IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 11 December 2022, accepted 27 December 2022, date of publication 29 December 2022,
date of current version 5 January 2023.

Digital Object Identifier 10.1109/ACCESS.2022.3233196

RESEARCH ARTICLE

A New Density Peak Clustering Algorithm With Adaptive Clustering Center Based on Differential Privacy

HUA CHEN, YUAN ZHOU, KEHUI MEI, NAN WANG, AND GUANGXING CAI
School of Science, Hubei University of Technology, Wuhan 430068, China

Corresponding author: Hua Chen (20070002@hbut.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61502156, in part by the Teaching and Research Project of Hubei Provincial Department of Education under Grant 282, and in part by the Doctoral Startup Fund of Hubei University of Technology under Grant BSQD13051.

ABSTRACT A new density peak clustering (DPC) algorithm with adaptive clustering center based on differential privacy was proposed to solve the problems of poor adaptability of high-dimensional data, inability to automatically determine clustering centers, and privacy problems in clustering analysis. First, to solve the problem of poor adaptability of high-dimensional data, cosine distance was used to measure the similarity between high-dimensional datasets. Then, aiming at the subjective problem of clustering center selection, from the perspective of ranking graph, the weight $(i-1)/i$ was introduced creatively, the slope trend of ranking graph was redefined to realize the adaptive clustering center. Finally, aiming at the privacy problem, the Laplacian noise of appropriate privacy budget was added to the core statistic (local density) of the algorithm to achieve the balance between privacy protection and algorithm effectiveness. Experimental results on both the synthetic and UCI datasets show that this algorithm can not only realize the automatic selection of clustering center, but also solve the privacy problem in clustering analysis, and improve the clustering evaluation index greatly, which proves the effectiveness of the algorithm.

INDEX TERMS Cosine distance, differential privacy, DPC algorithm, Laplacian noise, trend of slope change.

I. INTRODUCTION

With the continuous development of information technology, the data generated in all walks of daily life show explosive growth. It is very important to mine valuable information and models from massive data. In the era of big data, the data sharing mode based on data release [1] and data mining [2] has gradually taken shape. When various kinds of information are digitized, privacy leakage is becoming more and more serious, and privacy security is also getting more and more attention.

Clustering analysis is an important part of data mining and the basis of some data mining methods [3]. Its application scenarios are very wide, in computer science [4], [5], [6], biological [7], [8], chemistry [9], society [10] and other fields. Traditional clustering algorithms can be roughly divided into five types: partition-based algorithm, hierarchical algorithm, density-based algorithm, grid-based algorithm and model-based algorithm [11].

Partition based K-means algorithm [12] and its derivative algorithms are widely used in practical scenarios, but such algorithms are not suitable for arbitrary datasets, need multiple iterations and are sensitive to the initial cluster center. Density-based clustering algorithm has the advantages of clustering arbitrary shape datasets, insensitive to noise data, no need for iteration, and suitable for large-scale datasets. DBSCAN algorithm [13] is a typical algorithm based on density clustering algorithm, but it has many parameters and is sensitive to parameter values, so it is difficult to produce stable clustering results for datasets. The Density Peak clustering (DPC) algorithm [14] proposed in 2014 can quickly detect the density peak points, requiring fewer parameters,

The associate editor coordinating the review of this manuscript and approving it for publication was Jingen Ni.

which has good research value and application prospects. However, this algorithm also has some problems, such as poor adaptability of high-dimensional data, cutoff distance (denoted as d_c) and clustering center cannot be selected automatically, as well as privacy protection in clustering analysis.

In order to solve some problems existing in DPC algorithm, many scholars have improved it in many aspects. To solve the problem of poor adaptability of high-dimensional data, M. Du et al. [15] introduced PCA into the DPC-KNN algorithm to preprocess high-dimensional data. Yang [16] proposed to use weighted Euclidean distance to measure similarity between data. In view of the subjectivity of cutoff distance selection, some scholars adaptively determined the cutoff distance by constructing the function relation between cutoff distance and information entropy or Gini index [17], [18]. Wang et al. [19] proposed PLDPC based on mutual information criterion, which avoided manual pre-setting of parameters. Sun et al. [20] redefined the local density without setting truncation distance by introducing the nearest neighbor relationship between points, while Wang et al. [21] by introducing second-order k neighbors of nodes. Wang et al. [22] introduced local minimal spanning tree (LMST) to redefine the local density (denoted as ρ) and the center offset distance (denoted as δ) of each point. Du et al. [23] proposed DDPA-DP that all parameters could be adjusted adaptively based on the data-driven idea. Liu et al. [24] proposed SNNDPC based on the shared nearest neighbor theory to calculate ρ and δ using the shared nearest neighbor similarity. In view of the subjective problem of cluster center selection, Zhao et al. [25] determined the cluster center by constraining the redefined local density, while Ding [26] determined by restricting the values of local density and center offset distance respectively. Both literature [17], [18], and [27] selected clustering centers adaptively by defining the slope change trend of the sorting chart. Among them, literature [17] redefined the trend by introducing a new statistic, while literature [18] redefined the trend by introducing the weight $i-1$. To solve the problem of misplacement of data points, Yu et al. [28] proposed DPCSA algorithm by introducing weighted local density sequence and two-stage allocation strategy. In view of the possible privacy leakage caused by the reconstruction of DPC algorithm model results [29], the differential privacy technology proposed by Dwork et al. [30] in 2006 not only has strict definition and provability, but also provides a quantifiable level of privacy protection, overcoming the shortcomings of the traditional privacy protection model.

Chen proposed DP-CFSFDP by combining DPC algorithm with differential privacy technology. In order to solve the problem that noisy parameters may lead to deviation between the new center point and the correct center point, reachable center point is introduced to DP-RCCFSFDP [31]. Sun et al proposed DP-DPCSNNS based on shared nearest neighbor similarity [32], which used shared nearest neighbor similarity to calculate local density and detect cluster centers with neighborhood information, thus improving the accuracy of cluster center selection.

The main motivation of this paper is to avoid manual selection of clustering centers, improve clustering efficiency and reduce the risk of privacy leakage. Therefore, aiming at some problems existing in the DPC algorithm, a new DPC algorithm with adaptive clustering center based on differential privacy is proposed. The main innovations and contributions are summarized as follows:

1) Aiming at the problem of poor adaptability of high-dimensional data, cosine distance is used to measure the similarity between data in high-dimensional datasets.

2) Aiming at the subjective problem of clustering center selection, from the perspective of ranking graph, the weight $(i-1)/i$ is introduced creatively to redefine the trend of slope change of ranking graph to realize automatic clustering center selection.

3) For privacy protection, the Laplacian noise with appropriate privacy budget is added to the local density of the algorithm.

4) In order to verify the effectiveness and accuracy of the proposed algorithm, several experiments are performed on 6 synthetic datasets and 6 real UCI datasets, and the algorithm is evaluated by internal evaluation index CH and multiple external evaluation indexes ARI, AMI and FMI.

The rest of this paper is organized as follows. In Section II, we introduce the principle of differential privacy, the definition of DPC algorithm and evaluation metrics of clustering algorithm in detail. In Section III, we describe the principle, procedure and time complexity of the improved DPC algorithm with adaptive clustering center based on differential privacy. In the Section IV, the empirical analysis is made. Finally, a concise and comprehensive conclusion is made of the study in the Section V.

II. RELATED WORKS

In this section, we introduce some basic theory of differential privacy, DPC algorithm and evaluation metrics of clustering algorithm. Some of the symbols and their meanings used in this paper are shown in Table 1.

A. DPC ALGORITHM

The core ideas of DPC algorithm [14] are as follows :1) assume that the clustering center point of each class is the maximum local density point in this class, and the local density of other points is lower than the peak point and surrounding the center point. 2) The distance between different types of centers is relatively far. In order to find the cluster center which satisfies both these two conditions, the definition of local density and center offset distance is introduced.

Definition 1 (Local density): Let d_{ij} represent the distance between data points i and j, and $rank_i$ represents the ith index in ascending order of d_{ij}. d_c represents the cutoff distance determined by the cutoff percentage p, that is, $dc = rank_{p\times n}$.

TABLE 1. Notations and meanings.

