

Hierarchical DP-K Anonymous Data Publishing Model Based on Binary Tree

YUXIAO XIA^a, TAO ZHAO^a, YANLI LV^a, YUNZHAO LI^a, RUXIA YANG^{bc}

^a Big Data Center of State Grid Corporation of China

^b State Grid Key Laboratory of Information & Network Security, Nanjing 210003, China;

^c State Grid Smart Grid Research Institute Co., Ltd., Nanjing 210003, China

Corresponding Author:48838968@qq.com

Abstract— With the acceleration of data opening and sharing process in the power industry, the risk of sensitive data leakage is also gradually increasing. Privacy protection is one of the hot issues of privacy leakage control technology research in data release, and k-anonymity is the hot topic of privacy protection research in recent years. In this paper, we propose a hierarchical DP-K anonymous data release model based on binary tree clustering for the existing k-anonymity scheme and for minimizing the amount of information loss. A binary tree-based clustering algorithm (BTCA) is proposed to classify similar data records into the same equivalent class, which can improve the effect of clustering, reduce the information loss caused by anonymous data set release, and improve the availability of data. The clustered anonymous data sets redistribute different privacy budgets according to the privacy rights of the quasi-identifier attributes, and realize the hierarchical protection of the data with different degrees of sensitivity through the differential privacy noise increase mechanism, which enhances the privacy of the data.

Keywords— K-Anonymous; differential privacy; clustering; binary tree; privacy protection

I. INTRODUCTION

As a new factor of production in the era of digital economy, data realizes value creation in the flow and sharing, and at the same time, it also faces a huge risk of leakage. Power data is a vital tangible assets of enterprises, it involves regulation data, marketing data, production equipment data, human and property management data. With the continuous development of the company's data external service business, the power data application and value mining such as data release have become a rigid demand. Data security and privacy protection are the core issues of power data application and value mining process. Power grid data contains a large number of users' personal privacy information and commercial sensitive data, which will not only bring huge economic losses but also bear relevant legal responsibilities. Therefore, data need to be anonymized during data opening to protect sensitive data from leakage. With the development of ubiquitous power Internet of things, electric power data open in information perception, interconnection, open sharing will have a huge role, and more liquidity, big data, cloud computing technology give full play

to the advantages, power data energy and value will be more deeply mining and release [1], through open power data can maximize use power grid enterprise equipment, network, customer and platform advantage, power energy ecosystem sharing. But on the other hand, for the purpose of information sharing and data mining in the process of data release privacy is also increasingly prominent, if the attacker to released data reverse reasoning attack, restore the original data, the privacy of enterprise operation data, account information will leak or damage, cause huge economic losses to enterprises, so how to realize the information sharing at the same time, effectively protect private sensitive information not leakage is particularly important. Data publishers release data to protect sensitive information on the data set, and protect the private data without destroying the required data information and data characteristics. There are currently three main methods to protect privacy data: (1) K-anonymity. The most widely used technology today is k-anonymity, where companies hide unique identities such as ID numbers or names when publishing user data to protect users' privacy. However, this way cannot resist link attacks and background knowledge attacks; (2) Data disturbance. This method is to use scrambled, distortion, randomization and other technologies to disrupt the original data, so that the data loses the authenticity and integrity, the attacker can not obtain the real data, but the availability of the data is greatly reduced; (3) Data encryption. Combine cryptography with data security, and use homomorphic encryption and asymmetric encryption technology to form distributed security computing to support the work of privacy protection. For example, secure multi-party calculation, but the problem of this method is that it requires too much computing resources, and the cost is very high. Thus, the three methods have some problems, cannot effectively prevent data open the privacy in the process of leakage problem, as the country in recent years continue to strengthen data security and personal privacy regulation, released on September 1, 2021, the data security law, how to ensure data legal compliance use, prevent enterprise important data and personal privacy, is the current digital transformation, mining power data value facing major challenges.

II. RELATED STUDIES

K-anonymous Technology was created in 1998 by P. Samarati and L. Sweeney, it aims to protect private information when the data is released [2]. In 2002, L. Sweeney further proposed the K-anonymous privacy protection model [3], to realize this technology by generalization and hiding. Deng Jingjing [4] proposed a multi-dimensional data K anonymous solution based on R tree to meet the requirements of dynamic data, real-time response and other characteristics. However, how to achieve the privacy protection of published data more efficiently through anonymization technology is still an important topic of current research in the industry. k-anonymity is a typical representative of the anonymous data release method, which protects data privacy by grouping and generalizing the quasi-identifier attributes of published data, making each record indistinguishable from at least k-1 other records, etc. For different properties, or, different columns, for a dataset. According to its function, it can be roughly divided into three types, identifiers (ID), quasi-identifiers (QI), and sensitive data (SA). The general research goal of data privacy protection is to desensitization identification attributes on the basis of deleting identification attributes to maintain the balance between privacy and availability. Identifiers can uniquely determine a data, and combinations of different quasi-identifiers may also identify a data. As the name suggests, privacy data is the object of privacy protection. Take the January data of a power user in Nanjing as an example, see Table 1:

TABLE 1. JANUARY DATA OF A POWER USER IN NANJING

ID	sex	name	age	area	electricity level (One / two / three)	electricity consumption	assessment
1	man	Zhang San	46	Gulou district	two	1000 Degrees	1500

For example, in the table above table, ID number and name are identifiers, while gender, age, area and level of electricity can be used as quasi-identifiers. They have certain information and may determine the owner of this data after being combined. Electricity consumption, payable, is the private information that we need to protect.

The first step in k-anonymous data anonymity is to remove any explicit identifier that can directly define personal identity. Then, the generalization technique makes each record in the published data set indistinguishable from the set of k1 records sharing the same quasi-identifier values, called "equivalence classes." For example, in the table above (sex, age, area, electricity level) as the quasi identifier, so to (man, 46, Gulou district, two) as the condition, to the data set search,

eventually get at least k data, similar, with any other possible conditions to search, finally get the data the number of is greater than equal to k, so say the data set meet k anonymous. The K-Anonymous technique guarantees the following three points:

- (1) Anti-member reasoning attacks, the attacker cannot determine whether a target person is in publicly available data
- (2) Sensitive information protection, given a person an attacker cannot confirm whether he has a sensitive attribute
- (3) For the reverse analysis attack, the attacker cannot confirm the person corresponding to a piece of data.

The k-anonymity mechanism is the main method for the use of data desensitization in shared data release, which can effectively prevent background attacks and link attacks against private information. However, all existing schemes balance the privacy and availability of the data by finding optimal k-equivalent sets, and globally, k-equivalent sets are not necessarily optimal equivalent sets satisfying k-anonymity, and the usability optimization problem of the privacy mechanism remains unresolved for [5]. The k-anonymous model remains many drawbacks, such as vulnerability to homogeneity attacks when sensitive properties lack diversity. In 2006, Machanavajjhala [6] proposed the l-diversity model, which ensures that the sensitive properties of each equivalent class in a data table have at least l different values, thus enabling the published data sheet to defend against homogenous attacks. However, since l-diversity does not consider the semantic meaning of sensitive values, there may be equivalence sets of sensitive attribute value distribution very similar and susceptible to similarity attacks. For similarity attacks, in 2007, Li [7] proposed a t-closeness model that specifies that the distribution of sensitive attribute values in each equivalent class is as close as possible to the distribution in the original data table. The subsequent l-diversity and t-closeness anonymous models, which can protect the private information in the data to some extent, are only effective when assuming that the attacker does not have any background knowledge.

In order to enhance the protection of sensitive information, Cao Minzi [8] on k-anonymous and l-diversity of the anonymous model, and put forward a personalized (α, l) -diversity k-anonymous model, according to the different degree of sensitivity, the sensitive properties of value classification, set the corresponding constraints, provide personalized privacy protection for specific individuals. [9] et al. introduced the hierarchical classification of sensitive attributes and the concept of mutual information quantity, and proposed the (p, θ) k-anonymous model for equivalence attacks, which reduces the probability of sensitive information leakage.

In recent years, with the wide application of machine learning in various fields, the introduction of clustering algorithms into data release privacy protection technology has attracted wide attention to [10][11]. By clustering the data in the original data set first, and then anonymously processing it again, the loss of the data information can be effectively

reduced, and the availability of the data can be improved. The clustering-based SAC algorithm [12], can be used to overcome temporal attacks on published data to protect k-Anonymity-based privacy information. Compared with the traditional clustering method, the combination of density clustering and partition clustering eliminates some isolated point noises in the dataset, selects the initial clustering center according to the density characteristics of the dataset, and uses partition clustering iteration to achieve optimal clustering [13].

Differential privacy [14], as a publishing method based on data distortion, is widely studied and applied. Because it does not make any assumptions about the background knowledge of the attacker, and adds some available noise to disturb the published data, so as to provide a strong privacy guarantee. The combination of differential privacy and clustering algorithm for anonymous dataset publishing is a good privacy protection method. Shao Bozhang [15] and others introduced local differential privacy technology and proposed a trusted third-party K-mode clustering data privacy protection method for the real trusted third-party privacy protection model, which also has the risk of privacy disclosure. Nguyen [16] proposed a data release algorithm combining differential privacy and k-modes clustering method, which greatly improved the availability of type attribute data. However, because the initial clustering center is selected randomly selected, the clustering effect is not very good. In order to solve the problem of the initial center point [17], improve k-prototype clustering differential privacy mixed data release algorithm, combined with the local density and high density, differential privacy protection, improve the availability of data, but the privacy budget equalization principle, may cause differential privacy distribution of different properties, and reduce data availability. Yuan Quan et al. [18] aimed at the problem of excessive noise addition caused by the traditional differential privacy protection algorithm, combined with the spectrum clustering algorithm, clustered the social network weights into different clusters, randomly added noise for different clusters, reduced the amount of noise added, and determined the amount of noise added according to the size of the social network edge weight, realizing more balanced privacy protection. The above research only in part of the attributes of the data set used K-anonymous technology, the real data set attributes more attributes, the sensitivity of some attributes in the data set is low, the information release less impact on users, but hide all attributes, meet K-anonymous condition attributes to generalization, the greater the degree of generalization, the lower the accuracy of the data, will seriously affect the availability of data.

