Journal of Computers Vol. 32 No. 4, 2021, pp. 239-251
doi:10.53106/199115992021083204019

Cloud-side Collaborative Privacy Protection Based on Differential Privacy

Zhenjiang Zhang[1*], Desong Qin[1], Zihang Yu[1], Li He[2], Xiaohua Liu[2], Shutao Liu[2]

[1] Key Laboratory of Communication and Information Systems, Beijing Municipal Commission of Education, School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing 100044, China
{zhangzhenjiang, 18120117}@bjtu.edu.cn, Y245960175@163.com

[2] Westone Information Industry Co., LTD, Chengdu, 610041, China
{he.li, liu.xiaohua, liu.shutao}@westone.com.cn

Received 1 June 2021; Revised 1 July 2021; Accepted 8 August 2021

Abstract. With the rapid development of 5G and the Internet of Things, edge computing is playing an increasingly important role in real-world applications. But at the same time there is the risk of leaking user privacy information. Therefore, how to ensure users' personalized privacy requirements has become one of the hot issues in cloud-side collaborative computing scenarios. This paper focuses on the problems of the mean estimation and histogram estimation algorithms in the differential privacy protection model. The traditional personalized local differential privacy based on data derivation does not consider the influence of coding on the estimation error of histogram, and the data derivation algorithm has high algorithm complexity. This paper solves the above problems in the cloud-side collaboration scenario, and its main work is as follows: Established a distributed and personalized local differential privacy protection model for the privacy protection scenario of cloud-side collaboration. Under the premise of meeting the personalized privacy requirements of data at different edge nodes, the use of optimized unary coding reduces the mean square error of histogram estimation. Proposed an optimized personalized privacy data derivation algorithm based on optimized unary encoding. And confirmed the algorithm complexity of the data derivation algorithm is greatly reduced.

Keywords: differential privacy, edge computing, histogram estimation, mean estimation

1 Introduction

In recent years, with the exponential growth of the number of smart devices, edge computing and cloud computing have developed rapidly [10]. Especially for the upcoming 5G network, its low latency, high speed, and multi-antenna coverage characteristics will make the future development of the Internet of Things more rapid. In the development of the Internet of Things (IoT) and cloud computing, the time consumed in data transmission has become a major challenge that limits the quality of cloud computing services [11]. In order to solve the problems of network delay and high computing consumption in cloud computing services, distributed localized edgeq computing technology is proposed. Edge computing is a new computing paradigm that processes data at the edge of the network [1]. Therefore, it is more efficient to process data in an edge computing architecture close to the data generation end [2]. However, a huge amount of smart devices will generate huge amounts of information data, and this huge amount of information data may contain a large amount of user privacy information [12]. Due to the complexity, real-time nature of the edge computing service model, the heterogeneity of data from multiple sources, and the limited resources of the terminal, the data security and privacy protection mechanisms in the traditional cloud computing environment are no longer suitable for edge computing. Data privacy and location Security issues such as privacy and identity privacy have become more and more prominent [3]. With the increase of computing resources and the enhancement of computing power, people's demand

[*] Corresponding Author

for services tends to be more personalized, and the demand for privacy and security is also becoming more personalized [13]. How to make full use of user data while ensuring the user's personalized privacy needs will become a new problem.

2 Related Work

To protect private information, we must first clarify the definition of private information and its measurement method. Li Fenghua and others proposed a full life cycle model of private information [4]. As shown in Fig. 1, the model consists of 9 parts: privacy information generation, privacy perception, privacy protection, privacy release, privacy information storage, privacy exchange, privacy analysis, privacy destruction, and privacy recipients. Among them, privacy protection, privacy release/storage/exchange and privacy analysis have become the main research directions of privacy protection.

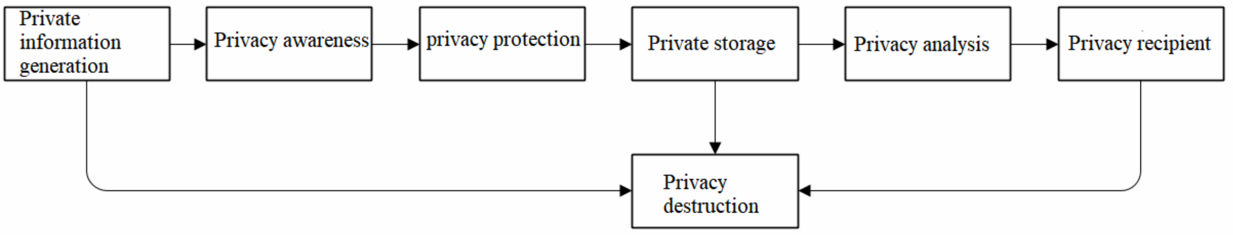

Fig. 1. The full life cycle of privacy information

At present, the k-anonymity model is the most widely used privacy protection model, and is often used in location privacy and identity privacy protection [5], but the k-anonymity model is vulnerable to background knowledge and homogeneity attacks. Afterwards, many scholars have proposed many optimized versions by studying the types of attacks on the k-anonymity model, but they all have other flaws. Differential privacy is a privacy protection mechanism based on a strict mathematical background. It provides a method to quantify, evaluate and prove the level of privacy [6]. Gu and others pointed out that different data should have different privacy levels, and solved the problem of distinguishing protection levels according to input content and reflecting different privacy requirements of different inputs. Developed and designed an IDUE mechanism based on unary encoding, and showed that the proposed mechanism to meet the MinID-LDP has a better effect than the local differential privacy mechanism [7].