Notations	Meanings
N	The sample size of the dataset
m	The dimensionality of the dataset
$X = [x_1, \cdots, x_i, \cdots, x_N]^T = \begin{bmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{N1} & \cdots & x_{Nm} \end{bmatrix}$	X represents the data set, x_i represents the ith sample of the data set, and x_{ik} represents the kth index of the ith sample of the data set
d_{ij}	The distance between data points i and j
p	The percentage of cut off distance
d_c	Cut off distance
$\rho = [\rho_1, \cdots, \rho_i, \cdots, \rho_N]^T$	ρ represents the local density of the dataset, and ρ_i represents the local density of the ith sample
$\delta = [\delta_1, \cdots, \delta_i, \cdots, \delta_N]^T$	δ represents the center offset distance of the dataset, and δ_i represents the center offset distance of the ith sample
$\gamma = [\gamma_1, \cdots, \gamma_i, \cdots, \gamma_N]^T$	γ represents the product of ρ and δ, and γ_i represents the composite index of the ith sample
D	Dataset
D'	The adjacent dataset of dataset D
$\rho' = [\rho_1', \cdots, \rho_i', \cdots, \rho_N']^T$	ρ' denotes the local density with noise, and ρ_i' denotes the ith local density with noise
$\delta' = [\delta_1', \cdots, \delta_i', \cdots, \delta_N']^T$	δ' denotes the center offset distance corresponding to ρ'
$\gamma' = [\gamma_1', \cdots, \gamma_i', \cdots, \gamma_N']^T$	γ' is the product of ρ' and δ'
$tend = [tend_1, \cdots, tend_N]^T$	$tend$ represents the slope change trend of the ranking graph drawn after γ is sorted in descending order, and $tend_i$ represents the slope change trend of ith points before γ is sorted

Local density of each point ρ_i can be defined in two ways according to the sample size and the type of data. Hard statistical method based on truncated kernel was used for datasets with large sample size and discrete data, which represents the number of points contained in the circle with x_i as the center and d_c as the radius. For continuous datasets with uniform distribution and small sample size, the soft statistical method based on Gaussian kernel is used to calculate the local density. The formula for calculating local density based on truncated kernel is shown in (1), where $\chi(d)$ means that when the variable value d is greater than 0, $\chi(d)$ takes 0, otherwise takes 1, and the formula is shown in (2). The formula for calculating local density based on Gaussian kernel is shown in (3).

$$\rho_i = \sum_j \chi(d_{ij} - d_c) \tag{1}$$

$$\chi(d) = \begin{cases} 0, & d > 0 \\ 1, & d \le 0 \end{cases} \tag{2}$$

$$\rho_i = \sum_j e^{-(\frac{d_{ij}}{d_c})^2} \tag{3}$$

The sample size of the experimental datasets in this paper are not large and the type of these datasets are continuous, so the method based on Gaussian kernel will be used to calculate the local density.

Definition 2 (Center deviation distance): When the local density of point set $\{x_j\}$ is greater than point x_i, the center deviation distance of each point δ_i represents the minimum distance between point set $\{x_j\}$ and point x_i. Otherwise, for the local maximum density point, the center deviation distance δ_i is the maximum distance between the two. Its calculation formula is shown in (4).

$$\delta_i = \begin{cases} \min\limits_j \{d_{ij}\}, & \rho_j > \rho_i \\ \max\limits_j \{d_{ij}\}, & otherwise \end{cases} \tag{4}$$

Definition 3 (Composite indicator): The DPC algorithm select the points with relatively large ρ_i and δ_i as the cluster center, and the point where ρ_i and δ_i are relatively large and the other is relatively small is regarded as the noise point. In order to avoid choosing noise points as clustering centers, a comprehensive index γ_i, the product of ρ_i and δ_i normalized by (5), is introduced to make the real clustering center point have a larger value of γ_i while the γ_i value of noise points is small. The calculation formula of Composite indicator is shown in (6).

$$\rho_i = \frac{\rho_i - \min(\rho_i)}{\max(\rho_i) - \min(\rho_i)} \tag{5}$$

$$\gamma_i = \rho_i \delta_i \tag{6}$$

B. DIFFERENTIAL PRIVACY [30]

Definition 4 (Differential privacy): Suppose M is a random algorithm, S is the set of all output results of the algorithm, $P[\cdot]$ is the probability of unknown set. For any two datasets D and D' with only one sample difference, if M satisfies to provide the datasets with differential privacy protection which value of privacy budget is ε, then:

$$P[M(D) \in S] \le e^{\varepsilon} P[M(D') \in S] \tag{7}$$

where, $\varepsilon > 0$ represents the differential privacy budget, which is used to describe the probability that a sample is added or reduced in the dataset and the algorithm outputs the same result, which can quantify the degree of privacy protection. It can be seen from (7) that the smaller ε is, the better the privacy protection effect is. When ε is equal to 0, the output distribution is indistinguishable from the actual results, but the availability of the original data is also lost, so the value of the privacy budget needs to be balanced between the degree of privacy protection and the availability of data.

C. LAPLACIAN NOISE MECHANISM

The differential privacy protection of the algorithm can be realized by adding a certain mechanism noise to the core statistic. Depending on the type of statistic to which noise is added, the type of mechanism for adding noise is decided: if the statistic is a continuity variable, the noise of the Laplacian mechanism [30] is added; Otherwise add exponential mechanism noise [33]. In this paper, the Laplace noise will be added to the continuity statistics.

The probability density function of the Laplace distribution is:

$$p(x|\mu, b) = \frac{1}{2b}\exp(-\frac{|x-\mu|}{b}) \tag{8}$$

where μ is the position parameter and b is the scale parameter. Let $f(D)$ be the core statistic of the algorithm, then the mechanism of adding Laplacian noise can be expressed as:

$$M(D) = f(D) + lap(\mu, b) \tag{9}$$

where $lap(\mu, b)$ is the random noise subject to Laplacian distribution. Under normal circumstances, μ is equal to 0, and the value of parameter b is determined by the sensitivity of statistics Δf [34] and privacy budget ε, which can be expressed as $b = \Delta f/\varepsilon$. The smaller ε is, the larger the scale parameter is, the larger the noise is, and the better the privacy protection degree is.

Δf describes the maximum impact of deletion or addition of any dataset on query results, its calculation formula is shown as follows:

$$\Delta f = \max_{D,D'} |f(D) - f(D')| \tag{10}$$

In this paper, the maximum difference of local density is 1, so Δf is equal to 1. Then the local density with Laplacian noise ρ_i' can be expressed as follows:

$$\rho_i' = \rho_i + lap(0, \frac{1}{\varepsilon}) \tag{11}$$

D. CLUSTER EVALUATION INDICATORS

Calinski-Harabasz Index (*CH*) [35], adjusted Rand Index (*ARI*) [36], adjusted mutual information (*AMI*) [36], Fowlkes and Mallows Index (*FMI*) [37] are used to evaluate the clustering results of the algorithm. Where *CH* is the internal evaluation method, the larger the value is, the closer within the class and the dispersed between the classes, that is, the better clustering result. *ARI*, *AMI* and *FMI* are external evaluation indicators. The larger the value is, the more consistent the clustering result is with the real situation, and the better the clustering effect is.

1) CH

CH measures the compactness of a class by calculating the square sum of the distance between each point in the class and the center of the class, and measures the separation of the dataset by calculating the square sum of the distance between each center point and the center of the dataset. *CH* is obtained by the ratio of separation degree and compactness. Its calculation formula is as follows:

$$CH(k) = \frac{SSB/(k-1)}{SSW/(n-k)} \tag{12}$$

$$SSB = \sum_{j=1}^{k} n_j |C_j - \bar{X}|^2 \tag{13}$$

$$SSW = \sum_{j=1}^{k} \sum_{x \in C_j} |x - C_j| \tag{14}$$

where, n represents the number of samples, k represents the number of clusters, *SSB* represents the sum of squares of deviations between classes, *SSW* represents the sum of squares of deviations within classes, n_j and C_j respectively represent the number of samples and the center point of class j, and $\bar{X}$ is the center point of the whole dataset.

TABLE 2. Contingency table.

		Predicted class		
		T	F	sum
Actual class	T	TP	FN	N_3
	F	FP	TN	N_4
	sum	N_1	N_2	N

2) FMI

FMI is the geometric mean of recall and precision when the two evaluation indexes in contradict each other. After clustering the data, a confusion matrix can be generated, as shown in the following table 2:

As is shown in Table 2, *TP* represents the number of sample pairs that are actually in a category and predicted to be in a category, *FN* represents the number of sample pairs that are actually in a category but predicted not to be in a category, *FP* represents the number of sample pairs that are actually not in a category but predicted in a category, *TN* represents the number of sample pairs that are not actually in a category and predicted not to be in a category, N_1 represents the number of sample pairs predicted in a category; N_2 represents the number of sample pairs that are not predicted in a category, N_3 represents the number of sample pairs that are actually in a category, N_4 is the number of sample pairs that are not actually in a category, and N is the number of any two samples in a class.

Precision (denoted as P) is the proportion of the samples predicted as a category that are actually in a category in terms of the predicted results. It can be expressed as:

$$P = \frac{TP}{TP + FP} \tag{15}$$

Recall (denoted as R) represents the proportion of the samples actually in a class that are predicted in a class. It can be expressed as:

$$R = \frac{TP}{TP + FN} \tag{16}$$

FMI which measures the geometric mean of P and R can be expressed by (17):

$$FMI = \sqrt{P * R} = \frac{TP}{\sqrt{(TP + FP)(TP + FN)}} \tag{17}$$

The value range of *FMI* is between [0,1]. The larger the value is, the more efficient the clustering result is.

