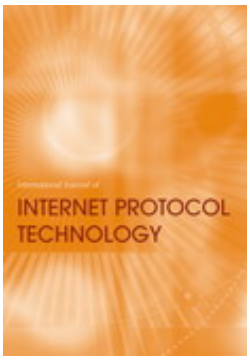


International Journal of Internet Protocol Technology

ISSN online: 1743-8217 - ISSN print: 1743-8209
https://www.inderscience.com/ijipt

Research on personalised privacy-preserving model of multi-sensitive attributes

Haiyan Kang, Yaping Feng, Xiameng Si, Kaili Lu

DOI: 10.1504/IJIPT.2023.10054907

Article History:

Received:	09 February 2021
Last revised:	12 August 2021
Accepted:	21 August 2021
Published online:	23 March 2023

Research on personalised privacy-preserving model of multi-sensitive attributes

Haiyan Kang*, Yaping Feng and Xiameng Si

School of Information and Management,
Beijing Information Science and Technology University,
Beijing 100192, China
Email: kanghaiyan@126.com
Email: fyp187@126.com
Email: sixiameng@bistu.edu.cn
*Corresponding author

Kaili Lu

School of Computer,
Beijing Information Science and Technology University,
Beijing 100192, China
Email: 2860250306@qq.com

Abstract: In order to protect user information from being leaked, it is imperative to improve the availability of published data and realise the safe and efficient information sharing. Aiming at the anonymous privacy-preserving of multi-sensitive attribute data release in logistics industry, this paper proposes a personalised privacy-preserving model of multi-sensitive attributes with weights clustering and dividing (PMSWCD) by analysing existing model. Firstly, according to the different needs of users, the corresponding weight is set for each sensitive attribute value to realise personalisation and then weighted clustering. Secondly, divide the records according to the weighted average value, and select records to establish a group that satisfies *l*-diversity. Finally, release data based on the idea of multi-dimensional bucket. Through experimental analysis, compared with WMBF algorithm, the release ratio of important data of PMSWCD algorithm proposed in this paper is significantly improved, reaching more than 95%, which improves the availability of data.

Keywords: multi-sensitive attributes; data release; personalised; privacy-preserving; weights clustering; dividing; multi-dimensional bucket; *l*-diversity.

Reference to this paper should be made as follows: Kang, H., Feng, Y., Si, X. and Lu, K. (2023) 'Research on personalised privacy-preserving model of multi-sensitive attributes', *Int. J. Internet Protocol Technology*, Vol. 16, No. 1, pp.58–67.

Biographical notes: Haiyan Kang is a senior member of the China Computer Federation (E200028533M), ACM Membership (9495204), and member of privacy protection committee of China Confidentiality Association. He received his PhD in computer application technology from Beijing Institute of Technology, China in 2005. His research interest fields include information system security, privacy preserving, and natural language processing (NLP). He is currently working as a professor at Department of Information Security, School of Information and Management, Beijing Information Science and Technology University, Beijing, China.

Yaping Feng is currently pursuing the MS degree in logistic engineering from Beijing Information Science and Technology University. Her research interests include privacy preserving in the field of logistic.

Xiameng Si received PhD in Communication and Information System from Beijing Jiao Tong University, China in 2011. Her research interest fields include information content security, social computing, and natural language processing (NLP). She is currently working as an associate professor at Department of Information Security, School of Information and Management, Beijing Information Science and Technology University, Beijing, China.

Kaili Lu is currently pursuing the MS degree in computer from Beijing Information Science and Technology University. Her research interests include privacy preserving in the field of logistic.

This paper is a revised and expanded version of a paper entitled 'An Enhanced Approach for Multiple Sensitive Attributes in Data Publishing' presented at 'Smart Grid and Internet of Things - 4th EAI International Conference', TaiChung, Taiwan PRC, 5–6 December 2020.

1 Introduction

With the development of data mining technology, more and more people are inseparable from the Internet. People browse news, shop, chat, register accounts and so on through the Internet every day, leaving traces on the network. Data is constantly generated, released, shared and used. However, while enjoying the convenience brought by data sharing, the risk of personal information being leaked is greatly increased.