This paper focuses on the principle and implementation mechanism of differential privacy protection, and designs a personalized privacy protection model for cloud-side collaboration scenarios. Aiming at the optimization problem of histogram estimation for personalized differential privacy in cloud-side collaboration scenarios, firstly, the error of histogram estimation is reduced by optimizing the encoding method. Through the analysis of the error function, optimized the calculation amount of the data derivation algorithm and the cloud-side node. And the optimization effect is verified by experimental simulation.

3 Overview of Personalized Differential Privacy Under Cloud-side Collaboration

3.1 System Overview

The existing differential privacy protection model is usually designed only with the two endpoints of the user and the third-party data collection center, ignoring the impact of the data transmission delay from the user to the data center on the personalized service [8]. To solve these problems, this paper designs a personalized differential privacy protection model in the cloud-side collaboration scenario, as shown in Fig. 2.

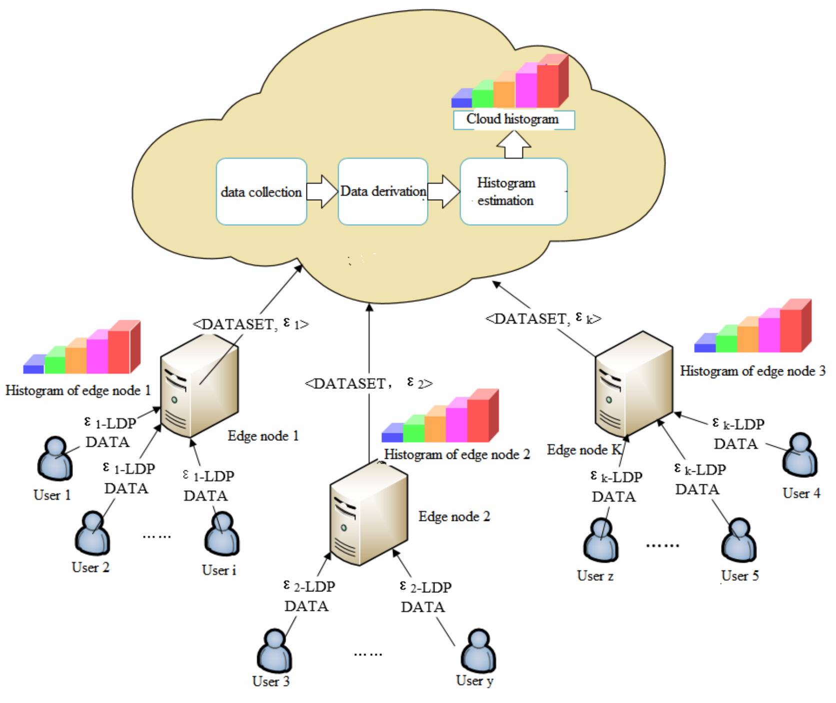

Fig. 2. Cloud-edge collaboration personalized local differential privacy model

In the cloud-edge collaboration scenario, with the advantage of being closer to the user side geographically and having higher computing power, edge computing nodes can provide more personalized and fast services to user groups within the node range. According to the local privacy protection budget, each user sends the protected data to the edge computing node within its own range. At this time, the edge computing node will immediately analyze the collected data to obtain statistical characteristics that are more similar with the user data in the node. These statistical results only exist locally at the edge node and provide personalized services to users based on the statistical results.

Cloud computing centers have more powerful computing capabilities, and are often used for statistical analysis of user data in a large range. However, what the cloud computing center collects is often the protection data generated based on different privacy protection levels. These data cannot be directly used for statistical analysis, because the statistical results will have particularly large errors, making the data lose use value. However, low-level privacy data often contains information with high-level privacy data. This can derive high-level privacy data from low-level privacy data, thereby increasing samples of higher-level privacy data and reducing the estimation error of statistical results.

Table 1 gives the description of each symbol used in the content of this chapter:

Table 1. Symbols description

symbol	Symbol description
T	Data value range
ε	Privacy budget set
N	amount of users
τ	The privacy protection level is τ
l_i	Privacy protection level of the i-th edge node
ε^τ	Privacy budget of level τ
G_t	Data collection with privacy budget ε^τ
X_u	Real binary vector of user u
Z_u	User u’s privacy data set
Z_u^τ	User u uses level τ plus noise privacy data
P	Histogram estimation of raw data
$\hat{P}$	Histogram estimation of raw data

3.2 Personalized Local Differential Privacy PLDP

Differential privacy protection ensures the indistinguishability between different inputs at the expense of certain data availability. Theoretically speaking, the higher the availability of data with low protection level, the higher the accuracy of providing services to users, but the data is easy to be leaked. On the contrary, the data with high protection level has low availability, but it can strictly limit privacy leakage risk. Therefore, in the real world, different user groups require different levels of protection for their own data. Some users want to reduce security in exchange for better service quality, and some users pay more attention to privacy and security. This requires the establishment of a personalized local differential privacy protection model for each user [9].