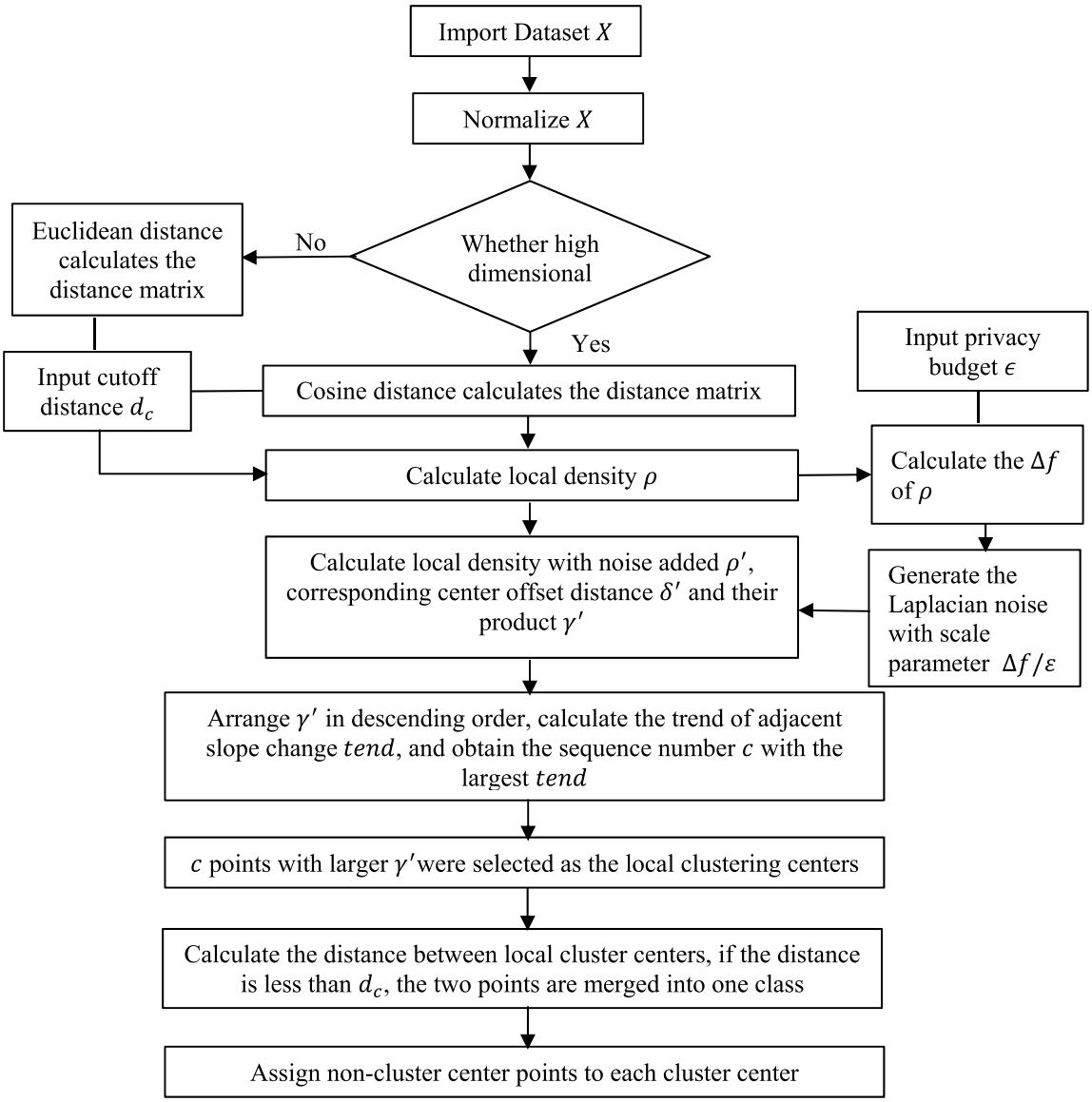


FIGURE 1. Flowchart of a new DPC algorithm with adaptive clustering center based on differential privacy.

3) *ARI* AND *AMI*

ARI and *AMI* are used to measure the degree of coincidence of two distributions, respectively obtained by rand index (RI) and mutual information index (MI) transformation, and their value ranges are between $[-1,1]$. The larger the value is, the more consistent the clustering effect is with the real situation, and the index is close to 0 when the clustering results are randomly generated.

The expression of RI is:

$$RI = \frac{TP + TN}{TP + FN + FP + FN} \tag{18}$$

Suppose $E(RI)$ represent the expectation of RI,the calculation formula of ARI is:

$$ARI = \frac{RI - E(RI)}{\max(RI) - E(RI)} \tag{19}$$

Suppose $H(u)$, $H(v)$ represent the information entropy of actual sample classification and sample prediction results, respectively, $MI(u, v)$ represent the mutual information between the two, and $EMI(u, v)\}$ represent the expectation of $MI(u, v)$, then AMI can be calculated as follows:

$$AMI = \frac{MI(u, v) - E\{MI(u, v)\}}{\max\{H(u), H(v)\} - E\{MI(u, v)\}} \tag{20}$$

III. IMPROVEMENT OF DPC ALGORITHM BASED ON DIFFERENTIAL PRIVACY

In view of the poor adaptability for high-dimensional data, the subjective selection of clustering centers in DPC algorithm and the privacy problems in clustering analysis, an improved DPC algorithm with adaptive clustering center based on differential privacy is proposed from the perspective of ranking graph.

Firstly, in view of the poor adaptability of the algorithm in high dimension, according to the classification method based on cosine distance proposed in the literature [38] can better identify the pollutant type than the method based on Euclidean distance, cosine distance is used to measure the similarity between data in high-dimensional datasets. Then, aiming at the subjectivity of cluster center selection and the problems existing in the slope change trend of the ranking graph defined in [18] and [27], the weight $(i-1)/i$ is introduced to redefine the trend of slope change of ranking graph creatively, and the threshold value of ranking graph statistics γ is obtained to realize automatic clustering center selection. Finally, aiming at the problem of privacy leakage, according to literature [31], the Laplacian noise of appropriate privacy budget is added to the core statistic of the algorithm (local density). In the whole process of the algorithm, as a distance measurement method, the calculation of local density and center offset distance of cosine distance is the same as the method using Euclidean distance, so the mechanism of adding Laplacian noise to local density is also consistent with that using Euclidean distance.

The specific improvement process is described in Figure 1.

A. COSINE DISTANCE

Euclidean distance cannot fully reflect the similarity between high-dimensional complex data. According to literature [38], the cosine distance is used to measure the similarity between data for high-dimensional datasets. Cosine distance measures the angle between two spatial samples rather than the amplitude difference between two spatial samples, and it is suitable for similarity measurement in high-dimensional data clustering.

Let θ denote the angle between x_i and x_j, then the cosine similarity $similarity(x_i, x_j)$ of the two samples can be expressed as follows:

$$similarity(x_i, x_j) = \cos(\theta) = \frac{x_i \bullet x_j}{|x_i||x_j|} = \frac{\sum_{k=1}^{m} x_{ik} \bullet x_{jk}}{\sqrt{\sum_{k=1}^{m} x_{ik}^2} \sqrt{\sum_{k=1}^{m} x_{jk}^2}} \tag{21}$$

where x_{ik} represents the kth index of the ith sample of the dataset, and m represents the dimension of the dataset. According to the definition of cosine similarity, when two samples are similar, the closer their cosine similarity is to 1, the closer the distance is. Otherwise, the closer the cosine similarity is to -1, the farther the distance is. Therefore, the cosine similarity can be transformed into a measure of cosine distance d_{ij} according to (22):

$$d_{ij} = 1 - similarity(x_i, x_j) \tag{22}$$

B. DPC ALGORITHM WITH ADAPTIVE CLUSTERING CENTERS BASED ON RANKING GRAPH

Aiming at the subjective problem of clustering center selection of DPC algorithm, an improved DPC algorithm with adaptive clustering center was proposed from the perspective of ranking graph.

The idea of selecting clustering centers based on ranking graph is to select the points with relatively large γ values in ranking graph as clustering centers. Literature [27] determines the critical point by searching the point with the largest slope change trend of the sorting graph, and then adaptively determines the clustering center. This algorithm is difficult to deal with the situation that the ranking graph drops suddenly from point 1 to point 2. In the slope variation trend of the ranking graph proposed in literature [18], the introduction of the weight $i-1$ can avoid the situation that the sorting graph may drop suddenly from point 1 to point 2 in actual practice. However, with the increase of the serial number of data points, the weight increases rapidly, thus making the judgment threshold wrong. In view of this situation, this paper introduces the weight $(i-1)/i$ to redefine the trend of slope change in the ranking graph, and determines the threshold of γ by searching the point with the largest trend of slope change, so as to realize the automatic selection of clustering centers. The specific steps are as follows:

1)Arrange γ values in descending order, and γ_i^* represents the ith value after descending order. (23), (24) are used to calculate the slope change trend of each point, respectively.

$$tend_i = \frac{i-1}{i} \cdot [(\gamma_i^* - \gamma_{i+1}^*)\text{-}(\gamma_{i+1}^* - \gamma_{i+2}^*)] \tag{23}$$

$$tend_i = \frac{i-1}{i} \cdot \frac{\gamma_i^* - \gamma_{i+1}^*}{\gamma_{i+1}^* - \gamma_{i+2}^*} \tag{24}$$

2)The maximum data point c of $tend$ value can be obtained through (25):

$$c = \arg\max_i \{tend_i, i = 1, 2, \cdots, n-2\} \tag{25}$$

3)The point $\gamma_i \geq \gamma_c^*$ is selected as the set ELC of potential clustering centers, which can be expressed as:

$$ELC = \{i | \gamma_i \geq \gamma_c^*, i = 1, 2, \cdots, n-1\} \tag{26}$$

4)Calculate the distance d_{ij} between ELC. If d_{ij} is greater than d_c, points i and j are determined as the actual clustering center; otherwise, points with low local density are excluded. In summary, the actual clustering center can be expressed as follows:

$$\begin{aligned} LC = &\{i, j | d_{ij} > d_c, i, j \in ELC\} \\ &\cup \{i | d_{ij} \leq d_c, \rho_i > \rho_j, i, j \in ELC\} \end{aligned} \tag{27}$$

The process of the improved DPC algorithm with adaptive clustering center (ADPC and UDPC) is shown in Algorithm 1, where ADPC represents the algorithm for calculating the slope change trend based on (23) and UDPC represents the algorithm for calculating the slope change trend based on (24):

C. THE IMPROVED DPC ALGORITHM WITH ADAPTIVE CLUSTERING CENTER BASED ON DIFFERENTIAL PRIVACY

Local density ρ_i is the core statistic of DPC algorithm. This paper adds the Laplacian noise of appropriate privacy budget

Algorithm 1 FADPC and UDPC

Input: Dataset X Ccutoff percentage p
Output: The clustering result
1: normalize X;
2: use Euclidean distance to calculate the distance matrix for low dimensional datasets, and use cosine distance to calculate the distance matrix for high dimensional datasets;
3: calculate the ρ_i of each point by (3), calculate the δ_i by (4), then normalize ρ_i and δ_i by (5), calculate the comprehensive index γ_i by (6);
4: γ_i^* is obtained after the descending order of γ_i. The slope variation trend of each point $tend_i$ is calculated according to (23) and (24), and the point c with the largest $tend_i$ is obtained through (25);
5: select the local clustering centers by (26);
6: calculate the distance between local cluster centers, if the distance is less than d_c, the two points are merged into one class by (27);
7: assign non-cluster center points to each cluster center;
8: return the clustering result.

