Information Sciences 646 (2023) 119400

Contents lists available at ScienceDirect

Information Sciences

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

Fuzzy granular anomaly detection using Markov random walk

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

[a] College of Computer Science, Sichuan University, Chengdu 610065, China
[b] School of Computing and Artificial Intelligence, Southwest Jiaotong University, Chengdu 611756, China
[c] Chengdu Ruibei Yingte Information Technology Co., Ltd, Chengdu 610054, China
[d] Sichuan Zhiqian Technology Co., Ltd, Chengdu 610065, China

ARTICLE INFO

Keywords:
Granular computing
Fuzzy information granulation
Anomaly detection
Outlier detection
Markov random walk

ABSTRACT

Fuzzy information granulation is an important mathematical model in the theory of granular computing that can effectively handle fuzzy or uncertain information. To address the deficiencies of existing anomaly detection techniques that are mainly suitable for certainty data, this study proposes an anomaly detection method based on fuzzy granules. This method uses the fuzzy information granulation model as a unified framework to gradually develop a fuzzy granular anomaly detection method for calculating the anomaly score of each sample. First, the fuzzy granular distance is defined by fusing single-attribute fuzzy granules to represent the dissimilarity between two samples. Second, a matrix is constructed using the fuzzy granular distance between the samples as the state transition matrix of the Markov random walk. Anomalies in the dataset are then detected using the stationary distribution generated by iterative calculations. Finally, the fuzzy granular anomaly score for each object is obtained by normalizing the stationary distribution. Experiments are conducted on public datasets to compare the proposed method with some state-of-the-art anomaly detection methods. The results indicate that the proposed method is effective. The code is publicly available online at https://github.com/BElloney/FGAS.

1. Introduction

Anomaly detection, which is an important task in knowledge discovery, aims to discover nonnormal objects from data. Anomalies typically signal a new perspective; therefore, discovering such data may be more meaningful than discovering normal data. The applications of anomaly detection include dangerous driving behavior detection [21], intrusion detection [23], fault diagnosis [17], and image processing [41]. Therefore, it is important to further explore and study more effective anomaly detection methods.

Many anomaly detection techniques have been proposed, which are broadly classified into the following categories: statistical-based, distance-based, density-based, and clustering-based methods. Statistical-based methods are the earliest anomaly detection methods, which regard samples appearing in low probability regions as outliers, such as parametric methods represented by Gaussian and regression models [32,37] and nonparametric methods represented by histograms [5]. Distance-based methods detect outliers directly by calculating the distance between two objects in the data, such as the k-Nearest Neighbor (kNN) [27] and Local Distance-based Outlier Factor (LDOF) [38]. Density-based methods compare the local density of a data sample with its neighborhood density to characterize the degree of outliers and thereby detect outliers. The most representative density-based method is the Local Outlier

* Corresponding author.
E-mail addresses: liuchangai@stu.scu.edu.cn (C. Liu), yuanzhong@scu.edu.cn (Z. Yuan), farstars@qq.com (B. Chen), hmchen@swjtu.edu.cn (H. Chen), pengdz@scu.edu.cn (D. Peng).

https://doi.org/10.1016/j.ins.2023.119400
Received 22 March 2023; Received in revised form 1 July 2023; Accepted 16 July 2023

Available online 20 July 2023
0020-0255/© 2023 Elsevier Inc. All rights reserved.

Factor (LOF) algorithm proposed by Breunig et al. [1]. Following the LOF algorithm, scholars have proposed many density-based algorithms, such as the Connectivity-based Outlier Factor (COF) [29], local correlation integral [25], INFLuenced Outlierness (INFLO) [13], and Local Outlier Probability (LoOP) [15]. The main idea of clustering-based methods is to detect anomalies by examining the relationship between data samples and clusters, including cluster-based local outlier factors [10] and relative outlier clustering factors [12].

Random walk methods have diverse applications in various fields, including image segmentation [3], clustering [22], classification [40], and knowledge acquisition [2,6]. For example, Lucas et al. [6] analyzed the learning curves of sequences generated from different random walk dynamics and network models with different topologies. The results indicated that true self-avoiding random walks exhibited the best coverage performance among the network models involved. However, outlier detection methods based on random walks have not been implemented effectively and still relatively few related studies exist [24,30,31]. Such methods construct a state transfer matrix based on the similarity or dissimilarity between data samples and then perform random walks through this matrix for outlier detection. For instance, Moonesinghe and Tan [24] first proposed the use of a random walk model for anomaly detection as an alternative to the aforementioned algorithm. Wang et al. [30] introduced a new model based on random walks called the virtual outlier score. Wang and Li [31] used a customized Markov random wandering process for outlier detection.

However, the aforementioned classical and random walk methods mainly focus on deterministic data and may not be able to effectively handle the anomaly detection of uncertain data with characteristics such as fuzziness. Fuzzy information granulation theory can handle large amounts of data with characteristics such as uncertainty, incompleteness, and fuzziness. Owing to these advantages, it can be applied to the learning tasks of fuzzy or uncertain data [36]. Good results have been achieved in feature selection [11,28], classification [4], and clustering [7,8]. For example, Guo et al. [8] proposed a three-way clustering method based on the *k*-means clustering algorithm and applied it to fuzzy large-group decision-making methods. However, fuzzy granular anomaly detection based on random walks has not been reported. In view of this, this study proposed a fuzzy granule-based anomaly detection method based on a random walk strategy that uses the fuzzy granular computing method to construct the fuzzy granular anomaly detection step-by-step to calculate the outlier score for each sample. First, a fuzzy granular distance is defined by fusing single-attribute fuzziness to represent the dissimilarity between two samples. Second, the fuzzy granular distances between the samples are used to construct a matrix that serves as the state transfer matrix of the Markov random walk. Outliers in the dataset are then detected by iteratively computing the generated smooth distribution. Finally, the Fuzzy Granular Anomaly Score (FGAS) of each sample is obtained by normalizing the smooth distribution. Experiments are conducted on publicly available datasets to compare the proposed method with state-of-the-art anomaly detection methods in terms of the Receiver Operating Characteristic (ROC) curve and Area Under Curve (AUC) indexes. The innovative aspects of this study are summarized as follows.