In order to solve this problem, this paper proposes a hierarchical DP-K anonymous data release model based on binary tree clustering, which aims to reduce the information loss caused by anonymous data set release, improve the availability of data, through the differential privacy noise mechanism to achieve different sensitivity of data hierarchical protection, enhance the privacy of data.

III. HIERARCHICAL DP-K ANONYMOUS MODEL BASED ON BINARY TREE CLUSTERING

The hierarchical DP-K anonymous model based on binary tree clustering is shown in Figure 1. Dividing similar data records into the same equivalent class through binary tree clustering algorithm can improve the clustering effect, reduce the information loss caused by anonymous data set release and improve the data availability. The clustered anonymous data sets are reallocated to different privacy budgets according to the privacy rights of the QI attribute, and the hierarchical protection of the data with different degrees of sensitivity can be realized through the differential privacy noise increase mechanism, which enhances the privacy of the data.

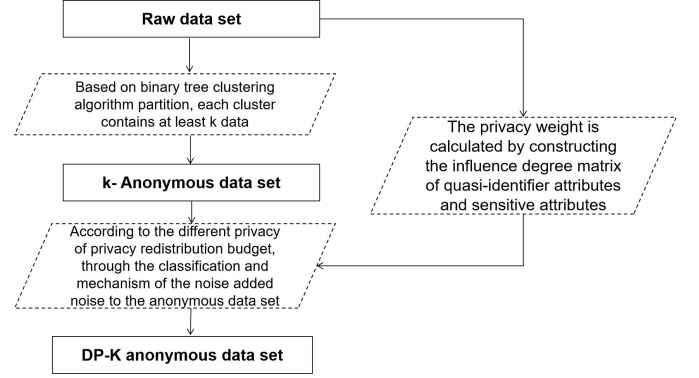


Figure 1. A hierarchical DP-K anonymous model based on binary tree clustering

A. Privacy right recalculation of QI attributes

1) The principle of privacy weight-computation:

Attackers infer sensitive attributes through a combination of quasi-identifiers, and different quasi-identifiers are also more related to sensitive attributes. Obviously, quasi-identifier properties that can reason about sensitive properties should have more privacy protection. This paper takes the correlation degree of quasi-identifier attributes and sensitive attributes as the basis for dividing the privacy protection degree of quasi-identifier attributes. The sensitive attributes are taken as the classification results, and the sensitive attributes are inferred through the combination of the quasi-identifier attributes. When there are multiple sensitive attributes in the data table, the correlation matrix of the quasi-identifier attributes and the sensitive attributes is established. In Table 2, for example:

TABLE 2. STATISTICS OF USER ELECTRICITY CONSUMPTION IN NANJING

sex	age	area	Electricity level	electricity consumption	power rate
man	55	Gulou district	one	100	<500
man	25	Gulou district	three	200	500~1000
woman	35	Xuanwu district	three	300	>1000

Break electricity charges as a sensitive attribute into three categories: ① low (<500) ② centre (500~1000) ③ tall (>1000), all quasi-identifiers are used to predict the classification of the classification accuracy, then delete each quasi-identifier in turn, and use the remaining quasi-identifiers to predict the electricity fee attribute classification to calculate the classification accuracy. The difference in the two classification accuracy indicates the extent to which the quasi-identifier attribute affects predicting the correct classification of the sensitive attribute.

2) Algorithm implementation:

The prediction of the sensitivity of the data field attributes was first made, and the random forest algorithm was used to calculate the effect of each attribute of the data on the sensitive attribute. Suppose a piece of data has n quasi-identifier QI attributes ($A_1, A_2, \dots, A_n$), m Sensitive attributes ($B_1, B_2, \dots, B_m$). First, calculate the average accuracy P_1 of predicting sensitive attribute B_1 through n QI attributes, and then calculate the average accuracy P_2 of predicting sensitive attributes with the remaining $n-1$ QI attributes after deleting the attribute. The difference between the two indicates the impact of the attribute on the prediction of sensitive attributes. The greater the difference, the greater the impact. Therefore, we can also be used to represent the sensitivity of this QI attribute to sensitive properties $\Delta P(A_i, B_j)$:

$$\Delta P(A_i, B_j) = P_1 - P_2 \quad (1)$$

Calculate the sensitive attribute B_j for QI attributes in turn. After sensitivity, normalization is performed to calculate the privacy weight w_{ij} of each QI attribute, that is:

$$w_{ij} = \frac{\Delta P(A_i, B_j)}{\sum_{i=1}^n \Delta P(A_i, B_j)} \quad (2)$$