Algorithm 2 FDP_ADPC and DP_UDPC

Input: Dataset X Ccutoff percentage p,privacy budget ε
Output: The clustering result
1: normalize X;
2: use Euclidean distance to calculate the distance matrix for low dimensional datasets, and use cosine distance to calculate the distance matrix for high dimensional datasets;
3: calculate the ρ_i of each point by (3); 4: select the appropriate privacy budget, generate the Laplacian noise according to the sensitivity of ρ_i, and add it to ρ_i to obtain the local density with noise (denoted as ρ_i') by (11), and calculate the corresponding center offset distance (denoted as δ_i') by (4), then normalize ρ_i' and δ_i' by (5), calculate the comprehensive index γ_i' by (6);
5: γ_i^* is obtained after the descending order of γ_i'. The slope variation trend of each point $tend_i$ is calculated according to (23) and (24), and the point c with the largest $tend_i$ is obtained through (25);
6: select the local clustering centers by (26);
7: calculate the distance between local cluster centers, if the distance is less than d_c, the two points are merged into one class by (27);
8: assign non-cluster center points to each cluster center;
9: return the clustering result.

to ρ_i to achieve the purpose of privacy protection. Based on the above analysis, the specific steps of the improved DPC algorithm with adaptive clustering center based on differential privacy are given:

1)Normalize the original data set, use Euclidean distance to calculate the distance between low dimensional dataset, or use cosine distance to calculate the distance matrix for high dimensional dataset;

2)Calculated local density ρ_i by (3);

3)Select the appropriate privacy budget, generate the Laplacian noise according to the sensitivity of ρ_i, and add it to ρ_i to obtain the local density with noise (denoted as ρ_i') by (11), and calculate the corresponding center offset distance (denoted as δ_i') by (4), then normalize ρ_i' and δ_i' by (5), calculate the comprehensive index γ_i' by (6);

4)Determine clustering center points adaptively according to Section III(B).

5)Cluster the non-central points and assign them to the class where the data points close to the central point and with greater local density are located.

The process of the improved DPC algorithm with adaptive clustering center based on differential privacy (DP_ADPC and DP_UDPC) is shown in Algorithm 2, where DP_ADPC represents the ADPC algorithm based on differential privacy, and DP_UDPC represents the UDPC algorithm based on differential privacy:

D. TIME COMPLEXITY ANALYSIS

This section will analyze the time complexity of the proposed algorithm. Compared with DPC algorithm, the improved DPC algorithm with adaptive clustering center based on differential privacy adds noise to the local density, calculates the comprehensive index γ_i of each sample and adaptively selects the clustering center. Its time complexity is mainly composed of the following parts:

1)Calculation of statistics. Standardize the dataset requires the time complexity is $O\left(N^2\right)$. The calculation of distance between samples requires the time complexity is $O\left(N^2\right)$. The time complexity of local density ρ_i of each sample is $O\left(N^2\right)$. The local density with noise ρ_i' requires the time complexity is $O(N)$. The center offset distance of each sample δ_i' corresponding to ρ_i costs $O\left(N^2\right)$. The time complexity of calculating γ_i' is $O(N)$. The total time of this part is $O\left(N^2\right) + O\left(N^2\right) + O\left(N^2\right) + O(N) + O\left(N^2\right) + O(N) \sim O\left(N^2\right)$.

2) Selection of clustering centers. Both ρ_i' and δ_i' are processed respectively in descending order to obtain the ordinals of all points after sorting. The time complexity is both $O(NlgN)$. To calculate the slope change trend of the ranking graph, the time complexity is $O(N)$. The point γ_i' larger than the threshold is selected as the local clustering center, and the time complexity is γ_i'. The distance d_{ij} between the local cluster centers is calculated., if d_{ij} is smaller than d_c, the two local cluster centers are grouped into one class, and the time complexity is $O\left(N^2\right)$. The total time complexity of this part is: $O(NlgN)+O(NlgN)+O(N)+O(N)+O\left(N^2\right) \sim O\left(N^2\right)$

3) Assign non-clustered center points. The non-clustering center points are allocated to the points with large local areas in the cluster according to the nearest neighbor principle. The time complexity of this part is the same as that of the DPC algorithm, both of which are $O(N)$.

The time complexity of DPC algorithm mainly comes from five parts :1) data standardization processing; 2) Calculate the

distance d_{ij} between samples; 3) Calculate the local density of each sample ρ_i; 4) Calculate the center offset distance of each sample δ_i; 5) Assign non-cluster center points. Before a few part time complexity is $O\left(N^2\right)$, the final step of time complexity is $O(N)$, so the total time of DPC algorithm is: $O\left(N^2\right)+O\left(N^2\right)+O\left(N^2\right)+O\left(N^2\right)+O(N) \sim O\left(N^2\right)$

To sum up, the total time complexity of the proposed algorithm is equal to that of the traditional DPC algorithm, which is $O\left(N^2\right)$.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

A. EXPERIMENTAL ENVIRONMENT AND DATA

In order to verify the effectiveness of the proposed algorithm, this paper runs on Jupyter based on python3.8.3 and adopts six synthetic datasets and six UCI datasets as test datasets. The characteristics of the data sets are shown in Table 3. The experimental environment is Windows10 system, the processor is Intel®Core (TM) i3-7130u CPU, the memory is 8.00gb, and the 64-bit operating system.

TABLE 3. Experimental datasets.

Dataset	Attributes	Size	Clusters	*Sources*
flame	2	240	2	Synthetic
spiral	2	312	3	Synthetic
compound	2	399	6	Synthetic
aggregation	2	788	7	Synthetic
R15	2	600	15	Synthetic
D31	2	3100	31	Synthetic
seeds	7	210	3	UCI
ecoli	7	336	8	UCI
movement	46	360	15	UCI
abalone	33	366	6	UCI

B. ANALYSIS OF EXPERIMENTAL RESULTS ON SYNTHETIC DATASETS

1) EFFECT OF THE IMPROVED DPC ALGORITHM WITH ADAPTIVE CLUSTERING CENTER

Each cluster evaluation index obtained by the proposed algorithm clustering on synthetic datasets is shown in the Figure 2, where GDPC represents the DPC algorithm of the slope change trend of the ranking graph defined in literature [18].

As can be seen from Figure 2, for flame dataset, set the cutoff percentage parameter p as 3, the *ARI*, *AMI* and *FMI* of ADPC algorithm are the best compared with the other two algorithms, the values of the three external evaluation indexes are all up to 1, the clustering result of ADPC algorithm is the closest to the standard classification result. However, the *CH* value of ADPC is smaller than that of the other two algorithms, which may be caused by the fact that the number of clustering of the other two algorithms is more than that of the standard classification result. For the dataset spiral, set p as 2, the *ARI*, *AMI*, *FMI* and *CH* of ADPC algorithm and UDPC algorithm are all optimal, the values of the three external evaluation indexes are all up to 1, *CH* is 6, and the *CH* of GDPC algorithm is 0 which is because its cluster label has only one class. For the dataset compound, when p is set to 6.5, GDPC algorithm and UDPC algorithm had better clustering effect than ADPC algorithm. For aggregation and R15 datasets, when p is set to 4, UDPC algorithm achieves the optimal value of each clustering evaluation, and the values of three external evaluation indexes are all about 1, indicating that the clustering results of UDPC algorithm are similar to the standard classification results, and the value of *CH* is also higher than the other two algorithms. For the dataset D31, set p to 1, the clustering effect of ADPC algorithm is the best, the clustering effect of UDPC algorithm is the second, and the clustering effect of GDPC algorithm is the worst.

According to the comprehensive analysis, ADPC algorithm performs better than the other two algorithms on the two datasets with fewer classification and the multi-classification dataset D31, while the UDPC algorithm performs better on the other three multi-classification datasets.

2) EFFECT OF THE IMPROVED DPC ALGORITHM WITH ADAPTIVE CLUSTERING CENTER BASED ON DIFFERENTIAL PRIVACY

The improved algorithm proposed in this paper combined with differential privacy technology is run in six synthetic datasets and its effect is observed. The value of privacy budget ε is gradually increased from 0.01 to 30, where ADPC algorithm is used for datasets with few categories and UDPC algorithm is used for datasets with multiple categories. The change of algorithm clustering effect with ε is shown in Figure 3.

