International Journal Of Engineering Research And Development
e- ISSN: 2278-067X, p-ISSN: 2278-800X, www.ijerd.com
Volume 21, Issue 2 (February 2025), PP 86-97

Differential Privacy-Preserving Fuzzy-C-Means Clustering Algorithm

Junxiang YANG[1], Xueping ZHANG[1], and Gazi Mohammad Ismail[1]
[1]School of Information Science and Engineering, Henan University of Technology, Zhengzhou City, Henan Province 450052, China

Abstract**：***Aiming to address the privacy leakage problem of the traditional Fuzzy-C-Means (FCM) clustering algorithm, the paper introduces a differential privacy mechanism into the FCM clustering algorithm. It proposes and designs a DP-FCM (Differential Privacy-FCM) clustering algorithm oriented towards differential privacy protection. The algorithm protects individual privacy information by perturbing the data. Simultaneously, to maintain high clustering performance, a balance between privacy protection and data utility is achieved by optimizing the selection of perturbation parameters and the allocation of privacy budget. Comparative experimental results show that the DP-FCM clustering algorithm can effectively discover the clustering structure in the dataset while protecting individual privacy. Compared with the traditional FCM algorithm, the DP-FCM algorithm demonstrates clear advantages in clustering accuracy and stability. In addition, the paper also explores the impact of the differential privacy budget on the performance of clustering algorithms. The results indicate that smaller privacy budgets enhance privacy protection but may also affect clustering performance to some extent.*

Key words**：***Privacy Preservation, Differential Privacy, FCM Clustering Algorithm, DP-FCM Clustering Algorithm.*

Date of Submission: 04-02-2025 Date of acceptance: 14-02-2025

I. Introduction

With the advent of the big data era, data clustering analysis plays an important role in various fields. Clustering algorithms aim to group similar data points into categories, revealing the inherent structure and patterns in the data. However, with the extensive collection and application of personal data, the issue of data privacy protection has become increasingly prominent. In many practical scenarios, the leakage of sensitive personal information can result in serious privacy violations and social risks. Therefore, determining how to effectively protect individual privacy during data clustering has become an important research task. Therefore, the research motivation of this paper is to introduce the concept of Differential Privacy (DP) while maintaining the accuracy and effectiveness of the clustering algorithm and to propose a Fuzzy-C-Means clustering algorithm for differential privacy protection. Differential Privacy, as a privacy-preserving technique, can provide a higher level of privacy protection for individuals by meaningfully analyzing and mining data while protecting their privacy.

Traditional privacy-preserving techniques in the field of data mining encounter challenges such as ineffective responses to background knowledge attacks and low usability. Differential Privacy (DP) is a privacy protection framework that offers a rigorous mathematical definition and quantifies the level of privacy protection. It is effective against various privacy attacks and inference methods.

Blum et al.[2] applied differential privacy to the k-means algorithm and introduced the DP-kmeans algorithm. However, their algorithm is not considered reliable due to noise. Li et al.[3] proposed the IDP k-means algorithm to ensure differential privacy, which relies on an initial center selection method. Nevertheless, the clustering efficiency and accuracy of their method are not sufficiently high. Song et al.[4] took a different approach from the traditional k-anonymous method by incorporating noise and randomization to address the limitation that at least k elements in a k-anonymous dataset must share the same quasi-identifier. However, their method is not suitable for preserving the privacy of numerical attributes with large ranges, and it results in slightly higher information loss compared to the traditional method. Yang et al.[5] proposed a clustering method based on Laplacian noise by incorporating the adaptive lattice method to preserve differential privacy. Zheng et al.[6] introduced a new approach. Privacy-preserving data sharing framework enables data sharers to share data on demand, utilizing differential privacy to ensure privacy preservation. Ping Xiong et al.[7] extensively explored the application of differential privacy in data publishing and data mining in a review paper. Paul Huang et al.[8] proposed the BDPK-means clustering algorithm by enhancing the K-means algorithm with a novel method for selecting appropriate initial centroids.

The research focus of this paper includes the following questions:

1. How to design a Fuzzy-C-Means clustering algorithm that can provide differential privacy protection while maintaining the accuracy and effectiveness of the clustering algorithm?
2. How to determine the privacy budget for differential privacy to control the risk of privacy leakage and the quality of clustering results?
3. How to handle data ambiguity effectively in differential privacy-preserving Fuzzy-C-Means clustering algorithms and improve the accuracy and stability of clustering results?
4. How to evaluate and quantify the privacy-preserving strength and clustering effect of differential privacy-preserving Fuzzy-C-Means clustering algorithms, and analyze them in comparison with traditional non-privacy-preserving algorithms?

Based on the previous research, this paper proposes a novel differential privacy-preserving Fuzzy-C-Means oriented (DP-FCM) clustering algorithm, whose innovations and differences with previous research are mainly reflected in the following aspects:

1. Dual optimization of privacy protection and clustering performance: the DP-FCM algorithm proposed in this paper not only focuses on privacy protection, but also pays attention to the optimization of clustering performance. Through the well-designed perturbation mechanism and privacy budget allocation strategy, the algorithm maintains the accuracy and stability of the clustering results as much as possible while protecting the data privacy, which is not yet common among the existing differential privacy clustering algorithms.
2. Dynamic privacy budget allocation strategy: unlike the practice of fixing the privacy budget in previous studies, this paper proposes a method of dynamically adjusting the privacy budget according to the characteristics of the dataset and the needs of the clustering task, in order to achieve the best balance between privacy protection and clustering performance.
3. Adaptive adjustment of fuzzy parameters: in the Fuzzy-C-Means clustering algorithm, the choice of fuzzy parameter m has an important impact on the clustering results. In this paper, a fuzzy parameter adaptive adjustment mechanism based on data distribution is proposed, which enables the algorithm to automatically adjust the fuzzy parameters according to the characteristics of the data, and improves the adaptability and accuracy of the clustering algorithm.
4. Comprehensive experimental evaluation and analysis: this paper comprehensively evaluates the performance of the DP-FCM algorithm through experiments on several real data sets. It not only compares the traditional FCM algorithm and other differential privacy clustering algorithms, but also deeply analyzes the influence of privacy budget, fuzzy parameters, etc. on the clustering performance, which provides a theoretical basis and practical guidance for the parameter selection of the algorithm.
5. Combination of theoretical analysis and empirical study: this paper not only theoretically proves that the DP-FCM algorithm satisfies ε-differential privacy protection, but also verifies the clustering performance of the algorithm under different levels of privacy protection through empirical studies, which provides a new perspective for the theoretical analysis and practical application of differential privacy clustering algorithms.