After the prediction of Sensitive attributes ($B_1, B_2, \dots, B_m$), all the sensitivities of QI attribute and sensitive attributes form the $m \times n$ privacy weight matrix, that is:

$$W = w_{ij} (m \times n) = \begin{bmatrix} w_{11} & w_{12} & \dots & w_{1n} \\ w_{21} & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ w_{m1} & \dots & \dots & w_{mn} \end{bmatrix} \quad (3)$$

B. Cluster algorithm based on the binary tree

1) Distance Definition:

Because the values of different numerical data attributes have different orders of magnitude, such as the power consumption grade, the interval is [1~3], and the power consumption interval is [0~10,000], in the distance measurement, the weight of the power consumption attribute leads to the shift of the cluster center, which affects the clustering effect. To solve this problem, we define any two

numerical attribute values A_i, A_j , and the distance between them as

$$d_N(A_i, A_j) = \frac{|A_i - A_j|}{A_{\max} - A_{\min}} \quad (4)$$

$A_{\max}, A_{\min}$ represent the maximum and minimum values in the finite numerical field of the continuous type.

In addition to numerical data, there are also sub-type data, which is usually a finite number or discrete properties, and there is no sequential relationship between the value of the attribute values, so the definition of the distance between numerical data is not applicable to sub-type data. However, there is some semantic correlation between most classification data values, such as {high school students, junior middle school students, primary school students} belong to {students}, and {students and teachers} can belong to the faculty of this group, as an example to make the semantic correlation better reflected that we usually use the classification tree to achieve.

According to the definition of the classification tree, we set D as the sub-type attribute domain, T_D representing the classification tree, $H(T_D)$ representing the height of the classification tree, $T_{(A_i, A_j)}$ representing the smallest public sub-tree of A_i and A_j in the classification tree, $H(T_{(A_i, A_j)})$ representing and the height of the smallest public sub-tree in the classification tree, so we define any two sub-type attribute values $A_i, A_j \in D$, the distance between A_i and A_j is

$$d_C(A_i, A_j) = \frac{H(T_{(A_i, A_j)})}{H(T_D)} \quad (5)$$

The distance between the two data records is equal to the sum of the distance between all their attributes. Assuming that a data table has n numerical attributes and m sub-type attributes, the data records $l_1 = \{A_1, A_2, \dots, A_{n+m}\}$ and $l_2 = \{B_1, B_2, \dots, B_{n+m}\}$, then the distance between these two data records l_1, l_2 is calculated as

$$d(l_1, l_2) = \sum_{i=1}^n d_N(l_1[A_i], l_2[B_i]) + \sum_{j=1}^m d_C(l_1[A_j], l_2[B_j]) \quad (6)$$

Where, $l_1[A_i]$ represents the value of the data recorded l_1 on the attribute A_i , $l_2[B_j]$ represents the value of the data record l_2 on the attribute.

2) Definition of information loss degree:

The principle of K-anonymous implementation is to generalize the QI attribute, and to protect the data on the basis of certain data loss, so that the attacker cannot be recognized. The clustered data set is designed to make the similarity of the divided equivalent classes higher, and thus reduce some k-

anonymous information loss. Therefore, in order to more reasonably measure the information loss degree of the proposed method in this paper, the information loss after anonymity is calculated from the numerical type and classification attributes according to the characteristics of different attribute types.

The generalization of numerical attributes is usually achieved by dividing the specific values to a certain range interval, such as the data where the value of the attribute age is 35 in Table 2, and the generalization value is [20,40]. We assume that the interval of the original data of a numerical attribute is $[A_{\min}, A_{\max}]$, after the value of the attribute on the equivalent class becomes $[a_{\min}, a_{\max}]$, and the loss of information in this data record is

$$IL = \frac{|a_{\max} - a_{\min}|}{|A_{\max} - A_{\min}|} \quad (7)$$

For example, in the age attribute in Table 2 above, the information loss after generalization is. $IL = \frac{40 - 20}{55 - 25} = \frac{2}{3}$

