DFNO: Detecting Fuzzy Neighborhood Outliers

Zhong Yuan, Peng Hu, Hongmei Chen, *Member, IEEE*, Yingke Chen, and Qilin Li

***Abstract*—Outlier Detection (OD) has attracted extensive research due to its application in many fields. The idea of neighborhood computing is one of the widely used methods in outlier analysis. Nevertheless, these methods mainly use certainty strategies to model outlier detection, so they cannot effectively handle the fuzzy information in the dataset. Moreover, they mainly focus on dealing with outlier detection in numerical data and cannot effectively find outliers in mixed-attribute data. Fuzzy information granulation theory is an effective granular computing model that allows objects to belong to a set to a certain extent (i.e., membership degree), which makes it possible to better handle uncertainty problems such as fuzziness. In this work, we propose an outlier detection model based on fuzzy neighborhoods. First, a hybrid fuzzy similarity is constructed to granulate the set of objects to form fuzzy information granules. Second, the fuzzy k-nearest neighbor is defined to describe the fuzzy local information. Then, the fuzzy neighborhood density is defined to indicate the degree of aggregation of each object. The smaller the fuzzy neighborhood density of an object, the more likely it is to be an outlier. Based on this idea, the fuzzy neighborhood deviation degree is defined to quantify the degree of outliers of objects. Finally, the fuzzy deviation degree on the set of conditional attributes is constructed to indicate the outlier scores of objects. Experimental comparisons with state-of-the-art methods show that the proposed method has a significant improvement on the AUC index and applies to three types of data.**

***Index Terms*—Granular computing, fuzzy information granulation theory, fuzzy neighborhood, outlier detection, mixed-attribute data.**

I. Introduction

FUZZY Information Granulation (FIG) theory is an information processing method based on fuzzy set theory [1], which aims to cope with problems and phenomena with ambiguity and uncertainty. The core idea of FIG theory is to divide the complex information space into several subspaces with different granularity and abstraction levels, so as to realize the simplification and generalization of information. FIG theory includes two main aspects: one is the establishment of FIG, i.e., how to select appropriate granularity, form, and structure to construct fuzzy subspaces according to different objectives and constraints; the second is the application of FIG, i.e., how to use fuzzy subspaces for effective reasoning, analysis, and decision making. FIG theory has a wide range of applications in several fields, such as clustering [2], freight volume forecasting [3], association rule mining [4], conflict analysis [5], etc. However, to the best of our knowledge, Outlier Detection (OD) methods based on FIG theory have been little studied [6], [7], [8], [9], [10], such as weighted fuzzy-rough density-based [7], multi-fuzzy granules-based [8], fuzzy rule-based [11], clustering-based fuzzy outlier [12].

Neighborhood is an important concept in topology that has been widely discussed and applied in fields such as machine learning and knowledge discovery. In recent years, neighborhood-based anomaly detection methods have attracted much attention from scholars. Neighborhood-based detection models identify outliers mainly based on the neighborhood information or similarity information of the data. Usually, the neighborhood calculation mainly includes two strategies: (1) Neighborhood radius and (2) Number of nearest neighbors. The first strategy forms a ε-neighborhood by computing all objects whose distance from an object is no greater than a positive number ε; while the second strategy forms a neighborhood by discovering the k closest neighbors to an object, i.e., k-Nearest Neighbors (kNN). Therefore, the first strategy can be used to construct ε-neighborhood-based detection model [13], [14], [15]. For example, Chen et al. [13] introduced the neighborhood model to construct a neighborhood detection method. In addition, scholars have proposed kNN-based detection models [16], [17].

The direct method is a common way of kNN detection model [16], which calculates the distance between a data object and its kth nearest neighbor as the anomaly determination criterion. The more distant an object is, the more anomalous it is. Based on the above idea, some other detection models have been proposed one after another, such as Local Distance-based Outlier Factor (LDOF) [18], Mean-shift OD (MOD) [19], and Local Gravitation-based OD (LGOD) [20]. Another method based on kNN is the density-based anomaly detection method, which mainly determines whether the data is anomalous based on the density of the neighborhood of the data, such as Local Outlier Factor (LOF) [17], Local Correlation Integral (LOCI) [21], Local Outlier Probabilities (LoOP) [22], Connectivity-based Outlier Factor (COF) [23], INFLuenced Outlierness (INFLO) [24], and Relative Density-based Outlier Score (RDOS) [25]. The

Received 28 September 2023; revised 2 July 2024; accepted 17 October 2024. Date of publication 21 October 2024; date of current version 26 November 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62306196 and Grant 62376230, in part by Sichuan Science and Technology Program under Grant 2023YFQ0020, and in part by the Fundamental Research Funds for the Central Universities under Grant YJ202245. Recommended for acceptance by Z. Wang. *(Corresponding author: Qilin Li.)*

Zhong Yuan and Peng Hu are with the College of Computer Science, Sichuan University, Chengdu 610065, China (e-mail: yuanzhong@scu.edu.cn; penghu.ml@gmail.com).

Hongmei Chen is with the School of Computing and Artificial Intelligence, Southwest Jiaotong University, Chengdu 611756, China (e-mail: hmchen@swjtu.edu.cn).

Yingke Chen is with the Department of Computer and Information Sciences, Northumbria University, NE1 8ST Newcastle upon Tyne, U.K. (e-mail: yingke.chen@northumbria.ac.uk).

Qilin Li is with State Grid Sichuan Electric Power Company, Chengdu 610045, China (e-mail: li_qi_lin@163.com).

The code is publicly available online at https://github.com/BELLoney/DFNO.

Digital Object Identifier 10.1109/TKDE.2024.3484448

lower the density of an object is, the fewer its nearest neighbors are, and the higher its degree of anomaly.

Neighborhood-based detection model described above focuses on numeric type data. However, in real data, the data usually exist in the form of mixed types, i.e., there is a coexistence of nominal and numerical type data. For anomaly detection of mixed data, such methods usually replace nominal data with different integers. It is meaningless to use euclidean distances to calculate the distances between these substitution values. As a result, they cannot effectively detect anomalies in mixed-attribute data. Besides, they build detectors based on certainty strategies and thus also cannot describe the fuzziness of objects in a fuzzy context.

To address the shortcomings of the present methods, this paper proposes the idea of the fuzzy neighborhood to detect the outliers in mixed data. First, a hybrid fuzzy similarity is defined to calculate the fuzzy similarity relation between data objects, which lays the foundation for the construction of OD model for mixed data. Second, the concept of Fuzzy kNN (FkNN) is proposed based on the concept of the fuzzy neighborhood to describe the fuzzy local information in the data. Further, to reduce statistical volatility, the fuzzy reachable similarity is defined as the minimum value between k-similarity and actual similarity. From this, the concepts of fuzzy neighborhood density and fuzzy neighborhood deviation factor are given in turn. Finally, the fuzzy neighborhood outlier score is defined to characterize the outlier degree of the object by the set of conditional attributes. The larger the outlier score of an object, the more likely it is to be an outlier. Specifically, the innovative aspects of this article are summarized as follows.

1) As mentioned earlier, existing methods often fail to handle mixed attribute data (nominal and numeric) efficiently. To address this issue, we define a hybrid fuzzy similarity measure. It granulates a set of objects into a fuzzy granular structure, making it suitable for uncertain mixed data scenarios.
2) The kNN method is widely used but lacks the capability to handle fuzziness in data. By extending kNN into FIG theory, we propose the FkNN method, which enhances the granulation process in fuzzy domains.
3) We construct a fuzzy neighborhood outlier detection model based on FkNN, which effectively identifies outliers in various data types (nominal, numerical, and mixed).
4) To implement the theoretical model in practical scenarios, we design a feasible algorithm for outlier detection.
5) We provide experimental results on three types of publicly available datasets, showing that the proposed detection model can effectively detect outliers.

This article is organized as follows. In the next section, we show some relevant preliminary knowledge and analyze the shortcomings of the current LOF. The third section proposes fuzzy neighborhood-based OD and designs the corresponding algorithm. The fourth section demonstrates the effectiveness of the proposed algorithm through extensive experimental comparison analyses. Finally, we conclude and propose future work.