In summary, the research in this paper not only enriches the connotation of differential privacy clustering algorithm theoretically, but also provides an effective solution for privacy-preserving clustering analysis in practice. Through this study, we expect to promote the development of privacy-preserving clustering algorithms and provide references and insights for research in related fields.

II. Definitions and Rationale

2.1 Differential privacy

The concept of differential privacy was first introduced by Dwork[1] in 2006. Its main goal is to address the problem of privacy leakage in statistical databases. Compared to the traditional approach, differential privacy offers a new concept of privacy where the outcomes of querying a database remain unaffected by alterations in individual records within the dataset. This definition requires protecting data privacy while still producing meaningful statistical results. From the perspective of privacy protection, individual users are considered the subjects of privacy, while specific attributes of a group of users are not deemed private. However, when aggregated information is released, there can be a risk of individual privacy being compromised. For instance, if a query for 100 patients in a hospital reveals 10 HIV-infected patients, and a query for 99 patients reveals 9 HIV-infected patients, it can be inferred that one person remains infected with HIV. This breach of privacy is known as a differential attack.

Differential privacy is implemented by adding random noise to the data. This process changes the query results from two specific values to two random variables that follow similar probability distributions, ensuring privacy at the individual level. In short, the differential privacy mechanism ensures that each individual in the dataset is not disclosed, but the outside world still has access to the statistical information of the dataset, such as the mean, variance, and other relevant data. Differential privacy is a concept that lacks a specific implementation and does not prescribe a particular perturbation method. The added noise can theoretically follow any distribution. The objective of the research is to optimize the accessibility of a dataset while ensuring the fulfillment of confidentiality through differential privacy.

Definition 1（Neighboring datasets）Suppose there are two datasets with the same attribute structure D and D', Their symmetry difference is denoted as $D\Delta D'$, $|D\Delta D'|$ is $D\Delta D'$ the number of records in the. If $|D\Delta D'|$ =1，then D and D' are said to be neighboring datasets**Error! Reference source not found.**.

Definition 2（ε-differential privacy）randomized algorithm M, P_M is M the set of all possible outputs. The input to a randomized algorithm M is any two neighboring data sets D and D' and the output is any subset S_M of P_M .If the algorithm satisfies

$$Pr[M(D) \in S_M] \leqslant e^{\varepsilon} \times Pr[M(D') \in S_M] \quad (1)$$

Then the algorithm M is said to provide ε-differential privacy protection, where the parameter ε is called the privacy-preserving budget in differential privacy[10].

Definition 3（global sensitivity）Having a function $f{:}D \to R^d$,the input is the data set, the output is a d -dimensional vector of real numbers. For two arbitrary neighboring data sets D and D' ,the global sensitivity [11] is shown in Eq(2).

$$\Delta f = \max_{D,D'} ||f(D) - f(D')|| \quad (2)$$

Global sensitivity is used to measure the maximum difference between the outputs of a function in a given dataset for any two neighboring datasets. It can be utilized to safeguard individuals' privacy by determining the level of noise added to the differential privacy mechanism. By limiting the global sensitivity, the risk of privacy leakage can be controlled to ensure the protection of individual sensitive information.

Definition 4（Laplace mechanism）With query function $f{:}D \to R^d$,its sensitivity is Δf, randomized algorithm [12] $M(D) = f(D) + Y$ Provides ε-differential privacy protection, where $Y = Lap\left(\frac{\Delta f}{\varepsilon}\right)$ is a random noise that follows a Laplace distribution with parameter $\frac{\Delta f}{\varepsilon}$. The magnitude of the ε value has an impact on the strength of privacy protection. When the value of ε is small, the probability density of the noise is average and the amount of added noise is large, thus providing stronger protection of data privacy. Conversely, when the value of ε is large, the probability density of the noise is uneven, the amount of added noise is small, and the strength of privacy protection is correspondingly weaker. The probability density function of the Laplace distribution is shown in equation (3).

$$p(x) = \frac{1}{2\lambda}\exp\left(-\frac{|x|}{\lambda}\right) \quad (3)$$

From Definition 4, it is clear that the selection of the privacy budget parameter is crucial when incorporating Laplace noise. A decrease in the privacy budget results in an increase in the amount of noise added, thereby enhancing privacy protection. And vice versa. In order to strike a balance between privacy protection and data utility, the privacy budget needs to be set appropriately when using differential privacy methods. This ensures that the privacy protection needs are met while maintaining the effectiveness of data analysis. The relationship between them can be seen from the Laplace distributions with different parameters (as shown in Fig. 1).

The Laplace distribution is a bimodal distribution with a probability density function that has a sharp peak at the mean and exponential decay on both sides. Graphically, it resembles a bell curve (e.g., Figure 1), but the Laplace distribution has heavier tails compared to the Gaussian distribution. The exact shape depends on the mean and scale parameters of the distribution.

In the Laplace mechanism, a smaller lambda parameter results in a smaller added noise. The scale parameter $\frac{\Delta f}{\varepsilon}$ of the Laplace distribution is proportional to the inverse of lambda, so the smaller lambda is, the larger the scale parameter $\frac{\Delta f}{\varepsilon}$ is, and the larger the magnitude of the noise. Specifically, when lambda is smaller, the scale parameter $\frac{\Delta f}{\varepsilon}$ of the Laplace distribution is larger and the magnitude of the noise increases. This means that the magnitude of the noise that perturbs the original data increases during the privacy preservation process. A larger magnitude of noise results in a more significant perturbation of the data, thus providing stronger privacy protection.