The classification attribute data generates a classification tree according to its generalization mechanism, *LeafNode* represents the number of contents of the attribute, that is, the number of leaf nodes of the classification tree. Using *AfterLeafNode* represents the number of leaf nodes in the subtree with the generalized value as the root node, the information loss of the subtype attribute is

$$IL = \frac{AfterLeafNode}{LeafNode} \quad (8)$$

A cluster generated after clustering is an equivalent class. The method to calculate the sum of information loss in a cluster is to add up the information loss of all quasi-identifier attributes and divide it by the number of quasi-identifiers. The information loss of all equivalent classes is the sum of information loss of the data table. We assume that the data table has n quasi-identifier attributes, then the sum of information loss is

$$SumIL = \sum_{j=1}^m m \cdot \frac{\sum_{i=1}^n IL_i}{n} \quad (9)$$

3) Algorithm implementation:

In this paper, the data structure of binary trees is used to divide the data set, with each partition with the minimum information loss, thus reducing the information loss caused by k -anonymity. The similarity between data records is calculated based on the distance between the data defined above, and divided according to the principle of distance minimization to ensure that the data within the agreed equivalence are as similar as possible to improve the clustering effect. To this end, we set two constraints during the binary tree partitioning process:

(1) The number of data records in each equivalent class is

greater than k ;

(2) In each division, the information loss of the left and right subtree sets is not greater than the root node data set.

We use D_0 to represent the original dataset, $D_{i-1child}$, $D_{i-1rchild}$ representing the left and right subsets of the dataset division, $|D_i|$ represents the number of data records in D_i . The constraints of the algorithm can be expressed as

$$C = \{|D_{i-1child}| \geq k, |D_{i-1rchild}| \geq k, IL(D_{i-1child}) + IL(D_{i-1rchild}) \leq IL(D_i)\} \quad (10)$$

First treat all the records in the data set as an equivalence set and minimize all the data attributes in the equivalence. Then, according to the distance calculation formula, select the two farthest two data records as the cluster center of the left and right equivalent subset, and then calculate the distance between other data records and the two cluster centers, that is, converted into a 2-means classification problem. After the division, it is necessary to judge whether the constraint is met. If not, the division is stopped and the constraint condition is continued. The binary tree clustering algorithm (BTCA) process is shown in Figure 2:

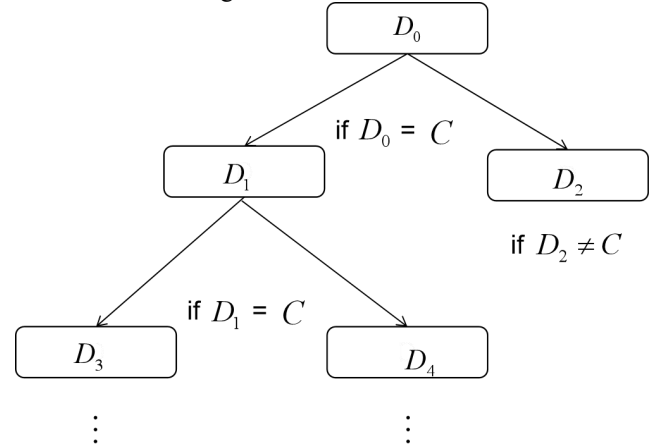


Figure 2. Implementation process of the BTCA algorithm

We perform the analysis of the availability of the above algorithms. The constraint (1) ensures that each equivalent class generated after clustering is at least k bars not separable on the quasi-identifier attribute. An attacker wants to obtain private data from a particular user. The probability of a successful attack is only $1/k$, thus ensuring the privacy of the data. Constraints (2) considering the availability of the data, classification according to the principle of minimum distance, and using a greedy strategy in the process of judging that the information loss of the left and right subsets must not be greater than the information loss of the root set each time. To obtain the optimal partition strategy, minimize the sum of information loss during generating the anonymous data set. Thus, the proposed algorithm is well available.

C. Classified noise-increasing mechanism based on differential privacy:

1) Differential privacy definition:

Differential privacy is a privacy protection method based on data distortion, which is based on a solid mathematical foundation, strictly defines privacy protection and provides a quantitative evaluation method, which makes the level of privacy protection provided by data sets under different parameter processing comparable. The basic principle is that the user queries the data set D through the query function F , and the random algorithm K perturbs the condition of differential privacy protection by disturbing the query function F , while ensuring the data availability and ensuring that the attacker cannot obtain the privacy information of individual users.

If the adjacent data set D_1 and D_2 at most between a data record, given the random algorithm K provides differential privacy, $Range(K)$ represents the random algorithm K value range, $Pr[*]$ represents the probability of a query result of S after the data set plus the same random noise, for only a difference between two data sets D_1 and D_2 , to query the probability of obtaining the same value is very close, so that even if the attacker has enough background knowledge, also cannot find out the individual user privacy data in the final result. The output $S \in Range(K)$ of algorithm K on D_1 D_2 satisfied:

$$Pr[K(D_1) = S] \leq e^\epsilon \times Pr[K(D_2) = S] \quad (11)$$

Among them, ϵ is the privacy protection budget factor, which is used to measure the intensity of privacy protection. The magnitude of the ϵ values is directly proportional to the protection effect and inversely proportional to the degree of data distortion.