(1) The state transfer matrix of the Markov random walk is constructed using the fuzzy granular distance between samples to efficiently handle uncertain information with fuzziness;
(2) A customized Markov random walk model is constructed and applied to the anomaly detection task;
(3) The experimental results show that the method is an effective scheme for anomaly detection.

The remainder of this paper is organized as follows. In Section 2, we present the relevant preparatory knowledge. Section 3 proposes the fuzzy granular outlier detection model and designs the corresponding algorithm. Section 4 demonstrates the effectiveness of the FGAS algorithm through an extensive experimental comparison analysis. Finally, we summarize the text and propose directions for future research.

2. Preliminaries

2.1. Fuzzy information granule

A data table without decision attributes is described by a two-tuple $\mathcal{D} = \langle \mathcal{O}, \mathcal{A} \rangle$, where $\mathcal{O} = \{o_1, o_2, \ldots, o_n\}$ is a non-empty finite set of objects; $\mathcal{A} = \{a_1, a_2, \ldots, a_m\}$ is a non-empty finite set of conditional attributes; $o \in \mathcal{O}$ and $a \in \mathcal{A}$, $a(o)$ denotes the value of o with respect to attribute a.

Given $\mathcal{B} \subseteq \mathcal{A}$, the fuzzy relation $R_{\mathcal{B}}$ with respect to $\mathcal{B}$ on $\mathcal{O}$ is denoted as

$$R_{\mathcal{B}} : \mathcal{O} \times \mathcal{O} \to [0, 1]. \tag{1}$$

For any $o, p \in \mathcal{O}$, if $R_{\mathcal{B}}$ satisfies 1) reflexivity ($R_{\mathcal{B}}(o, o) = 1$) and 2) symmetry ($R_{\mathcal{B}}(o, p) = R_{\mathcal{B}}(p, o)$), then $R_{\mathcal{B}}$ is a fuzzy similarity relation. $R_{\mathcal{B}}(o, p)$ denotes the similarity between o and p and $R_{\mathcal{B}}(o, p) = \min_{a \in \mathcal{B}} R_a(o, p)$.

$R_{\mathcal{B}}$ can induce a fuzzy granular structure of $\mathcal{O}$, denoted as [26]

$$G(R_{\mathcal{B}}) = \{[o_1]_{\mathcal{B}}, [o_2]_{\mathcal{B}}, \ldots, [o_n]_{\mathcal{B}}\}, \tag{2}$$

where $[o_i]_{\mathcal{B}} = (R_{\mathcal{B}}(o_i, o_1), R_{\mathcal{B}}(o_i, o_2), \ldots, R_{\mathcal{B}}(o_i, o_n))$. The cardinality of $[o_i]_{\mathcal{B}}$ is $|[o_i]_{\mathcal{B}}| = \sum_{j=1}^{n} R_{\mathcal{B}}(o_i, o_j)$. For any $o \in \mathcal{O}$, a parameterized fuzzy information granule associated with o is denoted as

$$[o]_{\mathcal{B}}(p) = \begin{cases} 0, & R_{\mathcal{B}}(o, p) < \sigma; \\ R_{\mathcal{B}}(o, p), & R_{\mathcal{B}}(o, p) \geq \sigma, \end{cases} \tag{3}$$

where σ is an adjustable parameter and $\sigma \in [0, 1)$.

2.2. Markov random walk

Intuitively, a stochastic process is a mathematical model for understanding the dynamics of complex random phenomena, which is defined formally as follows.

Let $(\Omega, \mathscr{F}, P)$ be a probability space, where $\Omega, \mathscr{F}, P$ denote the sample space, the set of random events and the probability, respectively. Given an indicator set T, if $X = \{X_t | t \in T\}$ is a family of random variables on $(\Omega, \mathscr{F}, P)$, then X is said to be a random process.

Given a stochastic process $X = \{X_t | t \in \{0, 1, \dots\}\}$. When $t > 1$, if $X(t)$ depends only on $X(t-1)$, but not on other random variables $\{X_0, X_1, \dots, X_{t-2}\}$ in the past, that is, it satisfies the following conditional probability distribution

$$p(X_t | X_0, X_1, \dots, X_{t-1}) = p(X_t | X_{t-1}), \tag{4}$$

then X is said to be a Markov chain in discrete time.

The stochastic process defined above can be described by the state transfer probability $p(X_t | X_{t-1})$ as a conditional probability. It means the probability that the state at the previous moment is X_{t-1} and the state X_t is transferred to the next moment. For a Markov chain containing n states, this conditional probability value can be represented by the state transfer matrix $\mathbf{P}$, which is denoted as

$$\mathbf{P} = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ p_{n1} & p_{n2} & \cdots & p_{nn} \end{bmatrix}. \tag{5}$$

By iteratively evolving the probability distribution of the states using the state transfer matrix, an interesting property of Markov chains can be found. That is, regardless of the value of the initial state probability distribution, the final state probability distribution will converge to a stable value as long as the Markov chain satisfies certain conditions. As a result, the concept of equilibrium distribution can be further defined. For a Markov chain with state transfer matrix $\mathbf{P}$, if there exists a probability distribution π satisfying

$$\pi = \pi \mathbf{P}, \tag{6}$$

then this distribution π is said to be a smooth distribution. The meaning is that if the state at the current moment obeys this distribution, it also obeys this distribution after transferring to a moment.

3. Proposed approach

3.1. Related definition

A key issue of Markov random walks is how to construct the state transfer matrix. In order to efficiently handle uncertain information with fuzziness, a Fuzzy Granular Distance (FGD) is first defined as follows.

Given a data table $D = \langle \mathcal{O}, \mathcal{A} \rangle$. For any $o, p \in \mathcal{O}$, the fuzzy granular distance between o and p is computed by

$$FGD(o, p) = \sum_{a \in \mathcal{A}} \left(\frac{|[o]_a|}{|\mathcal{O}|} - \frac{|[p]_a|}{|\mathcal{O}|} \right)^2. \tag{7}$$

It is not difficult to prove that the distance function FGD is a qualified distance measure, called the fuzzy granular distance in the fuzzy information granulation theory. The fuzzy granular distance replaces the value of each sample with the information granular probability, which can well describe the difference between the two samples. Further, the above distance function can be represented by a matrix $\mathbf{D} = [FGD(o, p)]_{n \times n}$.