Personalized Local Differential Privacy, PLDP: Let $T=\{t_1,t_2,\ldots,t_k\}$ be a discrete and limited data attribute range, the user u has his own data $t\in T$ and the required differential privacy budget e^{ε^u} , among them, $\varepsilon^u\in\varepsilon=(\varepsilon^1,\varepsilon^2,\ldots,\varepsilon^m)$, $u\in(1,2,\ldots,n)$. If for any $t\in T$ of user u, the random algorithm M satisfies the following conditions:

$$Pr[M(t)=Z_u]\le e^{\varepsilon^u}\times Pr[M(t')=Z_u] \tag{1}$$

Then the random algorithm M satisfies ε_u -personalized local differential privacy for user u, where Z_u contains all private data of user u. Traditional LDP sets a global privacy protection budget for all users. The difference from traditional LDP is that the differential privacy budget in PLDP is determined by the user, This means that the control of privacy is returned to the user. In addition, personalized differential privacy has the properties of serial combination and post-processing invariance.

4 Key Technology Research

4.1 Personalized Differential Privacy Protection Based on OUE

According to the histogram estimation principle, if the histogram estimation is to be performed, the original data needs to be encoded first. There are many existing encoding methods, such as Direct Encoding (DE), Histogram Encoding (HE), Unary Encoding (UE), Binary Local Hashing (BLH), etc. In this paper, use optimized unary encoding (Optimized Unary Encoding, OUE), which is an optimized scheme of binary encoding. Compared with other encoding schemes, OUE has a smaller variance.

During the data encoding process, the data value $T=\{t_1,t_2,\ldots,t_k\}$ is usually discrete and finite, but it can also be a bounded continuous value. In this case, it is necessary to discretize the continuous data first, because the ultimate goal of data collection is to effectively statistical data distribution histogram, so the length of the histogram unit interval is taken as the discretization step size of the data value range and it will not affect the validity of the data and the accuracy of the estimates. For the sake of simplicity without loss of generality, only the case where the user takes the value of discrete data is discussed here.

First, for any edge node, assuming that the privacy protection level used by all users in its collection range is τ, the algorithm for the privacy protection output of any user within the node range is shown in Table 2.

The following will prove that OUE coding satisfies ε^τ -PLDP:

$$\frac{Pr[M(t)=Z_u^\tau]}{Pr[M(t')=Z_u^\tau]}=\prod_{j=1}^{k}\frac{Pr[M(t)[j]=Z_u^\tau[j]]}{Pr[M(t')[j]=Z_u^\tau[j]]}\le\frac{\frac{1}{2}\times(1-\frac{1}{e^{\varepsilon^\tau}+1})}{(1-\frac{1}{2})\times(1-\frac{1}{e^{\varepsilon^\tau}+1})}=e^{\varepsilon^\tau} \tag{2}$$

That is:

$$Pr[M(t)=Z_u^\tau]\le e^{\varepsilon^u}\times Pr[M(t')=Z_u^\tau] \tag{3}$$

Therefore, OUE satisfies e^{ε^τ} -PLDP for any user u.

Table 2. OUE algorithm based on personalize local differential privacy

```
Algorithm 1. OUE coding based on personalized local differential privacy
Input: User u takes value $t_i \in T = \{t_1, t_2, \ldots, t_k\}$, privacy protection level τ
Output: User u's encoded data $Z_u^\tau$
1. Initialize a k-dimensional all-zero vector $X_u = [0, 0, \ldots, 0]_k$
2. Set the i-th position of the vector to 1, $X_u[i] = 1$
3. for j = 1 to k
    $Pr[Z_u^\tau[j] = 1] = \begin{cases} p = \frac{1}{1}, & if\ X_u[j] = 1 \\ q = \frac{1}{e^{\varepsilon^\tau} + 1}, & if\ X_u[j] = 0 \end{cases}$
4. end for
5. Output $Z_u^\tau$
```

4.2 Edge node Histogram Estimation

After the edge node collects the user's private data within the range of the node, the edge node needs to analyze the privacy data in order to provide more accurate and personalized services to the users within the range. The data analysis here uses the method of histogram estimation.

Assuming that there are *s* users in any edge node, the user value range is: $T = \{t_1, t_2, \ldots, t_k\}$, the actual data distribution of these users is: $F = [f_1, f_2, \ldots, f_k]$, Where f_k represents the proportion of users whose data is t_k to the total number of users in the node. The statistical result of the data collected by the edge node is $S = [s_1, s_2, \ldots, s_k]$, Where s_k collects the number of "1" at the k-th position in the data. Obviously, the statistical result of the private data is not an unbiased estimate of the statistical result of the real data, which needs to be revised.

Take the statistical result s_k with the corrected data value t_k of as an example, for any user u:

$$Pr[Z_u^\tau[k] = 1] = f_k \times p + (1 - f_k) \times q \tag{4}$$

$$Pr[Z_u^\tau[k] = 0] = f_k \times (1 - p) + (1 - f_k) \times (1 - q) \tag{5}$$

The above ratio is not an unbiased estimate of the true ratio. Now the result is corrected by the maximum likelihood function. First, the maximum likelihood function should be constructed:

$$L = [f_k \times p + (1 - f_k) \times q]^{s_k} \times [f_k \times (1 - p) + (1 - f_k) \times (1 - q)]^{s - s_k} \tag{6}$$

Take the logarithm of the likelihood function to get:

$$\begin{aligned} \ln L = s_k & \ln(f_k \times p + (1 - f_k) \times q) \\ & + (s - s_k) \ln(f_k \times (1 - p) + (1 - f_k) \times (1 - q)) \end{aligned} \tag{7}$$

Derivation of f_k on both sides of the above equation:

$$\frac{d(\ln L)}{d(f_k)} = \frac{s_k \times (p - q)}{f_k \times p + (1 - f_k) \times q} + \frac{(s - s_k) \times (p - q)}{f_k \times (1 - p) + (1 - f_k) \times (1 - q)} \tag{8}$$