The application of Laplace distribution in privacy preservation usually involves the concept of Differential Privacy (DP). In the Laplace mechanism, the addition of noise is based on the Laplace distribution. Specifically, suppose there is a dataset for which we wish to privacy protect individual data points and protect the privacy of the overall data distribution. Using the Laplace mechanism, we can add a noise from the Laplace distribution for each original data point. The scale parameter of the noise is controlled by the Privacy Budget, with larger scale parameters indicating stricter privacy protection. By adding Laplace noise, the specific values of individual data points are perturbed, thus protecting individual privacy. At the same time, because the Laplace distribution is characterized by spiking and rapid decay, the effect of the noise is mainly concentrated around the mean and has less impact on the tails of the distribution. This means that the characteristics of the overall data distribution are still retained, and only a certain degree of randomness is introduced at individual data points. The trade-off between privacy protection and data availability can be balanced by controlling the scale parameter of the Laplace noise. A larger scale parameter introduces stronger privacy protection but may lead to distortion of the data or reduced data availability. Therefore, choosing an appropriate scale parameter is an important consideration for privacy preservation in differential privacy.

In practice, differential privacy also has two key properties that satisfy the requirements of the algorithm given the budget parameter ε.

characteristic 1（serial nature）For a dataset D and a randomized algorithm $M_1(D), M_2(D), \ldots, M_n(D)$ satisfying D ,these randomized algorithms combine to satisfy $\left(\sum_{i=1}^{n} \varepsilon_i\right)$-differential privacy[9]。

characteristic 2（parallelism）There are disjoint datasets $D_1, D_2, \ldots, D_n$,randomized algorithms $M_1(D_1), M_2(D_2), \ldots, M_n(D_n)$ with privacy budgets $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$,respectively, and the combined algorithm $M(M_1(D_1), M_2(D_2), \ldots, M_n(D_n))$ of these randomized algorithms satisfies $(\max \varepsilon_i)$- differential privacy[9].

Differential Privacy Mechanism In contrast to other privacy protection mechanisms, differential privacy focuses on protecting the privacy of an individual rather than just anonymizing or desensitizing the data. It provides mathematically provable privacy protection that protects the privacy of individual data even when the attacker has background knowledge. Adjustable strength of differential privacy protection Differential privacy provides a parameter ε (epsilon) for the strength of privacy protection, which can be adjusted as required. A smaller value of ε indicates stronger privacy protection but may result in lower data quality; a larger value of ε indicates weaker privacy protection but can provide higher data quality. The differential privacy mechanism is forward privacy in nature, i.e., it protects the privacy of the data by adding random noise and ensures that individual data cannot be reduced or reconstructed from the original data. This protection is implemented before the release of the data, rather than a subsequent restoration of the released data. In contrast to some traditional privacy-preserving methods, the differential privacy mechanism maintains the availability of useful statistical queries on the data to a certain extent. It allows for some degree of statistical analysis and data mining of the data while protecting individual privacy.

Differential privacy highlights several performance metrics compared to traditional privacy protection methods. Privacy Protection Strength: By adjusting the ε parameter, the privacy protection strength of the differential privacy mechanism can be controlled. A smaller value of ε indicates stronger privacy protection and a larger value of ε indicates weaker privacy protection. Data Quality: Measures the extent to which the differential privacy mechanism affects the quality of the data while protecting privacy. A smaller value of ε may lead to a larger loss of data quality, requiring a trade-off between privacy protection and data availability. Forward privacy protection: the differential privacy mechanism adds noise to the data before it is released to protect the privacy of individual data and ensure that the data cannot be reduced or reconstructed from the original data. Probability of Differential Privacy Protection: Differential privacy provides mathematically provable privacy protection, ensuring that the privacy of individual data remains protected despite the attacker's background knowledge. Data availability: a measure of how much the differential privacy mechanism affects the availability of data while protecting privacy. There is a need to balance privacy protection with data availability to ensure that the data still has some statistical analysis and data mining capabilities. These performance metrics can be used to evaluate the performance and benefits of differential privacy mechanisms in privacy - preserving tasks.

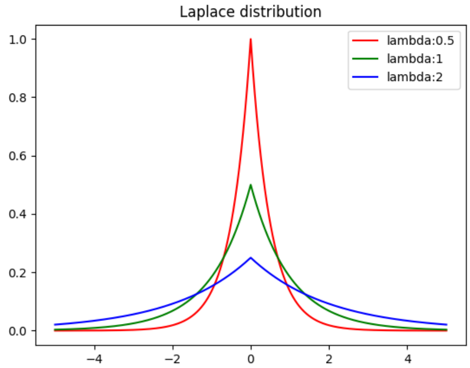

Fig.1 Laplace probability density function

2.2 FCM clustering

The Fuzzy-C-Means (FCM) algorithm is a classical soft clustering algorithm. Unlike hard clustering algorithms, the FCM algorithm does not strictly classify samples into a particular cluster but allows samples to belong to multiple clusters with a certain degree of affiliation. In the FCM algorithm, the degree of affiliation of each sample assigned to each cluster is a value between 0 and 1, indicating how similar the sample is to each cluster. It is defined as follows:

Given a data set containing n data points $X=[x_1,x_2,...,x_n]$, the FCM algorithm aims to classify these data points into c fuzzy clusters. In the FCM clustering algorithm, each data point x_i is assigned an affiliation degree belonging to each cluster, indicating the probability that the data point belongs to each cluster. Also, each cluster is represented by a center vector.

The goal of the algorithm is to minimize the following objective function:

$$J_m(U,V)=\sum_{i=1}^{n}\sum_{j=1}^{c}u_{ij}^{m}\cdot||x_i-v_j||^2 \qquad (4)$$

Where: u_{ij} denotes the affiliation of data point x_i to the j th cluster; v_j denotes the center vector of the j th cluster; m is a fuzzy parameter, usually taken as a real number greater than or equal to 1, which is used to control the degree of fuzzy clustering.

