Appl Intell (2016) 45:1179–1191
DOI 10.1007/s10489-016-0813-z

Outlier-eliminated *k*-means clustering algorithm based on differential privacy preservation

Qingying Yu[1,2] · Yonglong Luo[1,2] · Chuanming Chen[2] · Xintao Ding[2]

Published online: 11 July 2016

Abstract Individual privacy may be compromised during the process of mining for valuable information, and the potential for data mining is hindered by the need to preserve privacy. It is well known that *k*-means clustering algorithms based on differential privacy require preserving privacy while maintaining the availability of clustering. However, it is difficult to balance both aspects in traditional algorithms. In this paper, an outlier-eliminated differential privacy (OEDP) *k*-means algorithm is proposed that both preserves privacy and improves clustering efficiency. The proposed approach selects the initial centre points in accordance with the distribution density of data points, and adds Laplacian noise to the original data for privacy preservation. Both a theoretical analysis and comparative experiments were conducted. The theoretical analysis shows that the proposed algorithm satisfies ε-differential privacy. Furthermore, the experimental results show that, compared to other methods, the proposed algorithm effectively preserves data privacy and improves the clustering results in terms of accuracy, stability, and availability.

Keywords Differential privacy (DP) preservation · *k*-means clustering · Outlier · OEDP

✉ Yonglong Luo
ylluo@ustc.edu.cn

Qingying Yu
ahnuyuq@mail.ahnu.edu.cn

Chuanming Chen
ccm_0@163.com

Xintao Ding
accessdxt123@163.com

[1] School of Territorial Resources and Tourism, Anhui Normal University, Wuhu, 241003, China

[2] School of Mathematics and Computer Science, Anhui Normal University, Wuhu, 241003, China

1 Introduction

Individual data privacy may be compromised during the process of data mining for valuable information. With an increasing emphasis on security, privacy has become an important issue [2]. Owing to the fact that most privacy-preserving methods use data transformation to preserve data privacy, doing so while maintaining data availability has become an important research field in data mining and information security [19].

Currently, privacy-preserving models based on equivalence classes have been widely researched. For example, Sweeney [32] proposed the *k*-anonymity method to reduce data according to the granularity. Their method ensures that any given record cannot be distinguished from at least the other *k*-1 records. However, their method cannot ensure the diversity of values in *k* records. Machanavajjhala et al. [27] proposed an *l*-diversity method to avoid the shortcomings of homogeneity. Their proposal strengthens the diversity of sensitive values within equivalent groups, such that the probability of privacy loss does not exceed $1/l$. However, such privacy-preserving models do not provide a strict approach to measuring the level of privacy, and they require continual improvements in response to new attacks, such as background-knowledge attacks [27] and synthetic attacks [16].

As a result, differential privacy has been proposed as an intriguing and new privacy-preserving model. Dwork [11, 14] proposed concepts and algorithms related

to differential privacy. The general idea is that, for any two similar datasets, a given differential-privacy algorithm is approximately the same. This model avoids attacks based on background knowledge, and realizes privacy preservation by adding random noise to the query or analysis results. Unlike traditional privacy-preserving methods, differential privacy-preserving methods define a rigorous attack model, and they provide a quantitative representation and proof for the privacy-disclosure risk and the preserved data privacy while ensuring the availability of data. The amount of noise added to the results of the query or analysis is independent of the data size, and relatively little noise can achieve a high level of privacy.

Most research on differential privacy thus far has focused on the theoretical properties of the model, in terms of investigating its feasibility and infeasibility [13, 22]. Recently, several works have studied the use of differential privacy in practical applications. Research in related fields has resulted in a number of achievements related to differential privacy, including frequent-pattern-mining methods under a differential privacy model [5, 24, 36], differential privacy-preserving ID3 decision-tree classification [6, 15], and differential privacy-preserving logistic regression [7]. It is well known that clustering analysis is an especially important method for data mining. Moreover, it is the basis for many other mining methods. However, differential privacy models with specific applications for clustering are still in their infancy. As one of the most frequently used clustering methods, *k*-means algorithms are simple while offering high-speed clustering. Some literature has addressed related research. Blum et al. [6] proposed a differential privacy *k*-means approach. However, the availability of their clustering results is not robust to noise. Li et al. [26] proposed another differential privacy *k*-means method, along with one based on the initial centre, facilitating differential privacy with *k*-means clustering. However, their model selects the initial centres without considering the negative impact from outliers during the clustering process.

This paper presents an outlier-eliminated *k*-means clustering algorithm based on differential privacy preservation (namely, Outlier-eliminated Differential Privacy, or OEDP). Our proposed algorithm improves the scheme for selecting the initial clustering centres, and fully considers the level of privacy and clustering availability.

The contributions of this work can be summarized as follows:

1) An outlier detection method is proposed, which is used to select the initial cluster centres in order to avoid the negative impact of outliers on *k*-means clustering. Accordingly, the accuracy and efficiency of clustering is improved.
2) An outlier-eliminated dataset-partitioning algorithm (OEPT) is proposed, which is used to pre-process the dataset to improve the accuracy and availability of clustering.
3) An outlier-eliminated differential-privacy (OEDP) *k*-means clustering algorithm is proposed. It can maintain the availability of clustering while preserving privacy.
4) Several comparative experiments are performed to verify the effectiveness and efficiency of the proposed approach. The results show that our approach outperforms existing differential-privacy *k*-means algorithms.

The rest of this paper is organized as follows. In Section 2, we introduce related concepts and problems. In Section 3, we provide descriptions for the outlier-eliminated *k*-means clustering method. Section 4 introduces the OEDP *k*-means algorithm and discusses privacy preservation. Experimental results are presented in Section 5. Section 6 concludes the paper and provides future research directions.

2 Related concepts and problems

2.1 Differential privacy-preserving model

Compared to many other privacy-protecting methods [2, 27, 32], differential privacy-preserving technology is acknowledged as a rigorous and robust protection model. It provides formal privacy guarantees that do not depend on an adversary's background knowledge or computational power [15]. Formally, differential privacy is defined as follows:

Definition 1 (*Differential privacy*) [11, 12]: Assume K is a random function. $Range(K)$ represents the set of all possible outputs of K, and $Pr[Es]$ represents the disclosure risk of an event Es. The function K provides ε-differential privacy preservation for all datasets D and $D^{'}$ differing on at most one tuple, and all $S \subseteq Range(K)$, if K satisfies the following formula:

$$Pr[K(D) \in S] \leq exp(\varepsilon) \times Pr[K(D^{'}) \in S] \quad (1)$$

Here, $K(D)$ and $K(D^{'})$ represent the output of the function K, input with D and $D^{'}$, respectively, and ε is a parameter stipulating the level of privacy protection. The parameter ε is public, and its selection is matter of convention. In general, ε tends to be set within the range (0.01, 0.1), or in some cases ln2 or ln3 [12]. Lower values of ε provide stronger privacy, insofar as they limit any further influence of a record on the output of a calculation.

It can be seen from Definition 1 that differential privacy will guarantee that the outcome is not sensitive to any

particular record in the dataset. This definition is based on a theoretical point, and a noise mechanism is required to achieve differential privacy protection.

The noise mechanism is the main feature for achieving differential privacy protection. Laplacian and exponential mechanisms are two popular approaches to distributing noise. The magnitude of noise required to obtain ε-differential privacy depends on the sensitivity of the following function [10].

