IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 28 July 2023, accepted 16 August 2023, date of publication 21 August 2023, date of current version 29 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3307190

RESEARCH ARTICLE

Improved Denclue Outlier Detection Algorithm With Differential Privacy and Attribute Fuzzy Priority Relation Ordering

HUANGZHI XIA[1,2], (Member, IEEE), LIMIN CHEN[3], DONGYAN WANG[1,2], AND XIAOTONG LU[1,2]

[1]School of Mathematical Science, Mudanjiang Normal University, Mudanjiang 157011, China
[2]Institute of Applied Mathematics, Mudanjiang Normal University, Mudanjiang 157011, China
[3]School of Computer and Information Technology, Mudanjiang Normal University, Mudanjiang 157011, China

Corresponding author: Limin Chen (chenlimin_clm@126.com)

This work was supported in part by the Natural Science Foundation of Heilongjiang Province under Grant LH2019F051; and in part by the Science and Technology Innovation Projects of Mudanjiang Normal University under Grant kjcx2023-126mdjnu, Grant kjcx2023-125mdjnu, Grant kjcx2023-124mdjnu, and Grant kjcx2023-123mdjnu.

ABSTRACT Outlier detection is an important method in data mining. Although Denclue algorithm is particularly good at finding clusters of arbitrary shape and detecting outliers, it does not protect the user's privacy well in the operation process. In this paper, differential privacy technology is introduced into Denclue algorithm to ensure the privacy security in the application of Denclue algorithm and outlier detection. Firstly, the differential privacy technology is used to add the Laplacian noise to the density to realize the sensitive information hiding among the data objects. Secondly, in order to compensate for the decrease of outlier detection accuracy caused by noise, the information entropy weight distance was introduced to amplify the influence of important attributes in the algorithm, and the density function of entropy weight distance was used to calculate each data point. Finally, through the method of ordering fuzzy priority relation, a new measure index is defined by analogy to measure the degree of outliers among the attributes. According to the measure index, the attributes are reordered and the weight distance of information entropy is improved. A differential privacy the Denclue outlier detection algorithm based on attribute fuzzy priority ordering (EAF-DP-Denclue) is proposed. The numerical results of the experiment show that the performance of EAF-DP-Denclue is more than that of traditional algorithms, and the identification process of EAF-DP-Denclue protects sensitive privacy information, and is better than that of the DP-DBScan outlier detection algorithm.

INDEX TERMS Denclue, differential privacy, fuzzy priority relation order, information entropy weights, outlier detection.

I. INTRODUCTION

With the advent of the era of big data, the amount of network information has exploded and spread rapidly in engineering, finance, medical and other fields, and the problem of private information leakage has become more and more serious. In the field of privacy protection, the traditional privacy preserving methods include k-anonymity proposed by Samarati [1], l-diversity proposed by Machanavajjhala [2], and t-tightness proposed by Li [3]. These privacy protections rely on the background knowledge of the attacker, which has excellent limitations when applied in privacy protection projects. After that, differential privacy is proposed by Dwork [4], which regardless of the amount of background information. The public favors differential privacy because it provides users with higher quality privacy protection.

Outlier detection is an essential data preprocessing method. The standard outlier detection methods include distance-based, statistics-based, density-based, and clustering-based methods [5], [6]. Clustering-based techniques are often used

The associate editor coordinating the review of this manuscript and approving it for publication was Jerry Chun-Wei Lin.

in data mining [7], [8], [9]. These clustering techniques typically use metrics (such as Euclidean distance) to measure the similarity between data points and group similar data points into clusters with similar behavior. If the data points do not belong to clusters or are in clusters much smaller than other clusters, these data points are considered outliers. The identification of outliers relies on the assumption that typical values belong to large clusters. Meanwhile, outliers belong to small clusters or do not belong to any clusters. In addition, they use inter/intra-cluster distance thresholds and cluster widths. However, these parameters are not easy to select and must be chosen correctly to produce valid results [10]. Clustering techniques gather similar data points into clusters and merge some clusters to reduce computational costs.

Denclue (Density-based Clustering) is an effective clustering method proposed by Hinneburg [11], which is often used in the process of outlier detection [12]. It uses a specific Density function to distinguish data points according to a pre-set density threshold, so as to complete the work of outlier detection. The Denclue can work well in clustering high-dimensional data. The newly proposed VDenclue is able to find clusters with highly variable densities based on varying kernel density estimation [13]. Meanwhile, the MR-VDenclue solves the problem of the high computational overhead of the VDenclue [14]. Wang [15] discusses in detail the selection of kernel bandwidth parameters for the Denclue in his study. In order to make the Denclue well adapted to the high-noise environment, Huang [16] defined a new noise index and proposed a new denoising method by combining reinforcement learning to consolidate the high noise data structure. However, in the process of outlier detection using the Denclue, if there are privacy leakage problems, it will lead to unimaginable consequences, and the application of differential privacy technology will inevitably lead to the reduction of outlier detection accuracy. Then, is there an outlier detection method that can protect privacy by using differential privacy technology without reducing the accuracy of outlier detection? In this paper, a positive answer is given.

This paper combines the differential privacy in outlier detection algorithm, the sensitive information protection in the process of algorithm, the fuzzy mathematics in the field of fuzzy priority relation ordering method put forward a new method to improve the change in the distance measure is insufficient, the algorithm of outlier detection precision at a higher level elegant solves this problem.

II. RELATED WORK

A. ADVANCED OUTLIER DETECTION METHOD

As a primary method for data mining and data analysis, outlier detection has received extensive attention and research from many researchers. Meanwhile, there have been many novel and high-performance outlier detection algorithms that have been frequently used in practical problems [17], [18], [19]. Yuan [20] proposed an anomaly detection method based on fuzzy-rough density. Firstly, a fuzzy-rough density is used to characterize the degree of aggregation of the objects, and then fuzzy entropy is introduced to calculate the weights of each attribute. In addition, the density and fuzziness of the data samples are considered, and an anomaly score is constructed to characterize the degree of anomaly of the samples. Huang [21] proposed the concept of the Natural Outlier Factor (NOF) for measuring outliers and proposed an algorithm based on Natural Neighbor (NaN). This algorithm does not require parameters for calculating the NOF of an object and solves the problem that most algorithms have difficulty in selecting appropriate parameters. In the following year, the scholar proposed the Relative Outlier Cluster Factor (ROCF) [22] based on the concept of the mutual neighbor graph. Similarly, the ROCF algorithm does not require any parameters for checking outlier clusters. Yang [23] proposed a sliding window model based on efficient pruning and information entropy for outlier detection in high-dimensional data flow and dynamic data flow scenarios with insufficient accuracy. A new sliding window and sub-sequence measurement mechanism are designed to determine whether a data point is abnormal or not based on the distance between the target sequence and other sequences. In addition, a pruning strategy is integrated into the model to reduce the computational complexity of the algorithm. Abhaya [24] proposed a novel technique to solve the reconstruction error problem based on deep learning theory, which utilizes the Density Peak Clustering (DPC) to identify possible outliers. Meanwhile, the Self Organizing Map (SOM) is used as another clustering method targeted to improve the shortcomings of the DPC when setting thresholds. To solve the problem that current outlier detection algorithms cannot detect global outliers, local outliers, and outlier clusters effectively at the same time, Huang [25] proposed an outlier detection algorithm based on the outlier turning points. The main idea of this method is to find the outlier turning points in the dataset, then adaptively obtain the neighborhood parameter k through the natural neighborhood, and finally consider all the outlier turning points and their sparse neighbors as outliers.

B. CLUSTERING ALGORITHM WITH DIFFERENTIAL PRIVACY PRESERVING

At present, there are many results of clustering algorithms based on differential privacy. As early as 2005, Blum [26] proposed a k-means clustering method using differential privacy protection technology, but Blum did not explain how to set the privacy budget in the iteration process of the algorithm, which may reduce the usability of the clustering results. Therefore, Dwork analyzed the sensitivity calculation method of each query function in the differential privacy k-means algorithm in detail, and proposed the allocation method of privacy budget corresponding to two cases. Li [27] modified the selection method of initial clustering centers and proposed IDPk-means algorithm to solve the problem of poor availability of k-means algorithm results after introducing differential privacy technology. After that,

Hu [28] proposed the DPk-means-up algorithm, which improved the clustering effect compared with the former scholars under the same level of privacy protection. In 2015, Wu [29] proposed the DP-DBScan algorithm by combining density clustering the DBScan algorithm with differential privacy protection technology. In 2018, Wang [30] proposed an improved the DP-OPTICS differential privacy protection algorithm and verified the feasibility of this method. The method adopts a clustering algorithm to pre-process personal privacy information to realize differential privacy protection, which can reduce the noise accumulation phenomenon caused by directly releasing histogram data, and at the same time reduce the reconstruction error caused by different merging methods of histograms. The OPTICS algorithm based on density clustering is applied to the DP-DBScan differential privacy algorithm, which is sensitive to the input of data parameters. Zheng [31] proposed a spectral clustering algorithm based on differential privacy by combining spectral clustering algorithm and differential privacy protection model. The algorithm is based on the differential privacy model and uses the cumulative distribution function to generate random noise that satisfies the Laplace distribution. The noise is added to the function of sample similarity calculated by the spectral clustering, which interferes with the weight values between sample individuals. The information hiding between sample individuals is realized to achieve the purpose of privacy protection. Li [32] integrated multiple heterogeneous clustering algorithms by Stacking, and used the k-means clustering, hierarchical clustering, spectral clustering and Gaussian mixture clustering as the first round clustering, and combined the silhouette coefficient to weight the clustering results generated by the first round clustering algorithm into the original data. Then, the k-means algorithm was used as the secondary clustering algorithm to cluster the expanded data set. For categorical data clustering, Nguyen [33] introduced privacy protection technology into the k-modes clustering, so that the operation process of the algorithm was protected. Its research presents several scenarios in both interactive and non-interactive environments. It is proved that the newly proposed mechanism satisfies differential privacy and operates linearly on a large amount of data.