TABLE I
MAIN NOTATIONS OF THIS PAPER

Notation	Meaning
$\mathcal{D}$	Set of objects
o	$\forall o \in \mathcal{D}$
$\mathcal{A}, \mathcal{B}$	Set of conditional attributes
$R_{\mathcal{B}}$	Fuzzy relation
$G(\mathcal{B})$	Fuzzy granular structure
$[o]_{\mathcal{B}}$	Fuzzy information granule
$\mathcal{B}^{\text{nom}}, \mathcal{B}^{\text{num}}$	Nominal and numerical attribute subsets
HDM_{B}	Hybrid distance metric
$OM_{\mathcal{B}^{\text{nom}}}$	Overlap metric
$ED_{\mathcal{B}^{\text{num}}}$	Euclidean distance
$[o]_{\mathcal{B}}^{k}$	FkNN
$sim_{\mathcal{B}}^{k}$	k-similarity
$\text{reachsim}_{\mathcal{B}}^{k}$	Reachablility similarity
$FND_{\mathcal{B}}^{k}$	Fuzzy neighborhood density
$FNDD_{\mathcal{B}}^{k}$	Fuzzy neighborhood deviation degree
$FNOS^{k}$	Fuzzy neighborhood-based outlier score

II. PRELIMINARIES

Fuzzy binary relations are a common information granulation strategy, which refines the object into a set of fuzzy information granules [26]. For the convenience of readers, the main notations of this paper are listed in Table I.

Let $\mathcal{D} = \{o_1, o_2, \ldots, o_n\}$ denote a non-empty set of objects and $\mathcal{A} = \{a_1, a_2, \ldots, a_m\}$ denote a non-empty finite set of conditional attributes. $\forall \mathcal{B} \subseteq \mathcal{A}$, $\mathcal{B}$ can induce a fuzzy relation $R_{\mathcal{B}}$ on $\mathcal{D}$, which is defined as $R_{\mathcal{B}} : \mathcal{D} \times \mathcal{D} \to [0, 1]$. For any $o, p \in \mathcal{D}$, 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 called a fuzzy similarity relation. It can be written in matrix form, i.e., $M(R_{\mathcal{B}}) = (r_{ij}^{\mathcal{B}})_{n \times n}$, where $r_{ij}^{\mathcal{B}} = R_{\mathcal{B}}(o_i, o_j)$.

A fuzzy similarity relation can produce a family set of fuzzy information granules from the data, called a fuzzy granular structure. $\forall \mathcal{B} \subseteq \mathcal{A}$, a fuzzy granular structure with respect to $\mathcal{B}$ on $\mathcal{D}$ is defined as $G(\mathcal{B}) = \{[o_1]_{\mathcal{B}}, [o_2]_{\mathcal{B}}, \ldots, [o_n]_{\mathcal{B}}\}$, where $[o_i]_{\mathcal{B}}$ is a fuzzy information granule induced by $\mathcal{B}$. Obviously, $[o_i]_{\mathcal{B}}$ is also a fuzzy set with respect to $R_{\mathcal{B}}$ on $\mathcal{D}$ and $|[o_i]_{\mathcal{B}}| = \sum_{j=1}^{n} r_{ij}^{\mathcal{B}}$.

In response to the existing work that treats OD as a binary property, Breuning et al. [17] proposed a density-based detection scheme. In this detection scheme, the local outlier factor of an object $o \in \mathcal{D}$ is defined to characterize its outlier degree, which is defined as follows.

Definition 1: Given a positive integer k, the local outlier factor of o is defined as

$$LOF_k(o) = \frac{1}{|N_k(o)|} \sum_{p \in N_k(o)} \frac{lrd_k(p)}{lrd_k(o)}, \tag{1}$$

where $N_k(o)$ denotes k-nearest neighbors of o and $lrd_k(o)$ denotes local reachability density of o, which is calculated by

$$lrd_k(o) = \frac{|N_k(o)|}{\sum\limits_{p \in N_k(o)} \text{reach-dist}_k(o, p)},$$

where $\text{reach-dist}_k(o, p)$ denotes reachability distance of an object o with respect to object p.

Although LOF is an effective method for discovering outliers, it is specifically designed for numerical data. In many cases, both categorical and numerical attributes usually exist in the same dataset, i.e., mixed-attribute data. Therefore, it is necessary to study an OD model applicable to mixed data. Moreover, the above LOF is defined based on the k-nearest neighbors of o. For any $p \in \mathcal{D}$, there are only two cases of p with respect to $N_k(o)$, i.e., $p \in N_k(o)$ or $p \notin N_k(o)$. This reflects the certainty of LOF. Obviously, such a strategy may not accurately describe the fuzziness between data. Therefore, LOF method also cannot deal with fuzzy data effectively.

FIG theory uses the basic concept of fuzzy sets or the theory of continuous membership functions to granulate sets of objects. It allows elements to have partial membership degrees instead of only binary memberships of 0 or 1. This allows the theory to better handle uncertainty problems such as fuzziness. For this purpose, this paper defines FkNN to construct outlier scores of objects.

III. Detecting Fuzzy Neighborhood-Based Outliers

In this section, we give a hybrid fuzzy similarity, a definition of outliers, and a corresponding detection algorithm is designed.

A. Hybrid Fuzzy Similarity

Most real-life data exists in hybrid form (i.e., nominal and numerical attributes). Let $\mathcal{B} = \mathcal{B}^{\text{nom}} \cup \mathcal{B}^{\text{num}}$ and $\mathcal{B}^{\text{nom}} \cap \mathcal{B}^{\text{num}} = \emptyset$, where $\mathcal{B}^{\text{nom}}$ and $\mathcal{B}^{\text{num}}$ denote nominal and numerical attribute subsets, respectively. To handle these data efficiently, a hybrid distance metric is defined as

$$HDM_{\mathcal{B}}(o,p) = OM_{\mathcal{B}^{\text{nom}}}(o,p) + ED_{\mathcal{B}^{\text{num}}}(o,p), \tag{2}$$

where $OM_{\mathcal{B}^{nom}}(o,p) = |\{a \in \mathcal{B}^{nom} | a(o) \neq a(p)\}|$ denotes Overlap Metric between o and p w.r.t. nominal attribute subsets $\mathcal{B}^{\text{nom}}$. and $ED_{\mathcal{B}^{\text{num}}}(o,p)$ denotes Euclidean Distance between o and p w.r.t. numerical attribute subsets $\mathcal{B}^{\text{num}}$.

Further, the hybrid fuzzy similarity $R_{\mathcal{B}}(o,p)$ is calculated by

$$R_{\mathcal{B}}(o,p) = 1 - \frac{HDM_{\mathcal{B}}(o,p)}{|\mathcal{B}|}, \tag{3}$$

where $|\mathcal{B}|$ denotes the cardinality of the attribute subset $\mathcal{B}$. Before performing experiments, numerical attribute data are usually normalized to [0,1]. Therefore, the range of fuzzy similarity calculated by Eq. (3) is [0,1], which satisfies the definition of fuzzy relation.

In the above Eqs. (2)–(3), a hybrid distance metric is defined by fusing the overlap metric $OM_{\mathcal{B}^{nom}}$ and the euclidean distance $ED_{\mathcal{B}^{\text{num}}}$. Further, the fuzzy relation $R_{\mathcal{B}}$ is computed by the hybrid distance metric $HDM_{\mathcal{B}}$. Finally, a fuzzy granular structure $G(\mathcal{B})$ can be induced through the fuzzy relation. It can be seen that it is suitable for dealing with mixed data when $\mathcal{B}^{\text{nom}} \neq \emptyset$ and $\mathcal{B}^{\text{num}} \neq \emptyset$; it is suitable for dealing with nominal data when $\mathcal{B}^{\text{nom}} \neq \emptyset$ and $\mathcal{B}^{\text{num}} = \emptyset$; it is suitable for dealing with numerical data when $\mathcal{B}^{\text{nom}} = \emptyset$ and $\mathcal{B}^{\text{num}} \neq \emptyset$. This lays a theoretical foundation for the construction of subsequent outlier detection models for nominal, numerical, and mixed-attribute data.

B. Definition of Outliers

As described in Section II, each object is assigned to a class based on the classes of its kNN in the feature space. The object either belongs to the k-nearest neighbors of a query point or it does not. Obviously, such a strategy does not describe the fuzziness between the data well. For this reason, we introduce the idea of kNN into the FIG theory and propose the following idea of FkNN.

Definition 2: For any $k \in N_+$, the FkNN of o w.r.t. $\mathcal{B}$ is defined by

$$[o]^k_{\mathcal{B}}(p) = \begin{cases} R_{\mathcal{B}}(o,p), & R_{\mathcal{B}}(o,p) \geq \text{sim}^k_{\mathcal{B}}(o) \text{ and } p \neq o; \\ 0, & R_{\mathcal{B}}(o,p) < \text{sim}^k_{\mathcal{B}}(o) \text{ or } p = o, \end{cases} \tag{4}$$

where $\text{sim}^k_{\mathcal{B}}(o)$ denotes the k-similarity of o, which satisfies the following condition.