Presently, logistics industry has been achieving new growth. Especially since 2011, Chinese express delivery business volume has achieved nine consecutive years of growth, and the increase is more than or close to 50%, which also makes Chinese express logistics market to become the first in the world. In the end of 2020, the State Post Bureau announced that China's express delivery had reached 83.36 billion pieces, ranking the first in the world's express business for seven consecutive years. The rapid development of the express delivery industry has improved people's lives to a certain extent and provided great convenience for people's lives. But it will bring some concerns about the privacy and security of express delivery. In the process of sending and receiving express, users' address, name, telephone number and other information will be involved. This sensitive information belongs to users' private information. Therefore, if the relevant information is leaked, it will not only affect users' personal safety and property safety, but also cause some social problems. At present, a kind of grey industrial chain has even been formed in the society, that is, the express bill information is peddled with clear price, which will also produce some malicious advertising promotion, harassing phone calls and telephone fraud, which will bring very serious social consequences. But to make the best use of big data, data sharing is imperative. However, in the mass data shared, it often involves personal information or sensitive data. Personal privacy information may be disclosed if data is shared directly. Therefore, in order to ensure the security of personal sensitive information, privacy-preserving should be carried out at the time of data release.

Although researchers are constantly trying to prevent information leakage, personal information is still leaked in various unexpected ways. Most of the existing data privacy-preserving technologies are for single sensitive attribute data. In practical application, the released data involves multi-sensitive attributes, and there are often some specific links between many data. When releasing some information, it is equivalent to indirectly releasing other information. Because the releasing method of single sensitive attribute privacy data is completely different from the multi-sensitive attribute method, for such associated information sensitive data, using the single sensitive attribute releasing method is likely to have the problem of information disclosure. Through the research and study of the existing data (Li et al., 2018; Lu et al., 2017; Wang and Lu, 2019; Zhou et al., 2020), this paper further studies the privacy-preserving of the multi-sensitive attributes in the logistics field data release, and propose a personalised privacy-preserving model of Multi-Sensitive attributes with Weights Clustering and Dividing (PMSWCD).

2 Related research

From the research process, the privacy-preserving in data releasing has long been concerned by scholars, and many classical privacy protection models have been formed. Sweeney (2002) firstly proposed k-anonymity model. This model makes it impossible for an attacker to identify a user by a quasi-identifier with a confidence higher than $1/K$. However, this model only cut off the connection between identifiers and sensitive attributes in records, but did not make corresponding requirements for multi-sensitive attributes, which is prone to homogeneity attacks Homogeneous attack and background knowledge attack. Subsequently, scholars at home and abroad continue to perfect this model, and *l*-diversity (Machanavajjhala et al., 2007), *p*-sensitive *k*-anonymity model (Truta and Vinary, 2006; Sun et al., 2009), (a, k)-anonymous model (Wong et al., 2006), *t*-closeness model (Li et al., 2007) and other models were proposed successively. In 2016, Hasan et al. (2016) proposed to protect personal privacy by exchanging equivalent sensitive values. Invalid records generated by random exchange are eliminated by negative correlation rules. In 2018, Gunawan and Mambo (2018) improved the replacement algorithm and proposed a new data anonymisation scheme, called brother suppression II. This method uses multiple groups of opponent knowledge to replace the items in the opponent knowledge category with other items in the category, and can ensure the availability of data at the least cost.

With the deepening of research, researchers found that using clustering method can effectively find similar records to form equivalent groups, so as to reduce the information loss of data in the process of anonymity. Pramanik et al. (2016) proposed a novel clustering method to solve the problem that some data cannot be clustered due to general clustering. This method can measure the centre of the cluster class through the defined *n*-closeness, so that all records can be incorporated into the record cluster on the premise of low information loss. Jiang et al. (2017) proposed a greedy clustering anonymity

strategy. The strategy points out that the existing methods do not consider the classification of quasi-identifier generalisation. By using different generalisation methods for different quasi-identifier attribute types, their information loss is measured respectively. Li et al. (2017) improved the privacy protection and data availability of mixed attribute data through the combination of clustering algorithm and differential privacy protection.