2) Privacy budget allocation:

We adopt added noise to achieve the differential privacy of the anonymous data set, by first assigning an appropriate privacy protection budget to the protected quasi-identifier properties according to the need of privacy weight. Let the total privacy protection budget be ϵ . Due to the different sensitivity degree of each quasi-identifier, the privacy weight

of the quasi-identifier attribute B_j is calculated according to the privacy weight matrix W

$$W_j = \frac{\sum_{i=1}^m w_{ij}}{m} \quad (12)$$

In privacy protection, the smaller the ϵ , the more protection. The larger W_j we calculate, the higher the degree of protection required, so we need to make some transformation to $W_{QI} = \{W_1, \dots, W_n\}$, and here we take the reciprocal method to convert large values to small values, that is

$W_\epsilon = \{\frac{1}{W_1}, \dots, \frac{1}{W_n}\}$. The smaller the value in W_ϵ , the

higher the degree of privacy protection required, so that you can set your own allocation rules according to your own protection needs and privacy protection budget, and allocate the privacy budget for different quasi-identifier attributes, that is $\epsilon \rightarrow \{\epsilon_1, \epsilon_2, \dots, \epsilon_m\}$.

3) Graded noise-adding mechanism:

Differential privacy different noise addition methods are used for different types of data. The numerical type in the data use Laplace mechanism to add noise with a certain probability, while the exponential mechanism is actually used for differential privacy. We first need to define a sensitivity Δf , which means that for a mapping function function $f: D \rightarrow R^d$ representing a data set D to a dimensional space d , its maximum range of changes, such as the number of queries, the sensitivity is 1. We assume that the original data set is D , and f is a query function. To add the random variable x satisfying Laplace distribution into $f(D)$, where, the probability density function of the random variable x is

$$p(x | \mu, \lambda) = \frac{1}{2\lambda} e^{-\frac{|x-\mu|}{\lambda}} \quad (13)$$

As for the position parameter μ , generally the default is 0; as the scale parameter $\lambda > 0$, the added Laplace noise meets:

$$Lap(\lambda) = \frac{\Delta f}{\epsilon} \quad (14)$$

In order to enhance privacy protection, this paper combines the differential privacy model to conduct hierarchical noise addition processing of raw data. According to the privacy weight, we assign different privacy budget to each quasi-identifier attribute $\{\epsilon_1, \epsilon_2, \dots, \epsilon_m\}$, so as to achieve the purpose of hierarchical protection of differential privacy protection of anonymous data set.

The Laplace noise is used to process the numerical data, define the query function $f_i(D)$ to represent the record i in the returned data set. For $f_i(D_\epsilon)$ is the data record after noise processing, the quasi-identifier attributes in the record are graded and noise added, and the added Laplace noise meets:

$$f_i(D_\epsilon) = f_i(D) + (Lap_1(\frac{\Delta f}{\epsilon_1}), \dots, (Lap_m(\frac{\Delta f}{\epsilon_m}))^T \quad (15)$$

The exponential mechanism is used to add noise to classified data, obtain attribute values from a cluster, and select values based on the probability of the occurrence of each attribute value in the cluster. According to the input data, differential privacy parameters and probability selection values, the selection increases exponentially with the occurrence probability[19].

4) Algorithm implementation:

Algorithm input: Anonymous data set D , privacy protection budget ϵ_i ($i=1,2,\dots, m$);

Algorithm output: DP-K anonymized data set D_ϵ .
step:

① Query function f_i , query the article i record in the anonymous data set D ;

② Return the query result $f_i(D)$;

③ For numerical data, add Laplace noise to $f_i(D)$, that is $f_i(D) + (Lap_1(\frac{\Delta f}{\epsilon_1}), \dots, (Lap_m(\frac{\Delta f}{\epsilon_m}))^T$;

④ For categorical data, it generates a collection of attribute values for each cluster, and each cluster selects the output value using an exponential mechanism according to its respective attribute values.

⑤ After adds noise to each piece of data, the DP-K anonymous data set D_ϵ is obtained;

To prove the effectiveness of the algorithm, we satisfy the ϵ differential privacy in the perturbation data set obtained by adding noise to the original data set, and the principle of differential privacy is to query the real data and add noise. According to the parallel rule of privacy protection, the entire data set satisfies the ϵ -differential privacy if any two disjoint records satisfy the ϵ -differential privacy. Thus we obtain the condition that the DP-K anonymous data set satisfies the differential privacy.