1) There are at least k objects $p \in \mathcal{D} - \{o\}$ such that $R_{\mathcal{B}}(o,p) \geq sim^k_{\mathcal{B}}(o)$;
2) There are at most $(k-1)$ objects $p \in \mathcal{D} - \{o\}$ such that $R_{\mathcal{B}}(o,p) > sim^k_{\mathcal{B}}(o)$,

The $[o]^k_{\mathcal{B}}$ of an object o collect objects whose similarity to it is greater than or equal to $sim^k_{\mathcal{B}}(o)$ and their corresponding degrees of membership, which can be used to indicate the degree of aggregation of the object. The smaller the cardinality of an object's FkNN, the more dispersed and more likely the object is to be an outlier.

Definition 3: The fuzzy reachability similarity of p w.r.t. o regarding $\mathcal{B}$ is determined by

$$\text{reachsim}^k_{\mathcal{B}}(o \leftarrow p) = \min\{\text{sim}^k_{\mathcal{B}}(o), R_{\mathcal{B}}(p,o)\}. \tag{5}$$

In the above definition, if an object p is far from o, i.e., the similarity between them is small, the fuzzy reachable similarity between them is computed as their actual similarity; however, if they are very similar, the fuzzy reachable similarity between them is replaced by $\text{sim}^k_{\mathcal{B}}(o)$. The reason for the above strategy is that the statistical fluctuations of all $R_{\mathcal{B}}(o,p)$ can be reduced.

So far, we have defined FkNN and fuzzy reachable similarity. To detect outliers based on fuzzy neighborhoods, the fuzzy reachable similarity is utilized to define the fuzzy neighborhood density in which an object is located.

Definition 4: The fuzzy neighborhood density of o w.r.t. $\mathcal{B}$ is computed by

$$FND^k_{\mathcal{B}}(o) = \frac{1}{|[o]^k_{\mathcal{B}}|} \sum_{[o]^k_{\mathcal{B}}(p) \neq 0} \text{reachsim}^k_{\mathcal{B}}(o \leftarrow p). \tag{6}$$

According to the above definition, we can see that the fuzzy neighborhood density is calculated by the ratio of the sum of reachable similarities of o with respect to FkNN to its cardinality. Obviously, the smaller the density of an object, the more likely it is to be an outlier.

Next, the degree of deviation is defined to characterize the degree of anomaly of an object with respect to $\mathcal{B}$.

Definition 5: The fuzzy neighborhood deviation degree of o w.r.t. $\mathcal{B}$ is defined as

$$FNDD^k_{\mathcal{B}}(o) = \frac{1}{|[o]^k_{\mathcal{B}}|} \sum_{[o]^k_{\mathcal{B}}(p) \neq 0} \frac{FND^k_{\mathcal{B}}(p)}{FND^k_{\mathcal{B}}(o)}. \tag{7}$$

Algorithm 1: DFNO.

Input: A data table $\mathcal{D}$ and k.
Output: $FNOS^k$.
1 Initialize $FNOS^k \leftarrow \emptyset$;
2 Compute $M(R_{\mathcal{A}})$;
3 **for** $i \leftarrow 1$ *to* $|\mathcal{D}|$ **do**
4 　Record $sim^k_{\mathcal{A}}(o_i)$;
5 　Compute $[o_i]^k_{\mathcal{A}}$;
6 **end**
7 **for** $i \leftarrow 1$ *to* $|\mathcal{D}|$ **do**
8 　**for** $j \leftarrow 1$ *to* $|\mathcal{D}|$ **do**
9 　　Compute reachsim$^k_{\mathcal{A}}(o_i \leftarrow o_j)$;
10 　**end**
11 **end**
12 **for** $i \leftarrow 1$ *to* $|\mathcal{D}|$ **do**
13 　Compute $FNRD^k_{\mathcal{A}}(o_i)$;
14 **end**
15 **for** $i \leftarrow 1$ *to* $|\mathcal{D}|$ **do**
16 　Compute $FNDD^k_{\mathcal{A}}(o_i)$;
17 **end**
18 Compute $FNOS^k(o_i)$;
19 **return** $FNOS^k$.

$FNDD^k_{\mathcal{B}}(o)$ integrates the sum of the ratio of the density of FkNN of p to the density of o w.r.t. $\mathcal{B}$. It can be seen that the lower the density of o and the higher the density of FkNN of p, the higher the $FNDD^k_{\mathcal{B}}$ value of o.

For each $\mathcal{B} \subseteq \mathcal{A}$, we can compute the degree of deviation of an object o. Thus, we can get $2^{|\mathcal{A}|}$ of $FNDD^k_{\mathcal{B}}(o)$ to perform information fusion to get the object's anomaly score. However, this is not desirable due to the fact that such a strategy will make the time complexity of the algorithm exponential. For this reason, we consider only the set of conditional attributes $\mathcal{A}$ to compute the outlier scores of objects. The experimental part will confirm that the strategy yields superior detection results in most cases.

Definition 6: The fuzzy neighborhood-based outlier score of o is computed as

$$FNOS^k(o) = FNDD^k_{\mathcal{A}}(o). \tag{8}$$

The larger the value of an object's anomaly score, the greater the possibility of it becoming an outlier.

Based on the above concept of outlier scores, we next give the definition of fuzzy neighborhood outliers as follows.

Definition 7: A threshold μ is given. For each $o \in \mathcal{D}$, o is judged as a fuzzy neighborhood outlier if $FNOS^k(o) > \mu$.

C. Detection Algorithm

In this subsection, we propose the corresponding algorithm DFNO and analyze its complexity.

In Algorithm 1, we first initialize $FNOS^k$ to be empty. Then, the fuzzy relation matrix $M(R_{\mathcal{A}})$ is computed and its execution time is $(|\mathcal{D}| \times |\mathcal{D}|)$. Next, the reachable similarity is computed in Steps 7-11, and its execution time is also $(|\mathcal{D}| \times |\mathcal{D}|)$. Further, the fuzzy neighborhood density and fuzzy neighborhood deviation are computed sequentially, and they are both executed for $|\mathcal{D}|$ times. Finally, the fuzzy neighborhood anomaly score is calculated and $FNOS^k$ is returned. Thus, the time complexity of Algorithm 1 is $O(|\mathcal{D}|^2)$.

IV. Experimental Study

This section presents comparative experimental details and results of the proposed algorithm DFNO to evaluate its effectiveness and adaptability.

A. Experimental Preparations

We downloaded 27 OD datasets from the public Web[1] for validating the detection performance of the proposed model. These datasets are widely used in comparative studies of related OD. Table II summarizes an overview of the dataset information, including the name of the dataset, the number of samples, the number of attributes, the number of true outliers, and the type of data. As can be seen in Table II, the size, dimensionality, and type of the dataset vary, where the number of samples varies from 111 to 9172, the maximum dimensionality of the data is 279, and the type of the dataset includes three types, namely nominal, numeric, and mixed. These characteristics meet the comparison requirements of the proposed detection model. For the missing values in the dataset, the maximum frequency method is employed to fill them. In addition, for the numerical data in the dataset, the min-max normalization method is employed to normalize them to [0,1].

For a comprehensive comparison, we used two different evaluation indexes, namely the Receiver Operating Curve (ROC) and the Area Under Curve (AUC) [7], [8]. The ROC index is one of the commonly used evaluation indexes in anomaly detection. It is popular for its monotonicity and ease of interpretation. The higher the performance of a detection algorithm, the closer its ROC curve will be to the top left corner of the graph. However, it is difficult to determine which algorithm performs better when two algorithms have similar ROC curves. To address this issue, the AUC index was proposed as a measure of the overall effectiveness of the algorithm. The AUC value varies between 0 and 1, with values closer to 1 indicating better performance.