C. OUTLIER DETECTION ALGORITHM WITH CLUSTERING

The outlier detection algorithm based on the Denclue is often used in practical problems. Tramacere [34] used the Denclue for galaxy exploration. This study first uses the DBScan algorithm to extract groups of neighboring pixels with significant fluxes from CCD frames, and then applies the Denclue algorithm to separate the contributions of overlapping sources based on the localization of local maxima patterns. Through iterative computation, each pixel is finally associated to the nearest local maximum. Heaster [35] used the Denclue for medical research. This study applied the Denclue algorithm to metabolic autofluorescence measurements to identify heterogeneity in tumor cell culture at the cellular level. This is due to the ability of the Denclue algorithm to enable better distribution of cell clusters in samples with known heterogeneity. Yin [36] used the Denclue for anomaly detection. This study proposes an anomalous working condition filtering method based on cue density clustering, and the main work consists of two stages: working condition filtering and wind deflection calculation. First, the Denclue algorithm is used to filter the data for working conditions and filter out the abnormal working condition data for power generation. Then, the data are further processed, including data smoothing and wind speed Nimbin. Finally, a regression model is established based on the relationship between the output power of the unit and the wind deflection angle yaw angle to obtain the wind deflection angle of the unit. Rehioui [37] improved Denclue algorithm and combined it with the k-means to apply it to social networks. The new algorithm absorbed the effectiveness of the improved Denclue algorithm in clustering and the accuracy of the k-means algorithm for the number of clusters. It serves the emotion discovery of Twitter users well, but ignores the privacy issues of users. Jin [38] applied the Denclue algorithm to community discovery. Firstly, the network data were mapped into a low-dimensional Euclidean space that could preserve the structural characteristics of nodes through spectral analysis technology, and then the Denclue algorithm was used to detect the community in the network, but the risk of privacy leakage was not considered.

To sum up, most of the clustering-based outlier detection algorithms do not consider the problem of privacy leakage. Xia [39] introduces differential privacy into the Denclue, but the algorithm with added noise significantly decreases the accuracy in outlier detection. How to prevent the privacy leakage of their outlier detection and ensure the accuracy of outlier detection is the focus of this paper. The work done in this paper mainly has the following two points.

- This paper makes the first attempt to introduce differential privacy technology into Denclue algorithm. Laplacian noise is added to the Denclue outlier detection algorithm to hide sensitive information between data objects when calculating the density function of each data point, and the Euclidean distance in kernel density estimation is replaced by the improved information entropy weight distance. Thus, the influence of important attributes in the algorithm is amplified.
- The information entropy weight distance is improved by the fuzzy priority relation ordering method, and a new measurement method is defined to describe the pros and cons of the outlier degree between the outlier attributes, which highlights the important attributes while weakening the unimportant attributes.

III. THEORETICAL BASIS

A. DIFFERENTIAL PRIVACY

The differential privacy mechanism ensures that every data individual is not exposed, which maximizes the availability

of query results, and the overall statistical characteristics of the data set can be understood by the outside world.

Definition 1: (Sibling data sets) Data set D' and D'' are sibling data sets, if and only if the number of data points in the set after symmetric difference operation with D' and D'' is 1, i.e. $\left|D' \oplus D''\right| = 1$.

Definition 2: (Differential privacy) Algorithm Q satisfies ε-differential privacy if and only if

$$P\left[Q\left(\mathrm{D}'\right) = O\right] \leq \exp\left\{\varepsilon * \left|\mathrm{D}' \oplus \mathrm{D}''\right|\right\} * P\left[Q\left(\mathrm{D}''\right) = O\right] \tag{1}$$

where, ε denotes the privacy budget, $P[X]$ denotes the occurrence risk of events X, and O denotes the result vector output by the algorithm Q.

Definition 3: (Laplace mechanism) The global sensitivity of given query function $f : D \to R^m$ is

$$\Delta f = \max_{D', D''} \left\| f\left(\mathrm{D}'\right) - f\left(\mathrm{D}''\right) \right\|_1 \tag{2}$$

where, $\|*\|$ denotes the first-order norm.

Definition 4: (Global sensitivity) The random algorithm Q is created by adding Laplace noise to the output function value of f. It goes as follows.

$$Q(\mathrm{D}) = f(\mathrm{D}) + \ Laplace\ \left(\frac{\Delta f}{\varepsilon}\right) \tag{3}$$

Then the stochastic algorithm Q satisfies differential privacy, and the density function of Laplace noise distribution is ε-differential privacy. For independent variable x, the density function of Laplace noise distribution is

$$P(x, b) = \frac{1}{2 * b} \exp\left\{-\frac{|x|}{b}\right\} \tag{4}$$

where parameter b is determined by global sensitivity Δf and privacy budget ε simultaneously, satisfying $b = \Delta f / \varepsilon$.

Theorem 1: (Sequence combinatorial property) Given a data set D, set a set of algorithms $Q_1(D), Q_2(D), \ldots, Q_s(D)$ on D, each algorithm satisfies ε_i-differential privacy ($i = 1, 2, \ldots, s$). The random processes of any two algorithms are independent of each other, then the combined algorithm Q of these algorithms satisfies $\sum_{i=1}^{s} \varepsilon_i$- differential privacy.

B. DENCLUE ALGORITHM

Denclue algorithms can find different size and shape of the clusters, the core idea of outlier detection using influence function on points in data space density estimate, the formation of a surface density, surface of the local density of high density point of interest, the corresponding attract area become a cluster, the density function value is less than the density of data points threshold output as outliers. Gaussian functions are usually used as influence and density functions to estimate the density of a given data point.

Definition 5: (Gaussian influence function) Select data points x and y in the m dimensional attribute space A^m, and the Gaussian influence function of data point y on x is Equation(5).

$$f_B^y : A^m \to R_0^+,\ f_B^y = \exp\left\{-\frac{d^2(x, y)}{2 * \sigma^2}\right\} \tag{5}$$

where $d(x, y)$ is the distance between x and y, and σ is a parameter in the Denclue algorithm, representing the kernel bandwidth.

Definition 6: (Global density function) For a given data set $D = \{x_1, x_2, \ldots, x_n\}$, $\forall x \in D$. Then the global density function of x is Equation(6).

$$f_B^D(x) = \sum_{i=1}^{n} f_B^{x_i}(x) = \sum_{i=1}^{n} \exp\left\{-\frac{d^2\left(x, x_i\right)}{2 * \sigma^2}\right\} \tag{6}$$

Definition 7: (Gradient) The gradient of the global density function $f_B^D(x)$ is denoted as Equation(7).

$$\nabla f_B^D(x) = \sum_{i=1}^{n} (x_i - x) f_B^{x_i}(x) = \sum_{i=1}^{n} (x_i - x) \exp\left\{-\frac{d^2\left(x, x_i\right)}{2 * \sigma^2}\right\} \tag{7}$$

when the density $f_B^D(x^*)$ of data point $x^* \in near(x^*)$ is the local maximum, x^* is called the density attraction point. There exists a parameter $\delta > 0$ hat controls the convergence rate and satisfies $x^0 = x^*, x^j = x^{j-1} + \delta \frac{\nabla f_B^D(x^{j-1})}{\left\|\nabla f_B^D(x^{j-1})\right\|}, (j = 1, 2, \ldots, k)$. If $d\left(x^*, x^j\right) \leq \sigma, (j = 1, 2, \ldots, k)$, the algorithm puts x into the cluster generated by x^*, x is said to be density attracted by x^*.

Definition 8: (Centers clusters and Arbitrary shape clusters) Given the density threshold parameter ξ and the subset D_1 of dataset D, if $f_B^D(x^*) \geq \xi$, D_1 is said to be the cluster determined for the center of x^*; If any two densities attract points $x_1^*, x_2^* \in V$. The density function for each data point x along the path between them has $f_B^D(x) \geq \xi$, then D_1 is called a cluster of arbitrary shape determined by the set V of density attraction points. When $f_B^D(x^*) < \xi$, all the points attracted by the density of x^* are treated as outliers.

The algorithm process of Denclue is given in Algorithm 1.

Algorithm 1 Denclue

Require: The continuous data set D, kernel bandwidth σ, density threshold ξ.

Ensure: Outlier data points.

1: According to Equation(6), the influence of each data point in the neighborhood is calculated;
2: Identify the points with the highest local density and use these points as the points of density attraction;
3: Associate each data point with a point of density attraction along the direction of maximum density growth;
4: The data points to be analyzed are assigned to the clusters represented by the density attraction points to form clusters with the center determined and clusters of arbitrary shape;
5: Merge data points with connected clusters;
6: The rest of the data whose density was lower than the density threshold ξ were output as outliers.

C. INFORMATION ENTROPY WEIGHT DISTANCE

Information entropy weight distance, also known as entropy-weight distance, is an extended form of traditional Euclidean distance, which was proposed by Wang [40]. It introduces the concepts of information entropy and weight based on Euclidean distance, changes the traditional measurement method between data points, and plays the role of highlighting important attributes.

Definition 9: (Information entropy) Given a dataset $D = \{x_1, x_2, \ldots, x_n\}$. *Its attribute set is* $A(D) = \{A_1, A_2, \ldots, A_m\}$, *m is the number of attributes of the data set, then the information entropy is Equation*(8).

$$E(A_i) = -\sum_{x_j \in I(A_i)}^{m} p\left(x_j\right) \log p\left(x_j\right) \tag{8}$$

where $I(A_i)$ *denotes the set of all possible values in attribute* $A_i (i = 1, 2, \ldots, m)$. $p\left(x_j\right)$ *denotes the probability that* x_j *occurs.*

Information entropy describes the degree of uncertainty of a signal. The larger the entropy value is, the greater the degree of uncertainty of information transmission.