Definition 2 *(L_1 Sensitivity)* [11, 12]: Assume a function $f : D \rightarrow R^d$, for which the input is a dataset D and the output is a d-dimensional real vector. For all datasets D and $D^{'}$ differing on at most one tuple, the L_1 sensitivity of the function f is defined as follows:

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

where $\| f(D) - f(D^{'}) \|_1$ represents the 1-norm distance between $f(D)$ and $f(D^{'})$. Note that the L_1 sensitivity is a property of the function itself, and that it is independent of the dataset.

Definition 3 (*Probability density function*) [35]: Let $Lap(b)$ represent a Laplacian noise function, for which the position parameter is 0 and the scale parameter is b, such that $Lap(b) = exp(-|x|/b)$. The probability density function is defined as follows:

$$P(x|b) = \frac{1}{2b} exp(-\frac{|x|}{b}) \tag{3}$$

where $b = \Delta f/\varepsilon$. The value of ε generally ranges between 0 and 1. Taking $b = 1/\varepsilon$, the density at z is proportional to $e^{-\varepsilon|z|}$. Such a distribution reaches its maximum density at 0, for any z and $z^{'}$. If $|z - z^{'}| \leq 1$, the density at z is at most e^{ε} times the density at $z^{'}$. By decreasing ε, the distribution is flatter. That is, the smaller the value of ε, the higher the level of privacy.

Theorem 1 [12]*: For any function $f : D \rightarrow R^d$, the algorithm K that adds independently generated noise with $Lap(\Delta f/\varepsilon)$ to each output term in D satisfies ε-differential privacy, as shown in the following formula:*

$$K(D) = f(D) + (Lap_1(\Delta f/\varepsilon), Lap_2(\Delta f/\varepsilon), \ldots, Lap_d(\Delta f/\varepsilon)) \tag{4}$$

where any Laplacian variable $Lap_i(\Delta f/\varepsilon)(1 \leq i \leq d)$ is independent from the others, such that the noise depends exclusively on the sensitivity Δf and the parameter ε. These two values are independent of the number of rows in the dataset. Thus, even if the dataset is very large, the errors from typical queries that satisfy differential privacy are relatively few.

2.2 k-means clustering method based on differential privacy

Clustering analysis refers to the process of partitioning n data points in d dimensions into k clusters. Data points within each cluster are highly similar, and there is low similarity between different clusters. One clustering method is the k-means algorithm, which forms k clusters by associating each point in d dimensions with the closest cluster centre. The centre is the mean of each cluster, and this is updated according to some iterative rule, until a convergence criterion is reached or until a fixed number of iterations have been applied. More specifically [6]:

Given a dataset of points $\{p_1, \ldots, p_n\} \subset R^d$ and the initial cluster centres $\mu_1, \ldots, \mu_k$:

1) Partition the sample points $\{p_i\}$ into k sets $C_1, \ldots, C_k$, where each p_i is associated with the nearest μ_j;
2) For $1 \leq j \leq k$, set $\mu_j^{'} = \sum_{i \in C_j} p_i/|C_j|$. That is, the mean of sample points associated with μ_j is used as the new cluster centre.

During the process of k-means clustering, private data may be exposed. Privacy-preserving techniques in clustering analysis commonly include data disturbance and data transformation [2, 19, 25, 30, 31]. These methods preserve privacy by building privacy-preserving data-disturbance models for clustering. However, they fail to balance data availability with privacy-preserving strength.

Differential-privacy clustering algorithms are aimed at ensuring that nothing private is disclosed as a result of changes to the centre or to the quantity of records when any record from the dataset is deleted. Nissim et al. [29] proposed a k-means clustering publishing method that satisfies differential privacy. Their method provides sensitivity and error metrics. In addition, Dwork [9] presented two allocation schemes for a fixed privacy budget ε. Both methods satisfy ε-differential privacy, but they are unsuitable for practical applications, because of the difficulty in selecting k. Another ε-differential privacy k-means algorithm [26] preserved privacy by adding Laplacian noise to both the sum and the number of each subset. However, a random selection of the initial centres results in low clustering accuracy.

2.3 Outliers and their impact on k-means clustering

Outliers are data objects that differ significantly from other objects. Generally, outliers are classified into three categories: global outliers, contextual outliers, and collective outliers. Among these, global outliers refer to data objects that deviate significantly from the rest of the objects in the dataset. Contextual outliers refer to data objects that deviate significantly from the other objects given a specific context

(time, location, and other possible factors). Collective outliers refer to a subset that deviates significantly from the entire dataset. Considering that this paper mainly focuses on relevant numerical data privacy, rather than the analysis of specific situations and group behaviour, an "outlier" in this paper henceforth refers to global outliers that are detected by calculating the *r*-nearest-neighbour area density.

Outliers affect *k*-means clustering algorithms that depend in large part on the initial centre points. This can destabilise the clustering results, restricting the applications of such algorithms. Therefore, outliers must be detected and eliminated for clustering [20]. Hautamäki et al. [17] presented an outlier-removal clustering algorithm consisting of two stages. The first stage is a pure *k*-means process, while the second stage iteratively removes the outliers. Acs et al. [1] proposed a differentially private histogram-publishing method based on *k*-means clustering that considerably improves the accuracy of range queries. However, their method cannot be applied for processing outliers.

Let DT be a dataset, the correlational outliers for which are defined as follows:

Definition 4 (*r-nearest-neighbour area*): In DT, the region made up of object o and its r nearest neighbour is called the r-nearest-neighbour area for object o, denoted by *rNNA*, as shown in Fig. 1.

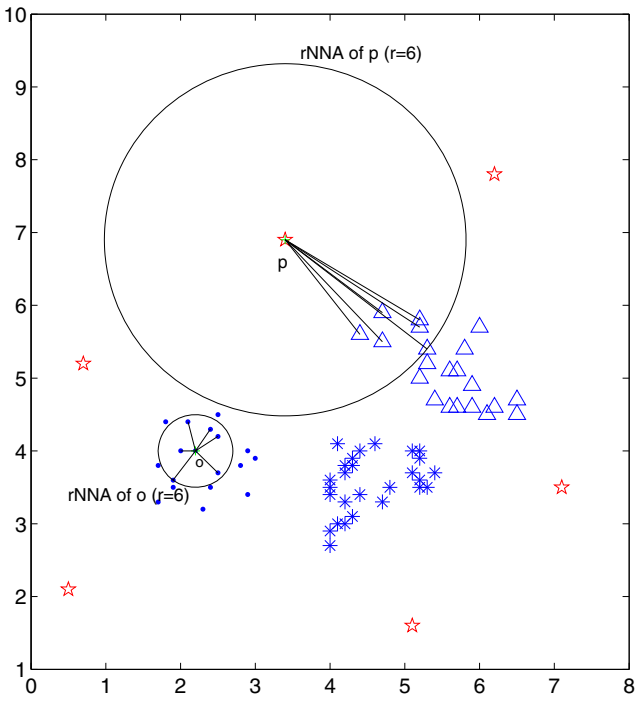


Fig. 1 $rNNA$ of objects o and p $(r = 6)$

Definition 5 (*r-nearest-neighbour distance*): Assuming that o is an object in DT, $dist(o, i)(1 \leq i \leq r)$ represents the Euclidean distance of o and r points in its *rNNA*. The r-nearest neighbour distance of o is defined as follows:

$$dist\_rNNA(o, DT) = \frac{\sum_{i=1}^{r} dist(o, i)}{r} \tag{5}$$

Definition 6 (*rNNA density*): Assuming that o is an object in DT, the *rNNA* density of o is defined as follows:

$$dens\_rNNA(o, DT) = \frac{1}{dist\_rNNA(o, DT)} \tag{6}$$

Definition 7 (*the k-th maximum distance*): Let $dist_{ij}$ be the distance between the i-th object and the j-th object in the dataset, and let $dist\_M$ be an $n \times n$ matrix consisting of $dist_{ij}(i, j = 1, 2, \ldots, n)$. The k-th maximum distance is the k-th maximum value of $dist\_M$, denoted by $dist_{ij}^{(k)}$.

