This may be due to the fact that COD leverages only a few unlabeled data (e.g., 100) to learn the outlier behavior, which makes it difficult to obtain a complete distribution of data in some scenarios, thus limiting detection performance.

Furthermore, most semi-supervised detectors perform better than the best unsupervised method IForest when only a few labeled outliers are available. For instance, DeepSVD, DevNet, FEAWAD and achieve higher AUC-ROC scores than IForest on 13 out of 20 datasets at 30 labeled outliers, PReNet (COD) beats IForest on 15 (16) out of 20 datasets at the same level of supervision. COD's advantage is due to its adoption of classification consistencies between attributes and decisions and its use to guide the outlier scoring. This allows COD to integrate unlabeled outlier factors and supervised guidance to construct outlier scores with the very limited labeled data.

5.5.4. Detection performances w.r.t. the number of candidate negative instances

As COD introduces a heuristic method to collect normal objects, the number of candidate negative instances $N\_$ may have an influence on its detection performance. This part investigates the detection performances of COD w.r.t. the number of candidate negative instances. We tune $N\_$ among $\{10, 50, 100, 150, 200\}$ and report both AUC-ROC and AUC-PR at 5 labeled outliers. Fig. 3 depicts the results of COD. As the change of $N\_$, both COD's AUC-ROC and AUC-PR keep stable or slightly fluctuate in most datasets. Notably, the performance on the Musk dataset is distinct, with the two curves completely overlapping across all parameter settings. This occurs because COD achieves maximum scores for both metrics (i.e., 1) throughout all tested values of $N\_$. Therefore, COD is robust to the choice of the number of selected negative instances, and it does not require a large number of negative instances to achieve good performance.

6. Conclusion

This paper proposes a novel consistency-guided outlier detection method (COD) for mixed data with the FRS theory. COD characterizes outliers with the fuzzy similarity class and introduces the classification consistency to guide the scoring of outliers, which improves the accuracy and efficiency of outlier detection with very limited labeled data. The label-informed fuzzy similarity relation designed in COD enables a downstream task-guided approach to determine the optimal fuzzy radius for the representation of heterogeneous data. The proposed algorithm may have the potential to advance the application of the FRS theory in real-world scenarios. Experimental results on various types of public datasets demonstrate the effectiveness of COD in handling mixed attribute data. Our future work will resort to identifying the types of outliers and further utilizing the implicit knowledge from available data to improve the model performance.

CRediT authorship contribution statement

Baiyang Chen: Writing – review & editing, Writing – original draft, Software, Methodology, Formal analysis, Conceptualization. **Zhong Yuan:** Writing – review & editing, Supervision, Project administration, Data curation. **Dezhong Peng:** Writing – review & editing, Supervision, Resources, Funding acquisition. **Xiaoliang Chen:** Writing – review & editing, Validation, Investigation. **Hongmei Chen:** Writing – review & editing, Resources.