Matrix $\mathbf{D}$ can be used to generate the state transfer matrix, which directly determines the wanderer's wandering behavior. $\mathbf{P}$ can be calculated by the following normalized formula.

$$\mathbf{P} = \mathbf{E}^{-1} \mathbf{D}, \tag{8}$$

where $\mathbf{E}$ =diag$(e_1, e_2, \dots, e_n)$ is a diagonal matrix with the ith diagonal element $e_i = \sum_{j=1}^{n} FGD(o_i, o_j)$. The normalization performed above ensures that each row of the state transfer matrix sum to 1, which is a fundamental characteristic of Markov chains.

After computing the transition matrix $\mathbf{P}$, it is necessary to make $\mathbf{P}$ both integrable and acyclic so that it to converge to a unique stationary distribution. In this paper, the strategy in [24] is used to restart the random walks by retaining a low probability value during each iteration. Thus, the modification of Formula (6) can be expressed as

$$\pi^{(t+1)} = d + (1 - d)\pi^{(t)} \mathbf{P}, \tag{9}$$

where $\pi = [\pi_1, \pi_2, \dots, \pi_n]$ indicates a probability distribution. d is a damping factor indicating the probability that a walker will choose a node from the probability distribution vector to make a jump. To ensure that the model can better portray the structure of the data itself, it is set to a smaller value of 0.10 in this paper. In addition, when $t = 0$, let $\pi^{(0)} = \left[\frac{1}{|\mathcal{O}|}, \frac{1}{|\mathcal{O}|}, \dots, \frac{1}{|\mathcal{O}|} \right]$.

The proposed detection model uses the fuzzy granular distance between nodes to define the state transfer probability matrix of a Markov random walk process. Each element p_{ij} in $\mathbf{P}$ can be calculated as $p_{ij} = \frac{FGD_{(o_i,o_j)}}{\sum_{j=1}^{n} FGD_{(o_i,o_j)}}$, where the denominator is fixed, and when the relative distance between the two samples is large, the larger the numerator, i.e., the larger the corresponding p_{ij}. According to this characteristic, anomalous objects tend to have a higher probability of being visited. We employ the power iteration method to solve the steady-state distribution vector of this model. When the Markov random walk process reaches a stable distribution state, each element in the steady-state distribution vector can represent the probability of a random wanderer staying on each sample, i.e., the Anomaly Score (AS) of each sample. The larger the AS of each sample, the greater its probability of being an anomaly.

The values in the steady-state distribution vector are all fractional. Their differences may not be significant, resulting in not easily distinguishing the degree of abnormality for each sample. To make the AS of each sample more significant, they are subjected to min-max normalization. Thus, the AS for each sample is calculated as

$$AS(o_i) = \frac{\pi_i - \min(\pi)}{\max(\pi) - \min(\pi)}. \tag{10}$$

Given an anomaly threshold ν, for any $o_i \in \mathcal{O}$, the object o_i is said to be a fuzzy granular anomaly if $AS(o_i) > \nu$.

Algorithm 1: FGAS.

Input: A data table $D = \langle \mathcal{O}, \mathcal{A} \rangle$, parameter σ.
Output: Anomaly score (AS).

```
1  Initialize AS = ∅;
2  for every a ∈ A do
3  |  Compute G(R_a) according to Eqs. (2) and (3);
4  end
5  for every o ∈ O do
6  |  for every p ∈ O do
7  |  |  Compute FGD(o, p) according to Eq. (7);
8  |  end
9  end
10 Construct D;
11 Compute P = E^{-1}D according to Eq. (8);
12 Initialize t = 0;
13 Initialize π^(t) = [1/|O|, 1/|O|, ..., 1/|O|];
14 repeat
15 |  π^(t+1) = d + (1 − d)π^(t)P according to Eq. (9);
16 |  t = t + 1;
17 until (||π^(t+1) − π^(t)||_1 ≤ 10^{-3});
18 Compute AS(o_i) according to Eq. (10);
19 return AS.
```

3.2. Related algorithm

In Algorithm 1, the anomaly score $AS = \emptyset$ is first initialized. Next, the fuzzy granular structure $G(R_a)$ of a single attribute is computed. Then, the fuzzy granular distance $FGD(o, p)$ between any two objects is calculated based on the fuzzy granular structure. Further, $\mathbf{D}$ is constructed, $\mathbf{P}$ is computed, and t and π^t are initialized in turn. Finally, when the stable distribution condition is reached, each element of the steady-state distribution vector is used to indicate the AS of each sample and returns the AS after normalization. In Algorithm 1, the number of loops in Steps 2-4 is $|\mathcal{A}|$, the number of loops in Steps 3 is $|\mathcal{O}| \times |\mathcal{O}|$, and the number of loops in Steps 5-9 is $|\mathcal{O}| \times |\mathcal{O}|$, so the total number of loops for algorithm one is $|\mathcal{A}| \times |\mathcal{O}| \times |\mathcal{O}| + |\mathcal{O}| \times |\mathcal{O}|$. Therefore, in the worst case, the time complexity of Algorithm 1 is $\mathcal{O}(|\mathcal{A}||\mathcal{O}|^2)$.

4. Experiments

We conduct an extensive experimental comparison with existing algorithms. Here, the preparations for the experiments are presented. The experimental results are then compared and analyzed. Hypothesis testing is performed to verify the statistical validity of the proposed algorithm. Furthermore, sensitivity analysis of the experimental parameters is performed.

4.1. Experimental preparation

A comparative study is conducted on publicly available datasets[1] [2] to evaluate the performance of FGAS. These datasets are commonly used in outlier detection research to assess the efficacy of various detection methods. Table 1 provides a description of

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

Table 1
Description of datasets.