Definition 10: (Outlier attributes and Attribute weights) Attribute A_i *is considered as an outlier attribute if the information entropy of attribute* A_i *satisfies the following inequality*

$$E(A_i) \geq \frac{1}{m} \sum_{j=1}^{m} E\left(A_j\right) \tag{9}$$

Then attribute $A_i (i = 1, 2, \ldots, m)$ *is given the corresponding attribute weight after judging by Equation*(9), *and the weight is given by Equation*(10) *as follows.*

$$\omega_i = \begin{cases} c, & A_i \text{ is the outlier attribute}, \\ 1, & A_i \text{ is not an outlier attribute}. \end{cases} \tag{10}$$

Among them, $\omega_i (i = 1, 2, \ldots, m)$ *is the weight corresponding to attribute* A_i, *which can highlight the degree of differentiation between the data points in the cluster and the outliers.*

Definition 11: (Entropy weight distance) The entropy weight distance is a weighted Euclidean distance based on information entropy, given a dataset $D, x, y \in D$, *then the entropy weight distance between them is denoted by Equation*(11).

$$Ed(x, y) = \sqrt{\sum_{i=1}^{m} \omega_i \left(g_{A_i}(x) - g_{A_i}(y)\right)^2} \tag{11}$$

where $g_{A_i}(x)$ *is the value of data point x on attribute* A_i, *and* $g_{A_i}(y)$ *is the value of data point y on attribute* A_i. *According to Equation*(6) *and Equation*(11), *the global density function estimation method based on entropy weight distance of data points is obtained as Equation*(12).

$$f_B^D(x) = \sum_{i=1}^{n} f_B^{x_i}(x) = \sum_{i=1}^{n} \exp\left\{-\frac{Ed^2(x, x_i)}{2 * \sigma^2}\right\} \tag{12}$$

D. FUZZY PRIORITY RELATION ORDERING

Fuzzy priority relation ordering method is a theoretical method used to determine the membership function of fuzzy sets in the field of fuzzy mathematics, through which the target objects can be ranked according to certain rules.

Definition 12: (Fuzzy priority relation) Let $Z = \{z_1, z_2, \ldots, z_n\}$, *establish a fuzzy relation* $\tilde{C} \in \mathscr{F}(Z \times Z)$ *on X according to some property. If the fuzzy relation* $\tilde{C}$ *atisfies the fuzzy priority relation, then the fuzzy priority matrix* $C = (c_{ij})$ *corresponding to the fuzzy relation is expressed as Equation*(13).

$$\begin{cases} c_{ii} = 0, \\ c_{ij} + c_{ji} = 1 \ (i \neq j). \end{cases} \tag{13}$$

where c_{ij} *represents the component in which* z_i *is superior to* z_j *when* z_i *is compared to* z_j.

Definition 13: (Cross sectional relation)Take a fixed threshold $\lambda \in [0, 1]$. *The* λ *Cross sectional relationship* C_λ *is denoted as Equation*(14).

$$C_\lambda = \left(c_{ij}^{\lambda}\right), c_{ij}^{\lambda} = \begin{cases} 1, & \text{if } c_{ij} \geq \lambda, \\ 0, & \text{if } c_{ij} < \lambda. \end{cases} \tag{14}$$

Definition 14: (Priority attribute)The threshold λ *is gradually lowered from* 1. *When* C_λ *appears for the first time, so that all the elements in row* i_1 *except the diagonal elements are* 1, x_{i_1} *is considered to be the first priority attribute (may not be unique). Then, the row* i_1 *and column* i_1 *are deleted from the fuzzy priority matrix C to obtain a new* $n - 1$ *order fuzzy priority matrix. The second priority attribute can be obtained by the same method. Recursively, all attributes can be sorted.*

IV. METHODOLOGY

A. USING DIFFERENTIAL PRIVACY DENCLUE ALGORITHM

The Denclue algorithm calculates the density function of each data point through Gaussian kernel density estimation method and combines it into a smooth density surface in space. However, if the density function relationship between any two data points x_i and x_j $(i, j = 1, 2, \ldots, n)$ is known by the attacker, the attacker may be based on the density function value of the data point. The sensitive information between two data points is inferred from the known kernel bandwidth σ, which leads to the leakage of sensitive information in the process of anomaly detection. In this paper, the differential privacy technology is used to calculate the density function value of each data point, and a certain amount of noise is added to each attribute for disturbance, and the noise term follows the Laplace$(\Delta f / \varepsilon)$ distribution, so as to achieve the purpose of privacy protection.

In order to make up for the decline in the accuracy of the outlier detection algorithm caused by noise addition, considering that the correlation degree of each attribute will affect the final outlier detection result when Denclue algorithm estimates the density function of data points, according to the different influence degree of each attribute on the outlier

detection result, when calculating the density function of data points, entropy weight distance is introduced to replace the Euclidean distance used in the calculation of Gaussian kernel density function in the Denclue algorithm, and more weight ω_i $(i = 1, 2, \ldots, m)$ is given to important attributes to highlight the outlier degree of the data points in the cluster and increase the degree of differentiation between the data points in the cluster and the outliers.

For the original data set $D = \{x_1, x_2, \ldots, x_n\}$ according to Equation(12), the influence of each data point x in the data set D in the neighborhood is determined, and the global density function estimate of x based on entropy weight distance is expressed by Equation(15).

$$EWd = \sum_{i=1}^{n} \exp\left\{-\frac{Ed^2(x, x_i)}{2 * \sigma^2}\right\} \tag{15}$$

Then, the noise following the Laplace$(\Delta f/\varepsilon)$ distribution is added to Equation (15), and the added noise density of any data pointis obtained as Equation (16).

$$EWd' = \sum_{i=1}^{n} \exp\left\{-\frac{Ed^2(x, x_i)}{2 * \sigma^2}\right\} + \text{Laplace}\left(\frac{\Delta f}{\varepsilon}\right) \tag{16}$$

The Denclue algorithm using differential privacy technology and measuring distance through Equation (16) is named as using differential privacy Denclue algorithm (DP-Denclue).

B. SYSTEM MODEL

The computing system model of the Denclue algorithm using differential privacy is shown in Fig. 1, including users and third-party servers. Now assume that the user has the original data set D, a trusted collector collect user information, and then calculated by the Equation (15), several sites, EWd' using differential privacy technology to the data point density add noise, disturbance EWd' are obtained by Equation (16), and will be released after the disturbance EWd' to the third-party server; Finally, the third-party server calculates the outlier detection results.

Furthermore, the traditional information entropy weight distance measurement method can magnify the influence degree of outlier attributes on the algorithm. However, this method assigns equal weight c to outlier attributes with different influence degrees, and highlighting outlier attributes is too general, ignoring the relationship between the degree of outlier among attributes. Here, the fuzzy priority relation sequence theory is introduced to improve the existing entropy weight distance, and the following method is proposed.

C. FUZZY PRIORITY RELATION ORDERING METHOD BASED ON INFORMATION ENTROPY

In this part, a fuzzy priority relation ordering method based on information entropy is proposed.

Definition 15: (Information entropy matrix) Let the set of attributes of a dataset D be $A(D) = \{A_1, A_2, \ldots, A_m\}$. *The information entropy matrix is defined as Equation (17).*

$$C = (c_{ij}) = \begin{pmatrix} c_{11} & c_{12} & c_{13} & \cdots & c_{1m} \\ c_{21} & c_{22} & c_{23} & \cdots & c_{2m} \\ c_{31} & c_{32} & c_{33} & \cdots & c_{3m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_{m1} & c_{m2} & c_{m3} & \cdots & c_{mm} \end{pmatrix} \tag{17}$$

where $c_{ij} = \frac{E(A_i)}{E(A_i)+E(A_j)}$, $c_{ji} = \frac{E(A_j)}{E(A_i)+E(A_j)}$ $(i, j = 1, 2, \ldots, m)$, $c_{ij} + c_{ji} = 1$ $(i \neq j)$. *Then* c_{ij} *represents the information entropy of* A_i *more than* A_j *when* A_i *is compared with* A_j, *and definition* $c_{ii} = 0$.

Definition 16: (Contrast matrix) Matrix $F_\alpha = (f_{ij})(i, j = 1, 2, \ldots, m)$ *is the contrast matrix,* $f_{ij} = \begin{cases} 1, c_{ij} \geq \alpha, \\ 0, c_{ij} < \alpha. \end{cases}$, $\alpha \in [0, 1]$.

Definition 17: (Uniform Outlier Value, UOV) According to Definition 15 and Definition 16, the sorted sequence of all attributes according to the outlier degree is obtained. According to this sorting, if the outlier degree of the first outlier attribute A_{i_1} *exceeds the threshold* α_1 *compared with other attributes, the Uniform Outlier Value (UOV) of attribute* A_{i_1} *is* α_1. *Similarly, the outlier degree of the second outlier attribute* A_{i_2} *exceeds the threshold* α_2 *compared with other remaining attributes (except the first outlier attribute). And so on, a dataset with d attributes will get* $d - 1$ *consistent outliers.*

According to the fuzzy priority relation ordering theory, the threshold α is used to divide the attributes with different outlier degrees, and these attributes are arranged in turn according to the order of division. The consistent outlier degree is used to measure the relationship between the outlier degrees of attributes and attributes. The threshold α is obtained one by one according to all elements C of the information entropy matrix c_{ij} from large to small, so as to determine the corresponding contrast matrix F_α. When there is a contrast matrix F_α, making the line i_1 elements in addition to the diagonal elements is equal to 1, is that A_{i_1} is the first group of attributes (can have multiple attributes at the same time as the first from the group of attributes), in the information entropy of matrix C delete line i_i and i_i in the first column, updated $m - 1$ order information entropy matrix $C^{(1)}$ is obtained, and the second outlier attribute can be obtained similarly. The above process is repeated, and all the attributes are sorted according to the degree of outlier. The fuzzy priority relation ordering process is shown in Fig. 2 below.