To conduct a fair experiment, DFNO is compared with 15 popular OD algorithms, which are described in Table III. From Table III, we can see that there are five types of comparison algorithms included, namely neighborhood computing, ensembles, probabilistic, representation learning, and rough computing strategies. In the experiments, LOF, LPOD, NC, VOS, and DCROD algorithms mainly involve the parameter k, so the optimal results are obtained by adjusting k from 2 to 60 in steps of 1. The number of base estimators for IForest is set to 100. For ITB, WDOD, and ODGrCR, the Fuzzy C-Mean (FCM) clustering discretization method is used to discretize the numerical attribute values with the number of discretization intervals of 3. For MIX, the optimal value is obtained by iterating 10 times. VarE is mainly concerned with parameter λ, and its optimal results are obtained on parameter set

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

TABLE II
DESCRIPTION OF EXPERIMENTAL DATA

No.	Datasets	Abbrs.	Number of attributes	Number of Samples	Number of anomalies	Types
1	Audiology_variant1	Audio	69	226	57	Nominal
2	Breast_cancer_variant1	Breast	9	286	85	Nominal
3	Chess_nowin_145_variant1	Chess1	36	1814	145	Nominal
4	Chess_nowin_87_variant1	Chess2	36	1756	87	Nominal
5	Lymphography	Lymph	18	148	6	Nominal
6	Monks_0_12_variant1	Monks1	6	240	12	Nominal
7	Monks_0_25_variant1	Monks2	6	253	25	Nominal
8	Mushroom_p_85_variant1	Mush	22	4293	85	Nominal
9	Tic_tac_toe_negative_26_variant1	Tic	9	652	26	Nominal
10	Ionosphere_b_24_variant1	Iono	34	249	24	Numeric
11	Iris_Irisvirginica_11_variant1	Iris	4	111	11	Numeric
12	Letter	Letter	32	1600	100	Numeric
13	pageblocks_1_258_variant1	Page	10	5171	258	Numeric
14	Sonar_M_10_variant1	Sonar	60	107	10	Numeric
15	Spambase_spam_56_variant1	Spam	57	2844	56	Numeric
16	Vowels	Vowel	12	1456	50	Numeric
17	Wbc_malignant_39_variant1	Wbc	9	483	39	Numeric
18	yeast_ERL_5_variant1	Yeast	8	1141	5	Numeric
19	Arrhythmia_variant1	Arrh	279	452	66	Mixed
20	Bands_band_6_variant1	Band	39	318	6	Mixed
21	CreditA_plus_42_variant1	Credit	15	425	42	Mixed
22	German_1_14_variant1	Germ	20	714	14	Mixed
23	Heart270_2_16_variant1	Heart	13	166	16	Mixed
24	Horse_1_12_variant1	Horse	27	256	12	Mixed
25	Sick_sick_35_variant1	Sick1	29	3576	35	Mixed
26	Sick_sick_72_variant1	Sick2	29	3613	72	Mixed
27	Thyroid_disease_variant1	Thyroid	28	9172	74	Mixed

TABLE III
DESCRIPTION OF THE COMPARISON ALGORITHMS

No.	Algorithms	Descriptions	Strategies
1	LOF (2000) [17]	Local Outlier Factor	Neighborhood computing
2	IForest (2012) [27]	Isolation Forest	Ensembles
3	ITB (2012) [28]	Information Theory-Based	Probabilistic
4	WDOD (2014) [29]	Weighted Density-based OD	Rough computing
5	ODGrCR (2015) [30]	OD based on GrC and Rough set	Rough computing
6	NOF (2016) [31]	Natural Outlier Factor	Neighborhood computing
7	LPOD (2018) [32]	Local Projection-based OD	Representation learning
8	NC (2018) [33]	Reverse uNreaChability	Representation learning
9	MIX (2019) [34]	A joint learning-based outlier detector in MIXed-Type Data	Neighborhood computing
10	VOS (2019) [35]	Virtual Outlier Score	Neighborhood computing
11	COPOD (2020) [36]	COPula-based Outlier Detector	Probabilistic
12	VarE (2020) [37]	Weighted Neighbourhood Information Network-based OD	Neighborhood computing
13	WNINOD (2021) [14]	Weighted Neighbourhood Information Network-based OD	Neighborhood computing
14	DCROD (2022) [38]	Directed density ratio Changing Rate-based OD	Neighborhood computing
15	ECOD (2022) [39]	Empirical Cumulative-based OD	Probabilistic
16	DFNO (Ours)	Detecting Fuzzy Neighborhood Outliers	Fuzzy Neighborhood computing

$\{10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}\}$. For DFNO, k is tuned from 2 to 60 to obtain the best results.

B. Analyses on ROC

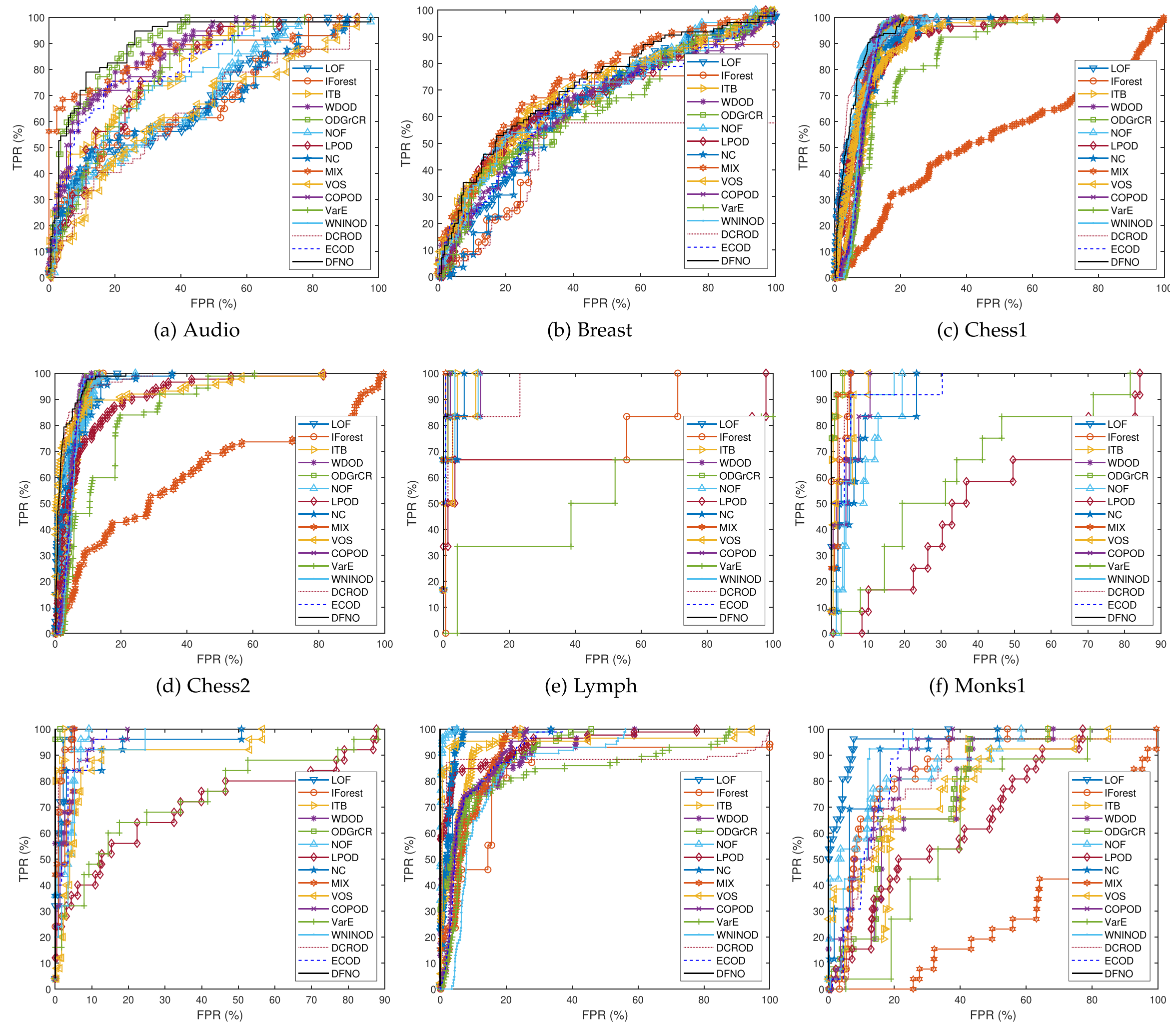
The ROC curves for the three types of datasets are shown in Figs. 1, 2, and 3, where the black curve indicates the proposed algorithm DFNO. The specific analyses are as follows.

1) Comparing the ROC curves for the three types of datasets, we can see that the ROC curve for each algorithm is indeed monotonically non-decreasing. This is consistent with the characteristics of the ROC curves.
2) The ROC curve of DFNO is closest to the upper left corner of the first quadrant on most datasets, and in particular shows excellent performance on datasets Lymph, Monks1, Monks2, Mush, Tic, Iono, Iris, Letter, Sonar, Yeast, and Sick1. This indicates optimal detection performance on these datasets.
3) On datasets Audio, Chess1, and Wbc, the ROC curves for both algorithms showed similar patterns, indicating that it is difficult to determine which algorithm is superior in this case.