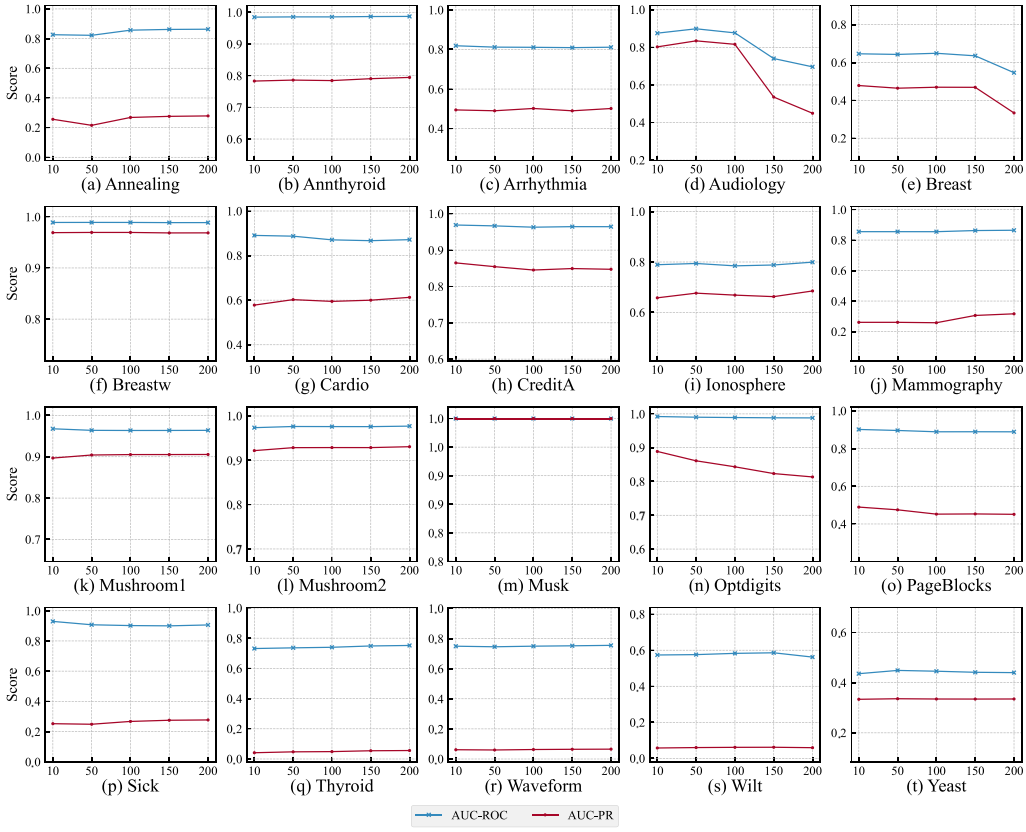

Fig. 3. COD's performances w.r.t. the number of selected negative instances.

Declaration of competing interest

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

Data availability

I have shared the link to data/code in the manuscript.

Acknowledgments

This work was supported by National Natural Science Foundation of China (62306196, 62372315, and 62376230), Sichuan Science and Technology Program, China (2023YFQ0020, 2023ZYD0143, 2024NSFTD0049, 24ZDZX0007, 2024YFHZ0144, 2024YFHZ0089, and MZGC20240057), and the Fundamental Research Funds for the Central Universities, China (YJ202245).

References

[1] T. Pourhabibi, K.-L. Ong, B.H. Kam, Y.L. Boo, Fraud detection: A systematic literature review of graph-based anomaly detection approaches, Decis. Support Syst. 133 (2020) 113303.
[2] F. Jiang, X. Yu, D. Gong, J. Du, A random approximate reduct-based ensemble learning approach and its application in software defect prediction, Inform. Sci. 609 (2022) 1147–1168.
[3] B. Wang, Z. Mao, Outlier detection based on Gaussian process with application to industrial processes, Appl. Soft Comput. 76 (2019) 505–516.
[4] D.M. Hawkins, Identification of Outliers, Springer, 1980.
[5] M.M. Breunig, H.P. Kriegel, R.T. Ng, J. Sander, LOF: identifying density-based local outliers, ACM SIGMOD Rec. 29 (2) (2000) 93–104.
[6] 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.
[7] 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.
[8] K. Zhang, M. Hutter, H. Jin, A new local distance-based outlier detection approach for scattered real-world data, in: Pacific-Asia Conference on Knowledge Discovery and Data Mining, PAKDD, 2009, pp. 813–822.
[9] 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.
[10] F. Jiang, Y.M. Chen, Outlier detection based on granular computing and rough set theory, Appl. Intell. 42 (2) (2015) 303–322.