As an example, consider a dataset D with three attributes, attribute set $A(D) = \{A_1, A_2, A_3\}$. Its information entropy matrix is

$$C = \begin{pmatrix} 0 & 0.7 & 0.45 \\ 0.3 & 0 & 0.9 \\ 0.55 & 0.1 & 0 \end{pmatrix}$$

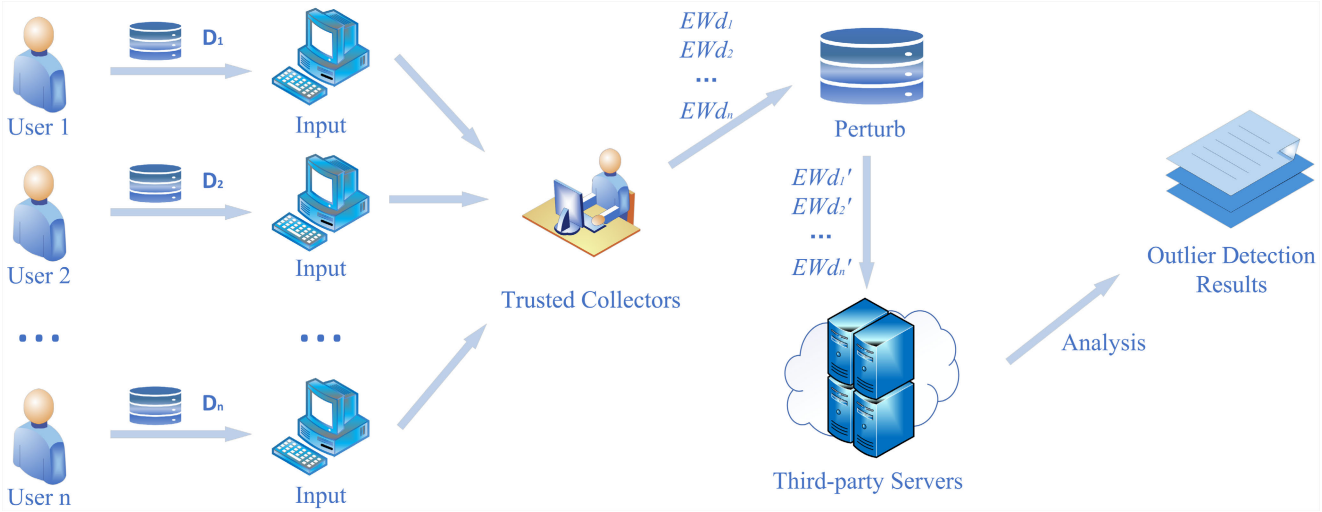

FIGURE 1. System model.

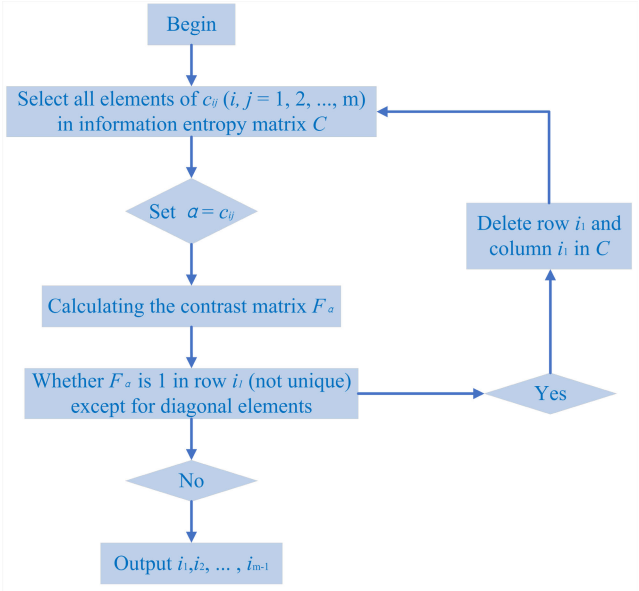

FIGURE 2. Fuzzy priority relation ordering method.

Let the threshold α be obtained from all c_{ij} in descending order, and the contrast matrix is obtained as follows.

$$F_{0.9} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}, \quad F_{0.7} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix},$$

$$F_{0.55} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}, \quad F_{0.45} = \begin{pmatrix} 0 & 1 & 1 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}.$$

When the threshold α is set to 0.45, it is the first time in the contrast matrix $F_{0.45}$ that the first row is all 1 except the diagonal elements. Therefore, attribute A_1 is identified as the first outlier attribute, which indicates that the outlier degree of attribute A_1 is more than 0.45 than other remaining attributes, and the consistent outlier degree of attribute A_1 is 0.45. Next, the information entropy matrix C is deleted from row 1 and column 1, and the updated information entropy matrix of 2 order is obtained as follows.

$$C^{(1)} = \begin{pmatrix} 0 & 0.9 \\ 0.1 & 0 \end{pmatrix}$$

Repeating the above steps, the contrast matrix is obtained as follows.

$$F_{0.9}^{(1)} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$$

Thus, attribute A_2 is identified as the second outlier attribute. And so on, the final outlier attribute influence degree is $A_1 > A_2 > A_3$, then the size relationship of attribute weight assignment is $\omega_1 > \omega_2 > \omega_3$.

After sorting all the attributes according to the degree of outlier, the specific weight ω_1 is given according to the permutation order. The weight of the attributes in the top order is adjusted up, and the weight of the attributes in the bottom order is adjusted down appropriately to obtain the improved entropy weight distance assignment scheme. The Algorithm 2 is the specific procedure of the using information entropy attribute fuzzy priority relation ordering method (EAF).

The Algorithm 2 can be for all attributes of the original data set from the group of degree to make sorting, according to the degree of different properties from the group gives the weights of different attributes, and then calculating the point density weighted attribute data, makes the Algorithm 2 when calculating the density of each data point can be more reasonable, and then put forward a kind of based on the difference of privacy modified the Denclue outlier detection algorithm. It is named as using attribute information entropy fuzzy priority relation ordering differential privacy the Denclue outlier detection algorithm (EAF-DP-Denclue). The reasonability of the algorithm lies in that the algorithm not only perturb the density value of data points estimated by the density function, so that the outlier detection process

Algorithm 2 EAF

Require: The continuous data set D.
Ensure: Attribute UOV, attribute order.

Each attribute of dataset D is discretized with equal interval;
Calculate the information entropy of each attribute;
Create a sequential array of attributes Order=[];
The information entropy matrix C is calculated from Equation (17);
while the number of rows in C is greater than 0 **do**
    Sort the elements of matrix C in descending order and save them into a one dimensional array sorted_C;
    Create a contrast matrix F of the same shape as matrix C;
    for $i = 0$ to len(sorted_C): // len(sorted_C)is the length of array sorted_C **do**
        Let α be equal to sorted_C[i];
        Calculate the contrast matrix F based on density thresholds α and Definition 15;
        Gets a row-indexed array index_array of matrix F that is equal to 1 in all but the diagonal positions;
        if len(index_ array) > 0 **then**
            Save index_array to Order;
            Delete the rows and columns corresponding to index_array in matrix C;
        end if
    end for
end while
return Order

is protected, but also reasonably magnify the effect of outlier attributes on outlier detection, so as to offset the accuracy loss of outlier detection after adding noise. The specific process is as follows the Algorithm 3.

D. SECURITY ANALYSIS

In this section, the privacy protection ability of the EAF-DP-Denclue algorithm is theoretically proved. Select sibling datasets D' and D'', and the privacy budget provided by the user is ε.

Proof: Let the Denclue algorithm estimate the density function using the improved Gaussian density function be f.

The results for data sets D' and D'' are $f(D') = (x'_1, x'_2, \ldots, x'_m)^T$ and $f(D'') = (x''_1, x''_2, \ldots, x''_m)^T$, respectively.

According to Definition 3, the global sensitivity can be expressed as

$$\Delta f = \max_{D',D'} \left\| f\left(\mathrm{D}'\right) - f\left(\mathrm{D}''\right) \right\|_1 = \max_{D',D'} \left(\sum_{i=1}^{m} \left| x'_i - x''_i \right| \right)$$

In this paper, the position of adding Laplace noise is in each attribute of the density function estimated by Denclue algorithm, so the density vector calculated by Denclue algorithm to add noise is specifically $O = (y_1, y_2, \ldots, y_m)^T$.

Algorithm 3 EAF-DP-Denclue

Require: The continuous data set D, kernel bandwidth σ, density threshold ξ, privacy budget ε.
Ensure: Outlier data points.

According to Equation (15), the influence of each data point in the neighborhood is calculated;
The attributes are sorted according to their UOV by Algorithm 2, and assign weight ω_i to each attribute;
The density EWd of each data point was calculated, and the privacy budget ε is allocated for each attribute of the point density. The added noise density is calculated according to Equation (16);
Identify the points with the highest local density and use these points as the points of density attraction;
Associate each data point with a point of density attraction along the direction of maximum density growth;
The data points to be analyzed are assigned to the clusters represented by the density attraction points to form clusters with the center determined and clusters of arbitrary shape;
Merge data points with connected clusters;
The rest of the data whose density was lower than the density threshold ξ were output as outliers.

Therefore, the Denclue algorithm calculates the density vector O obtained from the dataset D' as the risk of occurrence is

$$P\left[Q\left(\mathrm{D}'\right) = O\right] = \prod_{i=1}^{m} \frac{\exp\{-\frac{|y_i - x'_i|}{b}\}}{2 * b}$$

Similarly, the risk of computing a data set D'' to obtain a density vector of O occurs is

$$P\left[Q\left(\mathrm{D}''\right) = O\right] = \prod_{i=1}^{m} \frac{\exp\{-\frac{|y_i - x''_i|}{b}\}}{2 * b}$$

Combined with the triangle inequality, it follows that

$$\begin{aligned}
\frac{P\left[Q\left(\mathrm{D}'\right) = O\right]}{P\left[Q\left(\mathrm{D}''\right) = O\right]} &= \frac{\prod_{i=1}^{m} \frac{\exp\{-\frac{|y_i - x'_i|}{b}\}}{2*b}}{\prod_{i=1}^{m} \frac{\exp\{-\frac{|y_i - x''_i|}{b}\}}{2*b}} \\
&= \prod_{i=1}^{m} \exp\{-\frac{|y_i - x'_i| - |y_i - x''_i|}{b}\} \\
&= \exp\left\{ \frac{1}{b} * \sum_{i=1}^{m} \left(|y_i - x'_i| - |y_i - x''_i| \right) \right\} \\
&= \exp\left\{ \frac{\varepsilon}{\Delta f} * \sum_{i=1}^{m} \left(|y_i - x'_i| - |y_i - x''_i| \right) \right\} \\
&\le \exp\left\{ \frac{\varepsilon}{\Delta f} * \sum_{i=1}^{m} |x'_i - x''_i| \right\} \\
&\le \exp\left\{ \frac{\varepsilon}{\Delta f} * \Delta f \right\} \\
&= \exp(\varepsilon)
\end{aligned}$$

$P\left[Q\left(\mathrm{D}'\right)=O\right] \leq \exp(\varepsilon) * P\left[Q\left(\mathrm{D}''\right)=O\right]$. It shows that the process of adding Laplace noise to each attribute of the density function satisfies ε-differential privacy. According to Theorem 1, it is concluded that the outlier detection process of EAF-DP-Denclue algorithm satisfies ε-differential privacy, and the proof is completed. □

E. TIME COMPLEXITY ANALYSIS

Let the data size is n. The time complexity of the EAF-DP-Denclue algorithm consists of the following two main components. (1) Calculate the UOV of the all attributes of the dataset and sort them according to the UOV. The time complexity of this part is $O(n\log n)$. (2) The data points are modeled by the density function to find the density attractors and their correlation points. The time complexity of this part is $O(n\log n)$. Therefore, the time complexity of the EAF-DP-Denclue is $O(n\log n) + O(n\log n) \approx O(n\log n)$.