As shown in Fig. 1, the greater the *rNNA* density of an object, the smaller the nearest-neighbour distance of this object. Therefore, outliers can be detected in DT by setting the appropriate values for r and the density threshold α.

3 Outlier-eliminated *k*-means clustering method

The accuracy of the *k*-means algorithm depends largely on the choice of the initial centres. To improve the accuracy and availability of clustering results, and to reduce the disadvantages that result from a *k*-means algorithm with randomly selected initial centres, this paper presents the outlier-eliminated dataset partitioning (OEPT) algorithm to obtain the initial k subsets, with which a traditional *k*-means algorithm can be improved. The OEPT algorithm first detects and eliminates outliers. Second, it partitions the entire dataset into k subsets in accordance with the *rNNA* density. Angiulli et al. [4] understand rare classes as those with less than 5 % of the data points in the dataset. Therefore, in our algorithm, let top_n be the number of outliers. The outlier density threshold α is assigned the mean of the top_n distances, where $top\_n = |DT| \times 0.05$.

The following Algorithm 1 can be used to pre-process the dataset for clustering:

Based on the k subsets $\{C_j | j = 1, \ldots, k\}$ obtained from the OEPT algorithm, the improved *k*-means method with eliminated outliers (the OE *k*-means method) first calculates the sum and number of the set $C_j(1 \leq j \leq k)$ using the formula $sum_j = \sum_{i \in C_j} p_i$ and $num_j = |C_j|$, setting $\mu_j = sum_j / num_j$ as the initial centre of C_j. Then, the traditional *k*-means method (see Section 2.2) is applied for clustering.

Algorithm 1 OEPT: outlier-eliminated dataset partitioning

Input:
DT = $\{p_1,\ldots,p_n\}$ (a dataset of n points in d dimensions), r (the number of points in $rNNA$), k (the number of subsets)
Output:
k subsets $C = \{C_1,\ldots,C_k\}$
1: Calculate the r-nearest-neighbour distance $dist\_rNNA(p_i,DT)$ of each point $p_i(1 \le i \le n)$ using Eq.(5).
2: Calculate the $rNNA$ density $dens\_rNNA(p_i,DT)$ of every $p_i(1 \le i \le n)$ using Eq.(6).
3: Calculate $d^{'} = \{dist_{ij} | dist_{ij} \ge dist_{ij}^{(top\_n)}, (i,j = 1,2,\ldots,n)\}$ and $\alpha = k/mean(d^{'})$.
4: For each $p_i(1 \le i \le n)$, if $(dens\_rNNA(p_i,DT) < \alpha)$, then update DT to DT-$\{p_i\}$.
5: Sort data points in DT according to $dens\_rNNA$ in ascending order.
6: Based on the sorting results, equally partition the ordered DT into k subsets $C_j(1 \le j \le k)$ in sequential order.

To be clear, in order to include all of the points in the clustering results, outliers can ultimately be addressed using different methods, depending on the specific practical application. For example, they might be treated separately as a category, or they can be assigned to the nearest cluster according to the Euclidean distance of the outliers to each cluster centre. In our analysis, outliers do not affect cluster measurements.

4 OEDP k-means clustering algorithm

The differential-privacy k-means algorithm described in Section 2 is not sufficiently accurate, owing to the fact that the initial centres are selected randomly. To improve this k-means method, Li et al. [26] proposed an IDP k-means clustering method, which selects the initial centres after partitioning the dataset into k subsets, thus reducing deviations from the centres and offering higher clustering availability than the DP k-means method. However, in terms of selecting the initial centres, their method does not consider the negative impact from having several outliers when partitioning the dataset. In addition, partitioning the initial set into equal partitions is conceptually fuzzy. That is, the specific details regarding the partitioning process are not described with sufficient clarity. Therefore, this paper proposes the OEDP k-means clustering algorithm to reduce the negative impact from outliers, and to optimize the choice of the initial centres. A schematic diagram for the algorithm is provided in Fig. 2.

To improve the availability of clustering while preserving privacy, the initial centre-selection mechanism based on Algorithm 1 is used to avoid the interference of outliers, and Laplacian noise is added to preserve the privacy of the data. The proposed algorithm is described as follows:

The addition of Laplacian noise in the above algorithm is $Lap(b) = exp(-|x|/b)$, where $b = \Delta f/\varepsilon$. For each iteration, according to [9], the value of the privacy budget ε is halved.

Similar to the OE k-means method, outliers can be assigned to the nearest cluster according to the Euclidean distance of the outliers to each cluster centre, in order to maintain the integrity of the clustering data.

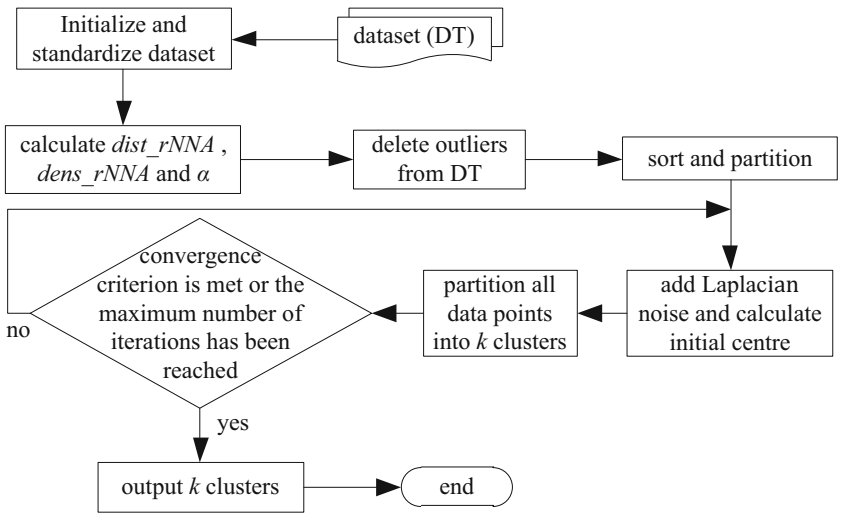


Fig. 2 Schematic diagram for the proposed OEDP k-means algorithm

Algorithm 2 OEDP k-means: outlier-eliminated differential-privacy k-means clustering

Input:
DT = $\{p_1, \ldots, p_n\}$ (a dataset of n points in d dimensions), r (the number of points in $rNNA$), k (the number of clusters), ε (the differential-privacy parameter)

Output:
k clusters $C = \{C_1, \ldots, C_k\}$

1: Partition DT into k subsets $C_j (1 \leq j \leq k)$, using Algorithm 1.
2: Calculate the sum and number of the set $C_j (1 \leq j \leq k)$ using the formula $sum_j = \sum_{i \in C_j} p_i$ and $num_j = |C_j|$, after adding Laplacian noise to sum_j and num_j, respectively, resulting in $sum_j^{'}$ and $num_j^{'}$. Then, set $\mu_j = sum_j^{'}/num_j^{'}$ as the centre point of C_j.
3: Partition each sample point $p_i (1 \leq i \leq n)$ into k sets $C_1, \ldots, C_k$. Each p_i is associated with the nearest $\mu_j (1 \leq j \leq k)$.
4: If the convergence criterion is not met and if the maximum number of iterations has not been reached, then return to Step 2.
5: Assign outliers to the nearest cluster according to the Euclidean distance of the outliers to each cluster centre.
6: Output k clusters $C = \{C_1, \ldots, C_k\}$.