IV. EXPERIMENTAL PROOF

The experiment of this paper was conducted on the windows10 operating system, written in pycharm using the python language, taking the Adult data set in the UCI machine learning database. It is the most commonly used standard test data set in privacy protection and data mining research. The attribute variables of this data set include 14 important information, including age, type of work, education background and occupation, among which 8 categories belong to category discrete variables, and the other 6 categories belong to numerical continuous variables. The data set is a categorical data set used to predict whether annual revenue exceeds \$50k. We randomly selected 1000 pieces of data, took the annual income as a sensitive attribute, and the remaining 14 variable attributes as the quasi-identifier attributes. We used the random forest algorithm to predict the weight of the sensitive attributes, and averaged the data 20 times before and after deleting the quasi-identifier properties.

Number	Quasi Identifier Attribute	$\Delta P(A_i, B_j)$	w_j	w_ϵ
1	age	0.00751	0.134756863	7.420
2	workclass	0.00398	0.071415755	14.003
3	fwlght	0.00160	0.028709851	34.831
4	education	0.00008	0.001435493	696.625
5	education-num	0.00238	0.042705903	23.416
6	marital-status	0.00231	0.041449847	24.126
7	occupation	0.00366	0.065673784	15.227
8	relationship	0.00181	0.032478019	30.790
9	race	0.00139	0.024941683	40.094
10	sex	0.00034	0.006100843	163.912
11	capital-gain	0.01970	0.353490041	2.829
12	capital-loss	0.00609	0.109276871	9.1513
13	hours-per-week	0.00316	0.056701956	17.636
14	native-country	0.00172	0.03086309	32.401

Figure 3. Prediction of the quasi-identifier attribute sensitivity

Since only one sensitive attribute is set, so the weight matrix W is $1 \times n$, and calculated W_ϵ , the smaller the value of W_ϵ , the higher the degree of privacy protection required. After that, they can set their own allocation rules according to their own privacy protection needs ϵ and the total privacy protection budget, and allocate the privacy budgets for different quasi-identifier attributes, that is $\epsilon \rightarrow \{\epsilon_1, \epsilon_2, \dots, \epsilon_{14}\}$, to meet the hierarchical data protection of the differential privacy hierarchical noise increase mechanism.

The proposed binary tree-based clustering algorithm BTCA is compared with the k-anonymous classical algorithm KACA algorithm to measure the sum of information loss. Considering that the number of quasi-identifiers also affects the amount of information loss, the number of quasi-identifiers is set to eight for experiments, including four numerical attributes {age, education-num, capital-gain, capital-loss}, and four subtype attributes {workclass, sex, occupation, education}. The experimental results are shown in Figure 3.2. It can be seen that under the increasing K value, the BTCA algorithm proposed has less information loss, and the advantage is more obvious and the data availability is higher with the increase of K value.

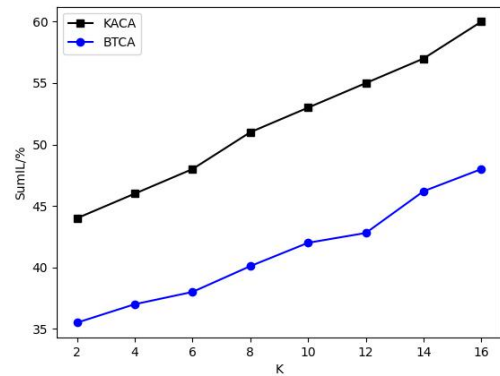


Figure 4. Comparison of information loss degree

V. CONCLUSIONS

In this paper, for the lack of attribute sensitivity considerations of existing k-anonymous schemes and how to minimize the amount of information loss, we propose a hierarchical DP-K anonymous data release model based on binary tree clustering. The division of similar data records into the same equivalent class by the binary tree-based clustering algorithm BTCA can improve the effect of clustering, reduce the information loss caused by anonymous data set release, and improve the data availability. The clustered anonymous data sets are reallocated to different privacy budgets according to the privacy rights of the QI attribute, and the hierarchical protection of the data with different degrees of sensitivity can be realized through the differential privacy noise increase mechanism, which enhances the privacy of the data. In the end, the proposed algorithm can effectively reduce the information loss and insufficient privacy protection in the generation process of anonymous data set. However, in this paper, only a single sensitive attribute in the data set is selected. For the data set of high-dimensional sensitive attribute is included, the efficiency and availability of the algorithm may be insufficient, which is also the place to be improved in the future. And, the biggest challenge of fusion difference privacy mechanism is the balance of privacy and availability, how to keep the result data privacy at the same time to reduce noise, based on the tradeoff between utility and privacy to adjust data privacy methods, and create metrics to evaluate the quality of measuring model information loss and privacy protection, is a challenge for future research.