Then the maximum likelihood estimate of f_k $\hat{f}_k$ satisfies the following equation:

$$\frac{s_k \times (p - q)}{\hat{f}_k \times p + (1 - \hat{f}_k) \times q} = \frac{(s - s_k) \times (p - q)}{\hat{f}_k \times (1 - p) + (1 - \hat{f}_k) \times (1 - q)} \tag{9}$$

Solutions have to:

$$\hat{f}_k = \frac{s_k - q \times s}{s \times (p-q)} = 2\frac{s_k\left(e^{\varepsilon^{\tau}}+1\right)-s}{s(e^{\varepsilon^{\tau}}-1)} \tag{10}$$

Prove that $\hat{f}_k$ is an unbiased estimate of f_k:

$$\begin{aligned} E(\hat{f}_k) &= \frac{1}{p-q}[-q+\frac{1}{s}\sum_{u=1}^{s} E[Z_u^{\tau}[k]]] \\ &= \frac{1}{p-q}[-q+f_k \times p + (1-f_k)\times q] = f_k \end{aligned} \tag{11}$$

In the same way, according to the statistical results of edge nodes, an unbiased estimate of the transmission probability of any data within the data value range can be obtained. Then the histogram estimation result in the edge node should be:

$$\begin{aligned} \hat{P} &= [\hat{f}_1, \hat{f}_2, \ldots, \hat{f}_k] \\ &= [2\frac{s_1(e^{\varepsilon^{\tau}}+1)-s}{s(e^{\varepsilon^{\tau}}-1)}, \frac{s_2(e^{\varepsilon^{\tau}}+1)-s}{s(e^{\varepsilon^{\tau}}-1)}, \ldots, 2\frac{s_k(e^{\varepsilon^{\tau}}+1)-s}{s(e^{\varepsilon^{\tau}}-1)}] \end{aligned} \tag{12}$$

4.3 Data Derivation Algorithm OUE-DRPP

In Data Recycle with Personalized Privacy (DRPP), the calculation of derived data is closely related to the disturbance mode of the encoding. In the data encoding method, OUE has a smaller variance. This article combines OUE and DRPP together, denoted as OUE-DRPP (Data Recycle with Personalized Privacy based on Optimized Unary Encoding). The specific steps of the OUE-DRPP algorithm are given in Table 3.

Table 3. OUE-DRPP Algorithm

Algorithm 2. OUE-DRPP algorithm

Input: privacy data set $G = [G_1, G_2, \ldots, G_{\tau}, \ldots, G_m]$ Privacy level set $[1, 2, \ldots, \tau, \ldots, m]$

Output: derived data set G^+

```
for $\tau$ = 1 to m :
    initialization: $G_{\tau}^{+} = G_{\tau}$
    for r > $\tau$ to m:
        $G_{\tau}^{+} = G_{\tau}^{+} \bigcup DR(G_{\gamma}, \tau)$
    end for
end for
return $G_{\tau}^{+}$
```

In the DRPP algorithm, the cloud computing center first groups the collected privacy data according to privacy levels, and establishes a privacy data set for each privacy level, denoted as $G = [G_1, G_2, \ldots, G_{\tau}, \ldots, G_m]$, and add all private data with privacy level τ to the data set G_{τ}. In addition, the derived data sets of each privacy level are recorded as $G^{+} = [G_1^{+}, G_2^{+}, \ldots, G_{\tau}^{+}, \ldots, G_m^{+}]$, $\alpha_{\tau} = p_{\tau} - q_{\tau} = \frac{1}{2} - \frac{1}{(e^{\varepsilon^{\tau}}+1)}$.

The following will introduce the process of data derivation algorithm through the example shown in Fig. 3:

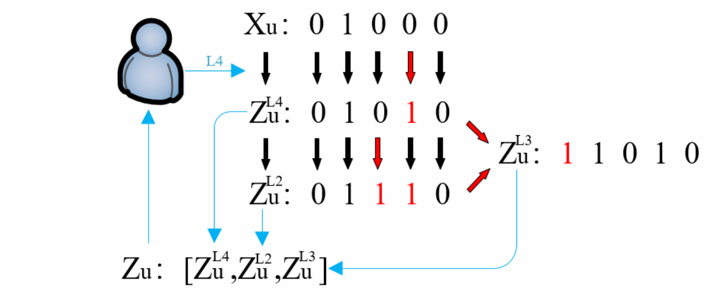

Fig. 3. The instance of data derivation

For users of edge computing nodes with a privacy protection level of L4, Firstly according to the privacy budget corresponding to the privacy protection level, the data is generated by OUE encoding locally, and the privacy data $Z_u^{L_4}$ with the protection level L4 is generated and sent to the edge computing node to which it belongs.

When the cloud computing center needs to count user information, it will obtain data from edge computing nodes. The edge computing node sends the collected privacy data and the privacy protection level of the node to the cloud computing center. After the cloud computing center collects the privacy data of different levels from each edge node, it uses a data derivation algorithm to improve the accuracy of the estimation, and derives a data version with a high privacy level from the data with a low privacy protection level.

4.4 Cloud Private Data Histogram Estimation and Estimation Error

Due to the requirements of differential privacy protection, only users with loose privacy protection requirements can provide valid additional data. Suppose that when the cloud collects private data, use n_τ to represent the number of samples provided by users with a privacy protection level of τ, and $N=\sum_{\tau=1}^{m} n_\tau$. Then the number of samples whose privacy protection level is τ after data derivation is: $n_\tau^+ = \sum_{i=\tau}^{m} n_i$.