No.	Datasets	Abbr.	Number of attributes	Number of objects	Number of outliers
1	Cardio	Card	21	1831	176
2	Diabetes_tested_positive_26_variant1	Diab	8	526	26
3	Ecoli	Ecoli	7	336	9
4	Ionosphere_b_24_variant1	Ionos	34	249	24
5	Iris_Irisvirginica_11_variant1	Iris	4	111	11
6	Musk	Musk	166	3062	97
7	Pageblocks_1_258_variant1	Page	10	5171	258
8	Pima_TRUE_55_variant1	Pima	9	555	55
9	Satellite	Sate	36	6435	2036
10	Sonar_M_10_variant1	Sonar	60	107	10
11	Thyroid	Thyro	6	3772	93
12	Wbc_malignant_39_variant1	Wbc	9	483	39
13	Wdbc_M_39_variant1	Wdbc	31	396	39
14	Wine	Wine	13	129	10
15	Wpbc_variant1	Wpbc	33	198	47

datasets, including sample sizes ranging from 107 to 6435 and dimensions ranging from 4 to 166. In the experiments, the maximum frequency method is used to fill in missing data.

FGAS algorithm is compared to 13 detection algorithms. Specifically, these detection algorithms are DIStance (DIS) [14], COF [29], Fast Angle-Based Outlier Detection (FastABOD) [16], INFLO [13], kNN [27], LDOF [38], LoOP [15], Outlier Detection using In-degree Number (ODIN) [9], Directed density ratio Changing Rate-based Outlier Detection (DCROD) [18], Empirical Cumulative-based Outlier Detection (ECOD) [19], Isolation Forest (IForest) [20], Outlier Ranking-a (OutRanka) [24], and Weighted Neighborhood Information Network-based Outlier Detection (WNINOD) [31], which employ various detection techniques and have demonstrated relatively superior performance.

In the experiments, COF, FastABOD, INFLO, kNN, LDOF, LoOP, ODIN, DCROD, and OutRanka mainly involved the parameter k; thus, we vary the value of k from 1 to 60 with step size 1 to determine the optimal setting. For IForest, we set the number of base estimators to 100. We set the parameter adjustment range of WNINOD to $[1, 10]$ with a step size of one. To facilitate the experimental comparison, we computed the outlier score of WNINOD by obtaining the reciprocal of all inlier scores in its output. In addition, it is not reasonable to classify outliers as binary outliers using a traditional distance approach. Hence, we define the distance AS to quantify the degree of outliers in each sample. We adjust the parameter σ for FGAS from 0 to 1 in steps of 0.05 to find the optimal result. In addition, we calculate the fuzzy similarity relation $R_a(o,p) = 1 - |a(o) - a(p)|$ for a.

Receiver Operating Characteristic (ROC) curve and Area Under Curve (AUC) index are used to evaluate the performance of the comparison algorithm comprehensively [34]. Specifically, all samples are arranged in descending order according to outlier scores to obtain the order number $\langle 1, 2, \ldots, ord, \ldots, |\mathcal{O}| \rangle$ for each sample. Given an order number ord, let $OS_d(ord)$ denote the set of detected outliers and OS_t denote the set of true outliers. The True Positive Rate (TPR) and False Positive Rate (FPR) are denoted as

$$TPR(ord) = \frac{|OS_d(ord) \bigcap OS_t|}{|OS_t|}, \tag{11}$$

$$FPR(ord) = \frac{|OS_d(ord) - OS_t|}{|\mathcal{O} - OS_t|} = \frac{|OS_d(ord) - OS_t|}{|OS_t^c|}. \tag{12}$$

Of the aforementioned two metrics, the ROC curve is plotted with $FPR(ord)$ as the horizontal coordinate and $TPR(ord)$ as the vertical coordinate. The AUC index is further proposed to quantitatively evaluate the effectiveness of an algorithm, which is defined as

$$\mathrm{AUC} = \frac{1}{|OS_t||OS_t^c|} \sum_{o_i \in OS_t, o_j \in OS_t^c} P_{ij}, \tag{13}$$

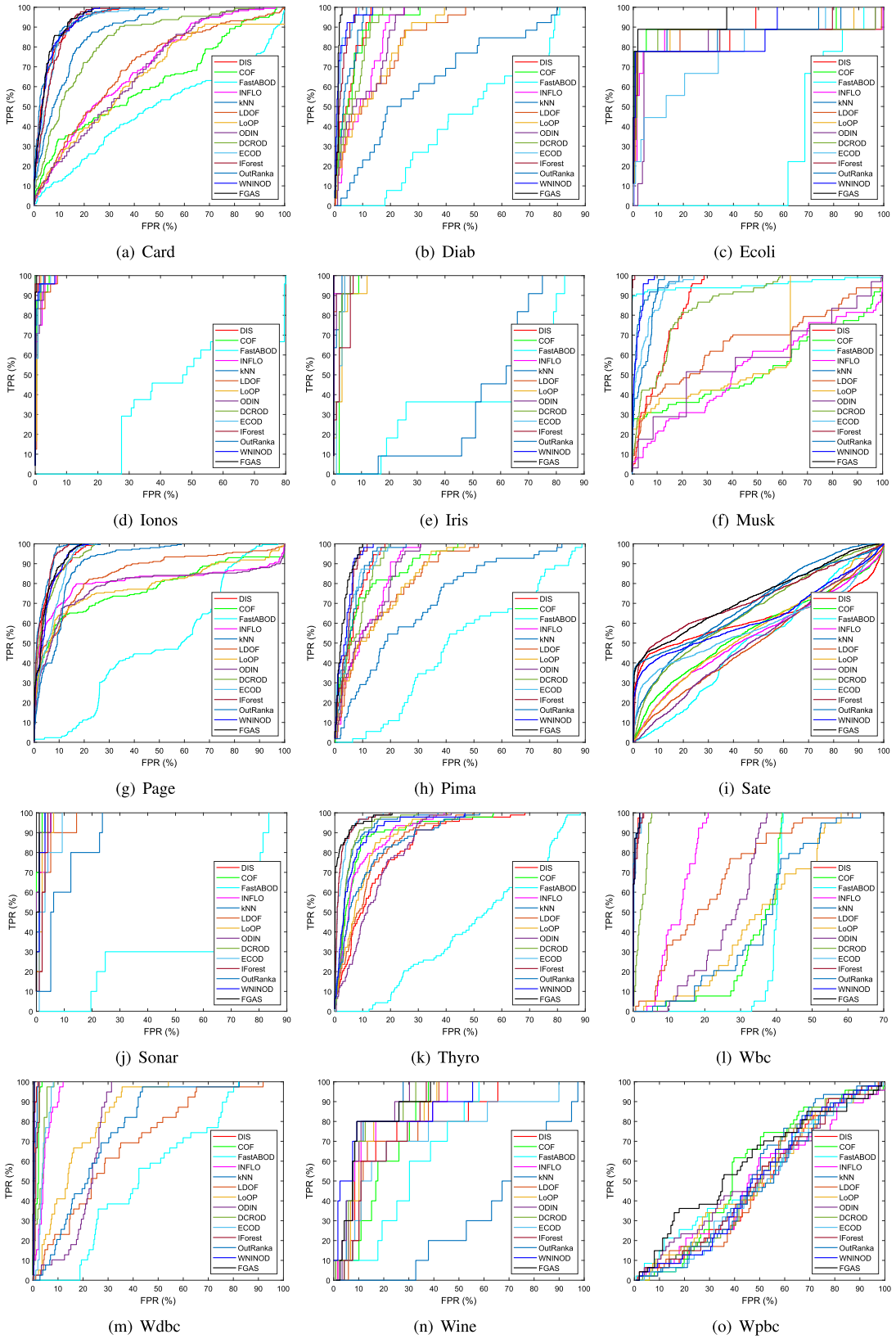
where $P_{ij} = 1$ if $AS(o_i) > AS(o_j)$; $P_{ij} = 0.5$ if $AS(o_i) = AS(o_j)$; $P_{ij} = 0$ if $AS(o_i) < AS(o_j)$.