The steps of the FCM algorithm are as follows:

Step 1. Initialize the affiliation of each data point to each cluster.

Step 2. Calculate the center vector of each cluster based on the current affiliation degree.

Step 3. Update the affiliation of each data point based on the center vector.

Step 4. Repeat steps 2 and 3 until the degree of affiliation no longer changes significantly or the maximum number of iterations is reached.

In each iteration, by adjusting the affiliation and clustering center vectors of the data points, the FCM algorithm tries to minimize the objective function to obtain fuzzy clustering results.

This soft clustering method can better reflect the fuzzy attribution of samples in different clusters rather than forcing them into a specific cluster. This soft clustering method is advantageous in dealing with the presence of fuzzy attribution relationships in a dataset, providing richer and more flexible clustering results.

III. The Importance of Privacy Protection for Data Clusters

Data clustering is the process of dividing a data set into groups or clusters with similar characteristics. In the field of data mining and machine learning, clustering algorithms are a commonly used data analysis technique that can help us discover patterns and structures in data and provide insights about data clusters. However, personal privacy protection has become increasingly important in modern society. Leakage of personal data may lead to problems such as exposure of personal identity, information misuse, targeted advertising or marketing, and may even lead to personal credit risk or identity theft. Therefore, protecting personal privacy becomes a key task in the data clustering process.

In data clustering analysis, privacy protection refers to ensuring the protection of individual sensitive information from unauthorized access and disclosure by adopting a series of techniques and measures. The importance of privacy protection for data clustering is reflected in the following aspects:

1. Individual privacy protection: in the process of analyzing data clusters, individual data may contain sensitive information, such as personal identity and health information. Privacy protection measures can ensure that the privacy of these individual data is adequately protected to avoid privacy leakage and abuse.

2. Compliance requirements: as data privacy regulations and standards continue to improve, the requirements for privacy protection during data processing are becoming more and more stringent. In data cluster analytics, it is crucial to ensure that privacy protection meets regulatory requirements.

3. Trust and cooperation: For data holders and parties involved in data cluster analysis, privacy protection is the basis for establishing a trust relationship. The cooperative relationship between data holders and data analysts can be enhanced through effective privacy protection measures.

4. Brand reputation: For companies and organizations, protecting user data privacy is key to maintaining brand reputation and user trust. Failure to effectively protect data privacy in data cluster analysis may damage the reputation and credibility of the organization.

The concept of differential privacy is introduced in the paper to protect the privacy of individuals in data clusters. Differential privacy is a privacy-preserving technique designed to meaningfully analyze and mine personal data while protecting individual privacy. Differential privacy hides personal information by adding noise or perturbation, making it impossible for an attacker to infer an individual's sensitive information from the clustering results. Applying differential privacy-preserving Fuzzy-C-Means clustering algorithms in data clusters ensures that the risk of privacy leakage during the clustering process is minimized. This means that even during the clustering process, an attacker cannot accurately determine which cluster or which individual's sensitive information a data point belongs to.

By protecting the privacy of individuals in data clusters, we can ensure that the results of data analysis and mining do not compromise the privacy interests of individuals. This is especially important for data clustering applications in sensitive domains (e.g., healthcare, finance, etc.), where data usually contains a large amount of sensitive information about individuals. Therefore, in the paper, privacy preservation is very important for data clustering, which ensures the confidentiality of individuals' privacy and at the same time provides a reliable solution for data analysis and mining.

IV. Differential Privacy DP-FCM Algorithm

4.1 DP-FCM algorithm

The article proposes Differential Privacy DP-FCM clustering algorithm which uses affiliation to represent the relationship between each data and also the algorithm is an objective function based algorithm. Given a dataset containing n data: $X=\{x_1,x_2,...,x_i,...,x_n\}$, X_i is the i th feature vector , and X_{ij} is the j th attribute of X_i. Each data sample contains d attributes, and the algorithm divides the dataset into C classes, C being a positive integer greater than 1, where the clustering centers of the C classes are $[v_1,v_2,...,v_n]$, respectively.

In the framework of differential privacy protection, privacy is achieved by adding noise to the cluster centroids. The original cluster centers are perturbed by adding noise vectors that satisfy the requirements of differential privacy to safeguard the privacy of the cluster centers.

The DP-FCM algorithm flow is specified as follows:

Inputs: sample dataset D, number of class clusters C, privacy budget ε, iteration threshold Max_iter, fuzzy parameter m (m>1).

Output: clustering results after perturbation.

1. The dataset D is normalized to map the range of values of each feature to between [0,1].
2. Initialize the affiliation matrix using random values in the range (0,1) $U=\{u(ij)\}$, and to satisfy the constraints $\sum_{j=1}^{k} u_{ij}=1, u_{ij}\in[0,1]$, where u_{ij} denotes the degree of affiliation of the data point x_i belonging to the clustering center v_j.
3. According to the affiliation matrix U ,Use equation (5) to calculate the clustering center v_j ,where m (m>1) denotes the fuzzy parameter.

$$v_j=\frac{\sum_{i=1}^{n} u_{ij}^m x_i}{\sum_{i=1}^{n} u_{ij}^m} \tag{5}$$

4. Cluster centers are added to Laplace noise, and for each cluster center v_j, a random noise vector y that obeys the Laplace distribution (Definition 5) is chosen to yield the perturbed cluster center $vj'=vj+y$.
5. The objective function is computed using equation (6), where d_{ij} is the distance between the sample point x_i and the clustering center v_j. The Euclidean distance $d_{ij}=||x_i-v_j||$ is used.

$$J(U,V) = \sum_{i=1}^{n} \sum_{j=1}^{k} u_{ij}^{m} d_{ij}^{2} \qquad (6)$$

6. Update the affiliation matrix and for each data point x_i ,compute the new affiliation matrix $U = \{u(ij)\}$,where

$$u_{ij} = \frac{1}{\sum_{c=1}^{k} \left(\frac{d_{ij}}{d_{ik}}\right)^{\frac{2}{m-1}}} \qquad (7)$$

7. Determine whether the number of iterations is greater than the threshold: if yes, go to step 8, otherwise go to step 3.
8. Based on the latest affiliation matrix, the data points are assigned to the class clusters with maximum affiliation to get the final clustering results.