Lemma 1 Each iteration in Algorithm 2 satisfies ε-differential privacy, the proof for which is as follows:

Proof Assume D and $D^{'}$ are two datasets differing on at most one tuple. The process of calculating k centres can be regarded as a query of the histogram in $[0, 1]^d$. According to Definition 2, the Δf of denominator num is 1, because the data points have been normalized, $p_i \in [0, 1]^d (1 \leq i \leq n)$. The maximum change to each dimension is thus 1, when adding or deleting one point in the d-dimensional dataset DT. Therefore, the maximum Δf of the numerator num is d. Assume $Clus(D)$ and $Clus(D^{'})$ represent the respective clustering results after adding noise to D and $D^{'}$. Let $Part$ denote an arbitrary clustering partition. The algorithm adds Laplacian noise to each output item with the parameter value $\Delta f/\varepsilon$. As the noise function $Lap(\Delta f/\varepsilon) = exp(-|x| \cdot \varepsilon/\Delta f)$, from Theorem 1, $Pr[Clus(D) = Part] \leq exp(\varepsilon) \times Pr[Clus(D^{'}) = Part]$. Therefore, according to Definition 1, the OEDP k-means algorithm satisfies ε-differential privacy. □

5 Experiments

To measure whether the proposed algorithm is effective, both the degree of privacy preservation and the high availability of the algorithm based on clustering results must be considered [36]. Therefore, it is necessary to coordinate the balance of these two aspects.

In this paper, we conducted a set of experiments with Matlab 8.3 on an Intel (R) Core (TM) 2 Duo CPU 3.3 GHz with 4 GB of RAM. The operating system was Windows 7. In order to demonstrate the effectiveness of the OEDP k-means algorithm, four datasets were run with the OEDP k-means, IDP k-means [26] and DP k-means algorithms based on differential privacy preservation. The results were compared and evaluated.

5.1 Dataset

Because the purpose of our experiments was to estimate the availability and time consumption of our privacy-preserving algorithm, while emphasizing on the premise of privacy preservation, we used UCI and synthetic data in our implementation. The experimental datasets comprised Ecoli, Iris, Wine, and Climate, which are typically used for clustering, outlier detection, and classification [21, 23, 33]. These datasets were generated at the University of California, Irvine, (UCI), and they are available at http://archive.ics.uci.edu/ml/datasets.html. The characteristics of these datasets are described in Table 1.

The Climate dataset was pre-processed based on the method outlined in [18]. As a result, the dataset contained 360 records, with 30 tuples (8 %) in one category and 320 tuples (92 %) in the other. Here, 30 tuples were regarded as outliers. We conducted the appropriate pre-treatment for each dataset before the experiment. We first removed duplicate tuples from each dataset and normalized the datasets.

Table 1 Dataset characteristics

Dataset	Number of records	Number of attributes	Number of clusters	Attribute type
Ecoli	336	8	8	Real
Iris	150	4	3	Real
Wine	178	13	3	Integer, Real
Climate	540	18	2	Real

The values of all attributes (except those for classification) were normalized to the interval [0,1]. The following normalization method was adopted [34]:

$$x' = \frac{x - min_A}{max_A - min_A} \tag{7}$$

where min_A and max_A are the minimum and the maximum values for attribute A, respectively. This method is referred to as min-max normalization, mapping a value x to x' in the range [0,1].

5.2 Evaluation methods

Taking into account the impact of noise on data availability is considerably important for preserving privacy. Generally, data availability can be evaluated in two ways: in theory and through application. For the former, (β, γ)-usefulness [28] is often used to measure the availability of a differential privacy algorithm. For the latter, popular availability metrics include the relative error, absolute error, the Euler function, and the F-measure [21]. Selecting a suitable metric depends on the specific data used.

In this paper, because the reference category is already provided by the selected datasets, we used the F-measure to evaluate the clustering performance. The F-measure (also known as the F-fraction) is a criterion for clustering availability associated with precision and recall for information retrieval. Compared to the other metrics, the result of F-measure is more pertinent. Assume n represents the size of a given dataset, i represents the right class label of the dataset, n_i and n_j represent the number of data points in class i and cluster C_j, respectively, and n_{ij} represents the number of data points at the intersection of class i and cluster C_j. The precision and recall are defined as follows:

$$prec(i, j) = max_{i,j}\{\frac{n_{ij}}{n_j}\}, rec(i, j) = max_{i,j}\{\frac{n_{ij}}{n_i}\} \tag{8}$$

For a given class i and cluster C_j, the F-measure is defined as follows:

$$Fmeas(i, j) = \frac{(\beta^2 + 1) \cdot prec(i, j) \cdot rec(i, j)}{\beta^2 \cdot prec(i, j) + rec(i, j)} \tag{9}$$

We set $\beta = 1$ to obtain the same weight for $prec(i, j)$ and $rec(i, j)$. The entire F-measure for a dataset of size n is computed as follows:

$$F = \sum_i \frac{n_i}{n} max_j\{Fmeas(i, j)\} \tag{10}$$

The range of the F-measure values is [0,1]. A higher value means that the algorithm has more clustering availability.

5.3 Experimental results

5.3.1 Parameter allocation

Three parameters were used in the experiments: r (the number of points in the *rNNA*), k (the number of clusters), and ε (the differential privacy parameter).

1) ε: Reasonable budget-allocation strategies are required to facilitate the life-cycle of ε such that it survives as