Based on the results of the ROC curve analysis above, DFNO shows superior performance in most cases, but on some data sets the performance is similar to other algorithms and it is difficult to determine the superiority or inferiority.

C. Analyses on AUC

In this subsection, we further give the results of the comparison of the AUC of 16 detection algorithms. The experimental



(a) Audio (b) Breast (c) Chess1

(d) Chess2 (e) Lymph (f) Monks1

(g) Monks2 (h) Mush (i) Tic

Fig. 1. ROC curves for nominal datasets.

comparison of AUC performance is summarised in Table IV, where the best result is highlighted in bold and the second best result is underlined.

Based on these 27 publicly available datasets, the performance of the proposed detection algorithm is more effectively reflected by Table IV. Some of the relevant analyses are as follows.

1) DFNO achieves higher AUC values on most of the datasets. For example, on the dataset Audio, the AUC value of DFNO is 0.9028, which is greater than those of all other algorithms.
2) DFNO performs well on most datasets, compared to other algorithms whose best results are only seen on a very small number of datasets.
3) The final average also provides better insight and validation of the comparative performance. As can be seen from Table IV, DFNO obtains the best value of 0.9352, significantly greater than those of the other algorithms.

These datasets also include three types of attributes. Thus, the comparative analyses above also demonstrate that DFNO can effectively detect outliers in data with multiple attribute types.

D. Analyses on Hypothesis Testing

In this subsection, following the previous work [7], [8], Friedman test and Nemenyi post-hoc test are further performed to validate the statistically significant differences between the 16 comparison algorithms mentioned above.

According to Friedman test, AUC value of each algorithm on all datasets is sorted from low to high, and the sequence number is assigned $(1, 2, \ldots)$. If AUC value of the two algorithms is the same, the ordinal values are equally divided. Then, Friedman test is used to determine whether 16 comparison algorithms have the same performance.

Suppose M algorithms are compared on N datasets, and let r_i denote the average ordinal value of the ith algorithm, then Friedman test is calculated by

$$\tau_F = \frac{(N-1)\tau_{\chi^2}}{N(M-1)-\tau_{\chi^2}} \text{ and}$$

$$\tau_{\chi^2} = \frac{12N}{M(M+1)}\left(\sum_{i=1}^{M} r_i^2 - \frac{M(M+1)^2}{4}\right). \qquad (9)$$

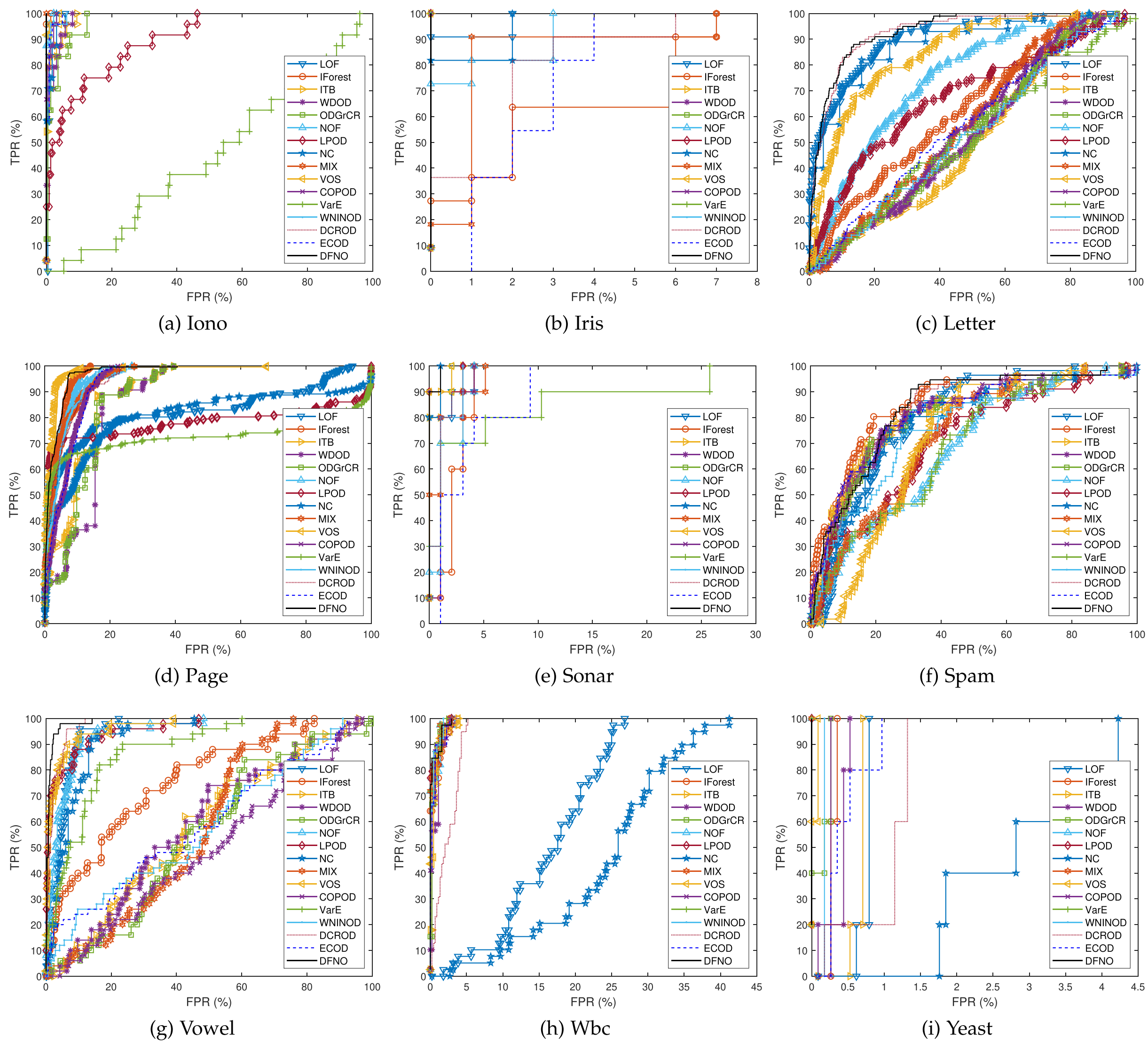


Fig. 2. ROC curves for numerical datasets.

τ_F obeys the F distribution with $(M-1)$ and $(M-1)(N-1)$ degrees of freedom. If the null hypothesis of “all algorithms have the same performance” is rejected, it means that the performance of the algorithms is significantly different. At this time, Nemenyi post-hoc test is used to further distinguish these algorithms. In Nemenyi test, the critical difference (CD) of the average ordinal value is calculated by