The following takes the histogram estimation under the privacy protection level τ as an example for illustration:

The sample size of τ after data derivation is:, The user's value range is still: $T=\{t_1, t_2, \ldots, t_k\}$, The distribution of the actual user data is: $P=[p_1, p_2, \ldots, p_k]$, Where p_k^τ represents the proportion of samples with data t_k to the current total number of samples n_τ^+ . The data statistics result is $S_\tau^+ = [s_1^\tau, s_2^\tau, \ldots, s_k^\tau]$, Where s_k^τ represents the number of samples whose k-th position is "1". According to the conclusion in the above, in the same way, the unbiased estimator of $P[j], j \in k$ can be obtained as:

$$\hat{P}_\tau[j] = \frac{s_\tau^+[j] - qn_\tau^+}{n_\tau^+(p-q)} = 2\frac{s_\tau^+[j](e^{\varepsilon^\tau}+1) - n_\tau^+}{n_\tau^+(e^{\varepsilon^\tau}-1)} \tag{12}$$

The histogram estimation result for the privacy protection level τ should be:

$$\begin{aligned}\hat{P}_\tau &= [\hat{p}_1^\tau, \hat{p}_2^\tau, \ldots, \hat{p}_k^\tau] \\ &= [2\frac{s_1^\tau(e^{\varepsilon^\tau}+1)-n_\tau^+}{n_\tau^+(e^{\varepsilon^\tau}-1)}, \frac{s_2^\tau(e^{\varepsilon^\tau}+1)-n_\tau^+}{n_\tau^+(e^{\varepsilon^\tau}-1)}, \ldots, 2\frac{s_2^\tau(e^{\varepsilon^\tau}+1)-n_\tau^+}{n_\tau^+(e^{\varepsilon^\tau}-1)}]\end{aligned} \tag{13}$$

The Mean-Square Error (MSE) is usually used to evaluate the estimation error of histogram estimation. The definition of the mean square error is as follows:

Mean square error: Mean square error is a measure of the degree of difference between the estimator and the estimator. Assuming that $\hat{\theta}$ is an estimator of the overall parameter θdetermined according to the sample, then the mathematical expectation of $(\hat{\theta}-\theta)^2$ is called the mean square error of the estimator $\hat{\theta}$, so the mean square error is expressed as follows:

$$MSE(\hat{\theta}) = E(\hat{\theta}-\theta)^2 \tag{14}$$

In the histogram estimation, the mean square error can be expressed in the following form:

$$MSE(\hat{P}) = E(\hat{P}-P)^2 \tag{15}$$

Among them, $\hat{P}=[\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_k]$ represents the histogram estimation result obtained from private data, $P=[p_1, p_2, \ldots, p_k]$ represents the sending of real data frequency. Then the mean square error of the histogram estimation at any privacy protection level τ is:

$$\begin{aligned} MSE(\hat{P}_\tau) &= E(\hat{P}_\tau - P)^2 = E[\| \hat{P}_\tau - P \|_2^2] = E[\sum_{j=1}^{k} (\hat{P}_\tau[j] - P[j])^2] \\ &= \sum_{j=1}^{k} (E[\hat{P}_\tau^2[j] - P^2[j])^2 = \sum_{j=1}^{k} Var[\hat{P}_\tau^2[j]] \end{aligned} \tag{16}$$

Since for any $j \in k, \hat{P}_\tau[j]$ is an unbiased estimate of $P[j]$, that is:

$$E[\hat{P}_\tau[j]] = E[\frac{s_\tau^+[j] - qn_\tau^+}{n_\tau^+(p-q)}] = P[j] = p_j \tag{17}$$

Then the variance of $\hat{P}_\tau[j]$ can be calculated as follows, because the value of each user is independent and uncorrelated, then:

$$\begin{aligned} Var[\hat{P}_\tau[j]] &= Var[\frac{s_\tau^+[j] - qn_\tau^+}{n_\tau^+(p-q)}] = \frac{n_\tau^+ Var[Z_u^t[j]]}{n_\tau^{+2}(p-q)^2} \\ &= \frac{[p \cdot p_j + (1-p_j) \cdot q][(1-p_j) \cdot p + p_j \cdot q]}{n_\tau^+(p-q)^2} \\ &= \frac{pq - (p_j{}^2 - p_j)(p-q)^2}{n_\tau^+(p-q)^2} \\ &= n_\tau^+[\frac{pq}{(p-q)^2}(p_j{}^2 - p_j)] \\ &= n_\tau^+[\frac{2(e^{\varepsilon^\tau}+1)}{(e^{\varepsilon^\tau}-1)^2} - (p_j{}^2 - p_j)] \end{aligned} \tag{18}$$

Then substituting formula 18 into formula 14, we get:

$$\begin{aligned} MSE(\hat{P}_\tau) &= \sum_{j=1}^{k} Var[\hat{P}_\tau[j]] \\ &= \sum_{j=1}^{k} \frac{1}{n_\tau^+}[\frac{2(e^{\varepsilon^\tau}+1)}{(e^{\varepsilon^\tau}-1)^2} - (p_j{}^2 - p_j)] \\ &= \frac{\frac{2k(e^{\varepsilon^\tau}+1)}{(e^{\varepsilon^\tau}-1)^2}}{n_\tau^+} + \frac{1 - \sum_{j=1}^{k} p_j{}^2}{n_\tau^+} \end{aligned} \tag{19}$$