V. Confirmation of privacy

Let the DP-FCM algorithm satisfy ε-differential privacy (Eq. 1),D and $D^{'}$ are neighboring datasets (Definition 1),$M(D)$ and $M(D^{'})$ are the clustering results of the algorithm for D and $D^{'}$, and P represents any one of the clustering results. From Definition 2:

$$Pr[M(D) = P] \leqslant e^{\varepsilon} \times Pr[M(D^{'}) = P] (8)$$

Assuming that the function f is to return true information at the centroid of the dataset, the randomized algorithm M is

$$M(D) = f(D) + Y, Y = Lap\left(\frac{\Delta f}{\varepsilon}\right) (9)$$

For the neighboring datasets D and $D^{'}$, the centroid returned by the function f is assumed to be an n -dimensional vector, i.e.

$$f(D) = (x_1, x_2, ..., x_n)^T \quad (10)$$

$$f(D^{'}) = (x_1^{'}, x_2^{'}, ..., x_n^{'})^T \quad (11)$$

$$f(D^{'}) = (x_1 + \Delta x_1, x_2 + \Delta x_2, ..., x_n + \Delta x_n)^T (12)$$

The sensitivity Δf is

$$\Delta f = \max\left(\sum_{i=1}^{n} |x_i - x_i^{'}|\right) = \max\left(\sum_{i=1}^{n} |\Delta x_i|\right) \qquad (13)$$

Let the output vector p be

$$P = (y_1, y_2, ..., y_n)^T \qquad (14)$$

For $M(D)$ and $M(D^{'})$ there are

$$Pr[M(D) = P] = \prod_{i=1}^{n} \frac{\varepsilon}{2\Delta f} e^{-\frac{\varepsilon}{\Delta f}|x_i - y_i|} (15)$$

$$Pr[M(D^{'}) = P] = \prod_{i=1}^{n} \frac{\varepsilon}{2\Delta f} e^{-\frac{\varepsilon}{\Delta f}|x_i + \Delta x_i - y_i|} (16)$$

$$\frac{Pr[M(D) = P]}{Pr[M(D^{'}) = P]} = e^{\frac{\varepsilon}{\Delta f}\sum_{i=1}^{n}(|x_i + \Delta x_i - y_i| - |x_i - y_i|)} \qquad (17)$$

due to

$$\sum_{i=1}^{n} ||x_i + \Delta x_i - y_i| - |x_i - y_i|| \leqslant \sum_{i=1}^{n} |\Delta x_i| \leqslant \Delta f (18)$$

obtainable

$$\frac{Pr[M(D) = P]}{Pr[M(D^{'}) = P]} = e^{\frac{\varepsilon}{\Delta f}\sum_{i=1}^{n}(|x_i + \Delta x_i - y_i| - |x_i - y_i|)} \leqslant e^{\varepsilon} (19)$$

From the above, the DP-FCM algorithm satisfies ε-differential privacy (Definition 2).

VI. Experimental results and analysis

6.1 Experimental setting and data

This study uses the Python language to conduct simulation experiments on the DP-FCM algorithm in a Windows 11 AMD Ryzen 7 5800H with Radeon Graphics 3.20 GHz, 16.0 GB of RAM, Python language version 3.8, and PyCharm 2021.3.1 as the development tool.

In this study, three representative datasets were carefully selected to evaluate the performance of the DP-FCM algorithm. Each dataset was chosen based on its relevance in real-world applications, the diversity of the data, and the challenging nature of privacy preservation. A detailed description of each dataset and the reasons for its selection are given below:

1. Iris dataset: The Iris dataset is one of the most well-known datasets in the field of machine learning and contains 150 samples divided into 3 classes with 4 features per sample. This dataset was selected for its simple structure, ease of understanding and wide availability. In terms of privacy preservation, although the Iris dataset does not contain obvious personally identifiable information, we included it in our study to demonstrate the clustering performance of the DP-FCM algorithm when dealing with non-sensitive data and its potential application in privacy preservation.

2. Breast Cancer Wisconsin dataset: this dataset contains 569 samples with 30 real-valued attributes each for breast cancer diagnosis. This dataset was chosen because of its practical application value in the medical field and the high demand for privacy protection. Medical data usually contains sensitive personal health information, so the need for privacy protection is particularly acute. In this study, we are challenged to effectively diagnose and analyze diseases without revealing patients' identities.

3. Aggregation dataset: this is a synthetic dataset containing 788 samples and 2 features organized into 7 categories. This dataset is characterized by insignificant boundaries between the categories, thus making it more challenging for clustering algorithms. We chose this dataset to test the performance of the DP-FCM algorithm when dealing with complex and ambiguous data boundaries. In addition, the use of synthetic datasets allows us to control the data generation process in order to better understand the behavior of the algorithm under different levels of privacy protection.

The privacy challenges posed by each dataset include, among others, the protection of personally identifiable information, especially in medical datasets such as Breast Cancer Wisconsin, where leakage of personal health information can lead to serious privacy violations. Sensitivity of data, certain datasets may contain sensitive information, such as disease diagnosis results, which, if leaked, may adversely affect individuals. Ambiguity of data, in clustering analysis, the ambiguity of data may lead to difficulties for privacy-preserving algorithms to accurately distinguish between individuals, thus affecting the accuracy of clustering results.

With the selection of these datasets, we aim to demonstrate the potential of the DP-FCM algorithm for application under different privacy challenges and to validate its ability to maintain efficient clustering performance while protecting individuals' private information. The combined use of these datasets provides us with a comprehensive testbed to evaluate and validate the practical effectiveness of the algorithm in diverse scenarios.

The sample datasets used in this experiment are from the artificial dataset and the UCI Knowledge Discovery Archive database.