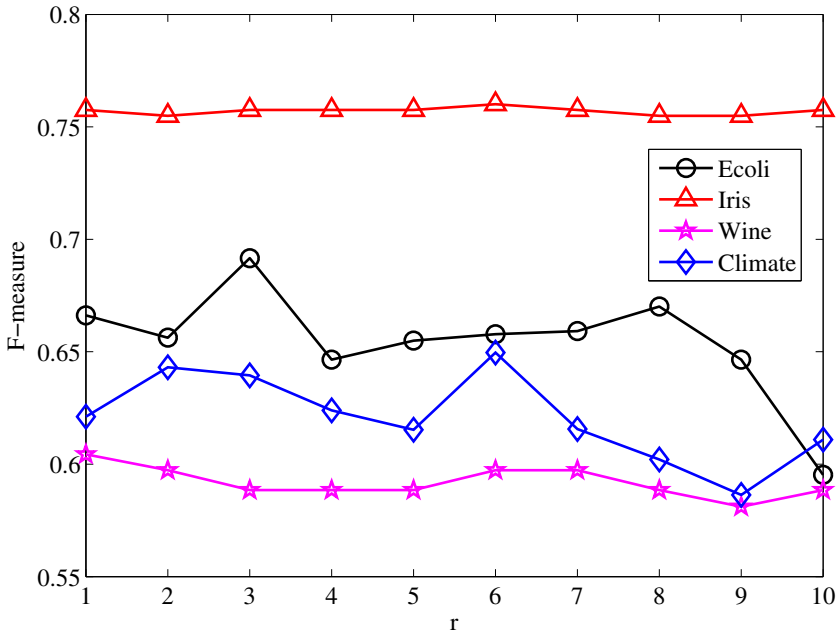
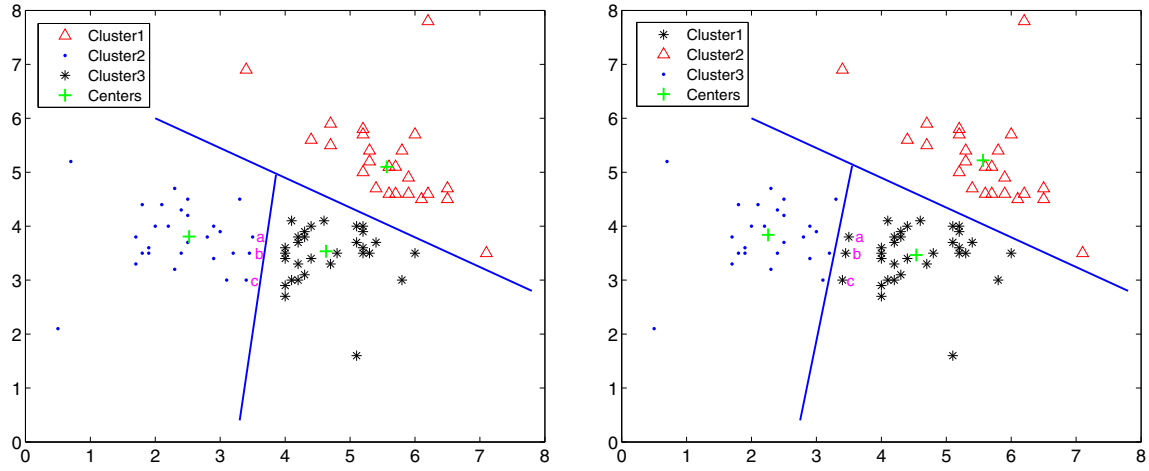


Fig. 3 Accuracy with various r-values



(a) OE *k*-means method (b) Traditional *k*-means method

Fig. 4 Comparison of clustering results from two *k*-means methods

long as possible. Popular distribution strategies include linear distributions, even distributions, exponential distributions, manual assignments, and mixed distributions [8]. As described in Section 2.1, ε generally tends to be understood as falling within the range of (0.01, 0.1). Therefore, a linear distribution in the interval [0,1] is generally selected for the allocation of ε in the experiments.

2) k: Because the aim of this paper is to apply differential-privacy technology to the k-means method, rather than applying simple clustering, we optimized the selection of the initial centres. Therefore, we selected k in accordance with the number of reference categories provided by the UCI dataset.
3) r: Insofar as it represents the number of points in the *rNNA*, r is an important parameter. The algorithm is sensitive to this parameter. Following the method proposed by Angiulli et al. [3], we modified r depending on the experimental results of the OE k-means method, which are shown in Fig. 3. With a different r, the OE k-means method was repeated multiple times in order to obtain the optimal r based on the accuracy of the clustering.

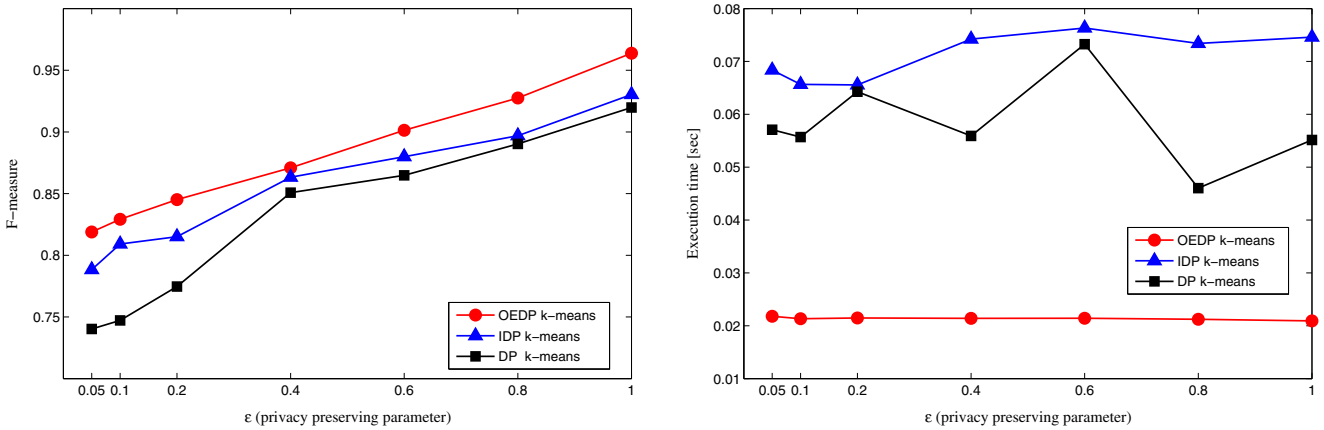
As shown in Fig. 3, the optimal value of r is 3 for the Ecoli dataset, 6 for the datasets Iris and Climate, and 1 for the Wine dataset. Therefore, in the OEDP algorithm, these respective values were used for the four datasets.

5.3.2 OE k-means method

We first conducted a simulation to compare the results of the OE k-means method with the traditional k-means clustering method on a synthetic dataset DS containing 82 data points in two dimensions (including outliers). The parameters were as follows: $r = 6$ (the number of points in the *rNNA*) and $k =$

Table 2 Comparing the accuracy and efficiency of two *k*-means methods

Dataset	Number of records	Dimensions (excluding the label)	Clustering method	*F*	Execution time [sec]
DS	82	2	OE *k*-means	0.8781	0.00115
			k-means	0.8415	0.09545
Ecoli	336	7	OE *k*-means	0.7598	0.08626
			k-means	0.6366	0.14326
Iris	150	3	OE *k*-means	0.9134	0.00914
			k-means	0.8719	0.03573
Wine	178	12	OE *k*-means	0.6044	0.01790
			k-means	0.6030	0.05638
Climate	540	17	OE k-means	0.6496	0.00022
			k-means	0.6433	0.14112



(a) F-measure (b) Execution time

Fig. 5 Running results comparison on dataset Ecoli

3 (the number of clusters). A visual comparison is provided in Fig. 4.