V. EXPERIMENT

A. EXPERIMENTAL ENVIRONMENT AND DATA

In this paper, Python is used to complete the algorithm comparison experiment. The experimental environment is shown in Table 1.

To test the performance of the algorithm in this paper under various complex data distributions. In this paper, experiments are conducted using six two-dimensional synthetic datasets (S1-S6) shown in Fig. 3, where the outliers are the points represented by ''o''. The information description of the synthetic datasets is shown in Table 2. Among them, S1 and S3 datasets are artificial datasets containing a number of clusters and the density between the class clusters has a difference, S2, S4, S5 and S6 are datasets containing a variety of complex data distributions of non-spherical class clusters, and the selection of the six datasets shown in Fig. 3 can be a more comprehensive test of the algorithm of this paper in a variety of complex data distributions of the detection effect of outlier points. The ten real-life datasets used in this paper are all from the UCI (University of California Irvine) dataset, the dimensions of the datasets range from 5-166, and the percentage of outliers ranges from 0.34% to 35.90%. Table 3, Table 4 and Table 5 demonstrate the feature information of the datasets. In the data preprocessing phase, the data were normalized using Equation (18).

$$X_{\text{norm}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{18}$$

where X_{norm} denotes the normalized data, and $X_{\max}$ and $X_{\min}$ denote the maximum and minimum values on each dimension, respectively. For missing null values, we use the mean value of the corresponding attribute of the dataset instead.

In the iris dataset experiment, the iris dataset was divided into three equal parts of iris1, iris2 and iris3 according to the classification attribute class as the sample dataset, and 5 pieces of data in iris2 were randomly selected as abnormal outlier samples and added to iris1 to form the experimental dataset iris12. Similarly, three experimental datasets, namely iris12, iris13 and iris23, will be obtained by integration. In the yeast dataset experiment, the CYT with the largest number of classes is taken as the normal sample dataset, and the outlier sample dataset with 5% of the normal sample in the remaining samples is randomly selected and added to the normal sample dataset to form the experimental dataset yeast1 and yeast2. The content of this experiment mainly

TABLE 1. Experiment environment.

Equipment	Parameter
CPU	2.30GHz Intel(R) i7-11800H
Hard Disk	512.0GB
Internal Storage	32.0GB
IDE	PyCharm
Computing Environment	Python3.8

TABLE 2. Synthetic data set.

Data set	Instances	Number of outliers	Proportion(%)
S1	1035	28	2.71
S2	1256	43	3.42
S3	1000	85	8.50
S4	876	77	8.79
S5	2042	64	3.13
S6	2259	159	7.04

TABLE 3. Real-life data set.

Data set	Instances	Dimension	Number of outliers	Proportion(%)
iris	150	5	10	6.67
yeast	1484	9	5	0.34
ecoli	168	7	25	14.88
wbc	223	9	10	4.48
ionosphere	351	34	126	35.90
wdbc	390	30	33	8.46
vowels	1456	12	50	3.43
musk	5852	166	241	4.12
sonar	120	60	9	7.50
rbds	372	105	38	10.22

TABLE 4. Class distribution of iris data set.

Class Name	Instances
iris-setosa	50
iris-versicolor	50
iris-virginica	50

TABLE 5. Class distribution of yeast data set.

Class Name	Instances
CYT	463
NUC	429
MIT	244
ME3	163
ME2	51
ME1	44
EXC	37
VAC	30
POX	20
ERL	5

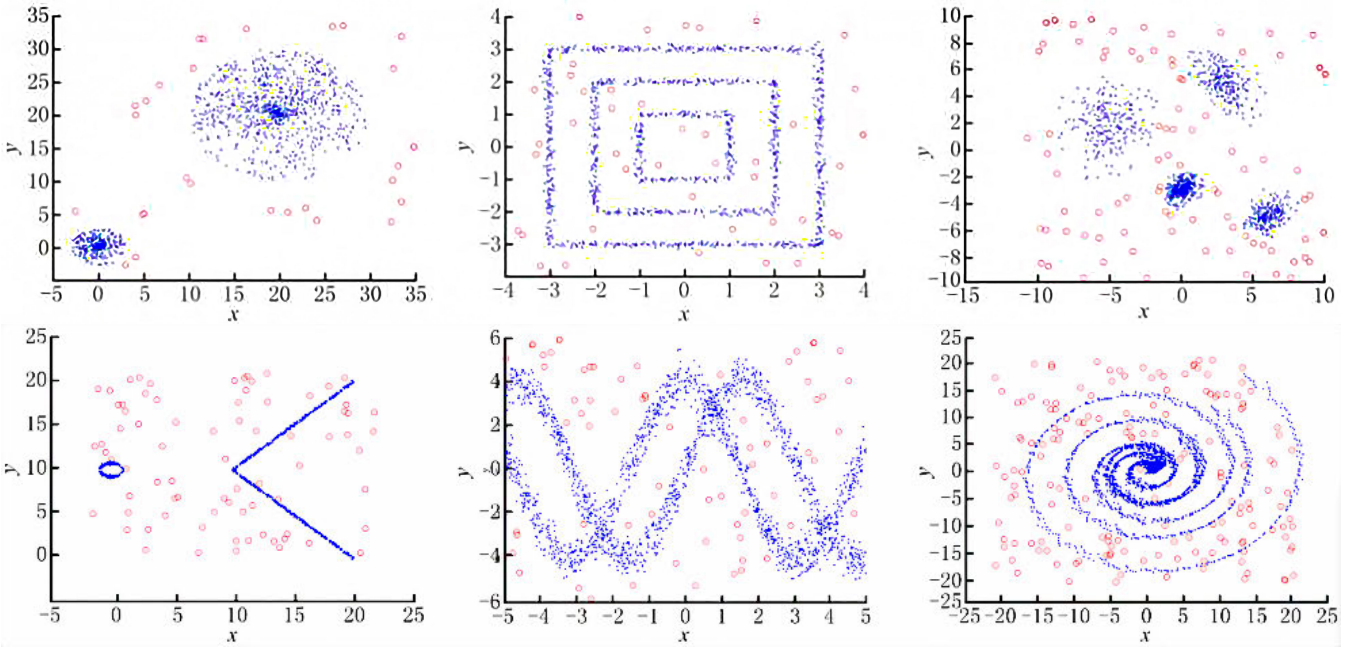

FIGURE 3. Two-dimensional scatter plot of the synthetic dataset S1-S6.

consists of two parts: the attribute consistent outlier degree ranking and the accuracy of outlier detection results. The experiment is carried out 30 times, and the average accuracy is finally taken as the experimental result. The preprocessed data description is shown in Table 6.

TABLE 6. Preprocessed Data Description.

Preprocessed data set	Number of normal data points	Number of outlier data points	Percentage of outlier data points
iris12	50	5	10%
iris13	50	5	10%
iris23	50	5	10%
yeast1	463	23	5%
yeast2	463	23	5%

B. EVALUATION INDICATORS

In the experimental part, the *Precision*, *Recall* and $F_1 - Score$ are selected as the evaluation criterion of the outlier detection accuracy of the algorithm. The $F_1 - Score$ combines numerical results for *Precision* and *Recall*. The *Precision* and $F_1 - Score$ are between 0 and 1, and the larger the value, the better the outlier detection performance. The calculation methods of the three evaluation indicators are shown in Equation (19), Equation (20) and Equation (21) respectively.

$$Precision = \frac{TP}{TP + FP} \tag{19}$$

$$Recall = \frac{TP}{TP + FN} \tag{20}$$

$$F_1 - Score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{21}$$

where TP is true positives, denoting the number of points that the algorithm correctly detects as outliers, FP is false positives, denoting the number of points that the algorithm incorrectly detects as outliers, and FN is false negatives, denoting the number of points that the algorithm detects as normal outliers.

C. ATTRIBUTE ORDERING AND WEIGHTING

If the kernel bandwidth σ is too large or too small, the density distribution in the high-density area will be too smooth, or the density distribution of the data points in the low-density area will fluctuate sharply, resulting in a large number of noise estimates and affect the experimental results, respectively [41]. In this paper, $\sigma = 0.1$ is set to ensure that the kernel bandwidth is at a moderate size.

In order to select the appropriate parameter c, a pre-experiment is carried out on the yeast data set with a large amount of data to observe the fit of parameter c in the traditional Denclue algorithm, as shown in Fig. 4, the algorithm has the highest accuracy when [1.05, 1.15], so the value of parameteris determined to be 1.10.

According to the method shown in Algorithm 2, the arrangement order and calculation results of the UOV of each data set are displayed, as shown in Fig. 5 and Fig. 6.

According to Fig. 5, the UOV sorting order of attributes in the dataset iris12 is $A_2 > A_1 > A_4 > A_3$, indicating that the attribute A_2 is the first outlier attribute, and the consistent outlier degree of the attribute A_2 compared to other attributes is 0.5189, so the highest weight should be given in, and the consistent outlier degree UOV of the attribute A_1 compared to attributes other than the attribute A_2 is 0.5697. The next

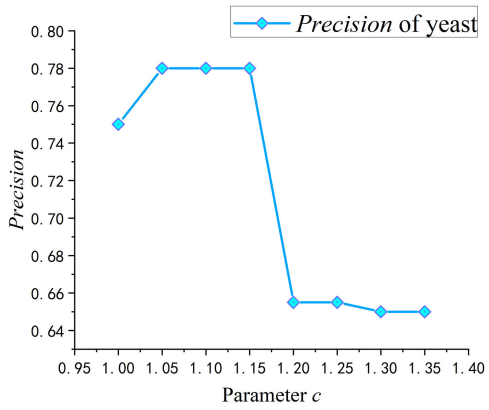

FIGURE 4. The selection of parameter c.