The specific information of the datasets used in this experiment is shown in Table 1.

Table 1 Experimental data information

Data sets	Tuples	Dims	Type	nick-name
Iris	151	4	Real	D1
Breast Cancer Wisconsin	569	30	Real	D2
Aggregation	788	2	Real	D3

VII. Experimental evaluation indicators

The F-measure is a metric used to evaluate the performance of a classification or clustering algorithm by combining Accuracy (AC) and Recall (RE). It provides a comprehensive assessment of classification or clustering results and is particularly useful for unbalanced datasets or tasks where both precision and completeness are important. F-measure combines accuracy and recall, providing a comprehensive performance metric by balancing precision and completeness. In unbalanced datasets, relying solely on accuracy or recall may yield misleading results. The F-measure, on the other hand, strikes a balance between the two metrics and is a commonly utilized evaluation criterion. This research algorithm impacts the accuracy of the clustering results following perturbation, and it is particularly crucial to ensure that the clustering results are reliable. So, it was decided to use the F-measure value to assess the effectiveness of the clustering results. The specific definition is shown in Equation (20).

$$F = \frac{2AC \times RE}{AC + RE} \tag{20}$$

The F-measure value falls within the range of [0,1], where 1 signifies the best performance and 0 indicates the worst performance. A higher value indicates better clustering and increased usability.

VIII. Experimental parameter setting

The experiments are conducted using the DP-FCM algorithm proposed in this study in comparison with the DPK-means algorithm[2], the DP-rcCFSFDP algorithm[17], and the IDP K-means algorithm[3] on the dataset.

Extensive pre-experiments and literature review are conducted in determining the privacy budget (ε), the fuzzy parameter (m) and the number of iterations (Max-iter). The privacy budget ε is a key parameter in differential privacy that controls the strength of privacy protection. We chose a range of ε values, from 0.05 to 1, to observe the effect of the level of privacy protection on the clustering performance. The fuzzy parameter m is the core of the FCM algorithm, which determines the degree of fuzzy affiliation. We chose m=2, a value commonly used in FCM algorithms to balance the fuzziness and clarity of clustering, based on the recommendations of previous studies and the results of the pre-experiments. The number of iterations Max-iter is the key to the convergence of the algorithm, and we set a large number of iterations (e.g., 1000) to ensure that the algorithm has enough time to converge to the optimal solution, and a small iteration threshold (e.g., 0.005) to avoid over-iteration.

In order to verify the reasonableness of the parameter choices, we performed a sensitivity analysis. This involves varying the value of the privacy budget ε and observing its effect on the clustering performance. We found that smaller values of ε, while providing stronger privacy protection, may also have a negative impact on the clustering performance. Therefore, we choose a compromise value of ε to strike a balance between privacy protection and clustering performance.

To ensure the reliability of the results, we conducted several experiments for each parameter setting and took the average value as the final result. This helps to minimize the effect of random noise on the experimental results and provides a robust estimate of the performance of the algorithms.

The variation of the F-measure value versus the privacy budget ε value for these four algorithms across the three different datasets is recorded, and the results are shown in Figure 2-Figure 4.

IX. Analysis of experimental results

The DP-FCM algorithm proposed in this study aims to achieve efficient clustering performance while protecting data privacy. With the experimental results shown in Fig. 2, Fig. 3 and Fig. 4, we can see that the DP-FCM algorithm outperforms the existing DPK-means, DP-rcCFSFDP and IDP K-means algorithms in terms of F-measure values on different datasets. These results show that the DP-FCM algorithm strikes a better balance between privacy preservation and clustering accuracy. A full explanation and discussion of these results is given below:

1. Privacy-preserving mechanism of the algorithm: the DP-FCM algorithm achieves differential privacy preservation by introducing Laplace noise in the computation of the clustering centers. This mechanism ensures that the output of the algorithm does not change significantly even if a single data point is removed or added to the dataset, thus protecting the privacy of individual data. The effectiveness of this protection mechanism is demonstrated in Fig. 2, Fig. 3 and Fig. 4, where the algorithm maintains a high clustering performance even with a low privacy budget.
2. Adaptive tuning of fuzzy parameters: the fuzzy parameter m in the DP-FCM algorithm is adaptively tuned, which enables the algorithm to automatically adjust the degree of fuzzy affiliation according to the characteristics of the data. This adaptive adjustment mechanism improves the adaptability of the algorithm to different data distributions, thus improving the accuracy of clustering results while maintaining privacy protection.
3. Reasonable allocation of privacy budget: experimental results show that by reasonably allocating the privacy budget, the DP-FCM algorithm is able to maintain stable clustering performance under different privacy protection levels. In Fig. 2, Fig. 3 and Fig. 4, the F-measure value gradually increases with the increase of the privacy budget ε, which indicates that the algorithm is able to increase the strength of privacy protection while reducing the impact on the clustering performance.
4. Stability of the algorithm: the DP-FCM algorithm shows good stability over many iterations. In the experiments, the algorithm obtains consistent clustering results in different runs, which shows that the algorithm is robust to changes in the initial conditions. This stability is crucial for privacy-preserving clustering analysis in practical applications.

5. Comparison with other algorithms: compared to existing differential privacy clustering algorithms, the DP-FCM algorithm provides higher clustering accuracy while maintaining privacy preservation. This may be due to the fact that the DP-FCM algorithm is designed with the optimization of clustering performance in mind, not just privacy preservation. In addition, the adaptive tuning mechanism and privacy budget allocation strategy in the algorithm also provide support for improving the clustering performance.

In summary, the DP-FCM algorithm outperforms other algorithms in the experiments mainly due to its unique privacy-preserving mechanism, its ability to adaptively adjust the fuzzy parameters, and its reasonable privacy budget allocation strategy. These features enable the algorithm to effectively protect individual privacy while maintaining high clustering performance when dealing with privacy-sensitive data. Future work will focus on further optimizing the computational efficiency of the algorithm, as well as extending the algorithm to handle larger datasets.