[11] Z. Li, Y. Zhao, N. Botta, C. Ionescu, X. Hu, COPOD: Copula-based outlier detection, in: 2020 IEEE International Conference on Data Mining, ICDM, 2020, pp. 1118–1123.
[12] Y. Almardeny, N. Boujnah, F. Cleary, A novel outlier detection method for multivariate data, IEEE Trans. Knowl. Data Eng. 34 (9) (2022) 4052–4062.
[13] K. Li, X. Gao, S. Fu, X. Diao, P. Ye, B. Xue, J. Yu, Z. Huang, Robust outlier detection based on the changing rate of directed density ratio, Expert Syst. Appl. 207 (2022) 117988.
[14] Z. Yuan, H.M. Chen, C. Luo, D.Z. Peng, MFGAD: Multi-fuzzy granules anomaly detection, Inf. Fusion 95 (2023) 17–25.
[15] X. Su, Z. Yuan, B. Chen, D. Peng, H. Chen, Y. Chen, Detecting anomalies with granular-ball fuzzy rough sets, Inform. Sci. (2024) 121016.
[16] G. Pang, L. Cao, L. Chen, H. Liu, Learning representations of ultrahigh-dimensional data for random distance-based outlier detection, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, ACM, 2018, pp. 2041–2050.
[17] Y. Zhao, M.K. Hryniewicki, XGBOD: Improving supervised outlier detection with unsupervised representation learning, in: 2018 International Joint Conference on Neural Networks, IJCNN, 2018, pp. 558–565.
[18] G. Pang, C. Shen, A. van den Hengel, Deep anomaly detection with deviation networks, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, ACM, 2019, pp. 353–362.
[19] G. Pang, C. Shen, H. Jin, A. van den Hengel, Deep weakly-supervised anomaly detection, in: Proceedings of the 29th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, ACM, 2023, pp. 1795–1807.
[20] L. Ruff, R.A. Vandermeulen, N. Görnitz, A. Binder, E. Müller, K.-R. Müller, M. Kloft, Deep semi-supervised anomaly detection, in: International Conference on Learning Representations, ICLR, 2020.
[21] C. Huang, F. Ye, P. Zhao, Y. Zhang, Y. Wang, Q. Tian, ESAD: End-to-end semi-supervised anomaly detection, in: The 32nd British Machine Vision Conference, 2021, pp. 1–14.