The AUC index accepts values in the range of [0,1]. The higher the value of the AUC of an algorithm, the better its performance. In addition, it does not require additional parameters and is more distinguishable than the ROC curve.

4.2. Experimental result

Fig. 1 shows the ROC curves of 14 algorithms, where the black curve represents the FGAS algorithm. Based on the ROC curves shown in Fig. 1, FGAS shows superior performance on many datasets, such as Diab, Ecoli, Iris, Musk, and Pima. Notably, the curves for FGAS on Iris and Musk overlapped completely with the coordinates, indicating excellent performance. Moreover, the ROC curves of FGAS exhibited superior performance on most datasets, including Card, Pima, and Wpbc. However, certain algorithms exhibit nearly overlapping curves on certain datasets, such as Page, Wdbc, and Wine, rendering it difficult to determine the best algorithm. To further analyze the performance, we compare the AUC indices of the 14 detection algorithms.

The AUC index comparison results from the experiment are presented in Table 2, where the best results are highlighted in bold.



(a) Card (b) Diab (c) Ecoli

(d) Ionos (e) Iris (f) Musk

(g) Page (h) Pima (i) Sate

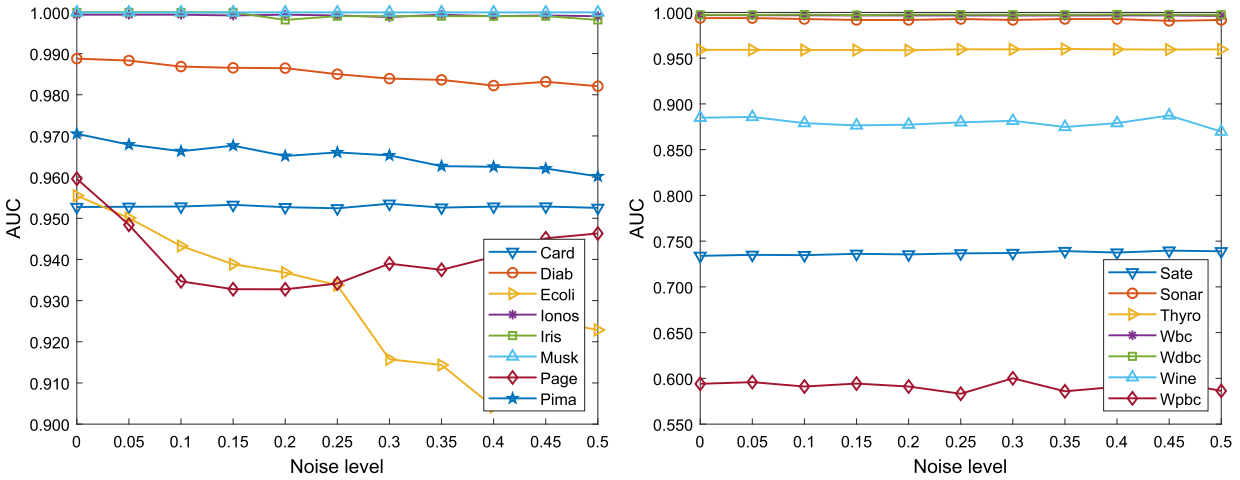
(j) Sonar (k) Thyro (l) Wbc

(m) Wdbc (n) Wine (o) Wpbc

Fig. 1. Comparison results on ROC.

Table 2
Experimental results on AUC.