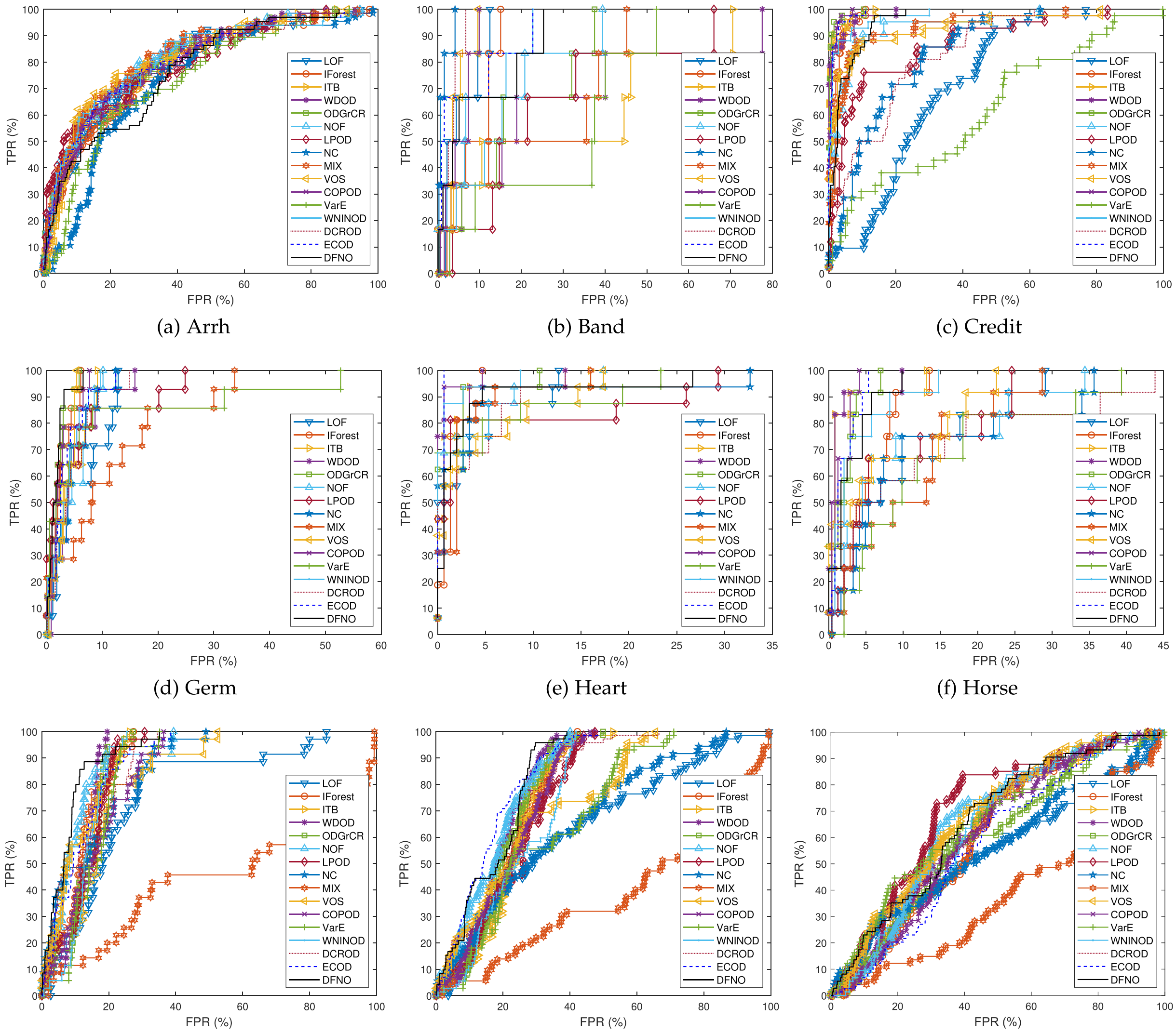
$$CD_\alpha = q_\alpha \sqrt{\frac{M(M+1)}{6N}}, \tag{10}$$

where q_α is the critical value of Tukey’s distribution, which can be found in [40].

Nemenyi test figure can be used to more intuitively represent the significant differences between the two algorithms [8]. In Nemenyi test figure, for each algorithm, a dot is used to show its average ordinal value, and a horizontal line segment with the dot as the center is used to indicate the size of CD. If a group of algorithms is connected by horizontal line segments, then it means that there is no significant difference between this group of algorithms.

According to Table IV, we can obtain $M = 16$ and $N = 27$, so τ_F distribution has 15 and 390 degrees of freedom. According to Friedman test, when $\alpha = 0.05$, the value of $\tau_F = 5.4799$ is greater than the critical value 1.6921. Therefore, we should reject the null hypothesis that “all algorithms have the same performance”. It shows that the performance of 16 OD algorithms is significantly different. At this time, a post-hoc test needs to be used to further distinguish them.

For significance level $\alpha = 0.05$, we can obtain $CD_{0.05} = 4.4393$. Finally, Nemenyi test figure on AUC is shown in Fig. 4. From Fig. 4, we can see that DFNO is statistically significantly different from most other algorithms. For example, it can be seen from Fig. 4 that DFNO is not connected to ECOD, COPOD, VOS, IForest, WNINOD, DCROD, WDOD, ITB, LOF, MIX, NC, LPOD, and VarE with horizontal line segments, which indicates that DFNO is statistically significantly



(a) Arrh (b) Band (c) Credit

(d) Germ (e) Heart (f) Horse

(g) Sick1 (h) Sick2 (i) Thyroid

Fig. 3. ROC curves for mixed datasets.

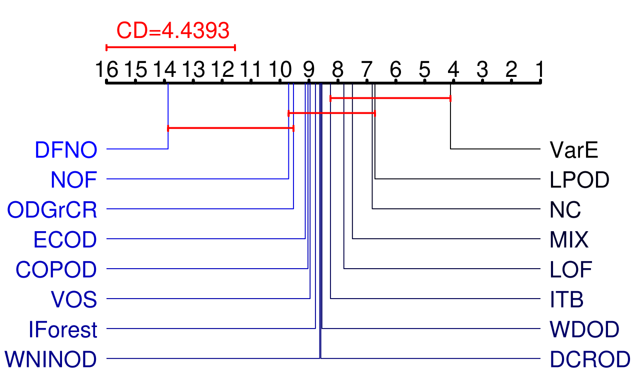


Fig. 4. Nemenyi test figures on AUC.

different from them. However, there is no consistent evidence to indicate the statistical differences among DFNO, NOF, and ODGrCR.

E. Analyses on Parameter

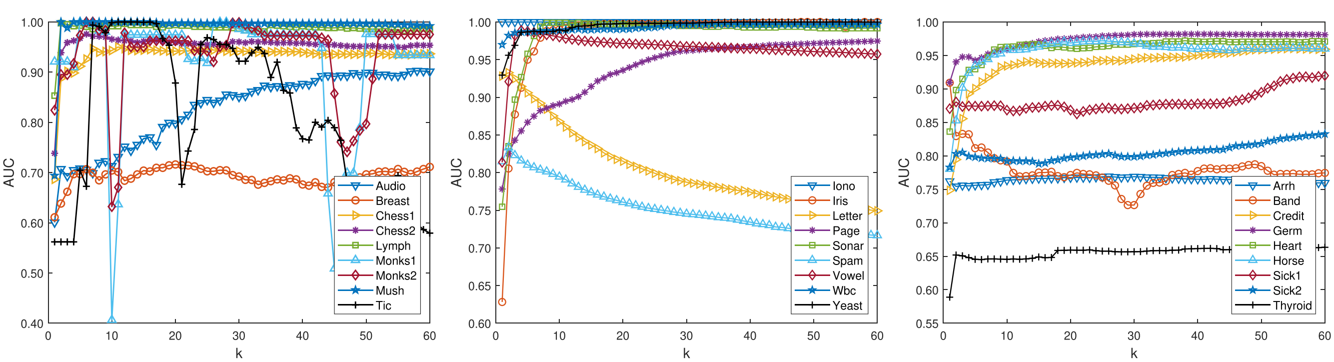
The plots of AUC with respect to parameters k for nominal, numerical, and mixed data are drawn in Fig. 5(a)–5(c). From Fig. 5(a)–5(c), we have the following analyses.

1) AUC values on most of the data sets first increased and then gradually levelled off as k increased. This indicates that the sensitivity of DFNO to the parameter k decreases when the parameter k reaches a certain value.
2) On some datasets, such as Breast, Monks1, Monks2 and Tic, the AUC values fluctuated significantly with increasing parameter k, indicating that DFNO is very sensitive to these datasets.
3) DFNO can achieve better AUC values with appropriate values of k, which indicates the need for adaptive parameter tuning to obtain optimal results.

In summary, DFNO has some sensitivity to the parameter k. Therefore, how to determine the effective optimal parameters is a major issue to improve the detection performance.