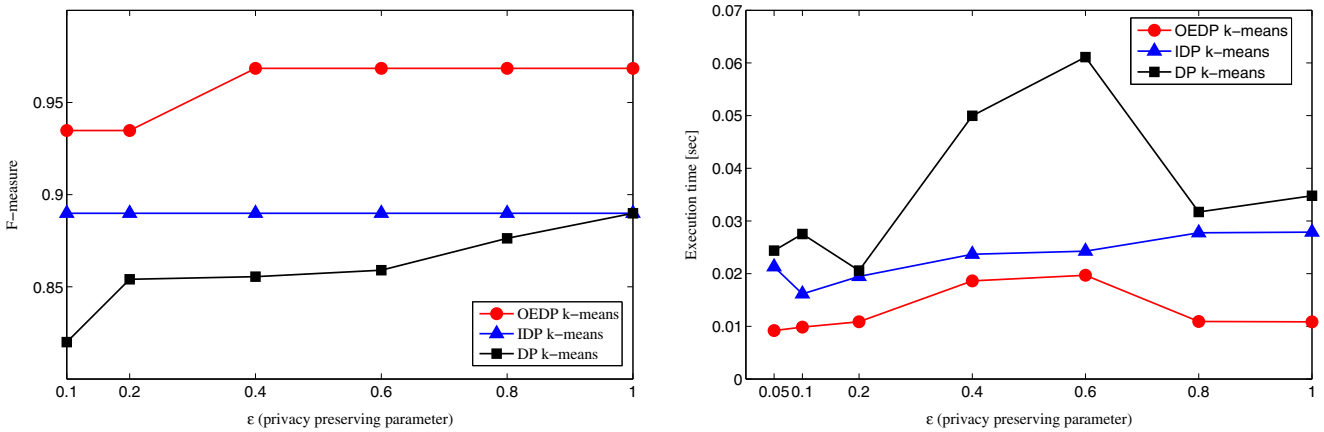
Figure 4 shows that, owing to the difference in how the initial centre is selected, the clustering results for data points a, b, and c are significantly different. It can be concluded from the results in Fig. 4 that the OE k-means method is more accurate than the traditional k-means method.

Second, we conducted a series of experiments comparing the accuracy and execution time of the two algorithms. For each dataset, with randomly generated initial centres, the k-means method was run 20 times. We noted the average F value from these results. As shown in Table 2, for all datasets, the OE k-means method is more accurate and requires less execution time. The main reason for this is the elimination of outliers before clustering. By eliminating outliers beforehand, we avoid the negative impact they have on the selection of initial centres.

5.3.3 OEDP k-means method

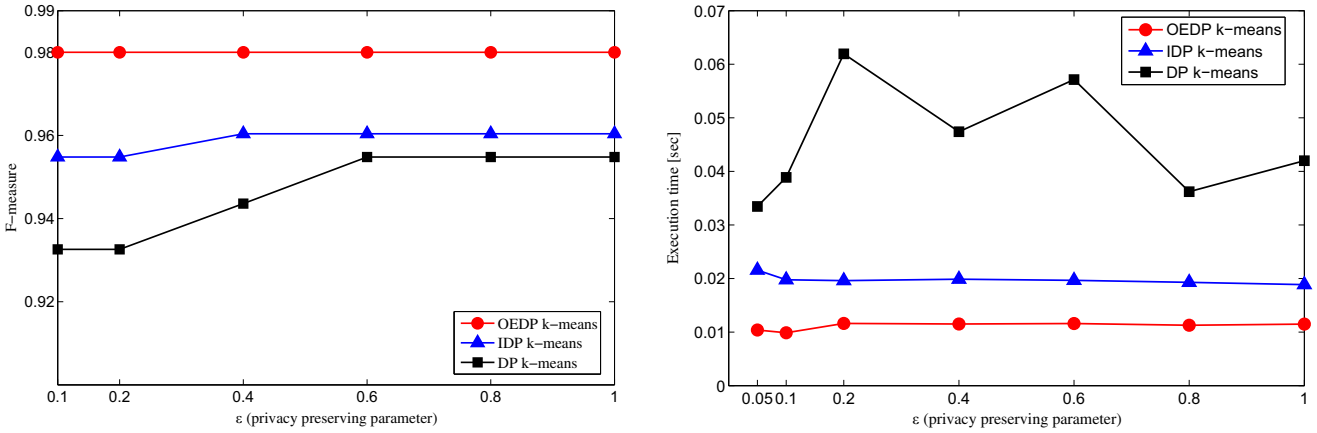
In this section, we first describe an experiment in which we ran the OEDP, IDP and DP k-means algorithms based on four datasets with classification labels already available. Apart from the classification, the datasets were normalized. We changed the value ε, varying it between 0 and 1, and the program was run ten times each time this value was changed. The results shown are the average F-measure and the execution time from these ten trials for each value of ε. The execution time refers to the time required for clustering after the initial centres have been selected. Figures 5(a)–8(a) and Figs. 5(b)– 8(b) respectively show a comparison of the F-measure values and the execution time from the three algorithms running on Ecoli, Iris, Wine, and Climate.

The greater the F-measure is, the more similar the clustering results are before and after adding noise and the



(a) F-measure (b) Execution time

Fig. 6 Running results comparison on dataset Iris



(a) F-measure (b) Execution time

Fig. 7 Running results comparison on dataset Wine

better the availability of the algorithm. Likewise, a shorter execution time increases the efficiency of the algorithm.

1) Analysis of clustering availability

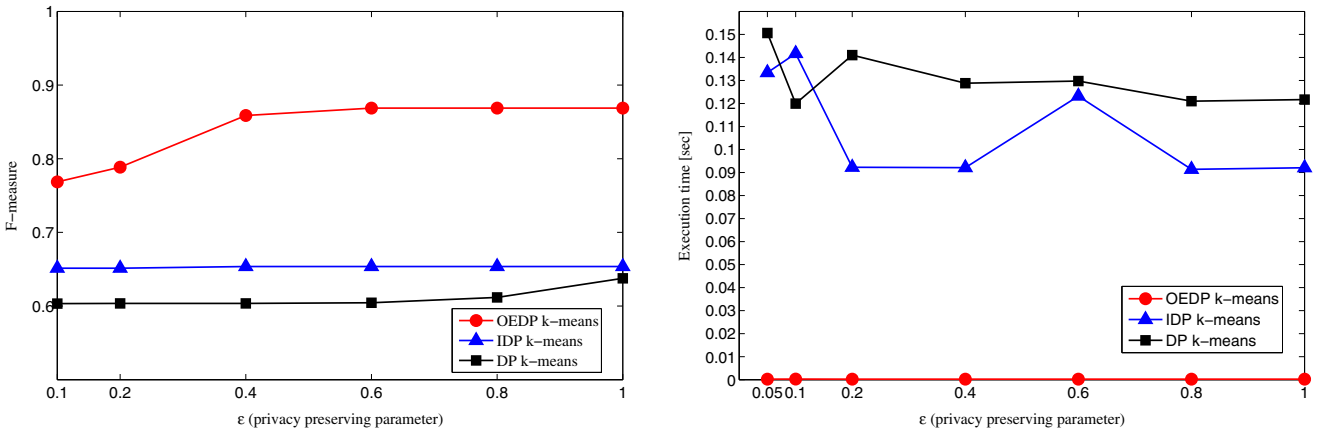
As can be seen from Figs. 5(a) – 8(a), the value of ε influences the F-measure. With the same ε, the proposed OEDP k-means algorithm resulted in a higher F-measure value compared to the other algorithms. Therefore, the clustering results from our algorithm are more similar to the original data, and they better maintain the clustering availability. As the differential privacy parameter ε increased, so too did the F-measure. This indicates that the clustering results improve as the level of privacy decreases.