Dataset	DIS	COF	FastABOD	INFLO	kNN	LDOF	LoOP	ODIN	DCROD	ECOD	IForest	OutRanka	WNINOD	FGAS
Card	0.946	0.620	0.482	0.698	0.884	0.698	0.667	0.686	0.834	0.935	0.941	0.950	0.949	**0.953**
Diab	0.952	0.942	0.450	0.902	0.958	0.863	0.870	0.898	0.935	0.979	0.976	0.693	0.981	**0.989**
Ecoli	0.899	0.902	0.378	0.862	0.891	0.863	0.886	0.835	0.875	0.781	0.867	0.879	0.874	**0.955**
Ionos	0.998	0.994	0.452	0.991	**1.000**	0.988	0.993	0.993	**1.000**	0.994	0.999	0.997	0.996	**1.000**
Iris	**1.000**	0.971	0.541	0.995	0.992	0.995	0.969	0.993	0.983	0.977	0.971	0.436	**1.000**	**1.000**
Musk	0.889	0.533	0.954	0.534	0.975	0.663	0.557	0.516	0.863	0.956	**1.000**	0.956	0.983	**1.000**
Page	0.956	0.776	0.513	0.807	**0.973**	0.849	0.791	0.784	0.953	0.938	0.970	0.904	0.956	0.960
Pima	0.937	0.908	0.492	0.894	0.943	0.855	0.860	0.889	0.928	0.947	0.957	0.745	0.963	**0.971**
Sate	0.620	0.562	0.498	0.536	0.694	0.510	0.559	0.515	0.662	0.583	0.723	0.697	0.631	**0.734**
Sonar	0.984	**0.996**	0.504	0.984	0.992	0.964	0.981	0.994	0.989	0.966	0.976	0.902	0.989	0.995
Thyro	0.862	0.926	0.472	0.920	0.951	0.887	0.909	0.863	0.955	0.977	0.980	0.896	0.943	**0.985**
Wbc	**0.997**	0.654	0.604	0.875	**0.997**	0.792	0.645	0.734	0.976	0.995	0.996	0.641	**0.997**	**0.997**
Wdbc	0.995	0.984	0.523	0.954	0.991	0.702	0.839	0.786	0.968	0.959	0.987	0.765	0.996	**0.998**
Wine	0.797	0.798	0.681	0.850	**0.895**	0.816	0.855	0.887	0.885	0.733	0.824	0.303	0.873	0.885
Wpbc	0.492	0.561	0.532	0.495	0.532	0.504	0.505	0.507	0.513	0.481	0.504	0.483	0.506	**0.594**
Average	0.888	0.808	0.538	0.820	0.911	0.797	0.793	0.792	0.888	0.880	0.911	0.750	0.909	**0.934**
1st order	2	1	0	0	4	0	0	0	1	0	1	0	2	**12**



(a) Datasts 1-8 (b) Datasts 9-15

Fig. 2. Effect of attribute noise on AUC.

Based on the 15 public datasets used in this study, Table 2 provides a better representation of the FGAS algorithm's performance. FGAS achieves the highest AUC scores in most cases, such as the Card dataset (0.953). However, the scores of the other algorithms are lower than that of FGAS, with scores ranging from 0.482 to 0.950. We rank the 14 algorithms in a dataset according to their AUC scores for easier comparison, wherein higher values led to higher rankings. In addition, we count the number of times each algorithm achieved the highest ranking among the 15 datasets. FGAS obtains the highest scores on 12 datasets statistically, whereas the other 13 algorithms achieve a superior rank in only a limited number of datasets. In other words, FGAS outperformed the other 13 algorithms on most datasets. The final average value further confirms the superior performance of FGAS. The average AUC of FGAS is 0.934, which is significantly better than those of the other algorithms.

4.3. Effect of attribute noise on AUC

We apply the strategy in [43] to evaluate the effect of attribute noise in terms of AUC. In this strategy, the sample $\lceil xn \rceil$ with a specific attribute value is replaced by a random value, and the numerical attribute of the sample is replaced by a random value between the maximum and minimum values, resulting in a dataset with a noise level of $x \times 100\%$. Here, $\lceil xn \rceil$ refers to n samples of the same attribute. The final result of the influence of the attribute noise level on AUC for this strategy is shown in Fig. 2. By observing the impact results on Musk, Iris, Sonar, Wdbc, and other datasets, we find that as the noise level increases, the AUC curve tends to fluctuate up and down in a small range, which indicates that FGAS is robust to attribute noise.

4.4. Statistical analysis

According to the strategy in [35], the Friedman test is first used to estimate whether a significant difference exists among all algorithms. Then, the Nemenyi test is used to further distinguish them. In the Nemenyi test figure, we plot the average ordinal values of the 14 different anomaly detection algorithms at the corresponding positions on the number axis. If a horizontal line segment connects a group of algorithms, then no significant difference exists among them.

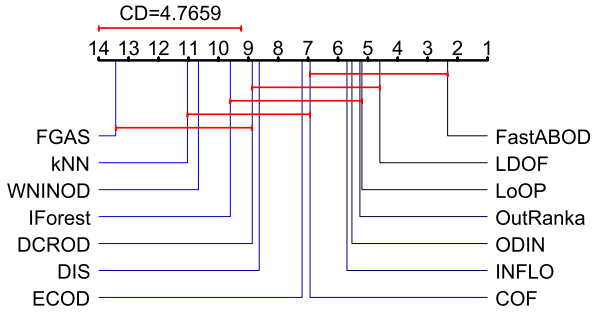
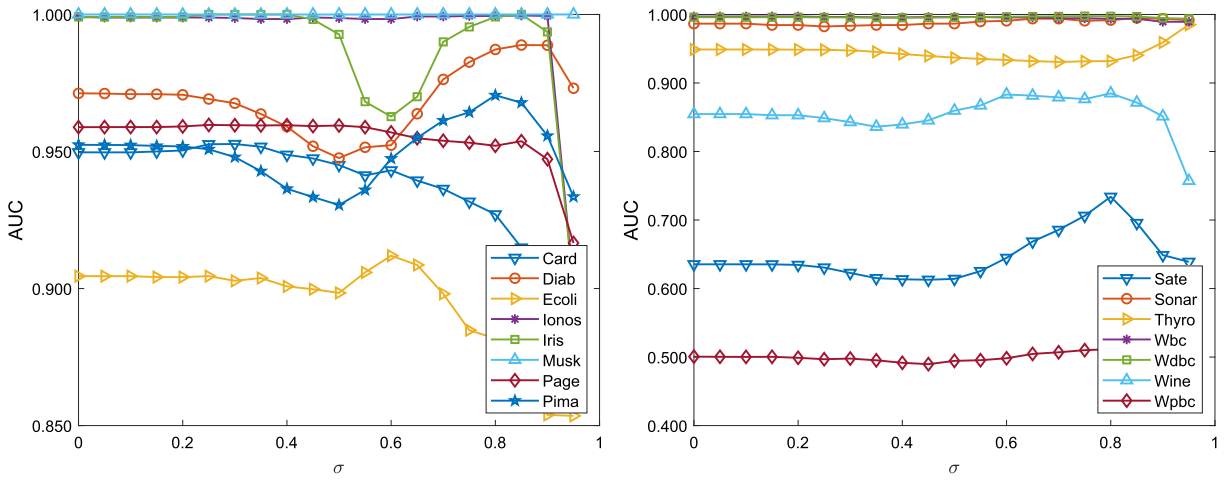


Fig. 3. Nemenyi's test figure on AUC.