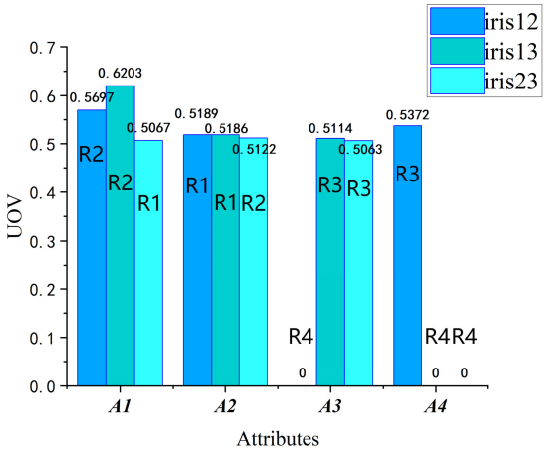

FIGURE 5. UOV of iris.

highest weight should be assigned, and the ranking of the attribute A_2 should be the lowest, and the weight assigned should be adjusted appropriately. The same is true for dataset iris13 and iris23.

In addition, it can be seen from Fig. 6 that the UOV sorting order of attributes in datasets yeast1 and yeast2

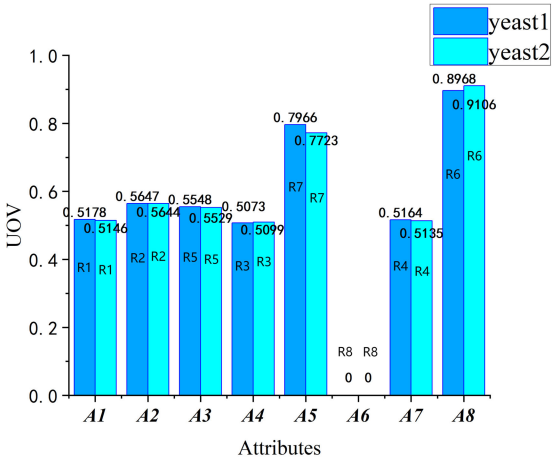

FIGURE 6. UOV of yeast.

is $A_1 > A_5 > A_5 > A_3 > A_7 > A_8 > A_4 > A_6$, indicating that the attribute A_1 is the first outlier attribute in datasets yeast1 and yeast2, and the consistent outlier degree of the attribute A_1 compared to other attributes in the two datasets is 0.5178 and 0.5146, respectively. Therefore, the highest weight is assigned, while the ranking of the attribute A_6 is the last one, and the assigned weight should be lowered appropriately.

In the iris dataset, the weight assignment of the four attributes was obtained from 1.10 down with a step size of 0.05 according to the UOV order, and set to 1.10, 1.05, 1.00 and 0.95, respectively. In the yeast dataset, the weight assignments for the eight attributes are obtained in the UOV order from 1.15 down to 1.10, 1.05, 1.00, 0.95, 0.90, 0.85, and 0.80, respectively.

D. COMPARATIVE ANALYSIS OF THE SAME TYPE OF ALGORITHMS

In this part of the experiment, the Denclue [11], the DP-Denclue [39] and the DB-DBScan [29] are selected as benchmark algorithms to verify the precision of the EAF-DP-Denclue in outlier detection. The DP-DBScan algorithm is a classical representative of clustering-based outlier detection algorithms with differential privacy preserving. There are two parameters *MinPts* and *Eps* in the DP-DBScan algorithm. 1/25 of the data scale is used as *MinPts*, the parameter *Eps* is gradually adjusted upward from 0.1 to 0.45 with a gradient of 0.05, and the privacy budget ε is set as a gradient of 0.1.

TABLE 7. Experimental results of iris dataset (Precision).

Benchmark Algorithms	iris12	iris13	iris23
DP-Denclue ($\xi = 5, \varepsilon = 0.1$)	0.81	0.68	0.96
DP-Denclue ($\xi = 5, \varepsilon = 0.5$)	0.82	0.73	0.99
DP-Denclue ($\xi = 10, \varepsilon = 0.1$)	0.90	0.91	0.85
DP-Denclue ($\xi = 10, \varepsilon = 0.5$)	0.91	0.92	0.86
EAF-DP-Denclue ($\xi = 5, \varepsilon = 0.1$)	**0.82**	**0.82**	**0.97**
EAF-DP-Denclue ($\xi = 5, \varepsilon = 0.5$)	**0.84**	**0.83**	**0.99**
EAF-DP-Denclue ($\xi = 10, \varepsilon = 0.1$)	**0.91**	**0.93**	**0.91**
EAF-DP-Denclue ($\xi = 10, \varepsilon = 0.5$)	**0.92**	**0.93**	**0.92**
Denclue ($\xi = 5$)	0.80	0.67	1.00
Denclue ($\xi = 10$)	0.91	0.91	0.82

According to the results in Table 7, when ξ is set to 5 and ε is set to 0.1, the precision of the DP-Denclue on the three data sets iris12, iris13 and iris23 reaches 0.81, 0.68 and 0.96 respectively. When ε is set to 0.5, the disturbance degree of noise on data decreases. The precision of the DP-Denclue is improved to 0.82, 0.73 and 0.99, respectively, and the precision is basically not lower than that of the Denclue. When ξ is set to 5 and ε is set to 0.1, the precision of the EAF-DP-Denclue is higher than that of the DP-Denclue on iris12, iris13 and iris23 datasets after weighting by the EAF algorithm, reaching 0.82, 0.82 and 0.97 respectively. When ξ is set to 5 and ε is set to 0.5, the precision of the EAF-DP-Denclue achieves 0.84, 0.83 and 0.99 respectively. When ξ is set to 10 and ε is set to 0.1 and 0.5, the fluctuation

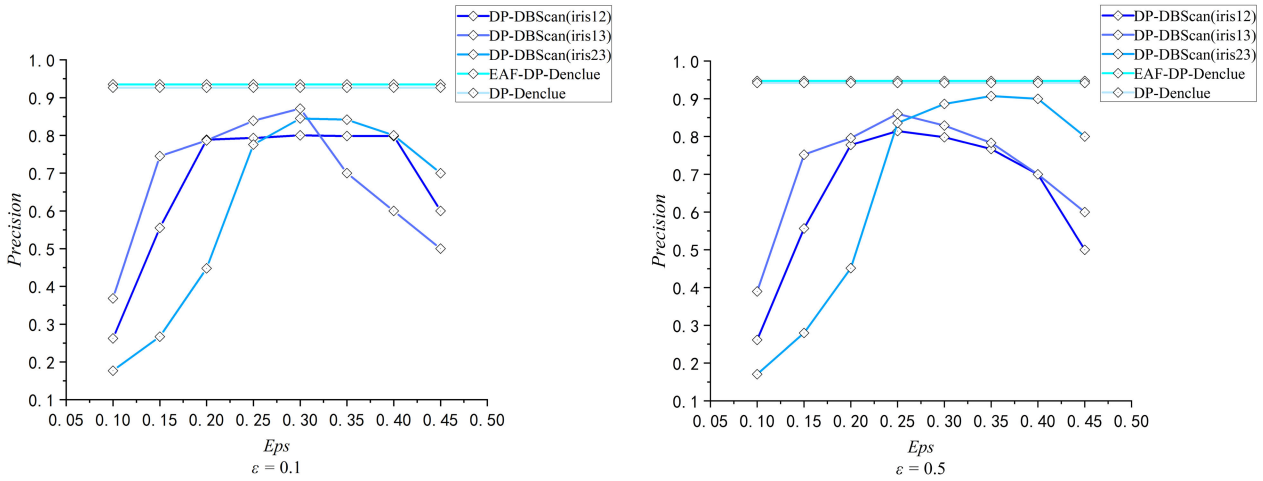

FIGURE 7. Precision of outlier detection (iris data set).

of the experimental precision affected by the change of ξ is also very small, and the results have the same trend, which is in line with the experimental expectation.

As shown in Fig. 7, the experimental effects of the DP-DBScan compared with the EAF-DP-Denclue on the three datasets iris12, iris13 and iris23 are shown. From the experimental results, it can be seen that when ε is set to 0.1 or 0.5, the change of precision is very sensitive with the increase of *Eps* due to the characteristic of the DP-DBScan that is very sensitive to parameters. The precision peak appears between [0.25, 0.4], which is about 80%. However, the average precision of the DP-Denclue on the three experimental data sets is still higher than the precision of the DP-DBScan in all cases. The average precision of the EAF-DP-Denclue is higher than that of the DP-Denclue. The experiments in this part also fully show the accuracy of the EAF-DP-Denclue.

Fig. 8 shows the experimental results compared from another perspective. It can be seen from the experimental results that with the increase of privacy budget ε, the precision of the DP-Denclue and the EAF-DP-Denclue is improved, and the reason is that the disturbance degree of noise on data decreases when the privacy budget increases. When $\varepsilon = 0.11$ or larger, the EAF-DP-Denclue exceeds the precision of Denclue on yeast dataset, indicating that the effect of outlier attribute on outlier detection exceeds the effect of noise on outlier detection. It achieves the expected effect of the algorithm. In addition, when the privacy budget of the EAP-DP-Denclue is $\varepsilon = 0.03$ and larger, the precision exceeds the Denclue, and the precision is greatly improved compared with the DP-Denclue, which also shows that the EAF algorithm proposed in this paper is effective.