Observing the mean square error function, we can find that since $\sum_{j=1}^{k} p_j = 1$, then $0 < 1 - \sum_{j=1}^{k} p_j{}^2 < 1$, Therefore, regardless of the distribution of the user's value, the range of the mean square error of the histogram estimation under a fixed sample size, privacy protection level and value range is $\frac{1}{n_\tau^+} \ll \frac{\frac{2k(e^{\varepsilon^\tau}+1)}{(e^{\varepsilon^\tau}-1)^2}}{n_\tau^+}$. Therefore, the decisive factors for the histogram estimation are the privacy protection level τ, the number of expanded samples of the privacy level τ, and the value range of user data.

4.5 Optimized Data Derivation Algorithm OUE-ODRPP

Because the data collected by the cloud computing center is subject to different levels of personalized differential privacy protection, if the overall privacy data is statistically analyzed in the cloud center, the histogram estimation can only be done in the samples of different privacy levels. The problem is that it is difficult to effectively use the effective information in the overall data, resulting in large estimation errors. In the above, the data derivation method is used to derive the corresponding high-privacy samples from the data with low privacy requirements, which increases the number of samples of all privacy levels except the lowest privacy level, and makes full use of the effective information of the overall data. However, even if use the data derivation method, the histogram estimation can only be completed under each privacy level, and then the histogram estimation with the lowest mean square error is found as the histogram estimation result of the overall data.

Through the analysis of the mean square error function, it can be observed that in the mean square error of the histogram estimation of all privacy levels, there is an optimal privacy level, and the mean square error of the histogram estimation is the smallest under this privacy level. Through the analysis of the error function in the previous section, it is found that the term that plays a decisive role in the mean square error function is $\frac{\frac{2k(e^{\varepsilon^\tau}+1)}{(e^{\varepsilon^\tau}-1)^2}}{n_\tau^+}$. Where k represents the number of all possible values for the user, and has nothing to do with the privacy level, n_τ^+ represents the number of samples with a privacy level τ derived from data, that is, the total number of users whose privacy level is greater than τ. Therefore, before performing the data derivation algorithm, first calculate the $\frac{\frac{(e^{\varepsilon^\tau}+1)}{(e^{\varepsilon^\tau}-1)^2}}{n_\tau^+}$ value. Choosing the privacy level with the smallest value, and then only deriving data for this level, this can greatly reduce the amount of calculation of the data derivation algorithm and ensure the best histogram estimation error. Suppose there are a total of h edge nodes, and the number of users under each edge node is $u_i, i \in h$. Use OUE-ODRPP to represent the optimized algorithm. The algorithm steps are shown in Table 4:

In the optimization algorithm, the cloud computing center first collects the number of users and privacy budgets of each edge node, and then calculates the decision items that affect the estimation error to obtain the privacy protection level. This level can achieve the smallest estimation error. The cloud computing center collects data from all edge nodes whose privacy level is not less than *v*, and the user data of edge nodes whose privacy level is less than v cannot provide effective information for the histogram estimation under privacy level *v*, so there is no need to collect data, this way also reduces data Number of transfers.

After collecting the data in the cloud computing center, according to lines 8-19 in the OUE-ODRPP algorithm, all the collected data only needs to derive the privacy data with a privacy level of *v*. Finally, all available samples with a privacy level of *v* are obtained, and then the histogram estimation in the case of a privacy level of *v* is calculated.

According to the prior error analysis, the optimized algorithm greatly reduces the calculation amount of the data derivation algorithm. The calculation amount of the optimized algorithm is related to the selection of the privacy protection level to obtain the smallest mean square error. In the simulation experiment, the time spent by the algorithm is used as the algorithm Optimized metrics.

Table 4. OUE-DRPP algorithm

Algorithm 3. OUE-ODRPP

```
Each edge node sends the number of users and privacy level of the node $< u_i, l_i >$
for $\tau = \min(l_i, i \in h)$ to $\max(l_i, i \in h)$
    $n_\tau^+ = \Sigma(u_i \mid l_i > \tau)$
    $o_\tau = \frac{e^{\varepsilon^\tau} + 1}{n_\tau^+ (e^{\varepsilon^\tau} - 1)^2}$
end for
$v = \{\tau \mid o_\tau = \min(o_{\min(l_i, i \in h)}, ..., o_\tau, ..., o_{\max(l_i, i \in h)})$
The cloud center collects all edge node data Z with privacy level $l_i \geq v$
for $Z_u$ in Z
    $i = inf\{i \mid Z_u^i \in Z_u$ and $i \geq v\}$
    $Z_{sup} = Z_u^i$
  if $i = v$
      $G_v^+ \; add \; Z_u^v$
  else
    for d = 1 to k:
        $Z_u^v[d] = \begin{cases} Z_{sup}, & \frac{a_i + a_v}{2a_i} \\ 1 - Z_{sup}, & \frac{a_i - a_v}{2a_i} \end{cases}$
    end for
    $G_v^+ \; add \; Z_u^v, Z_u \; add \; Z_u^v$
  end if
end for
sum($G_v^+$) = $S_v^+$ = $[s_1^v, s_2^v, ..., s_k^v]$
for j = 1 to k :
$\hat{P}_v = \frac{s_j^v - qn_v^+}{n_v^+(p - q)} = 2\frac{s_j^v(e^{\varepsilon^v} + 1) - n_v^+}{n_v^+(e^{\varepsilon^v} - 1)}$
end for
return $\hat{P}_v = [\hat{p}_1^v, \hat{p}_2^v, ..., \hat{p}_k^v]$
```

5 Optimized Data Derivation Algorithm OUE-ODRPP

This article uses the Adult data set, one of UCI's most popular data sets. This data mainly records data such as job type, marital status, and education level. There are 48,842 instances in total. The system environment of the simulation experiment is Intel(R) Core(TM) i5-7300HQ CPU, 2.50 GHz, 16.0 G RAM, windows10 professional 64-bit operating system, and the simulation tool is JetBrains PyCharm 2019.2.2×64.