(a) Datasts 1-8 (b) Datasts 9-15

Fig. 4. The variation curve of AUC on σ.

From Table 2, we can observe that there are 14 algorithms and 15 datasets in the experiment, from which we can obtain F distribution with 13 and 182 degrees of freedom. According to the Friedman test, $\tau_F = 15.1347$ is greater than the critical value of 1.5610 when the significance level $\alpha = 0.1$. Therefore, the null hypothesis that "all algorithms perform equally" does not hold. This illustrates a significant difference in performance among these anomaly detection algorithms, necessitating a post-hoc test to further differentiate them.

For a significance level $\alpha = 0.1$, the corresponding critical distance $CD_{0.1} = 4.7659$ is calculated. Nemenyi test figure for AUC is shown in Fig. 3. As shown in Fig. 3, FGAS is statistically significantly different from most other algorithms. For example, based on the results presented in Fig. 3, we can observe that FGAS has no horizontal line segment connection with DIS, ECOD, FastABOD, LDOF, LoOP, OutRanka, ODIN, INFLO, and COF, indicating that FGAS is statistically significantly different from these algorithms. However, no consistent evidence exists for statistical differences among kNN, WNINOD, IForest, and DCROD.

4.5. Experimental parameter

The variation curve of AUC with regard to the parameter σ is shown in Fig. 4. Fig. 4 indicates that FGAS is robust to the parameter σ. For most datasets, the corresponding AUC scores tended to stabilize as parameter σ increased. For some datasets, such as Diab, Iris, and Pima, the AUC scores exhibited approximately the same trend. This indicates that these datasets have approximately the same distribution of data values.

5. Conclusion

This study proposes a novel anomaly detection model based on fuzzy granules using Markov random walks. The proposed model leverages the fact that fuzzy computing can effectively handle uncertain information such as fuzziness. First, the fuzzy granular distance between objects is calculated using fuzzy granules, and the corresponding distance matrix is constructed. Subsequently, the state transition matrix is defined according to the distance matrix. Finally, based on the state transition matrix, a customized Markov random walk process is used for anomaly detection. The corresponding FGAS algorithm is designed and analyzed based on the proposed detection method. The results show that FGAS is acceptable in terms of time complexity. FGAS is compared with

state-of-the-art algorithms and analyzed to validate its effectiveness. The results indicate that FGAS performs better in most cases in terms of ROC curves and AUC scores. In addition, hypothetical statistical experiments show that FGAS is significantly different from most algorithms. In future work, we will consider applying the ideas of three-way decisions [33,39,42] and true self-avoiding random walks [2,6] for anomaly detection.

CRediT authorship contribution statement

Chang Liu: Conceptualization, Methodology, Software, Investigation, Writing - original draft.
Zhong Yuan: Supervision, Funding acquisition, Structuralization, Validation, Writing - review & editing.
Baiyang Chen: Validation, Writing - review & editing.
Hongmei Chen: Supervision, Project administration, Funding acquisition
Dezhong Peng: Supervision, Project administration, Funding acquisition.

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

The datasets are publicly available online at https://github.com/BElloney/Outlier-detection.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (61976182), Sichuan Science and Technology Program (2023YFQ0020, 2023YFG0033, 2023ZHCG0016, 2022YFQ0014, 2022YFH0021), and Fundamental Research Funds for the Central Universities (YJ202245).

References

[1] 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.
[2] J.Q. Ding, T. d'Orsi, R. Nasser, D. Steurer, Robust recovery for stochastic block models, in: 2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS), IEEE, 2022, pp. 387–394.
[3] X.P. Dong, J.B. Shen, L. Shao, L. Van Gool, Sub-markov random walk for image segmentation, IEEE Trans. Image Process. 25 (2) (2015) 516–527.
[4] C. Fu, W. Lu, W. Pedrycz, J.H. Yang, Fuzzy granular classification based on the principle of justifiable granularity, Knowl.-Based Syst. 170 (2019) 89–101.
[5] M. Gebski, R.K. Wong, An efficient histogram method for outlier detection, in: Proceedings of the International Conference on Database Systems for Advanced Applications, Springer, Berlin, 2007, pp. 176–187.
[6] L. Guerreiro, F.N. Silva, D.R. Amancio, A comparative analysis of knowledge acquisition performance in complex networks, Inf. Sci. 555 (2021) 46–57.
[7] H.Y. Guo, L.D. Wang, X.D. Liu, W. Pedrycz, Information granulation-based fuzzy clustering of time series, IEEE Trans. Cybern. 51 (12) (2021) 6253–6261.
[8] L. Guo, J.M. Zhan, Z.S. Xu, J.C.R. Alcantud, A consensus measure-based three-way clustering method for fuzzy large group decision making, Inf. Sci. 632 (2023) 144–163.
[9] V. Hautamaki, I. Karkkainen, P. Franti, Outlier detection using k-nearest neighbour graph, Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004, vol. 3, IEEE, 2004, pp. 430–433.
[10] Z.Y. He, X.F. Xu, S.C. Deng, Discovering cluster-based local outliers, Pattern Recognit. Lett. 24 (9–10) (2003) 1641–1650.
[11] Q.H. Hu, Z.X. Xie, D.R. Yu, Hybrid attribute reduction based on a novel fuzzy-rough model and information granulation, Pattern Recognit. 40 (12) (2007) 3509–3521.
[12] J.L. Huang, Q.S. Zhu, L.J. Yang, D.D. Cheng, Q.W. Wu, A novel outlier cluster detection algorithm without top-n parameter, Knowl.-Based Syst. 121 (2017) 32–40.
[13] W. Jin, A.K. Tung, J.W. Han, W. Wang, Ranking outliers using symmetric neighborhood relationship, in: Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, Berlin, 2006, pp. 577–593.
[14] E.M. Knox, R.T. Ng, Algorithms for mining distance-based outliers in large datasets, in: Proceedings of the International Conference on Very Large Data Bases, Citeseer, 1998, pp. 392–403.
[15] H.P. Kriegel, P. Kröger, E. Schubert, A. Zimek, LoOP: Local outlier probabilities, in: Proceedings of the 18th ACM Conference on Information and Knowledge Management, ACM, New York, 2009, pp. 1649–1652.
[16] H.P. Kriegel, M. Schubert, A. Zimek, Angle-based outlier detection in high-dimensional data, in: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2008, pp. 444–452.
[17] Y.G. Lei, B. Yang, X.W. Jiang, F. Jia, N.P. Li, A.K. Nandi, Applications of machine learning to machine fault diagnosis: a review and roadmap, Mech. Syst. Signal Process. 138 (2020) 106587.
[18] K.S. Li, X. Gao, S.Y. Fu, X.P. Diao, P. Ye, B. Xue, J.H. Yu, Z.J. Huang, Robust outlier detection based on the changing rate of directed density ratio, Expert Syst. Appl. 207 (2022) 117988.
[19] Z. Li, Y. Zhao, X.Y. Hu, N. Botta, C. Ionescu, G. Chen, ECOD: Unsupervised outlier detection using empirical cumulative distribution functions, IEEE Trans. Knowl. Data Eng. (2023), https://doi.org/10.1109/TKDE.2022.3159580.
[20] 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.
[21] J. Liu, W. Huang, H. Li, S.G. Ji, Y.J. Du, T.R. Li. Slafusion, Attention fusion based on sax and lstm for dangerous driving behavior detection, Inf. Sci. 640 (2023) 119063.
[22] R.S. Liu, Z.C. Lin, Z.X. Su, Learning markov random walks for robust subspace clustering and estimation, Neural Netw. 59 (2014) 1–15.
[23] W.G. Ma, R.Q. Liu, K.H. Li, S. Yan, J. Guo, An adversarial domain adaptation approach combining dual domain pairing strategy for iot intrusion detection under few-shot samples, Inf. Sci. 629 (2023) 719–745.