Experimental results on six synthetic datasets show that the improved DPC algorithm with adaptive clustering center based on differential privacy has the following characteristics:

On the whole, with the slow increase of ε, all external clustering evaluation indexes of the algorithm on datasets show a trend of rising first and then reaching a stable state. By comparing the characteristics of the algorithm to reach the stable state on the six datasets, the dataset flame and compound have larger fluctuation range than that of other datasets, and the corresponding ε value of spiral dataset is significantly larger than that of other datasets after reaching the stable state.

The experimental results show that in a certain range, the clustering efficiency of the algorithm is better with the increase of ε. According to (7), the smaller ε is, the better the privacy protection effect is. In order to achieve the balance between privacy protection and effectiveness of the algorithm, the critical value of ε should be selected to make the algorithm reach the stationary state.

The noise added to the algorithm is random noise that obeys the Laplace distribution, so the clustering effect shows cyclic fluctuation changes after reaching the stationary state.

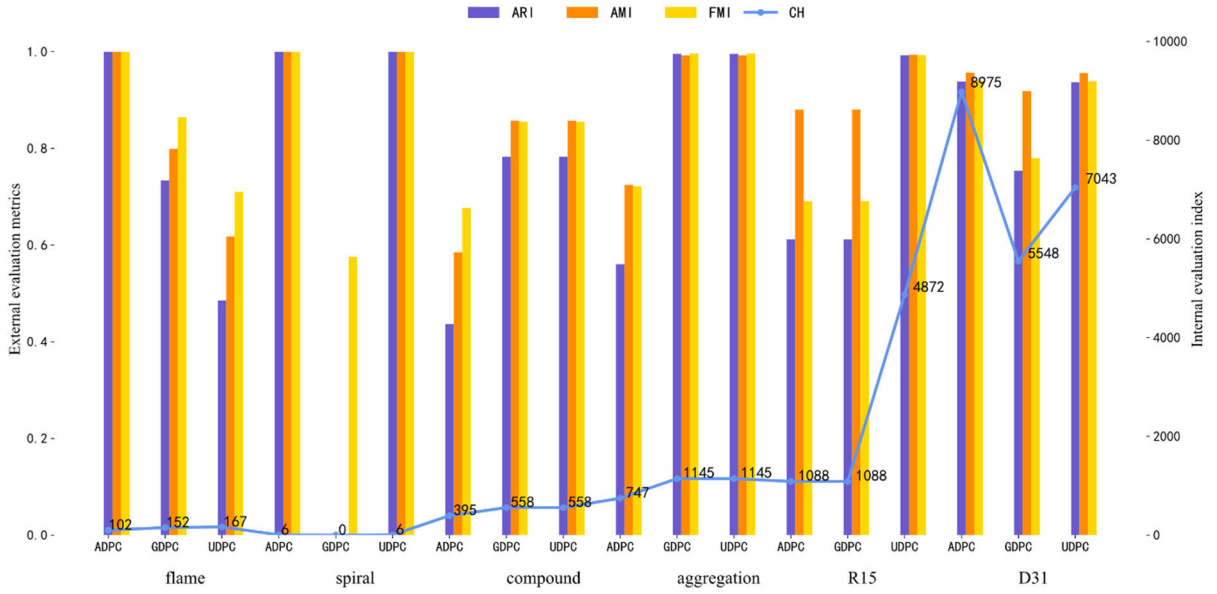


FIGURE 2. Comparison of ARIAMIFMI and CH for the algorithm running on synthetic datasets.

The average value of each evaluation index after 50 runs of the algorithm is taken, and the results on each synthetic dataset are shown in Table 4.

It can be seen from Table 4, for flame dataset, when ε is 10, DP_ADPC algorithm has the largest *ARI*, *AMI* and *FMI*, which indicates the better clustering effect than the other two algorithms. The *CH* of DP_UDPC algorithm is the largest, which may be caused by the excessive number of clusters. A clustering result of DP_ADPC algorithm on flame is shown in Figure 4(a). For spiral dataset, when ε is 30, the comparison results of several algorithms are the same as those of flame dataset. A clustering result of this data is shown in Figure 4(b). For compound dataset, when ε is 18, the values of four evaluation indexes of DP_GDPC are higher than those of the other two algorithms, and DP_UDPC has the second best effect. One of its clustering results is shown in Figure 4(c). For R15 dataset, when ε is 6, the evaluation index value of DP_UDPC is greater than that of the other two algorithms. A clustering result of this algorithm is shown in Figure 4(d). For aggregation dataset, when ε is set as 3, the comparison results of several algorithms are the same as those of R15 dataset. A clustering result is shown in Figure 4(e). When the dataset D31, ε is set at 2, the comparison results of several algorithms are the same as those of R15 and aggregation. A clustering result is shown in Figure 4(f).

On the whole, the appropriate ε for each dataset obtained from Figure 3 shows a rule: the larger the sample size is, the smaller the value of ε is. The smaller the ε is, the higher the degree of privacy protection is. Therefore, for the dataset with a larger sample size, the better privacy protection can be obtained by adding the Laplacian noise data with a smaller ε.

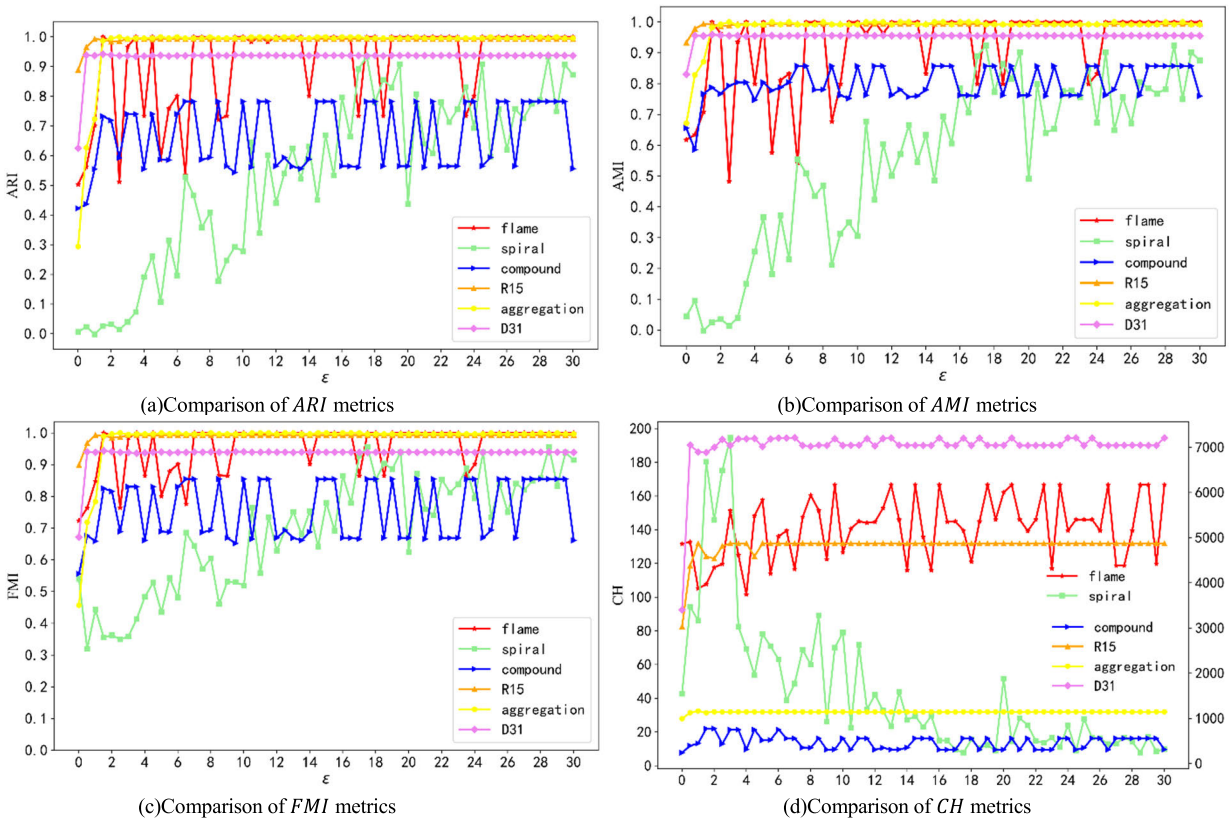
The next section we would verify this rule of the proposed algorithm on some UCI datasets.

C. ANALYSIS OF EXPERIMENTAL RESULTS ON UCI DATASETS

1) EFFECT OF THE IMPROVED DPC ALGORITHM WITH ADAPTIVE CLUSTERING CENTER

In this section, ADPC algorithm and UDPC algorithm are tested on six UCI datasets. Euclidean distance and cosine distance are respectively used to measure the similarity between data, and their effects are compared with k-means [12], DBSCAN [13] and SNNDPC [24]. The results are shown in Table 5. Where ''eu'' means the similarity between data measured by Euclidean distance, and ''cos'' means the similarity between data measured by cosine distance. The parameter k of k-means algorithm represents the number of clustering centers. The parameters *eps* and *minpts* of DBSCAN algorithm represent the clustering radius and density threshold, respectively, the maximum *ARI* criterion is used to determine these parameter values in this paper. The parameters *nc* and *kn* of SNNDPC algorithm respectively represent the number of clustering centers and the number of nearest neighbors, $kn = 2nc + 1$; Parameter p of the improved DPC algorithm with adaptive clustering center is also determined by the maximum *ARI* criterion. The values in bold in the table indicate the best experimental results.