TABLE IV
COMPARISON OF AUC VALUES OF 16 OD ALGORITHMS

Datasets	LOF	IForest	ITB	WDOD	ODGrCR	NOF	LPOD	NC	MIX	VOS	COPOD	VarE	WNINOD	DCROD	ECOD	DFNO
Audio	0.6975	0.6850	0.8147	0.8680	**0.9025**	0.6940	0.7992	0.6855	0.8626	0.6510	0.8599	0.7923	0.7784	0.6454	0.8326	**0.9028**
Breast	0.6541	0.6352	0.6699	0.6786	0.6521	0.6761	0.6717	0.6420	**0.7336**	0.6947	0.6322	0.6369	0.6707	0.6035	0.6555	$\underline{0.7162}$
Chess1	0.9364	0.9285	0.9115	0.9194	0.9135	0.9388	0.9080	0.9230	0.5306	0.9133	0.9160	0.8478	0.9194	**0.9506**	0.9150	$\underline{0.9502}$
Chess2	0.9679	0.9566	0.9494	0.9503	0.9487	0.9543	0.9140	0.9554	0.6118	0.9395	0.9514	0.8571	0.9478	$\underline{0.9697}$	0.9522	**0.9759**
Lymph	0.9941	0.7805	0.9906	0.9742	0.9953	0.9765	0.6725	0.9824	**0.9977**	0.9742	0.9941	0.5070	0.9918	0.9542	0.9965	**0.9977**
Monks1	0.9730	0.9836	0.9945	0.9759	**0.9956**	0.9176	0.5658	0.9322	0.9846	0.9686	0.9587	0.6798	0.9532	0.9859	0.9572	**1.0000**
Monks2	0.9889	0.9882	0.9960	0.9802	**0.9995**	0.9711	0.7337	0.9598	0.9777	0.9163	0.9668	0.7630	0.9511	0.9916	0.9691	**1.0000**
Mush	0.9778	0.8388	0.9504	0.9118	0.9240	**0.9976**	0.9460	0.9758	0.9223	0.9547	0.9283	0.8304	0.8664	0.9251	0.9220	**1.0000**
Tic	**0.9673**	0.8616	0.7571	0.7755	0.7756	0.8747	0.6656	0.9321	0.2796	0.7969	0.8725	0.6433	0.9058	0.8259	0.8814	**1.0000**
Iono	0.9911	0.9994	0.9885	0.9854	0.9759	0.9985	0.9004	0.9927	**1.0000**	0.9980	0.9954	0.4559	0.9965	**1.0000**	0.9943	**1.0000**
Iris	0.9982	0.9709	**1.0000**	**1.0000**	**1.0000**	0.9936	**1.0000**	0.9964	0.9864	**1.0000**	**1.0000**	**1.0000**	**1.0000**	0.9827	0.9773	**1.0000**
Letter	0.9062	0.6295	0.5163	0.5294	0.5382	0.7481	0.6802	0.9013	0.5795	0.8482	0.5596	0.5292	0.5374	$\underline{0.9281}$	0.5723	**0.9327**
Page	0.8241	0.9702	0.8962	0.8663	0.8814	0.9630	0.7794	0.8031	0.9560	**0.9809**	0.9386	0.7269	0.9560	0.9528	0.9376	$\underline{0.9747}$
Sonar	0.9938	0.9763	0.9959	0.9969	0.9979	0.9845	0.9969	$\underline{0.9985}$	0.9887	0.9979	0.9856	0.9546	0.9887	0.9887	0.9660	**0.9990**
Spam	0.7996	**0.8455**	0.8013	0.7935	0.7899	0.6801	0.7042	0.7673	0.7304	0.7127	0.8146	0.6847	0.7543	0.8253	0.8123	$\underline{0.8333}$
Vowel	0.9510	0.7610	0.5799	0.5896	0.5519	0.9444	0.9579	0.9269	0.5827	0.9675	0.4958	0.8781	0.5807	$\underline{0.9777}$	0.5929	**0.9888**
Wbc	0.8365	0.9957	0.9949	0.9932	0.9955	0.9966	0.9967	0.7559	0.9969	0.9947	0.9955	**0.9973**	0.9971	0.9764	0.9955	$\underline{0.9971}$
Yeast	0.9924	0.9970	0.9933	0.9961	0.9986	**1.0000**	**1.0000**	0.9697	**1.0000**	0.9996	0.9974	**1.0000**	0.9982	0.9898	0.9952	**1.0000**
Arrh	0.8004	0.7881	0.8015	0.8108	0.8127	$\underline{0.8160}$	0.7992	0.7337	**0.8264**	0.8135	0.8046	0.7385	0.8146	0.7961	0.8071	0.7693
Band	0.9455	0.9156	0.7062	0.7345	0.8237	0.8739	0.7468	**0.9824**	0.7719	0.9573	0.9476	0.7025	0.8825	$\underline{0.9690}$	0.9359	0.9092
Credit	0.7265	0.9825	0.9910	0.9868	**0.9942**	0.9410	0.8727	0.8371	0.9370	0.9343	$\underline{0.9921}$	0.6250	0.9787	0.8406	0.9903	0.9605
Germ	0.9410	0.9754	0.9595	0.9528	$\underline{0.9795}$	0.9551	0.9488	0.9548	0.8914	0.9680	0.9730	0.9190	0.9650	0.9546	0.9655	**0.9820**
Heart	0.9729	0.9838	0.9825	0.9871	0.9850	0.9738	0.9492	0.9685	0.9738	0.9638	$\underline{0.9929}$	0.9621	0.9917	0.9704	**0.9963**	0.9721
Horse	0.9023	0.9539	0.9809	$\underline{0.9853}$	0.9798	0.9081	0.9109	0.8851	0.8931	0.9293	**0.9857**	0.8689	0.9658	0.8777	0.9805	0.9720
Sick1	0.7667	0.8813	0.8591	0.8802	0.8698	$\underline{0.8936}$	0.8667	0.8533	0.4329	0.8621	0.8386	0.8416	0.8644	0.8750	0.8833	**0.9198**
Sick2	0.6443	0.7845	0.7772	0.8077	0.7942	0.8264	0.7661	0.6781	0.3793	0.7402	0.7804	0.6905	0.7539	0.7876	**0.8425**	$\underline{0.8327}$
Thyroid	0.5485	0.6137	0.6489	0.6352	0.6037	0.6620	**0.7083**	0.5644	0.3907	$\underline{0.6714}$	0.6110	0.6496	0.6396	0.6461	0.5808	0.6634
Average	0.8777	0.8771	0.8706	0.8728	0.8770	0.8948	0.8319	0.8725	0.7858	0.8944	0.8810	0.7697	0.8759	$\underline{0.8959}$	0.8855	**0.9352**



(a) Nominal (b) Numeric (c) Mixed

Fig. 5. The AUC change curve on k.

V. CONCLUSION

This paper proposes an OD method based on fuzzy neighborhoods. The proposed method is not only applicable to many types of data sets but also can effectively handle the fuzzy local information in the data. The corresponding DFNO algorithm is designed and analyzed for the proposed method, and the results show that the proposed algorithm is within the acceptable time complexity. In order to verify the effectiveness of the proposed algorithm, the proposed algorithm is compared and analyzed with some state-of-the-art algorithms. The results show that the proposed algorithm achieves better performance in most cases in terms of both ROC curves and AUC values. Further, hypothesis statistical tests are performed to verify the statistically significant differences between the proposed algorithm and the other algorithms. Finally, the parametric analysis shows the sensitivity of the proposed algorithm to the parameters. However, the proposed method only considers outlier scores on all conditional attributes, which may lead to inaccurate detection of outliers in some special distribution data. Therefore, in future work, we will further consider the idea of multi-granularity to construct more adaptive OD models.

REFERENCES

[1] L. A. Zadeh, "Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic," *Fuzzy Sets Syst.*, vol. 90, no. 2, pp. 111–127, 1997.

[2] H. Y. Guo, L. D. Wang, X. D. Liu, and W. Pedrycz, "Information granulation-based fuzzy clustering of time series," *IEEE Trans. Cybern.*, vol. 51, no. 12, pp. 6253–6261, Dec. 2021.

[3] S. Yin, Y. C. Jiang, Y. Tian, and O. Kaynak, "A data-driven fuzzy information granulation approach for freight volume forecasting," *IEEE Trans. Ind. Electron.*, vol. 64, no. 2, pp. 1447–1456, Feb. 2017.

[4] F. Li, Y. Q. Tang, F. S. Yu, W. Pedrycz, Y. M. Liu, and W. Y. Zeng, "Multilinear-trend fuzzy information granule-based short-term forecasting for time series," *IEEE Trans. Fuzzy Syst.*, vol. 30, no. 8, pp. 3360–3372, Aug. 2022.

[5] G. M. Lang, D. Q. Miao, and H. Fujita, "Three-way group conflict analysis based on Pythagorean fuzzy set theory," *IEEE Trans. Fuzzy Syst.*, vol. 28, no. 3, pp. 447–461, Mar. 2020.

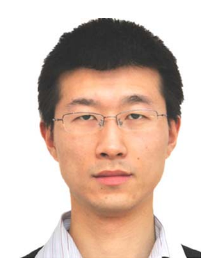
[6] N. Nikolova, R. M. Rodríguez, M. Symes, D. Toneva, K. Kolev, and K. Tenekedjiev, "Outlier detection algorithms over fuzzy data with weighted least squares," *Int. J. Fuzzy Syst.*, vol. 23, no. 5, pp. 1234–1256, 2021.