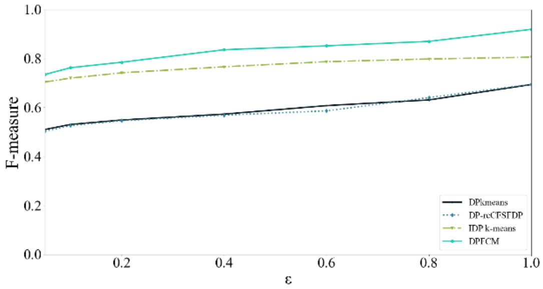

Fig.2 Comparison of clustering accuracy on dataset D1

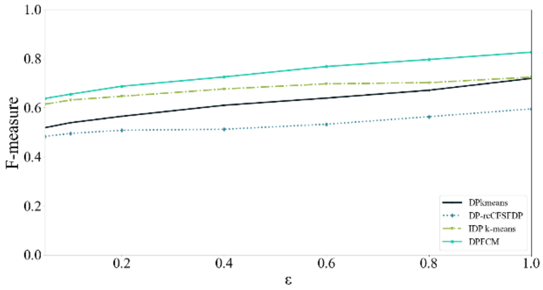

Fig.3 Comparison of clustering accuracy on dataset D2

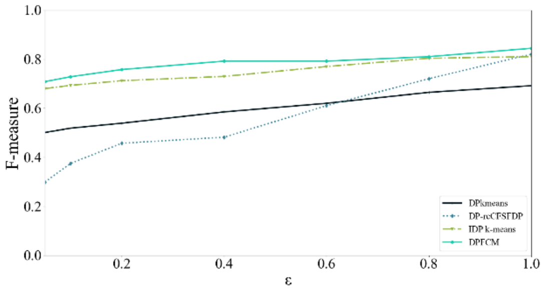

Fig.4 Comparison of clustering accuracy on dataset D3

X. Concluding remarks and outlook for the future

In order to solve the problem of low usability of traditional privacy protection techniques, the article proposes a differential privacy protection oriented DP-FCM algorithm based on differential privacy theory. The algorithm prevents the leakage of individual privacy information by introducing differential privacy mechanism and adding noise to perturb the data during the algorithm. The privacy protection level of differential privacy can be adjusted according to the privacy budget so as to find a balance between privacy protection and data utility. Also the algorithm shows good results in clustering performance. By combining the differential privacy mechanism with the traditional Fuzzy-C-Means clustering algorithm, the DP-FCM algorithm is able to effectively discover the clustering structure in the dataset while protecting privacy.

Differential privacy protects individual privacy by introducing random noise into the data processing process, where the privacy budget ε is a key parameter to control the amount of noise. a smaller value of ε indicates stronger privacy protection, but at the same time may have a greater impact on the accuracy and usability of the data. Therefore, finding an appropriate value of ε to strike a balance between protecting privacy and maintaining data accuracy is a central challenge in the design of differential privacy clustering algorithms.

In this study, we analyze in detail the impact of the privacy budget ε on the clustering performance of the DP-FCM algorithm through a series of experiments. The experimental results show that the F-measure value of clustering gradually increases as the value of ε increases, which indicates that the algorithm is able to obtain more accurate clustering results at a lower level of privacy protection. This phenomenon can be explained in two ways:

1. The effect of noise: when the value of ε is small, the algorithm adds a larger noise to the center of the clusters in order to provide a stronger privacy protection. This noise interferes with the clustering process and leads to an inaccurate estimation of the clustering centers, thus affecting the quality of the clustering results. On the contrary, when the value of ε is large, the added noise is smaller and the interference to the clustering center is reduced, thus the clustering performance is improved.
2. Data availability: the privacy budget ε not only affects the strength of privacy protection, but also determines the availability of data. A smaller value of ε means more random noise in the data, which may mask the true pattern of the data and reduce the accuracy of the clustering algorithm. Whereas a larger ε value allows the algorithm to be closer to the original data, thus maintaining high data availability and clustering accuracy.

In practice, choosing the ε value needs to consider the sensitivity of the data and the needs of the application scenario. For example, when dealing with highly sensitive medical data, it may be necessary to choose a smaller ε value to provide stronger privacy protection, even though this may sacrifice some clustering accuracy. In other scenarios with less stringent privacy requirements, a larger value of ε can be chosen for better clustering performance.

In conclusion, the choice of the privacy budget ε is a trade-off process that needs to be decided based on specific application requirements and data characteristics. This study experimentally verifies the performance of DP-FCM algorithm under different privacy budgets, which provides a valuable reference for the balance between privacy and accuracy in practical applications. Future work will further explore ways to adaptively adjust the privacy budget to achieve better privacy protection and clustering performance.

The DP-FCM algorithm proposed in this study achieves a certain balance between privacy protection and clustering performance, but there are still many issues that deserve further exploration. The following are a few potential directions for future research:

1. Algorithm performance on high-dimensional data: current research focuses on datasets with low and medium dimensions. Future work could explore the performance of the DP-FCM algorithm on high-dimensional data and how to optimize the algorithm to handle datasets with more features. High-dimensional data is often accompanied by dimensionality catastrophes, which may affect the accuracy of clustering and the scalability of the algorithm.
2. Algorithm performance on large-scale datasets: as the amount of data continues to grow, the performance and efficiency of algorithms on large-scale datasets becomes particularly important. Future research could focus on how to optimize the DP-FCM algorithm to handle large-scale datasets, including the use of distributed computing and optimizing the computational complexity of the algorithm.
3. Robustness analysis of the algorithm: in practical applications, data may contain noise and outliers. It is an important research direction to study the robustness of DP-FCM algorithms in the face of these challenges, and how to improve the algorithms to increase their resistance to noise and outliers.
4. Combination of Joint Learning and Differential Privacy: Joint learning is an emerging technique that allows multiple data sources to train models together without sharing the original data. Combining DP-FCM algorithms with joint learning may provide a more robust framework to improve model performance while preserving privacy.
5. Algorithm evaluation for cross-domain applications: while this study validated the performance of the DP-FCM algorithm on several datasets, future work could explore the application of the algorithm in different domains (e.g., finance, healthcare, social networking, etc.) and evaluate its effectiveness and applicability in these domains.
6. Interpretability of algorithms: interpretability is an important consideration in privacy preserving algorithms. Future research could focus on how to improve the interpretability of DP-FCM algorithms so that they can provide transparency to users and regulators while protecting privacy.
7. Real-time data processing: In certain application scenarios, such as financial market analysis and cyber security, real-time data processing is crucial. It is a challenging research direction to investigate how to adapt DP-