2) Analysis of the algorithm's efficiency

As can be seen from Figs. 5(b) – 8(b), with the same ε, the execution time for the OEDP k-means algorithm is significantly less than it is for the other two algorithms. Moreover, the curves of the execution time for the IDP and DP k-means algorithms show obvious fluctuations when changing the ε value. By contrast, our algorithm is relatively stable. These results demonstrate that our algorithm outperforms the other two algorithms, and this is mainly because of the optimized selection of the initial centres by the OEPT algorithm. Thus, due to the elimination of outliers, the OEDP k-means algorithm performs better, and the total execution time decreases.

In summary, the experimental results show that, at the same level of privacy, the proposed OEDP k-means clustering algorithm is superior to both the IDP k-means and DP k-means algorithms in terms of clustering effectiveness and efficiency.

Second, we present the experimental results on analysis of the algorithm's security. Dwork [9] proposed that differential privacy ensures privacy preservation, independent of whether any tuple in to, or out of, the dataset. The absence



(a) F-measure (b) Execution time

Fig. 8 Running results comparison on dataset Climate

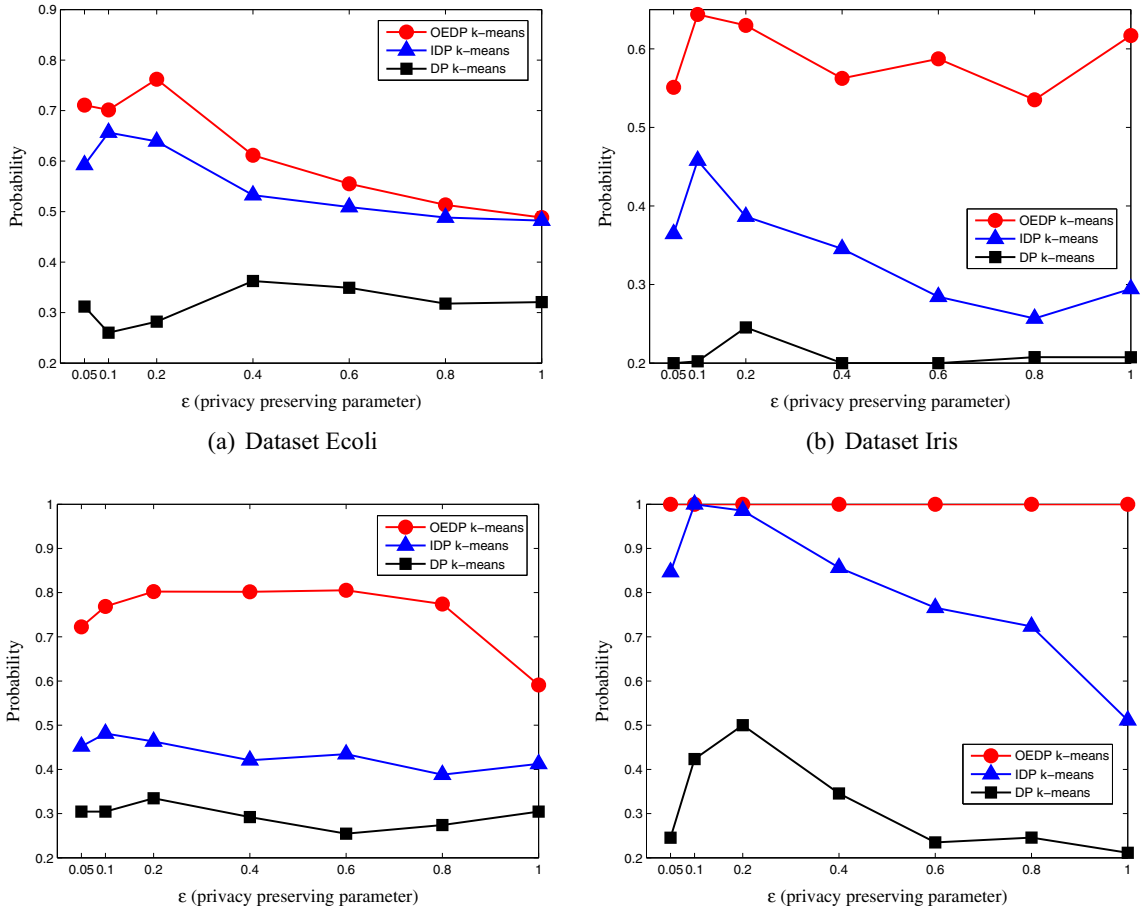


Fig. 9 Running results comparison of the probability with different datasets

of any tuple in the dataset will not significantly affect its chance of receiving coverage [9]. In order to test the effect of privacy preservation, we ran the OEDP, IDP, and DP k-means algorithms based on the above four datasets to conduct comparison on the Probability—that is, the possibility of centre points with no change before and after the deletion of any tuple in the dataset. The results provide an assistant confirmation of privacy preservation.

We changed the value ε, varying it between 0 and 1. The results shown are the average probabilities from all trials after respectively removing each tuple for each value of ε. Figure 9(a)–(d) respectively shows a comparison of the probability from the three algorithms running on Ecoli, Iris, Wine, and Climate.

As can be seen from Fig. 9, with the same ε, the probability of the OEDP k-means algorithm is significantly higher than it is for the other two algorithms. These results demonstrate that our algorithm outperforms the other two algorithms. Thus, privacy preservation was confirmed, considering $Pr[Clus(D) = Part] \leq exp(\varepsilon) \times Pr[Clus(D^{'}) = Part]$. The same result has also been proved by means of theoretical analysis in Lemma 1.

6 Conclusion

In this paper, we described applications for differential privacy with k-means clustering, and proposed an outlier-eliminated differential-privacy algorithm for k-means clustering. The proposed algorithm uses the densities of the data points in the r-nearest-neighbour area to eliminate outliers and increase the effectiveness and efficiency of clustering while better preserving privacy. Both theoretical analysis and experimental results show that the proposed OEDP k-means method provides differential privacy while expanding the scope of application for k-means algorithms. Compared with the DP k-means and IDP k-means methods, the proposed OEDP k-means method reduces the negative impact of outliers when selecting the initial centres. Furthermore, it promotes stable clustering results and significantly

improves the clustering availability. In future research, we plan to improve the security of our proposal using different tactics for allocating the privacy budget, and we shall explore further applications for the OEDP *k*-means method.

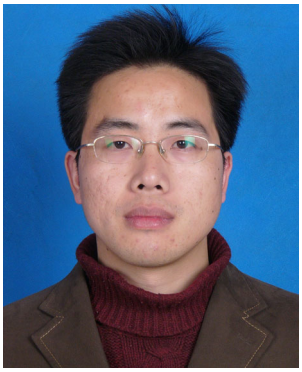
Acknowledgments The authors would like to thank the reviewers for their useful comments and suggestions for this paper. This work was supported by the National Natural Science Foundation of China (61370050) and the Natural Science Foundation of Anhui Province (1508085QF134).