For the dataset seeds, which contains 210 samples and 7 attributes with 3 categories, ADPC-eu has the largest *ARI*, *AMI* and *FMI* compared with other algorithms, but *CH* is lower than K-means algorithm and ADPC-cos algorithm. For ecoli dataset, which contains 336 samples and 7 attributes with 8 categories, the *AMI* and *FMI* of UDPC-cos algorithm are the best compared with other algorithms, the *ARI* of SNNDPC algorithm is the best, and the *CH* of K-means algorithm is the best. For movement dataset, which contains 360 samples and 90 features with 15 categories, the values of



(a)Comparison of *ARI* metrics (b)Comparison of *AMI* metrics

(c)Comparison of *FMI* metrics (d)Comparison of *CH* metrics

FIGURE 3. Comparison of clustering metrics on synthetic datasets after adding noise.

TABLE 4. Clustering effect of each algorithm based on differential privacy on synthetic dataset.

Algorithm	ϵ	*ARI*	*AMI*	*FMI*	*CH*	ϵ	*ARI*	*AMI*	*FMI*	*CH*
			flame					spiral		
DP_ADPC		**0.9663**	**0.9679**	**0.9834**	106.3737		**0.8503**	**0.8527**	**0.9006**	11.78586
DP_GDPC	10	0.8321	0.8665	0.9141	130.8034	22	0.5690	0.6256	0.7693	12.13237
DP_UDPC		0.4956	0.6122	0.7163	**144.1171**		0.5171	0.6327	0.6659	**37.05883**
			compound					R15		
DP_ADPC		0.4366	0.5853	0.6763	395.3581		0.2244	0.4262	0.4179	535.3412
DP_GDPC	18	**0.7678**	**0.8490**	**0.8432**	**571.198**	6	0.5238	0.8094	0.6311	1219.703
DP_UDPC		0.6786	0.8125	0.7663	444.3942		**0.9928**	**0.9938**	**0.9932**	**4871.864**
			aggregation					D31		
DP_ADPC		0.0898	0.1163	0.5064	119.4161		0.1353	0.5227	0.3181	2069.684
DP_GDPC	3	0.9093	0.9564	0.9306	1116.544	2	0.6843	0.8920	0.7298	5043.003
DP_UDPC		**0.9563**	**0.9777**	**0.9665**	**1139.618**		**0.9090**	**0.9520**	**0.9126**	**8036.337**

ARI, *AMI* and *FMI* of UDPC-cos are optimal, and ADPC-cos has the best *CH*. For dermatology dataset, which contains 366 samples and 33 features with 6 categories, the four cluster evaluation indexes of ADPC-cos algorithm are the best compared with other algorithms. For banknote dataset, which contains 1372 samples and 4 features with two classes, the *ARI*, *AMI* and *FMI* of ADPC-eu is the largest compared to other algorithms, and *CH* of K-means algorithm is the best. For abalone dataset, which contains 4177 samples and 8 features with 3 categories, the *AMI* and *FMI* of ADPC-eu are the best compared with other algorithms, and the *ARI* and *CH* of UDPC-cos are the best, these optimal values are not different from those of ADPC-cos.

In summary, the clustering effect of ADPC algorithm is better for seeds, dermatology, banknote and abalone, which are less categorical datasets; the clustering effect of UDPC algorithm is better for ecoli and movement, which are multi-classification datasets. Using Euclidean distance to measure low-dimensional datasets (such as datasets seeds, banknote, abalone) works well, while cosine distance is well for high-dimensional datasets movement, dermatology and low dimensional dataset ecoli.

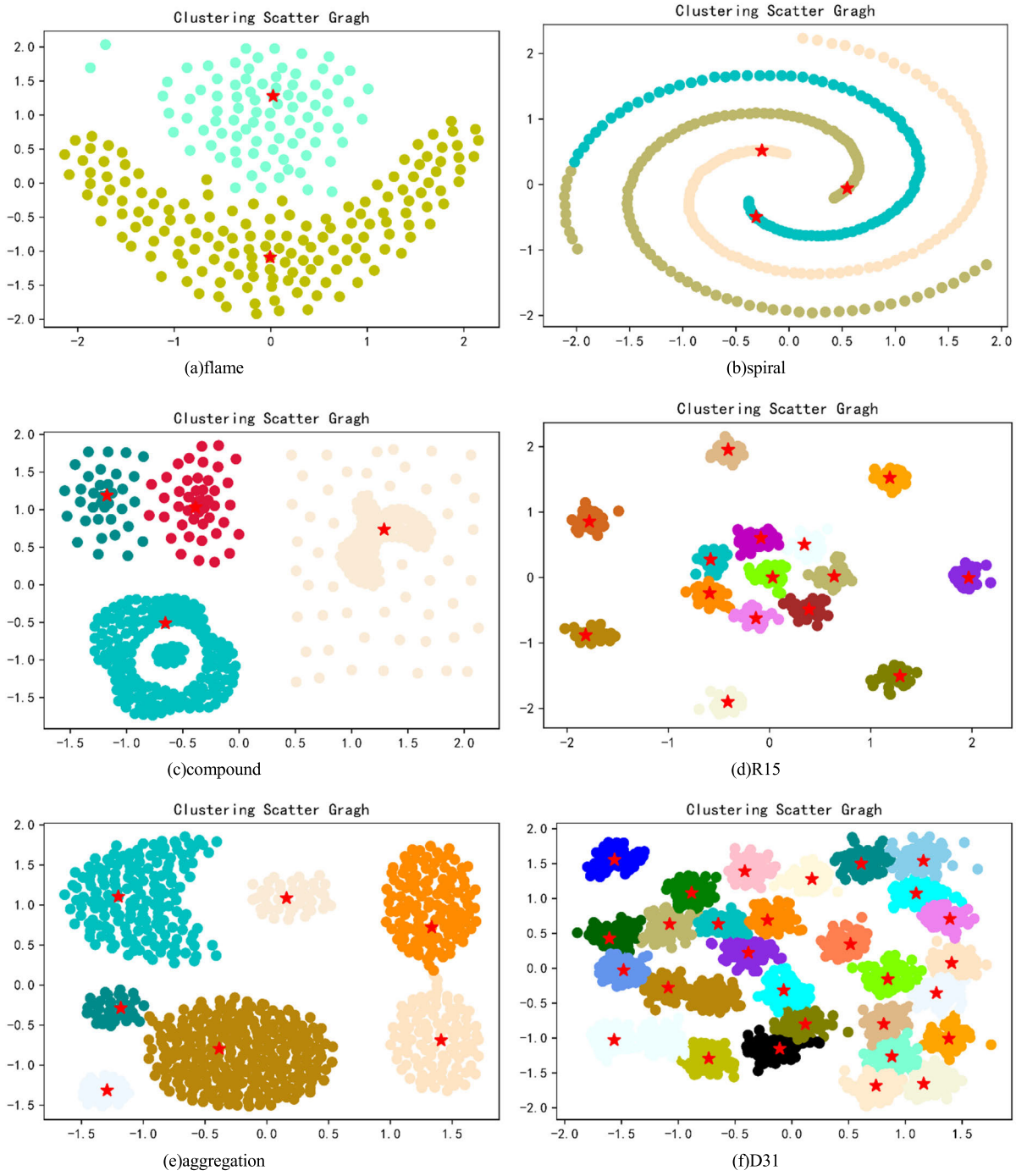
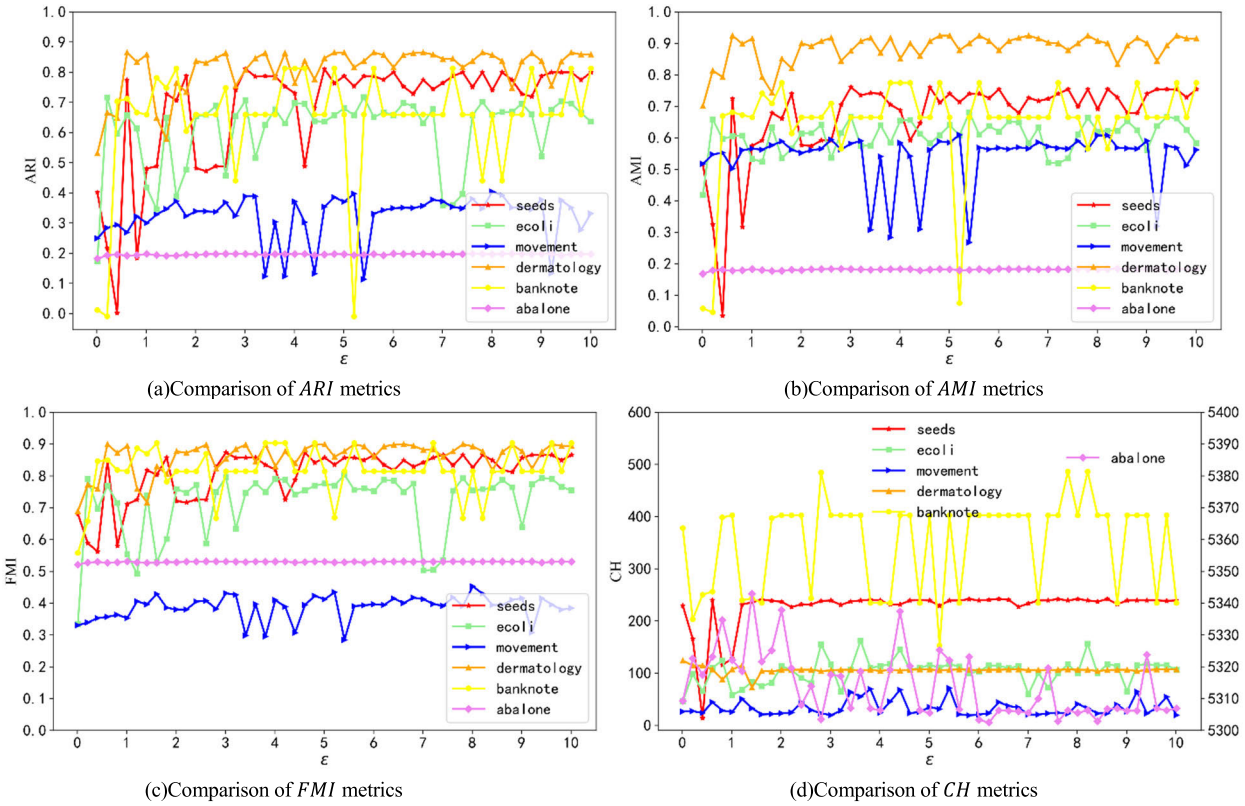


FIGURE 4. Scatter plot of each synthetic dataset after adding the appropriate privacy budget.