First of all, in order to verify the effectiveness of OUE coding for the enhancement algorithm, this paper sets the level of personalized differential privacy protection to 10 different levels from L1 to L10, L1 represents the highest level (minimum privacy protection budget), and L10 represents the lowest level (Privacy protection budget is the largest), and the privacy protection budget of level L1 is 0.1, and then the privacy protection budget of each privacy level increases by 0.1, so the privacy protection budget corresponding to these 10 privacy protection levels is 0.1~1. This experiment assumes that 10 privacy levels are uniformly selected by users, that is, the number of users under each privacy level is equal. This paper compares the mean square error of DRPP and OUE-DRPP under 10 privacy protection levels. The experimental results are shown in Fig. 4:

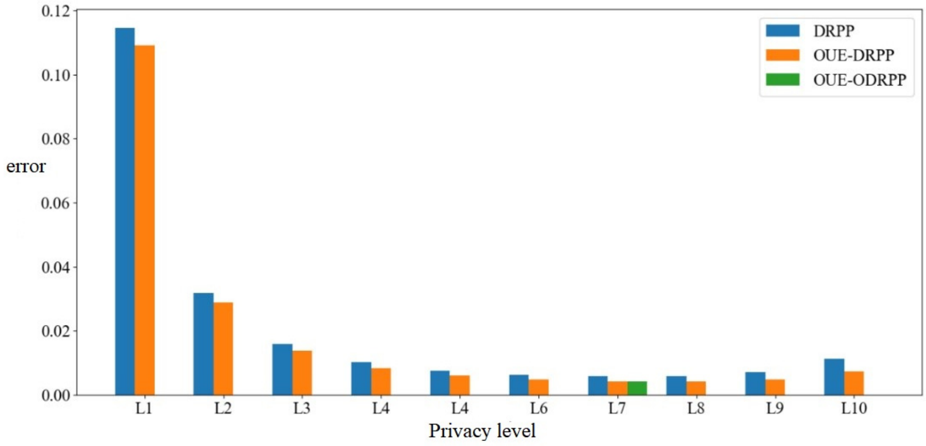

Fig. 4. Comparison of mean square error

It can be seen from the experimental results that at any level of privacy protection, the mean square error of the OUE-DRPP algorithm is always better than the mean square error of the DRPP algorithm. This is because the data disturbance obtained by OUE coding by solving the minimum value of the coding variance Probability, so the noise variance generated by coding is smaller than the noise variance generated by UE coding.

On the other hand, among the 10 privacy levels set, the mean square error of the histogram estimation under the L7 level is the smallest. This proves that after using the data to derive the extended sample, there must be a minimum in the mean square error of the histogram estimation under different differential privacy levels. Just find the privacy protection level corresponding to the minimum value first, and then only expand the data of this privacy level in the data derivation algorithm, which greatly reduces the amount of calculation.

In order to verify the ability of OUE-ODRPP to reduce the computational complexity of the algorithm, this article first uses random extraction to divide Adult data into 10 subsets arithmetic, the amount of data is {4884, 9768, 14652, 19536, 24420, 29304, 34188, 39072, 43956, 48840}, because the calculation amount of OUE-DRPP and DRPP algorithms are the same, in order to compare the calculation amount of the optimized algorithm and the original algorithm under the condition of obtaining the same minimum mean square error, in the experiment, DRPP algorithm, OUE-DRPP algorithm and OUE-ODRPP algorithm are used for comparison. The DRPP algorithm, the OUE-ODRPP algorithm, and the OUE-DRPP algorithm are used to estimate the histogram of these 10 subsets, and it is still assumed that the amount of users for each privacy level is the same. Under the condition that the DRPP algorithm, the ODRPP algorithm and the OUE-DRPP algorithm have the same minimum mean square error respectively, compare the time spent by the DRPP algorithm, OUE-ODRPP and DRPP. The experimental results are shown in Fig. 5.