[22] Y. Zhou, X. Song, Y. Zhang, F. Liu, C. Zhu, L. Liu, Feature encoding with autoencoders for weakly supervised anomaly detection, IEEE Trans. Neural Netw. Learn. Syst. 33 (6) (2022) 2454–2465.
[23] X. Zhang, C. Mei, D. Chen, J. Li, Feature selection in mixed data: A method using a novel fuzzy rough set-based information entropy, Pattern Recognit. 56 (2016) 1–15.
[24] D. Dubois, H. Prade, Rough fuzzy sets and fuzzy rough sets, Int. J. Gener. Syst. 17 (2–3) (1990) 191–209.
[25] B. Sang, W. Xu, H. Chen, T. Li, Active anti-noise fuzzy dominance rough feature selection using adaptive K-nearest neighbors, IEEE Trans. Fuzzy Syst. (2023) 1–15.
[26] Z. Yuan, B. Chen, J. Liu, H. Chen, D. Peng, P. Li, Anomaly detection based on weighted fuzzy-rough density, Appl. Soft Comput. 134 (2023) 109995.
[27] B. Chen, Y. Li, D. Peng, H. Chen, Z. Yuan, Fusing multi-scale fuzzy information to detect outliers, Inf. Fusion 103 (2024) 102133.
[28] A. Boukerche, L. Zheng, O. Alfandi, Outlier detection: Methods, models, and classification, ACM Comput. Surv. 53 (3) (2020).
[29] M. Jiang, C. Hou, A. Zheng, X. Hu, S. Han, H. Huang, X. He, P.S. Yu, Y. Zhao, Weakly supervised anomaly detection: A survey, 2023, arXiv:2302.04549.
[30] S. Han, X. Hu, H. Huang, M. Jiang, Y. Zhao, ADBench: Anomaly detection benchmark, in: Advances in Neural Information Processing Systems, NeurIPS, Vol. 35, 2022, pp. 32142–32159.
[31] L. Stradiotti, L. Perini, J. Davis, Semi-supervised isolation forest for anomaly detection, in: Proceedings of the 2024 SIAM International Conference on Data Mining, SDM, 2024, pp. 670–678.
[32] B. Micenková, B. McWilliams, I. Assent, Learning outlier ensembles: The best of both worlds–supervised and unsupervised, in: Proceedings of the ACM SIGKDD Workshop on Outlier Detection and Description under Data Diversity, ACM, 2014, pp. 51–54.
[33] L. Ruff, R. Vandermeulen, N. Goernitz, L. Deecke, S.A. Siddiqui, A. Binder, E. Müller, M. Kloft, Deep one-class classification, in: International Conference on Machine Learning, ICML, PMLR, 2018, pp. 4393–4402.
[34] J. Zhu, C. Ding, Y. Tian, G. Pang, Anomaly heterogeneity learning for open-set supervised anomaly detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2024, pp. 17616–17626.
[35] B. Tian, Q. Su, J. Yin, Anomaly detection by leveraging incomplete anomalous knowledge with anomaly-aware bidirectional GANs, in: Proceedings of the 31st International Joint Conference on Artificial Intelligence, IJCAI, 2022, pp. 2255–2261.
[36] J. Donahue, P. Krähenbühl, T. Darrell, Adversarial feature learning, in: International Conference on Learning Representations, ICLR, 2016.
[37] Z. Li, C. Sun, C. Liu, X. Chen, M. Wang, Y. Liu, Dual-MGAN: An efficient approach for semi-supervised outlier detection with few identified anomalies, ACM Trans. Knowl. Discov. Data 16 (6) (2022).
[38] S. Liu, M. Tian, Mutual information maximization for semi-supervised anomaly detection, Knowl.-Based Syst. 284 (2024) 111196.
[39] N. Görnitz, M. Kloft, K. Rieck, U. Brefeld, Toward supervised anomaly detection, Journal of Artificial Intelligence Research 46 (1) (2013).
[40] S. Das, W.-K. Wong, T. Dietterich, A. Fern, A. Emmott, Incorporating expert feedback into active anomaly discovery, in: IEEE International Conference on Data Mining, ICDM, 2016, pp. 853–858.
[41] D. Zha, K.-H. Lai, M. Wan, X. Hu, Meta-AAD: Active anomaly detection with deep reinforcement learning, in: IEEE International Conference on Data Mining, ICDM, 2020, pp. 771–780.
[42] G. Pang, A. van den Hengel, C. Shen, L. Cao, Toward deep supervised anomaly detection: Reinforcement learning from partially labeled anomaly data, in: Proceedings of the 27th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, ACM, 2021, pp. 1298–1308.
[43] C. Chen, D. Wang, F. Mao, J. Xu, Z. Zhang, Y. Yu, Deep anomaly detection via active anomaly search, in: Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems, 2024, pp. 308–316.
[44] Z. Yuan, H. Chen, T. Li, B. Sang, S. Wang, Outlier detection based on fuzzy rough granules in mixed attribute data, IEEE Trans. Cybern. 52 (8) (2022) 8399–8412.
[45] Q. Hu, D. Yu, Z. Xie, J. Liu, Fuzzy probabilistic approximation spaces and their information measures, IEEE Trans. Fuzzy Syst. 14 (2) (2006) 191–201.
[46] W. Pedrycz, X. Wang, Designing fuzzy sets with the use of the parametric principle of justifiable granularity, IEEE Trans. Fuzzy Syst. 24 (2) (2016) 489–496.
[47] Z. Yuan, H. Chen, P. Xie, P. Zhang, J. Liu, T. Li, Attribute reduction methods in fuzzy rough set theory: An overview, comparative experiments, and new directions, Appl. Soft Comput. 107 (2021) 107353.
[48] R. Jensen, Q. Shen, Fuzzy-rough sets assisted attribute selection, IEEE Trans. Fuzzy Syst. 15 (1) (2007) 73–89.
[49] Z. Li, Y. Zhao, X. Hu, N. Botta, C. Ionescu, G. Chen, ECOD: Unsupervised outlier detection using empirical cumulative distribution functions, IEEE Trans. Knowl. Data Eng. (2022) 1–1.
[50] C. Liu, Z. Yuan, B. Chen, H. Chen, D. Peng, Fuzzy granular anomaly detection using Markov random walk, Inform. Sci. 646 (2023) 119400.
[51] Y. Zhao, Z. Nasrullah, Z. Li, PyOD: A python toolbox for scalable outlier detection, J. Mach. Learn. Res. 20 (96) (2019) 1–7.
[52] F.T. Liu, K.M. Ting, Z.-H. Zhou, Isolation forest, in: 2008 Eighth IEEE International Conference on Data Mining, IEEE, 2008, pp. 413–422.
[53] H.-P. Kriegel, P. Kröger, E. Schubert, A. Zimek, Outlier detection in axis-parallel subspaces of high dimensional data, in: Advances in Knowledge Discovery and Data Mining, Springer, 2009, pp. 831–838.