[24] H. Moonesinghe, P.N. Tan, Outlier detection using random walks, in: 2006 18th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'06), IEEE, 2006, pp. 532–539.
[25] S. Papadimitriou, H. Kitagawa, P.B. Gibbons, C. Faloutsos, LOCI: Fast outlier detection using the local correlation integral, in: Proceedings of the 19th International Conference on Data Engineering, IEEE, New York, 2003, pp. 315–326.
[26] Y.H. Qian, J.Y. Liang, W.Z. Wu, C.Y. Dang, Information granularity in fuzzy binary grc model, IEEE Trans. Fuzzy Syst. 19 (2) (2010) 253–264.
[27] S. Ramaswamy, R. Rastogi, K. Shim, Efficient algorithms for mining outliers from large data sets, ACM SIGMOD Rec. 29 (2) (2000) 427–438.
[28] B.B. Sang, W.H. Xu, H.M. Chen, T.R. Li, Active anti-noise fuzzy dominance rough feature selection using adaptive k-nearest neighbors, IEEE Trans. Fuzzy Syst. (2023), https://doi.org/10.1109/TFUZZ.2023.3272316.
[29] J. Tang, Z. Chen, A.W.C. Fu, D.W. Cheung, Enhancing effectiveness of outlier detections for low density patterns, in: Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, Berlin, 2002, pp. 535–548.
[30] C. Wang, Z. Liu, H. Gao, Y. Fu, VOS: A new outlier detection model using virtual graph, Knowl.-Based Syst. 185 (2019) 104907.
[31] Y. Wang, Y.P. Li, Outlier detection based on weighted neighbourhood information network for mixed-valued datasets, Inf. Sci. 564 (2021) 396–415.
[32] X.W. Yang, L.J. Latecki, D. Pokrajac, Outlier detection with globally optimal exemplar-based gmm, in: Proceedings of the SIAM International Conference on Data Mining, SIAM, Philadelphia, 2009, pp. 145–154.
[33] Y.Y. Yao, Tri-level thinking: models of three-way decision, Int. J. Mach. Learn. Cybern. 11 (5) (2020) 947–959.
[34] Z. Yuan, B.Y. Chen, J. Liu, H.M. Chen, D.Z. Peng, P.L. Li, Anomaly detection based on weighted fuzzy-rough density, Appl. Soft Comput. 134 (2023) 109995.
[35] Z. Yuan, H.M. Chen, C. Luo, D.Z. Peng, MFGAD: Multi-fuzzy granules anomaly detection, Inf. Fusion 95 (2023) 17–25.
[36] L.A. Zadeh, Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic, Fuzzy Sets Syst. 90 (2) (1997) 111–127.
[37] J. Zhang, Advancements of outlier detection: a survey, ICST Trans. Scalable Inf. Syst. 13 (1) (2013) 1–26.
[38] K. Zhang, M. Hutter, H.D. Jin, A new local distance-based outlier detection approach for scattered real-world data, in: Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, 2009, pp. 813–822.
[39] R.T. Zhang, X.L. Ma, J.M. Zhan, Y.Y. Yao, 3WC-D: A feature distribution-based adaptive three-way clustering method, Appl. Intell. (2023) 15561–15579.
[40] X.D. Zhao, R. Tao, W. Li, H.C. Li, Q. Du, W.Z. Liao, W. Philips, Joint classification of hyperspectral and lidar data using hierarchical random walk and deep cnn architecture, IEEE Trans. Geosci. Remote Sens. 58 (10) (2020) 7355–7370.
[41] K. Zhou, J. Li, Y.T. Xiao, J.L. Yang, J. Cheng, W. Liu, W.X. Luo, J. Liu, S.H. Gao, Memorizing structure-texture correspondence for image anomaly detection, IEEE Trans. Neural Netw. Learn. Syst. 33 (6) (2022) 2335–2349.
[42] C.L. Zhu, X.L. Ma, C. Zhang, W.P. Ding, J.M. Zhan, Information granules-based long-term forecasting of time series via bpnn under three-way decision framework, Inf. Sci. 634 (2023) 696–715.
[43] X.Q. Zhu, X.D. Wu, Class noise vs. attribute noise: a quantitative study, Artif. Intell. Rev. 22 (3) (2004) 177–210.