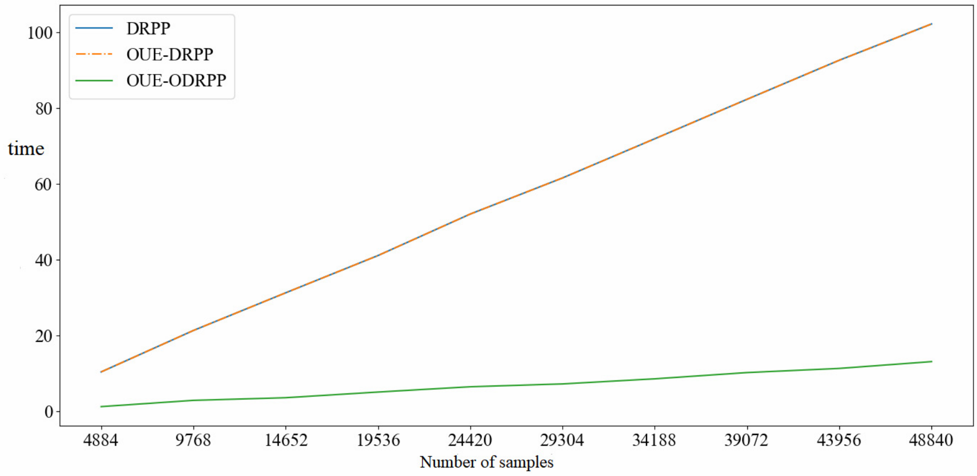

Fig. 5. Comparison of time spent

The experimental results show that the DRPP algorithm has the same amount of calculation as the OUE-DRPP algorithm, and the OUE-ODRPP algorithm is better than the OUE-DRPP algorithm in terms of calculation amount, especially when the number of samples is huge, the advantages of OUE-ODRPP are more obvious. The reason is that the OUE-DRPP algorithm and the DRPP algorithm need to derive the data of all privacy levels except the lowest privacy level, even if the estimation error under most privacy levels is large. The OUE-ODRPP algorithm calculates the decisive terms of the error function under different privacy levels to obtain the privacy protection level which can obtain the optimal error, so that in the data derivation algorithm, only derive the required data, avoiding the useless derivation or low-availability data occupies computing resources, which greatly reduces the overall calculation amount of the algorithm.

6 Summary

Aiming at the privacy protection problem in the cloud-side collaboration scenario, this paper proposes a distributed and personalized local differential privacy protection model, and uses the OUE scheme to optimize the noise variance in the data encoding to reduce the estimation error. In addition, in order to solve the problems of reduced data effectiveness and excessive estimation errors caused by different privacy protection levels in the overall estimation, introduce a data expansion strategy based on data derivation technology, which does not require additional information and does not reduce other privacy levels. Under the premise of the amount of data, the amount of data of a certain privacy level is expanded to increase the estimation accuracy under this level. By calculating the privacy level corresponding to the optimal error a priori, the calculation amount of the original data derivation algorithm can be reduced, and the optimization of the calculation amount is very obvious when the amount of data is large. Finally, the algorithm is simulated with the same data set as the control algorithm, which verifies the correctness of the theory and the optimization effect of the algorithm. In the future, with the development of technology, people will pay more attention to data privacy, and the demand for personalized privacy will become more diversified. How to meet the personalized privacy needs of users while obtaining better data statistics is still a research focus.

Acknowledgements

The research of the authors was supported by Industrial Internet innovation and development project of MIIT, China.

References

[1] W. Shi, J. Cao, Q. Zhang, Y. Li, L Xu, Edge Computing: Vision and Challenges, IEEE Internet of Things Journal 3(5)(2016) 637-646.

[2] Q. Miao, W. Jing, H. Song, Differential privacy–based location privacy enhancing in edge computing, Concurrency and Computation: Practice and Experience 31(8)(2019) e4735.

[3] J.L. Zhang, Y.C. Zhao, B. Chen, F. Hu, K. Zhu, Survey on data security and privacy-preserving for the research of edge computing, Journal on Communications 39(3)(2018) 1-21.

[4] F.H. Li, H. Li, Y. Jia, N.H. Yu, J. Weng, Privacy computing: concept, connotation and its research trend, Journal on Communications 37(4)(2016) 1-11.

[5] Z. Zhang, W. Zhang, H.-C. Chao, C.-F. Lai, Toward Belief Function-Based Cooperative Sensing for Interference Resistant Industrial Wireless Sensor Networks, IEEE Transactions on Industrial Informatics 12(6)(2016) 2115-2126.

[6] C. Dwork, Differential Privacy: A Survey of Results, in: Proc. International Conference on Theory and Applications of Models of Computation, 2008.

[7] X. Gu, M. Li, L. Xiong, Y. Cao, Providing Input-Discriminative Protection for Local Differential Privacy, in: Proc. 2020 IEEE 36th International Conference on Data Engineering (ICDE), 2020.

[8] Y. Qu, L. Cui, S. Yu, W. Zhou, J. Wu, Improving Data Utility through Game Theory in Personalized Differential Privacy, in: Proc. 2018 IEEE International Conference on Communications (ICC 2018), 2018.

[9] F. Li, H. Song, J. Li, Personalized Data Collection Based on Local Differential Privacy in the Mobile Crowdsensing, in: Proc. 2020 IEEE 6th International Conference on Computer and Communications (ICCC), 2020.

[10] W. Huang, S. Zhou, T. Zhu, Y. Liao, C. Wu, S. Qiu, Improving Laplace Mechanism of Differential Privacy by Personalized Sampling, in: Proc. 2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), 2020.

[11] J. Soria-Comas, J. Domingo-Ferrer, D. Sánchez, D. Megias, Individual Differential Privacy: A Utility-Preserving Formulation of Differential Privacy Guarantees, IEEE Transactions on Information Forensics and Security 12(6)(2017) 1418-1429.

[12] K. Nissim, S. Raskhodnikova, A. Smith, Smooth sensitivity and sampling in private data analysis, in: Proceedings of the 39th Annual ACM Symposium on Theory of Computing, 2007.

[13] F. Liu, Generalized Gaussian Mechanism for Differential Privacy, IEEE Transactions on Knowledge and Data Engineering 31(4)(2019) 747-756.