FCM algorithms to real-time data streams and provide fast and accurate clustering results while maintaining privacy protection.
8. Privacy protection issues in the field of deep clustering: with the wide application of deep learning in clustering tasks, deep clustering techniques show powerful performance. However, in the process of deep clustering, the issue of data privacy protection has gradually come to the fore. On the one hand, deep clustering models usually require a large amount of training data, which may contain sensitive information, such as personal identity, medical records, or trade secrets. How to ensure that these sensitive data are not leaked during the data collection and training phases is an important challenge in the field of deep clustering. On the other hand, the structure and parameters of deep clustering models may also become risk points for privacy leakage. For example, an attacker may infer certain features of the original data by analyzing the parameters or output of the model. Future research can delve into the privacy protection issue in the field of deep clustering, and study how to use encryption techniques, differential privacy mechanisms, or homomorphic encryption to effectively protect the training data and models without affecting the clustering performance. In addition, it is also possible to explore how to design privacy-preserving friendly deep clustering algorithms so that they can achieve efficient clustering while ensuring data privacy security when dealing with large-scale and high-dimensional data.

By exploring these future works, we expect to further advance the development of privacy-preserving clustering algorithms and provide more comprehensive and effective solutions for privacy preservation in practical applications.

Bibliography：

[1] Dwork, Cynthia. “Differential Privacy.” International Colloquium on Automata, Languages and Programming (2006).

[2] Blum A, Dwork C, McSherry F, et al. Practical privacy: the SuLQ framework. In: Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, 2005, pp.128–138.

[3] Li Y, Hao Z, Wen W, et al. Research on differential privacy preserving k-means clustering. Comput Sci 2013; 40(3): 287–290.

[4] Song, F., Ma, T., Tian, Y., & Al-Rodhaan, M. (2019). A New Method of Privacy Protection: Random k-Anonymous. IEEE Access, 7, 75434–75445.

[5] Yang D, Li S, Liu Z, et al. Differentially private geospatial data publication based on grid clustering. Int J Embed Syst 2019; 11(5): 613–623.

[6] Zheng X, Cai Z, Yu J, et al. Privacy-preserved data sharing towards multiple parties in industrial IoTs. IEEE J Select Areas Commun 2020; 38(5): 968–979.

[7] McSherry F D. Privacy integrated queries: an extensible platform for privacy-preserving data analysis[C]//Proceedings of the 2009 ACM SIGMOD International Conference on Management of data. 2009: 19-30.

[8] Dwork C. Differential privacy: A survey of results[C]//International conference on theory and applications of models of computation. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008: 1-19.

[9] Nissim K, Raskhodnikova S, Smith A. Smooth sensitivity and sampling in private data analysis[C]//Proceedings of the thirty-ninth annual ACM symposium on Theory of computing. 2007: 75-84.

[10] Yu Q, Luo Y, Chen C, et al. Outlier-eliminated k-means clustering algorithm based on differential privacy preservation[J]. Applied Intelligence, 2016, 45: 1179-1191.

[11] Kong Y, Qian Y, Tan F, et al. CVDP k-means clustering algorithm for differential privacy based on coefficient of variation[J]. Journal of Intelligent \& Fuzzy Systems, 2022, 43(5): 6027-6045.

[12] Chen H, Mei K, Zhou Y, et al. A density peaking clustering algorithm for differential privacy preservation[J]. IEEE Access, 2023.

[13] Wang B, Li H, Ren X, et al. An Efficient Differential Privacy-Based Method for Location Privacy Protection in Location-Based Services[J]. Sensors, 2023, 23(11): 5219.

[14] Zhang Z, Wu T, Sun X, et al. MPDP k-medoids: Multiple partition differential privacy preserving k-medoids clustering for data publishing in the Internet of Medical Things[J]. International Journal of Distributed Sensor Networks, 2021, 17(10): 15501477211042543.

[15] Zhu T, Ye D, Wang W, et al. More than privacy: Applying differential privacy in key areas of artificial intelligence[J]. IEEE Transactions on Knowledge and Data Engineering, 2020, 34(6): 2824-2843.

[16] Wei K, Li J, Ding M, et al. Federated learning with differential privacy: Algorithms and performance analysis[J]. IEEE transactions on information forensics and security, 2020, 15: 3454-3469.

[17] Chen Y, Du Y, Cao X. Density peak clustering algorithm based on differential privacy preserving[C]//Science of Cyber Security: Second International Conference, SciSec 2019, Nanjing, China, August 9–11, 2019, Revised Selected Papers 2. Springer International Publishing, 2019: 20-32.

[18] Min M, Xiao L, Ding J, et al. 3D geo-indistinguishability for indoor location-based services[J]. IEEE Transactions on Wireless Communications, 2021, 21(7): 4682-4694.

[19] Wang B, Chen Y, Jiang H, et al. Ppefl: Privacy-preserving edge federated learning with local differential privacy[J]. IEEE Internet of Things Journal, 2023.

[20] Deng J, Guo J, Wang Y. A Novel K-medoids clustering recommendation algorithm based on probability distribution for collaborative filtering[J]. Knowledge-Based Systems, 2019, 175: 96-106.

[21] Ma X, Guo D, Cui L, et al. SOM Clustering Collaborative Filtering Algorithm Based on Singular Value Decomposition[C]. Proceedings of the 2019 4th International Conference on Mathematics and Artificial Intelligence. ACM, 2019: 61-65.