ACKNOWLEDGMENT

This paper is supported by science and technology project of Big Data Center of State Grid Corporation of China :
 “ Research on key technologies of power data sharing protection based on generation antagonism privacy protection ”
 (Grand No.SGSJ0000HGJS2200033)

REFERENCES

- [1] Wang Yi, Chen Qixin, Zhang Ning, et al. Integration of 5G communication and ubiquitous Power Internet of Things: Application analysis and Research Outlook [J]. Power Grid Technology, 2019,43 (5): 1575-1585.
- [2] P. Samarati and L. Sweeney. Generalizing data to provide anonymity when disclosing information. In Proc. of ACM PODS, 1998.
- [3] L. Sweeney. k-anonymity: A model for protecting privacy. International Journal of Uncertainty , Fuzziness and Knowledge-Based Systems , 2002,10(5):557–570.
- [4] Deng Jingjing, Ye Xiaojun. Multidimensional K-anonymity algorithm based on R-tree [J]. Computer Engineering, 2008 (01): 80-82
- [5] Zhang Qiang, Ye Ayong, Ye Guohua, Deng Huina, Chen Aimin. The k-anonymous data privacy protection mechanism for optimal clustering [J]. Computer Research and Development, 2022,59 (07): 1625-1635.
- [6] Machanavajjhala, KIFERD , GEHRKE J.I-diversity : Privacy Beyond k-anonymity [C] // Proceedings of the 22nd International Conference

on Data Engineering Workshop.Washington D.C.USA : IEEE Press , 2006 : 24-35.

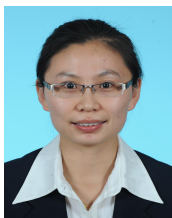
- [7] Li NH , LI TC , VENKATASUBRAMANIAN S.T-closeness privacy Beyond k-anonymity and l-diversity [C] //Proceedings of the 23rd International Conference on Data Engineering.Washington D.C.USA : IEEE Press , 2007 : 106-115.
- [8] Cao Minzi, Zhang Linlin, Bi Xuehua, Zhao Kai. Personalized (α , l) - Diversity k-anonymous privacy protection model [J]. Computer Science, 2018,45 (11): 180-186.
- [9] Fu Jingjing, Xu Xiaodong. A (p, θ) k-Anonymous model that resists equivalence attacks [J]. Computer and Digital Engineering, 2021,49 (08): 1619-1624.
- [10] Hussain S F,Haris M.A k-means based co-clustering (kCC) algorithm for sparse, high dimensional data[J].Expert Systems with Applications,2019,18(5):20-34.
- [11] R.Gangarde, A.Pawar and A.Sharma,Comparisons of Different Clustering Algorithms for Privacy of Online Social Media Network,2021 IEEE Pune Section International Conference (PuneCon), 2021:1-5.
- [12] C.N.Sowmyarani, L.G.Namya, G.K.Nidhi and P.Ramakanth Kumar, Enhanced k-Anonymity model based on clustering to overcome Temporal attack in Privacy Preserving Data Publishing,2022 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT), 2022 : 1-6.
- [13] Fang Kai, Shi icai, Jia Yuanyuan. The k-anonymized data release algorithm based on mixed clustering [J / OL]. Electronic Technology, 2022:1-7.
- [14] Cynthia Dwork and Rebecca Pottenger.Toward practicing privacy.Journal of the American Medical Informatics Association, 2013,20(1):102–108.
- [15] Zhang Shaobo, former Liu Jie, MAO Xinjun, Zhu Gengming. A K-modes clustering data privacy protection method based on local differential privacy [J]. Electronic Journal, 2022,50 (09): 2181-2188.
- [16] Nguyen H H.Privacy-preserving mechanisms for k-modes clustering [J].Computers & Security, 2018.
- [17] Zhang Xing, Zhang Xing. DCKPDP: Improve the differential privacy mixed attribute data publishing method for k-prototype clustering [J]. Computer Application Research, 2022,39 (01): 249-253.
- [18] Yuan Quan, Yan Feiyang, Wen Zhiyun, Zhang Zhenkang. Differential Privacy Protection algorithm for Social Network based on Spectral clustering [J]. Computer Engineering and Science, 2022,44 (02): 251-256.
- [19] Ding Jian, Du Tiantian. Clustering-based differential privacy of civil aviation passenger data release algorithm [J]. Computer Engineering and Design, 2022,43 (03): 608-615.



Yuxiao Xia born in 1990, received the Master degree in University of Chinese Academy of Sciences in 2015. Her research interests include Technology of Computer Application.



Tao Zhao born in 1986 , received the Master degree in Southwest jiaotong University in 2014. His research interests include Network Security and Data Security.



Yanli Lv born in 1980 , received the Doctor degree in Beihang University in 2016. Her research interests include Data Security and Data Compliance.



Yunzhao Li born in 1996, received the M.S. degree from Beijing University of Posts and Telecommunications in 2021. His main research interest include edge computing, Network Security and Data Security.



Ruxia Yang born in 1992 , received the Master degree in North China Electric Power University in 2016. Her research interests include Network Security and Data Security