E. COMPARATIVE ANALYSIS OF DIFFERENT TYPES OF ALGORITHMS

To further evaluate the performance of the EAF-DP-Denclue, we selected the local outlier factor (LOF) [42], the connective outlier factor (COF) [43], the local structure outlier factor (LSOF) [44] and the isolation forest (IForest) [45] for experiments on all datasets. The LOF is a classical density based outlier detection algorithm. The COF is an improved version of the LOF. The LOF and COF algorithms have the same time complexity with $O(n^2)$. The LSOF is an algorithm that introduces nearest-neighbor tree neighborhood relations to outlier detection, and the time complexity of the algorithm is $O(n\log n)$ after the KD-Tree structure is used to retrieve neighboring points. The IForest is one of the representatives of a classical outlier detection algorithm with linear time complexity and excellent detection performance. In this part of the experiment, based on the experimental numerical results in Table 7, we set the parameters $\xi = 10$ and $\varepsilon = 0.5$ of the EAF-DP-Denclue.

Table 8 and Table 9 record the precision and F_1-Score experimental results of the proposed algorithm in this paper

TABLE 8. Precision of different algorithms on synthetic data set.

Algorithms	S1	S2	S3	S4	S5	S6
LOF	0.75	**0.95**	0.59	0.44	0.81	0.56
COF	0.54	0.88	0.47	0.40	0.63	0.51
LSOF	0.39	0.72	0.38	0.27	0.52	0.30
IForest	**1.00**	0.49	0.68	0.71	0.34	0.38
Denclue	0.82	0.79	0.53	0.55	0.53	0.33
EAF-DP-Denclue	0.93	**0.95**	**0.82**	**0.87**	**0.84**	**0.67**

TABLE 9. F_1-Score of different algorithms on synthetic data set.

Algorithms	S1	S2	S3	S4	S5	S6
LOF	0.73	0.92	0.56	0.43	0.78	0.53
COF	0.53	0.85	0.46	0.38	0.63	0.50
LSOF	0.37	0.71	0.39	0.28	0.51	0.31
IForest	**0.97**	0.48	0.65	0.69	0.35	0.37
Denclue	0.80	0.79	0.51	0.53	0.52	0.34
EAF-DP-Denclue	0.92	**0.94**	**0.80**	**0.85**	**0.82**	**0.65**

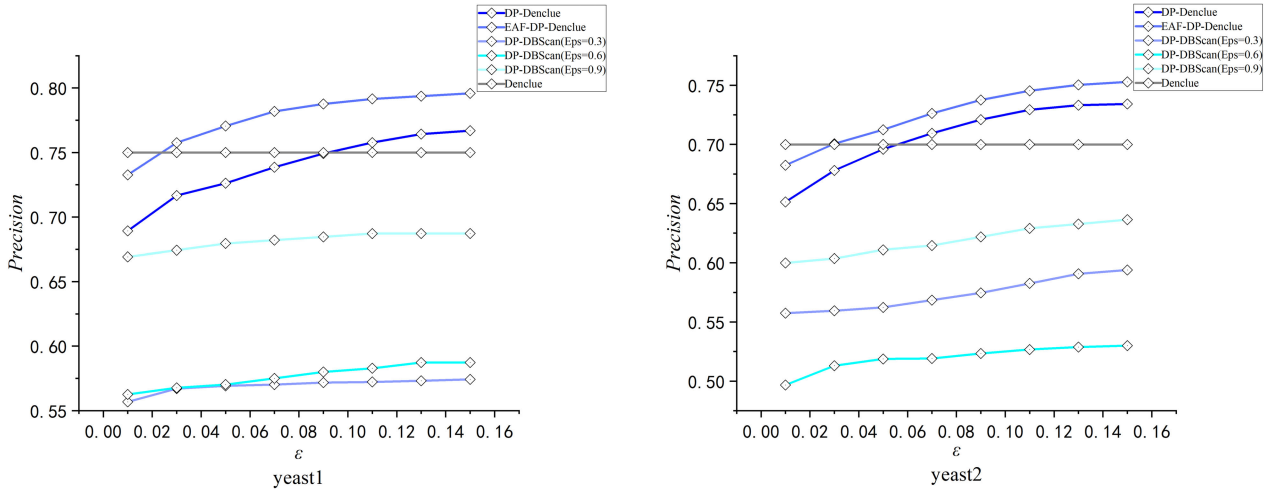

FIGURE 8. Precision of outlier detection (yeast data set).

TABLE 10. Precision of different algorithms on real-life data set.

Algorithms	iris	yeast	ecoli	wbc	ionosphere	wdbc	vowels	musk	sonar	rbds
LOF	0.60	0.80	0.84	0.10	0.81	0.85	0.54	0.76	**0.67**	0.53
COF	0.40	0.20	0.68	0.10	0.81	**0.91**	0.50	0.80	0.44	0.29
LSOF	0.10	0.40	0.84	0.20	0.79	0.67	0.48	0.73	0.56	0.58
IForest	0.60	0.80	0.88	**0.90**	0.65	0.52	0.14	0.79	0.22	**0.83**
Denclue	0.70	0.80	0.80	0.30	0.69	0.39	0.42	0.76	0.33	0.66
EAF-DP-Denclue	**0.90**	**1.00**	**0.92**	0.70	**0.87**	0.82	**0.72**	**0.86**	**0.67**	0.79

TABLE 11. F_1-Score of different algorithms on real-life data set.

Algorithms	iris	yeast	ecoli	wbc	ionosphere	wdbc	vowels	musk	sonar	rbds
LOF	0.48	0.69	0.73	0.15	0.78	0.77	0.46	0.71	0.44	0.50
COF	0.40	0.20	0.64	0.13	0.77	**0.86**	0.40	0.77	0.38	0.33
LSOF	0.13	0.40	0.78	0.24	0.76	0.63	0.41	0.72	0.32	0.55
IForest	0.55	0.53	0.84	**0.85**	0.62	0.47	0.16	0.78	0.15	**0.78**
Denclue	0.58	0.69	0.69	0.24	0.65	0.32	0.35	0.74	0.17	0.62
EAF-DP-Denclue	**0.72**	**1.00**	**0.86**	0.65	**0.84**	0.77	**0.67**	**0.85**	**0.53**	**0.78**

with other outlier detection algorithms on synthetic datasets. Table 10 and Table 11 record the precision and F_1-Score experimental results of the proposed algorithm in this paper with other outlier detection algorithms on real-life datasets.

According to Table 8, although the outlier detection precision of the EAF-DP-Denclue on S1 is slightly lower than that of IForest, it still reaches 92%. In addition, the EAF-DP-Denclue has the highest precision on S2, S3, S4, S5 and S6 datasets, and the outlier detection performance is better compared to other outlier detection algorithms. This is due to the fact that the EAF-DP-Denclue is based on information entropy weights, which calculate the importance of the attributes with respect to each other and amplify the weights of the outlier attributes. Further, the fuzzy priority relation ordering method is utilized to rank the data attributes with corresponding information entropy weights, which improves the applicability of the algorithm on datasets with different densities. According to Table 9, the F_1-Score and precision obtained from the experiments of each algorithm are similar. The EAF-DP-Denclue is lower than IForest only on S1, and has the best performance on all other synthetic datasets.

According to Table 10, the EAF-DP-Denclue got the highest precision on seven real-life datasets. The precision of EAF-DP-Denclue is higher than 85% on iris, ecoli, yeast, ionosphere and musk datasets and 100% on yeast dataset. In most cases, the precision of the EAF-DP-Denclue is higher than 70%, only on the sonar dataset the precision is lower at 67%, but also still in line with that of the LOF. This shows that the EAF-DP-Denclue has a stable outlier detection efficiency on real-life datasets with different sizes, different numbers of attributes and different ratios of outlier points. According to Table 11, the EAF-DP-Denclue achieves

an F_1-Score of 100% on the yeast dataset. Except for the wbc and wdbc datasets, the EAF-DP-Denclue has the largest F_1-Score on the other eight real datasets, which gives the EAF-DP-Denclue excellent performance compared to other outlier detection algorithms. In addition, the F_1-Score on the wbc and wdbc datasets remains no less than 70%, again with excellent performance. In general, according to the numerical results in Table 10 and Table 11, the introduction of the fuzzy priority relation ordering method in the Denclue compensates for the degradation of the outlier detection performance caused by the differential privacy technology, and the EAF-DP-Denclue has an F_1-Score improvement of about 30% compared to the Denclue. Combined with the numerical precision results, it further demonstrates that the EAF-DP-Denclue proposed in this paper has a stable outlier detection efficiency on different real-life datasets. Therefore, it can be proved that the EAF-DP-Denclue in this paper can efficiently and consistently detect outliers on real-life datasets.

VI. CONCLUSION

In this paper, aiming at the privacy leakage problem of the Denclue in the process of outlier detection and the problem of low accuracy of results caused by differential privacy technology, based on the differential privacy protection model and fuzzy precedence relation ordering method, we propose the EAF-DP-Denclue with better performance. Experiments and theoretical proof show that the algorithm can accurately realize the identification of outliers and satisfy ε-differential privacy. However, the disadvantage of the EAF-DP-Denclue is that the running speed of the algorithm itself is much slower than that of other outlier detection algorithms. Improving the running efficiency of the algorithm and making it better applied to various fields is a problem that needs to be studied in the future.