References

1. Acs G., Castelluccia C., Chen R. (2012) Differentially private histogram publishing through lossy compression. In: proceedings of IEEE 12th International Conference on Data Mining, ICDM, pp 1–10
2. Agrawal R, Srikant R (2000) Privacy-preserving data mining. ACM Sigmod Record 29(2):439–450
3. Angiulli F, Fassetti F (2009) DOLPHIN: an efficient algorithm for mining distance-based outliers in very large datasets. ACM Transactions on Knowledge Discovery from Data(TKDD) 3(1):4:1–57
4. Angiulli F, Pizzuti C (2002) Fast outlier detection in high dimensional spaces. In: proceedings of the 6th European Conference on the Principles of Data Mining and Knowledge Discovery, pp 15–27
5. Bhaskar R, Laxman S, Smith A, Thakurta A (2010) Discovering frequent patterns in sensitive data. In: proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD10). Washington, USA, pp 503–512
6. Blum A, Dwork C, Mcsherry F, Nissim K (2005) Practical privacy: the SuLQ framework. In: proceedings of the 24th ACM SIGMOD International Conference on Management of Data / Principles of Database Systems. New Yorks:ACM Press, pp 128–138
7. Chaudhuri K, Monteleoni C (2008) Privacy-preserving logistic regression. In: proceedings of the 22nd Annual Conference on Neural Information Processing Systems. Vancouver, Canada, pp 289–296
8. Chen R, Acs G, Castelluccia C (2012) Differentially private sequential data publication via variable-length n-grams. In: proceedings of the 2012 ACM Conference on Computer and Communications Security, pp 638–649
9. Dwork C (2011) A firm foundation for private data analysis. Commun ACM 54(1):86–95
10. Dwork C, McSherry F, Nissim K, Smith A (2006) Calibrating noise to sensitivity in private data analysis. In: proceedings of the 3rd Conference on Theory of Cryptography. New York, USA, pp 265–284
11. Dwork C (2006) Differential privacy. In: proceedings of the 33rd International Colloquium on Automata,languages and Programming. Springer, Berlin, pp 1–12
12. Dwork C (2010) Differential privacy in new settings. In: proceedings of ACM-SIAM Symposium on Discrete Algorithms (SODA), pp 174–183
13. Dwork C (2008) Differential privacy: a survey of results. In: proceedings of the 5th International Conference on Theory and Application of Models of Computation. Berlin Heidelberg, pp 1–19
14. Dwork C (2009) The differential privacy frontier(extended abstract). In: proceedings of the 6th Theory of Cryptography Conference(TCC09). Springer, Berlin, pp 496–502
15. Friedman A, Schuster A (2010) Data mining with differential privacy. In: proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington, USA, pp 493–502
16. Ganta SR, Kasiviswanathan S, Smith A (2008) Composition attacks and auxiliary information in data privacy. In: proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Las Vegas, USA, pp 265–273
17. Hautamäki V, Cherednichenko S, Kärkkäinen I, Kinnunen T, Fränti P (2005) Improving *k*-means by outlier removal. Lect Notes Comput Sci 3540:978–987
18. Hawkins S, He H, Williams G, Baxter R (2002) Outlier detection using replicator neural networks. In: proceedings of the 4th International Conference on Data Warehousing and Knowledge Discovery. Springer, Berlin Heidelberg, pp 170–180
19. Huang M, Ni W, Wang J, Sun F, Chong Z (2012) A logarithmic spiral based data perturbation method for clustering. Chin J Comput 35(11):2275–2282
20. Jiang F, Chen Y-M (2015) Outlier detection based on granular computing and rough set theory. Appl Intell 42(2):303–322
21. Jiang H, Yi S, Li J, Yang F, Hu X (2010) Ant clustering algorithm with k-harmonic means clustering. Expert Systems With Applications 37(12):8679–8684
22. Kasiviswanathan SP, Lee HK, Nissim K, Raskhodnikova S, Smith A (2009) What can we learn privately? Foundations of Computer Science Annual Symposium on 40(3):531–540
23. Knorr EM, Ng RT, Tucakov V (2000) Distance-based outliers: algorithms and applications. The VLDB Journal-The International Journal on Very Large Data Bases 8(3-4):237–253
24. Li N, Qardaji W, Su D, Cao J (2012) PrivBasis: frequent itemset mining with differential privacy. In: proceedings of the 38th International Conference on Very Large Data Bases(VLDB12). New Yorks:ACM, pp 1340–1351
25. Li X-B, Sarkar S (2010) Data clustering and micro-perturbation for privacy-preserving data sharing and analysis. In: proceedings of the International Conference on Information Systems (ICIS). Yamagata, Japan, pp 58–73
26. Li Y, Hao Z, Wen W, Xie G (2013) Research on differential privacy preserving *k*-means clustering. Comput Sci 40(3):287–290
27. Machanavajjhala A, Kifer D, Gehrke J, Venkitasubramaniam M (2007) l-diversity: privacy beyond k-anonymity. ACM Trans Knowl Discov Data 1(1):24
28. Nguyen HH, Kim J, Kim Y (2013) Differential privacy in practice. Int J Comput Sci Eng 7(3):177–186
29. Nissim K, Raskhodnikova S, Smith A (2007) Smooth sensitivity and sampling in private data analysis. In: proceedings of the 39th Annual ACM symposium on Theory of Computing – STOC 07, pp 75–84
30. Oliveira SRM, Zaïane OR (2004) Achieving privacy preservation when sharing data for clustering. In: proceedings of the International Workshop on Secure Data Management in a Connnected World. Toronto, Canada, pp 67–82
31. Parameswaran R, Blough DM (2008) Privacy preserving data obfuscation for inherently clustered data. Int J Inf Comput Secur 2(1):4–26
32. Sweeney L (2002) K-anonymity: a model for protecting privacy. Int J Uncertainty Fuzziness Knowledge Based Syst 10(5):557–570
33. Tung AKH, Xu X, Ooi BC (2005) CURLER: finding and visualizing nonlinear correlation clusters. In: proceedings of the International Conference on Management of Data, pp 467–478
34. Visalakshi NK, Thangavel K (2009) Impact of normalization in distributed *k*-means clustering. Int J Soft Comput 4(4):168–172
35. Xiong P, Zhu T, Wang X (2014) A survey on differential privacy and applications. Chin J Comput 37(1):101–122
36. Zhang X, Wang M, Meng X (2014) An accurate method for mining top-k frequent pattern under differential privacy. Journal of Computer Research and Development 51(1):104–114

Qingying Yu received her B.S. degree and M.S. degree from Department of Computer Science and Technology, Anhui University, Hefei, China, respectively in 2002 and 2005. Currently, she is a Ph.D. Candidate at Anhui Normal University, Wuhu, China. Her main research interests are spatial data processing and information security.

Yonglong Luo received his Ph.D. degree from the School of Computer Science and Technology, University of Science and Technology of China in 2005. Since 2007, he has been a professor in School of Mathematics and Computer Science, Anhui Normal University. Currently, he is the Ph.D. supervisor of Anhui Normal University. He is the Director of Engineering Technology Research Center of Network and Information Security. His research interests are information security and spatial data processing.



Chuanming Chen received his B.S. degree and M.S. degree from Department of Computer Science and Technology, Anhui University, Hefei, China, respectively in 2002 and 2005. Currently, he is a Ph.D. Candidate at Nanjing University of Aeronautics and Astronautics, Nanjing, China. His main research interests are data mining and intelligent computing.

Xintao Ding received his M.S. degree in computational mathematics from the Department of Mathematics, East China Normal University, Shanghai, China, in 2005, and his Ph.D. degree from Anhui Normal University, Wuhu, China in 2015. Currently, he is an associate professor at Anhui Normal University, China. His research interests are computer vision and machine learning.