2) EFFECT OF THE IMPROVED DPC ALGORITHM WITH ADAPTIVE CLUSTERING CENTER BASED ON DIFFERENTIAL PRIVACY

The optimal algorithm mentioned in the last section combined with differential privacy technology was run in the UCI datasets and its effect was observed. The value of privacy budget ε was slowly increased from 0.01 to 10, the clustering effect of the algorithm on each dataset changed with ε, as shown in the Figure 5.

According to the trend of clustering evaluation index after noise is added to each UCI dataset in Figure 5, an appropriate privacy budget is selected for each dataset, and the average of the four evaluation indexes obtained by running the algorithm for 50 times is shown in the Table 6, where the algorithm



(a)Comparison of *ARI* metrics (b)Comparison of *AMI* metrics

(c)Comparison of *FMI* metrics (d)Comparison of *CH* metrics

FIGURE 5. **Comparison of clustering metrics on UCI datasets after adding noise.**

TABLE 5. **Comparison of clustering results of UCI datasets.**

Algorithm	Parameter	*ARI*	*AMI*	*FMI*	*CH*	Parameter	*ARI*	*AMI*	*FMI*	*CH*
	seeds					ecoli				
K-means	3	0.7733	0.7255	0.8482	**249.784**	8	0.4998	0.6179	0.6203	**163.235**
DBSCAN	1.21，17	0.4084	0.4528	0.6477	96.50353	0.81，4	0.4634	0.4437	0.6383	39.27404
SNNDPC	3，7	0.7445	0.7366	0.8296	238.0984	8，17	**0.6738**	0.6298	0.7693	67.61783
ADPC-eu	**0.9**	**0.7997**	**0.7542**	**0.8659**	239.5727	0.5	0.3633	0.4218	0.6392	101.1801
UDPC-eu	6.5	0.7479	0.6852	0.8311	206.9717	1.7	0.6017	0.5612	0.7261	78.12206
ADPC-cos	0.5	0.4969	0.5793	0.7248	248.9427	0.1	0.6166	0.5780	0.7437	114.125
UDPC-cos	1.3	0.7209	0.6834	0.8137	236.6507	1.7	0.6714	**0.6326**	**0.7721**	113.7497
	movement					dermatology				
K-means	15	0.3300	0.5587	0.3775	57.98495	6	0.7018	0.8567	0.7612	92.76218
DBSCAN	2.81，1	0.2546	0.4787	0.2986	14.13968	4.01，4	0.4403	0.6309	0.5830	61.98288
SNNDPC	15，31	0.2784	0.4616	0.3417	33.60206	6，13	0.8251	0.8830	0.8681	86.14091
ADPC-eu	0.5	0.1511	0.3707	0.3108	56.35564	0.1	0.7766	0.7763	0.8201	63.80155
UDPC-eu	0.5	0.3469	0.5774	0.3886	30.42896	0.1	0.7750	0.7657	0.8193	48.46513
ADPC-cos	3.7	0.1822	0.3925	0.3327	**64.7497**	0.5	**0.8371**	**0.9003**	**0.8773**	**106.465**
UDPC-cos	2.1	**0.3937**	**0.5954**	**0.4355**	37.43286	1.7	0.7235	0.7875	0.7803	39.35878
	banknote					abalone				
K-means	2	0.0132	0.0105	0.5096	**807.2202**	3	0.1357	0.1631	0.4298	4941.174
DBSCAN	0.41，4	0.6109	0.5681	0.7835	100.7685	0.41,19	0.1493	0.1360	0.5010	2444.325
SNNDPC	2，5	-0.0032	0.0130	0.7039	39.8792	3，7	0.0313	0.1048	0.5381	836.9171
ADPC-eu	2.5	**0.9624**	**0.9316**	**0.9814**	316.8732	2.9	0.1951	**0.1846**	**0.5383**	4911.784
UDPC-eu	4.1	0.7482	0.7013	0.8669	297.7253	1.7	0.1549	0.1761	0.4772	1358.754
ADPC-cos	0.5	0.2670	0.3608	0.5479	631.297	17.7	0.1969	0.1825	0.5302	5306.91
UDPC-cos	15.3	0.2775	0.4512	0.5293	172.8368	19.3	**0.197**	0.1824	0.5304	**5319.57**

with "DP" represents each improved algorithm based on differential privacy.

As can be seen from Table 6, for datasets seeds and banknote, which are low dimensional and less categorical

TABLE 6. Clustering effect of each algorithm based on differential privacy on UCI dataset.

Algorithm	ϵ	*ARI*	*AMI*	*FMI*	*CH*	ϵ	*ARI*	*AMI*	*FMI*	*CH*
	seeds					ecoli				
DP_ADPC-eu	5	**0.7676**	**0.7269**	**0.8457**	231.3868	4.5	0.3781	0.4457	0.6371	106.5097
DP_UDPC-eu		0.6960	0.6746	0.7906	134.4915		0.5449	0.5439	0.6966	96.00133
DP_ADPC-cos		0.4887	0.5659	0.7184	**240.2988**		0.5237	0.5116	0.7052	**108.7815**
DP_UDPC-cos		0.6115	0.6193	0.7564	204.7486		**0.6470**	**0.6207**	**0.7485**	105.7548
	movement					dermatology				
DP_ADPC-eu	2	0.0967	0.2935	0.2830	49.32037	5	0.5331	0.5898	0.6280	48.2837
DP_UDPC-eu		0.2932	0.5459	0.3547	29.90634		0.5316	0.6403	0.6398	25.84898
DP_ADPC-cos		0.1732	0.3799	0.3295	**62.56932**		**0.8373**	**0.9013**	**0.8792**	**106.5613**
DP_UDPC-cos		**0.3673**	**0.5801**	**0.4162**	31.91866		0.4766	0.6370	0.5837	20.00002
	banknote					abalone				
DP_ADPC-eu	2	**0.8864**	**0.8594**	**0.9403**	86.92925	1	0.1762	0.1785	0.4948	1446.487
DP_UDPC-eu		0.7387	0.7074	0.8608	351.612		0.1543	0.1771	0.4756	1934.379
DP_ADPC-cos		0.3248	0.3599	0.5954	**639.6806**		**0.1950**	**0.1806**	**0.5291**	**5319.416**
DP_UDPC-cos		0.1135	0.3242	0.3240	181.7855		0.1896	0.1741	0.5127	4892.484

datasets, respectively set ε as 5 and 2, DP_ADPC-eu has the best *ARI*, *AMI* and *FMI*. For datasets ecoli and movement, which are high-dimensional and multi-classification datasets, respectively set ε as 4.5 and 2, the *ARI*, *AMI* and *FMI* of DP_UDPC-cos is better than others. For datasets dermatology and abalone, which are high-dimensional and less categorical datasets, respectively set ε as 5 and 1, *ARI*, *AMI*, *FMI* and *CH* of DP_ADPC-cos are better than those of other data sets, and the clustering effect is the best.

By analyzing the relationship between the sample size and the critical value of different types of datasets, we can get the same conclusion as the synthetic datasets: that is, for datasets with larger sample size, adding the Laplacian noise data with smaller ε can get good privacy protection.

V. CONCLUSION

Based on the study of the improved DPC algorithm with adaptive clustering center, this paper introduces differential privacy protection technology, which can eliminate the hidden danger of data being attacked by strictly defined attack model in the process of clustering, and effectively protect data privacy.

Firstly, in order to solve the problem of poor adaptability of the algorithm in high dimension, different methods were used to measure the similarity between datasets in different dimensions: cosine distance was used to measure the similarity of high-dimensional datasets, and Euclidean distance was used to measure the similarity of low-dimensional datasets. Then, aiming at the problem that the cluster center cannot be automatically selected, the weight $(i-1)/i$ was introduced to measure the slope change trend of the ranking graph. Finally, aiming at the differential privacy problem, the Laplacian noise with appropriate privacy budget is added to the local density to achieve the balance between privacy protection and clustering effectiveness.

Experimental results on 6 synthetic datasets and 6 UCI datasets show that:

1)Within a certain range, the clustering effectiveness of the algorithm is better with the increase of ε. In order to select the appropriate ε, it is necessary to find the critical value that makes each evaluation index reach the stable state.

2)For the same type of datasets, the larger the sample size, the smaller the privacy budget is, so as to achieve the purpose of clustering effectiveness of privacy protection.

3)The proposed algorithm in this paper can also get a better clustering effect on the basis of considering the privacy problem.