[7] Z. Yuan, B. Y. Chen, J. Liu, H. M. Chen, D. Z. Peng, and P. L. Li, "Anomaly detection based on weighted fuzzy-rough density," *Appl. Soft Comput.*, vol. 134, 2023, Art. no. 109995.
[8] Z. Yuan, H. Chen, C. Luo, and D. Z. Peng, "MFGAD: Multi-fuzzy granules anomaly detection," *Inf. Fusion*, vol. 95, pp. 17–25, 2023.
[9] Z. Yuan, H. M. Chen, T. R. Li, B. B. Sang, and S. Wang, "Outlier detection based on fuzzy rough granules in mixed attribute data," *IEEE Trans. Cybern.*, vol. 52, no. 8, pp. 8399–8412, Aug. 2022.
[10] H. Yadav, J. Singh, and A. Gosain, "Experimental analysis of fuzzy clustering techniques for outlier detection," *Procedia Comput. Sci.*, vol. 218, pp. 959–968, 2023.
[11] K. Kiersztyn and A. Kiersztyn, "Fuzzy rule-based outlier detector," in *Proc. 2022 IEEE Int. Conf. Fuzzy Syst.*, 2022, pp. 1–7.
[12] N. A. Yousri, M. A. Ismail, and M. S. Kamel, "Fuzzy outlier analysis a combined clustering-outlier detection approach," in *Proc. 2007 IEEE Int. Conf. Syst. Man Cybern.*, 2007, pp. 412–418.
[13] Y. M. Chen, D. Q. Miao, and H. Y. Zhang, "Neighborhood outlier detection," *Expert Syst. Appl.*, vol. 37, no. 12, pp. 8745–8749, 2010.
[14] Y. Wang and Y. P. Li, "Outlier detection based on weighted neighbourhood information network for mixed-valued datasets," *Inf. Sci.*, vol. 564, pp. 396–415, 2021.
[15] X. Zhang, Z. Yuan, and D. Miao, "Outlier detection using three-way neighborhood characteristic regions and corresponding fusion measurement," *IEEE Trans. Knowl. Data Eng.*, vol. 36, no. 5, pp. 2082–2095, May 2024.
[16] S. Ramaswamy, R. Rastogi, and K. Shim, "Efficient algorithms for mining outliers from large data sets," in *Proc. 2000 ACM SIGMOD Int. Conf. Manage. Data*, 2000, pp. 427–438.
[17] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying density-based local outliers," in *Proc. 2000 ACM SIGMOD Int. Conf. Manage. Data*, 2000, pp. 93–104.
[18] K. Zhang, M. Hutter, and H. D. Jin, "A new local distance-based outlier detection approach for scattered real-world data," in *Proc. Pacific-Asia Conf. Knowl. Discov. Data Mining*, Springer, 2009, pp. 813–822.
[19] J. W. Yang, S. Rahardja, and P. Fränti, "Mean-shift outlier detection and filtering," *Pattern Recognit.*, vol. 115, 2021, Art. no. 107874.
[20] J. Xie, Z. Y. Xiong, Q. Z. Dai, X. X. Wang, and Y. F. Zhang, "A local-gravitation-based method for the detection of outliers and boundary points," *Knowl.-Based Syst.*, vol. 192, 2020, Art. no. 105331.
[21] S. Papadimitriou, H. Kitagawa, P. B. Gibbons, and C. Faloutsos, "LOCI: Fast outlier detection using the local correlation integral," in *Proc. 19th Int. Conf. Data Eng.*, 2003, pp. 315–326.
[22] H.-P. Kriegel, P. Kröger, E. Schubert, and A. Zimek, "LoOP: Local outlier probabilities," in *Proc. 18th ACM Conf. Inf. Knowl. Manage.*, 2009, pp. 1649–1652.
[23] J. Tang, Z. Chen, A. W.-C. Fu, and D. W. Cheung, "Enhancing effectiveness of outlier detections for low density patterns," in *Proc. Pacific-Asia Conf. Knowl. Discov. Data Mining*, Springer, 2002, pp. 535–548.
[24] W. Jin, A. K. Tung, J. Han, and W. Wang, "Ranking outliers using symmetric neighborhood relationship," in *Proc. Pacific-Asia Conf. Knowl. Discov. Data Mining*, Springer, 2006, pp. 577–593.
[25] B. Tang and H. B. He, "A local density-based approach for outlier detection," *Neurocomputing*, vol. 241, pp. 171–180, 2017.
[26] Y. H. Qian, J. Y. Liang, W. Z. Wu, and C. Y. Dang, "Information granularity in fuzzy binary GrC model," *IEEE Trans. Fuzzy Syst.*, vol. 19, no. 2, pp. 253–264, Apr. 2011.
[27] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation-based anomaly detection," *ACM Trans. Knowl. Discov. Data*, vol. 6, no. 1, pp. 1–39, 2012.
[28] S. Wu and S. Wang, "Information-theoretic outlier detection for large-scale categorical data," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 3, pp. 589–602, Mar. 2013.
[29] X. W. Zhao, J. Y. Liang, and F. Y. Cao, "A simple and effective outlier detection algorithm for categorical data," *Int. J. Mach. Learn. Cybern.*, vol. 5, pp. 469–477, 2014.
[30] F. Jiang and Y. M. Chen, "Outlier detection based on granular computing and rough set theory," *Appl. Intell.*, vol. 42, no. 2, pp. 303–322, 2015.
[31] J. L. Huang, Q. S. Zhu, L. J. Yang, and J. Feng, "A non-parameter outlier detection algorithm based on natural neighbor," *Knowl.-Based Syst.*, vol. 92, pp. 71–77, 2016.
[32] H. W. Liu, X. L. Li, J. Y. Li, and S. C. Zhang, "Efficient outlier detection for high-dimensional data," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 48, no. 12, pp. 2451–2461, Dec. 2018.
[33] X. J. Li, J. C. Lv, and Z. Yi, "An efficient representation-based method for boundary point and outlier detection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 1, pp. 51–62, Jan. 2018.
[34] H. Z. Xu, Y. J. Wang, Y. J. Wang, and Z. Y. Wu, "MIX: A joint learning framework for detecting both clustered and scattered outliers in mixed-type data," in *Proc. 2019 IEEE Int. Conf. Data Mining*, 2019, pp. 1408–1413.
[35] C. Wang, Z. Liu, H. Gao, and Y. Fu, "VOS: A new outlier detection model using virtual graph," *Knowl.-Based Syst.*, vol. 185, 2019, Art. no. 104907.
[36] Z. Li, Y. Zhao, N. Botta, C. Ionescu, and X. Y. Hu, "COPOD: Copula-based outlier detection," in *Proc. 2020 IEEE Int. Conf. Data Mining*, 2020, pp. 1118–1123.
[37] X. Li, J. Lv, and Z. Yi, "Outlier detection using structural scores in a high-dimensional space," *IEEE Trans. Cybern.*, vol. 50, no. 5, pp. 2302–2310, May 2020.
[38] K. S. Li et al., "Robust outlier detection based on the changing rate of directed density ratio," *Expert Syst. Appl.*, vol. 207, 2022, Art. no. 117988.
[39] Z. Li, Y. Zhao, X. Y. Hu, N. Botta, C. Ionescu, and G. Chen, "ECOD: Unsupervised outlier detection using empirical cumulative distribution functions," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 12, pp. 12181–12193, Dec. 2023.
[40] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learn. Res.*, vol. 7, no. Jan., pp. 1–30, 2006.

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

Peng Hu received the PhD degree in computer science and technology from Sichuan University, China, in 2019. He is currently an associate research professor with the College of Computer Science, Sichuan University. From 2019 to 2020, he was a research scientist with Institute for Infocomm, Research Agency for Science, Technology, and Research (A*STAR) Singapore. His research interests mainly focus on multi-view learning, cross-modal retrieval, and network compression. On these areas, he has authored more than 30 articles in the top-tier conferences and journals.

Hongmei Chen (Member, IEEE) received the MSc degree from the University of Electronic Science and Technology of China, Chengdu, China, in 2000, and the PhD degree from the Southwest Jiaotong University, Chengdu, China, in 2013. She is currently a professor with the School of Computing and Artificial Intelligence, Southwest Jiaotong University. Her research interests include the areas of data mining, pattern recognition, fuzzy sets, and rough sets.



Yingke Chen received the PhD degree from Aalborg University, Denmark, in 2013. He has held postdoctoral research associate positions with Queen's University Belfast, U.K. and Georgia University, USA. Currently, he is an associate professor with the Department of Computer and Information Sciences, Northumbria University, Newcastle, U.K. His research interests include artificial intelligence, multiagent learning, and formal methods.

Qilin Li received the BSc degree in computer application, and the MSc and PhD degrees in computer software and theory and computer application from the University of Electronic Science and Technology of China, Chengdu, China, in 1996, 1999, and 2006. He is currently a professorate senior engineer with State Grid Sichuan Electric Power Company, Chengdu. His research interests include power system automation, granular computing, and neural networks.