REFERENCES

[1] P. Samarati and L. Sweeney, "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppresion," in *Proc. IEEE Symp. Res. Secur. Privacy*, 1998, pp. 384–393.
[2] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity," *ACM Trans. Knowl. Discovery From Data (TKDD)*, vol. 1, no. 1, p. 3, 2007.
[3] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and l-diversity," in *Proc. IEEE 23rd Int. Conf. Data Eng.*, Apr. 2007, pp. 106–115.
[4] C. Dwork, "Differential privacy: A survey of results," in *Proc. Int. Conf. Theory Appl. Models Comput.* Xi'an, China: Springer, Apr. 2008, pp. 1–19.
[5] H. Wang, M. J. Bah, and M. Hammad, "Progress in outlier detection techniques: A survey," *IEEE Access*, vol. 7, pp. 107964–108000, 2019.
[6] M. A. Samara, I. Bennis, A. Abouaissa, and P. Lorenz, "A survey of outlier detection techniques in IoT: Review and classification," *J. Sensor Actuator Netw.*, vol. 11, no. 1, p. 4, Jan. 2022.
[7] J. Santos, P. Leroux, T. Wauters, B. Volckaert, and F. De Turck, "Anomaly detection for smart city applications over 5G low power wide area networks," in *Proc. NOMS - IEEE/IFIP Netw. Oper. Manage. Symp.*, Apr. 2018, pp. 1–9.
[8] N. Nesa, T. Ghosh, and I. Banerjee, "Outlier detection in sensed data using statistical learning models for IoT," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2018, pp. 1–6.
[9] H. Ghallab, H. Fahmy, and M. Nasr, "Detection outliers on Internet of Things using big data technology," *Egyptian Informat. J.*, vol. 21, no. 3, pp. 131–138, Sep. 2020.
[10] H.-x. Tian, X.-j. Liu, and M. Han, "An outliers detection method of time series data for soft sensor modeling," in *Proc. Chin. Control Decis. Conf. (CCDC)*, May 2016, pp. 3918–3922.
[11] A. Hinneburg and D. A. Keim, "An efficient approach to clustering in large multimedia databases with noise," in *Proc. Knowl. Discovery Datamining (KDD)*, 1998, pp. 58–65.
[12] R. Bhuyan and S. Borah, "A survey of some density based clustering techniques," 2023, *arXiv:2306.09256*.
[13] M. S. Khader and G. Al-Naymat, "VDENCLUE: An enhanced variant of DENCLUE algorithm," in *Proc. SAI Intell. Syst. Conf.* Cham, Switzerland: Springer, 2021, pp. 425–436.
[14] G. Al-Naymat, M. Khader, M. A. Al-Betar, R. Hriez, and A. Hadi, "MR-VDENCLUE: Varying density clustering using MapReduce," in *Proc. SAI Intell. Syst. Conf.* Cham, Switzerland: Springer, 2022, pp. 771–788.
[15] H. Wang, "Optimal bandwidth selection for DENCLUE algorithm," in *Proc. 9th Int. Conf. Control, Autom. Robot. (ICCAR)*, Apr. 2023, pp. 245–249.
[16] T. Huang, Z. Cai, R. Li, S. Wang, and W. Zhu, "Consolidation of structure of high noise data by a new noise index and reinforcement learning," *Inf. Sci.*, vol. 614, pp. 206–222, Oct. 2022.
[17] C. Shao, S. Zheng, C. Gu, Y. Hu, and X. Qin, "A novel outlier detection method for monitoring data in dam engineering," *Expert Syst. Appl.*, vol. 193, May 2022, Art. no. 116476.
[18] C. Wu, J. Wang, and Y. Hao, "Deterministic and uncertainty crude oil price forecasting based on outlier detection and modified multi-objective optimization algorithm," *Resour. Policy*, vol. 77, Aug. 2022, Art. no. 102780.
[19] H. Yuan, N. Cui, C. Li, Z. Cui, and L. Chang, "Early stage internal short circuit fault diagnosis for lithium-ion batteries based on local-outlier detection," *J. Energy Storage*, vol. 57, Jan. 2023, Art. no. 106196.
[20] Z. Yuan, B. Chen, J. Liu, H. Chen, D. Peng, and P. Li, "Anomaly detection based on weighted fuzzy-rough density," *Appl. Soft Comput.*, vol. 134, Feb. 2023, Art. no. 109995.
[21] J. Huang, Q. Zhu, L. Yang, and J. Feng, "A non-parameter outlier detection algorithm based on natural neighbor," *Knowl.-Based Syst.*, vol. 92, pp. 71–77, Jan. 2016.
[22] J. Huang, Q. Zhu, L. Yang, D. Cheng, and Q. Wu, "A novel outlier cluster detection algorithm without top-n parameter," *Knowl.-Based Syst.*, vol. 121, pp. 32–40, Apr. 2017.
[23] Y. Yang, C. Fan, L. Chen, and H. Xiong, "IPMOD: An efficient outlier detection model for high-dimensional medical data streams," *Expert Syst. Appl.*, vol. 191, Apr. 2022, Art. no. 116212.
[24] A. Abhaya and B. K. Patra, "An efficient method for autoencoder based outlier detection," *Expert Syst. Appl.*, vol. 213, Mar. 2023, Art. no. 118904.
[25] J. Huang, D. Cheng, and S. Zhang, "A novel outlier detecting algorithm based on the outlier turning points," *Expert Syst. Appl.*, vol. 231, Nov. 2023, Art. no. 120799.
[26] A. Blum, C. Dwork, F. McSherry, and K. Nissim, "Practical privacy: The SuLQ framework," in *Proc. 24th ACM SIGMOD-SIGACT-SIGART Symp. Princ. Database Syst.*, Jun. 2005, pp. 128–138.
[27] Y. Li, Z. Hao, W. Wen, and G. Xie, "Research on differential privacy preserving k-means clustering," *Comput. Sci.*, vol. 40, no. 3, pp. 287–290, 2013.
[28] C. Hu, G. Yang, and Y. L. Bai, "Clustering algorithm in differential privacy preserving," *Comput. Sci.*, vol. 46, no. 2, pp. 120–126, 2019.
[29] W. Wei-Min and H. Huan-Kun, "A DP-DBScan clustering algorithm based on differential privacy preserving," *Comput. Eng. Science/Jisuanji Gongcheng yu Kexue*, vol. 37, no. 4, p. 830, 2015.
[30] H. Wang, L. Ge, S. Wang, L. Wang, Y. Zhang, and J. Liang, "Improvement of differential privacy protection algorithm based on optics clustering," *J. Comput. Appl.*, vol. 38, no. 1, p. 73, 2018.
[31] X. Zheng, D. Chen, Y. Liu, H. You, X. Wang, and L. Sun, "Spectral clustering algorithm based on differential privacy protection," *J. Comput. Appl.*, vol. 38, no. 10, p. 2918, 2018.
[32] S. Li, J. C. Chang, M. Z. LiLv, and K. J. Cai, "A stacking ensemble clustering algorithm based on differential privacy protection," *Comput. Eng. Sci./Jisuanji Gongcheng Yu Kexue*, vol. 44, no. 8, pp. 1402–1408, 2022.
[33] H. H. Nguyen, "Privacy-preserving mechanisms for k-modes clustering," *Comput. Secur.*, vol. 78, pp. 60–75, Sep. 2018.
[34] A. Tramacere, D. Paraficz, P. Dubath, J.-P. Kneib, and F. Courbin, "ASTErIsM: Application of topometric clustering algorithms in automatic galaxy detection and classification," *Monthly Notices Roy. Astronomical Soc.*, vol. 463, no. 3, pp. 2939–2957, Dec. 2016.

[35] T. M. Heaster, A. J. Walsh, B. A. Landman, and M. C. Skala, "Density-based clustering analyses to identify heterogeneous cellular subpopulations," *Proc. SPIE*, vol. 10043, pp. 67–74, Feb. 2017.
[36] X. Yin, Y. Shi, X. Xu, Y. Duan, Y. Jia, G. Chen, X. Zhang, and F. Yin, "Research on wind deviation detection based on denclue abnormal working condition filtering," *IOP Conf. Ser., Earth Environ. Sci.*, vol. 617, no. 1, 2020, Art. no. 012015.
[37] H. Rehioui and A. Idrissi, "New clustering algorithms for Twitter sentiment analysis," *IEEE Syst. J.*, vol. 14, no. 1, pp. 530–537, Mar. 2020.
[38] H. Jin, W. Yu, and S. Li, "A clustering algorithm for determining community structure in complex networks," *Phys. A, Stat. Mech. Appl.*, vol. 492, pp. 980–993, Feb. 2018.
[39] H.-Z. Xia, L.-M. Chen, F. Qi, X.-D. Mao, L.-Q. Sun, and F.-Y. Xue, "DP-denclue: An outlier detection algorithm with differential privacy preservation," in *Proc. IEEE 24th Int. Conf. High Perform. Comput. Commun.; 8th Int. Conf. Data Sci. Syst.; 20th Int. Conf. Smart City; 8th Int. Conf Dependability Sensor, Cloud Big Data Syst. Appl. (HPCC/DSS/SmartCity/DependSys)*, Dec. 2022, pp. 2264–2269.
[40] L. Wang, C. Feng, Y. Ren, and J. Xia, "Local outlier detection based on information entropy weighting," *Int. J. Sensor Netw.*, vol. 30, no. 4, pp. 207–217, 2019.
[41] Z. Zhang, W. Liu, Y. Zhang, Y. Deng, and M. Wei, "ERDOF: Outlier detection algorithm based on entropy weight distance and relative density outlier factor," *J. Commun.*, vol. 42, no. 9, pp. 133–143, 2021.
[42] W. Jin, A. K. Tung, J. Han, and W. Wang, "Ranking outliers using symmetric neighborhood relationship," in *Proc. Pacific–Asia Conf. Knowl. Discovery Data Mining*. Singapore: Springer, Apr. 2006, pp. 577–593.
[43] 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. Discovery Data Mining*. Taipei, Taiwan: Springer, May 2002, pp. 535–548.
[44] R. Wang and Q. Zhu, "LSOF: Novel outlier detection approach based on local structure," in *Proc. IEEE Int. Conf. Parallel Distrib. Process. With Appl., Big Data Cloud Comput., Sustain. Comput. Commun., Social Comput. Netw. (ISPA/BDCloud/SocialCom/SustainCom)*, Dec. 2019, pp. 838–846.
[45] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *Proc. 8th IEEE Int. Conf. Data Mining*, Dec. 2008, pp. 413–422.

HUANGZHI XIA (Member, IEEE) was born in Fuzhou, China, in 1999. He is currently pursuing the M.S. degree in applied mathematics with Mudanjiang Normal University, China. His research interests include machine learning and intelligent optimization algorithm.

LIMIN CHEN received the Ph.D. degree from Harbin Engineering University, China. She is currently a Professor with Mudanjiang Normal University. Her research interests include machine learning and data mining.

DONGYAN WANG was born in Nanyang, China, in 2000. He is currently pursuing the M.S. degree in applied mathematics with Mudanjiang Normal University, China. His research interests include machine learning and federated learning.

XIAOTONG LU was born in Suihua, China, in 2000. She is currently pursuing the M.S. degree in applied mathematics with Mudanjiang Normal University, China. Her research interests include machine learning and data mining.