The next step is to realize the adaptive selection of cutoff distance, because that the cutoff distance of the proposed algorithm is determined by maximum *ARI* criterion which would cost a lot of time.

REFERENCES

[1] X. Cheng and F. Qu, "Ocean data sharing based on blockchain," in *Proc. IEEE 6th Int. Conf. Big Data Analytics (ICBDA)*, Mar. 2021, pp. 155–159.

[2] J. Wu, N. Mu, X. Lei, J. Le, D. Zhang, and X. Liao, "SecEDMO: Enabling efficient data mining with strong privacy protection in cloud computing," *IEEE Trans. Cloud Comput.*, vol. 10, no. 1, pp. 691–705, Jan. 2022.

[3] M. A. Mahdi, K. M. Hosny, and I. Elhenawy, "Scalable clustering algorithms for big data: A review," *IEEE Access*, vol. 9, pp. 80015–80027, 2021.

[4] K. Lyu and H. Yan, "Identification method of dress pattern drawing based on machine vision algorithm," in *Proc. 3rd Int. Conf. Comput. Vis., Image Deep Learn. Int. Conf. Comput. Eng. Appl. (CVIDL ICCEA)*, May 2022, pp. 76–79.

[5] C. Kolluru, J. Lee, Y. Gharaibeh, H. G. Bezerra, and D. L. Wilson, "Learning with fewer images via image clustering: Application to intravascular OCT image segmentation," *IEEE Access*, vol. 9, pp. 37273–37280, 2021.

[6] C. Yang, S. Liu, L. Bruzzone, R. Guan, and P. Du, "A feature-metric-based affinity propagation technique for feature selection in hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 10, no. 5, pp. 1152–1156, Sep. 2013.

[7] Q. Wang, H. Li, T. Wang, C.-J. Wang, and X. Yin, "Using hierarchical clustering algorithm to detect community structure in traditional Chinese medicine formula network," in *Proc. IEEE 27th Int. Conf. Tools Artif. Intell. (ICTAI)*, Nov. 2015, pp. 132–138.

[8] J. Peng, L. Zhu, Y. Wang, and J. Chen, "Mining relationships among multiple entities in biological networks," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 17, no. 3, pp. 769–776, May 2020.

[9] F. Deng, W. Gu, W. Zeng, Z. Zhang, and F. Wang, "Hazardous chemical accident prevention based on K-means clustering analysis of incident information," *IEEE Access*, vol. 8, pp. 180171–180183, 2020.
[10] C. Li, H. Chen, T. Li, and X. Yang, "A stable community detection approach for complex network based on density peak clustering and label propagation," *Int. J. Speech Technol.*, vol. 52, no. 2, pp. 1188–1208, Jan. 2022.
[11] X. Q. Chen, L. J. Zhou, and Y. Z. Liu, "Review on clustering algorithms," *J. Integr. Technol.*, vol. 6, no. 3, pp. 41–49, 2017.
[12] C. Jie, Z. Jiyue, W. Junhui, W. Yusheng, S. Huiping, and L. Kaiyan, "Review on the research of K-means clustering algorithm in big data," in *Proc. IEEE 3rd Int. Conf. Electron. Commun. Eng. (ICECE)*, Dec. 2020, pp. 107–111.
[13] M. Ester, H. P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. 2nd Int. Conf. Knowl. Discovery Data Mining*. Sacramento, CA, USA: AAAI Press, 1996, pp. 226–231.
[14] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," *Science*, vol. 344, no. 6191, pp. 1492–1496, Jun. 2014.
[15] M. Du, S. Ding, and H. Jia, "Study on density peaks clustering based on K-nearest neighbors and principal component analysis," *Knowl.-Based Syst.*, vol. 99, pp. 135–145, May 2016.
[16] Y. Yang, "Improvements of the density-peak-based clustering algorithm," M.S. thesis, Dept. Comput. Eng., Xidian Univ., Xi'an, China, 2020, doi: 10.27389/d.cnki.gxadu.2020.001048.
[17] S. Ma, H. You, L. Tang, and P. He, "An adaptive density peak clustering algorithm," *J. Northeastern Univ. Natural Sci.*, vol. 43, no. 6, pp. 761–768, 2022.
[18] Z. Yang, H. Wang, and Y. Zhou, "A clustering algorithm with adaptive cut-off distance and cluster centers," *Data Anal. Knowl. Discovery*, vol. 2, no. 3, pp. 39–48, 2018.
[19] Y. Wang, W. Pang, and J. Zhou, "An improved density peak clustering algorithm guided by pseudo labels," *Knowl.-Based Syst.*, vol. 252, Sep. 2022, Art. no. 109374.
[20] L. Sun, X. Qin, W. Ding, and J. Xu, "Nearest neighbors-based adaptive density peaks clustering with optimized allocation strategy," *Neurocomputing*, vol. 473, pp. 159–181, Feb. 2022.
[21] D. Wang, S. Ding, and J. Zhong, "Research of density peaks clustering algorithm based on second-order K neighbors," *J. Frontiers Comput. Sci. Technol.*, vol. 15, no. 8, pp. 1490–1500, 2021.
[22] R. Wang and Q. Zhu, "Density peaks clustering based on local minimal spanning tree," *IEEE Access*, vol. 7, pp. 108438–108446, 2019.
[23] T. Du, S. Qu, and Q. Wang, "A data-driven parameter adaptive clustering algorithm based on density peak," *Complexity*, vol. 2018, pp. 1–14, Oct. 2018.
[24] R. Liu, H. Wang, and X. Yu, "Shared-nearest-neighbor-based clustering by fast search and find of density peaks," *Inf. Sci.*, vol. 450, pp. 200–226, Jun. 2018.
[25] L. Zhao, J. Wang, and H. Chen, "Density-peak clustering algorithm on decentralized and weighted clusters merging," *J. Frontiers Comput. Sci. Technol.*, vol. 16, no. 8, pp. 1910–1922, 2022.
[26] J. Ding, Z. Chen, X. He, and Y. Zhan, "Clustering by finding density peaks based on Chebyshev's inequality," in *Proc. 35th Chin. Control Conf. (CCC)*, Jul. 2016, pp. 7169–7172.
[27] Y. Wang and G. Zhang, "Automatically determine density of cluster center of peak algorithm," *Comput. Eng. Appl.*, vol. 54, no. 8, pp. 137–142, 2018.
[28] D. Yu, G. Liu, M. Guo, X. Liu, and S. Yao, "Density peaks clustering based on weighted local density sequence and nearest neighbor assignment," *IEEE Access*, vol. 7, pp. 34301–34317, 2019.
[29] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2017, pp. 3–18.
[30] C. Dwork, "Differential privacy," in *Proc. 33rd Int. Colloq. Automata, Lang. Program.* Venice, Italy: Springer, 2006, pp. 1–12.
[31] Y. Chen, Y. Du, and X. Cao, "Density peak clustering algorithm based on differential privacy preserving," in *Proc. Int. Conf. Sci. Cyber Secur.* Cham, Switzerland: Springer, 2019, pp. 20–32.
[32] L. Sun, S. Bao, S. Ci, X. Zheng, L. Guo, and Y. Luo, "Differential privacy-preserving density peaks clustering based on shared near neighbors similarity," *IEEE Access*, vol. 7, pp. 89427–89440, 2019.
[33] F. McSherry and K. Talwar, "Mechanism design via differential privacy," in *Proc. 48th Annu. IEEE Symp. Found. Comput. Sci. (FOCS)*, Oct. 2007, pp. 94–103.
[34] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Found. Trends Theor. Comput. Sci.*, vol. 9, nos. 3–4, pp. 211–407, 2014.
[35] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Commun. Stat., Theory Methods*, vol. 3, no. 1, pp. 1–27, 1974.
[36] A. Fahad, N. Alshatri, Z. Tari, A. Alamri, I. Khalil, A. Y. Zomaya, and S. Foufou, "A survey of clustering algorithms for big data: Taxonomy and empirical analysis," *IEEE Trans. Emerg. Topics Comput.*, vol. 2, no. 3, pp. 267–279, Sep. 2014.
[37] E. B. Fowlkes and C. L. Mallows, "A method for comparing two hierarchical clusterings," *J. Amer. Statist. Assoc.*, vol. 78, no. 383, pp. 553–569, 1983.
[38] S. Liu, H. Che, K. Smith, and T. Chang, "Contaminant classification using cosine distances based on multiple conventional sensors," *Environ. Sci., Processes Impacts*, vol. 17, no. 2, pp. 343–350, 2015.

HUA CHEN received the Ph.D. degree in applied mathematics from the School of Mathematics and Statistics, Wuhan University, in 2012. She has been an Associate Professor at the School of Science, Hubei University of Technology, since 2015. Her research interests include machine learning, information security, and cryptography.

YUAN ZHOU received the B.E. degree in applied statistics from the School of Science, Hubei University of Technology, in 2020, where she is currently pursuing the master's degree in applied statistics with the School of Science. Her research interests include data mining and privacy protection.

KEHUI MEI received the B.E. degree in mathematics and applied mathematics from the School of Mathematics and Computer Science, Jianghan University, in 2020. He is currently pursuing the master's degree in applied statistics with the School of Science, Hubei University of Technology. His research interests include data mining and privacy protection.

NAN WANG received the B.E. degree from the School of Science, Hubei University of Technology, in 2020. She is currently pursuing the master's degree. Her research interests include data mining and machine learning.

GUANGXING CAI has been a Professor with the Hubei University of Technology for many years. His research interests include mathematics, information and coding, and cryptography.