Yang et al. (2008) first studied the issue of multi-sensitive attribute data releasing in detail, inherited the idea of lossy join proposed a Multi-Sensitive Bucketisation (MSB) for releasing data with multi-sensitive attributes, which is a breakthrough in this field. They applied clustering and MSB techniques to release the microdata with multiple numerical sensitive attributes. Ye et al. (2017) proposed an anonymisation method combining anatomy and permutation for protecting privacy of the microdata with multiple sensitive attributes. This method includes two major steps: anatomising microdata and permutating quasi-identifier attributes. To realise the anonymisation method, they further proposed two algorithms, namely naive multi-sensitive bucketisation permutation algorithm (NMBPA) and closest distance multi-sensitive bucketisation permutation algorithm (CDMBPA). Reddy et al. (2018) proposed a privacy preserving data publishing model that manages personalisation for publishing the microdata with multiple sensitive attributes. The model uses the slicing technique supported by deterministic anonymisation for quasi-identifier attribute, i.e., generalisation for categorical sensitive attributes and fuzzy approach for numerical sensitive attributes based on diversity. Raju et al. (2019) proposed a novel dynamic KCi-Slice publishing prototype for retaining the privacy and utility of multiple sensitive attributes, which is an improvement of KC-Slice. Kanwal et al. (2019) proposed a privacy-preserving model for 1:M records dataset with multiple sensitive attributes, called (p, l)-Angelisation. Xiao et al. (2021) defined three security levels for different sensitive attribute values that have different sensitivity requirements, and given an L_{sl}-diversity model for multiple sensitive attributes. Can solve the problem that the information loss of MBF, MSDCF and MMDCF increases greatly with the increasing of sensitive attribute number.

To sum up, the existing data privacy-preserving methods for multi-sensitive attributes have some deficiencies in the aspects of algorithm efficiency and background knowledge attacks and others. Such as: 1) When grouping, only the attribute value of sensitive attribute is considered, but the range of quasi-identifier is not considered, resulting in large data loss; 2) The personalised privacy-preserving method requires users to set the weight of the bucket, which requires too much professionalism on users. Therefore, we propose a personalised privacy-preserving model of Multi-Sensitive attributes with Weights Clustering and Dividing.

3 Basic knowledge

3.1 Problem description

The use of the Internet and various applications helps third parties collect a large amount of user information. If these data are directly shared, it is likely to be obtained and used by attackers, resulting in the disclosure of personal privacy information and unexpected harm. Therefore, in order to ensure the security of personal sensitive information, we need to use some method to hide the original data set of user information. However, due to the different requirements of different users for the protection degree of sensitive information, we need to carry out personalised protection of user information according to user needs. However, there are few studies on personalised privacy protection, so we need to study and analyse personalised privacy protection.

User information often involves multiple sensitive attributes, such as logistics information. This requires the data owner to protect multiple sensitive attributes when publishing data. Through the research and analysis of the existing literature, at present, we use the multi-dimensional bucket grouping model to release data with multi-sensitive properties. The records in the data table are mapped to the corresponding buckets according to the value of each dimension of the multi-dimensional sensitive attribute, and then grouped according to some algorithm to make each group meet the multi-dimensional sensitive attribute l-diversity, and finally the data is released. But the multi-dimensional bucket grouping model has some shortcomings, such as (1) Low grouping efficiency, (2) Easy to lead to data suppression problems, and (3) Do not meet the user's personalised needs. In view of the above problems, this paper makes the corresponding improvement. To solve the above problems, we need to improve the privacy protection method of multi sensitive attributes and design a better algorithm to realise the privacy protection of multi sensitive attributes.

3.2 Related definition

Suppose there are n data records in the data table T $\{Q_1, Q_2, \ldots, Q_x, S_1, S_2, \ldots, S_y\}$, then $|T|=n$, and each data record is t_i ($1 \leq i \leq n$). Each record t_i includes x quasi-identifier attributes $Q1$, $Q2$, ..., Qx and y sensitive attributes S_1, S_2, ..., S_y.

Definition 1 (Multi-sensitive attributes). A record has multiple sensitive attributes. Expressed by S, Si ($1 \leq i \leq$ y) represents the i-dimensional sensitivity attribute.

Definition 2 (Grouping). A group is a subset of the data records in the data table T. Each data record belongs to only one group in the data table T. The group of the data table T is denoted as GT $\{G_1, G_2, \ldots, G_m\}$, and ($QI_i \cap QI_j = \varphi$) ($1 \leq i \neq j \leq m$).

Definition 3 (Multi-sensitive attributes' *l*-diversity) (Zhu et al., 2014). In the group *G*, if each dimension sensitive attribute value of all data records meet *l*-diversity respectively, then the group *G* satisfies *l*-diversity of multi-dimensional sensitive attributes. In other words, the group *G* satisfies multi-dimensional sensitive attributes' *l*-diversity.

Definition 4 (*d*-dimensional bucket) (Yang et al., 2008). If there are *d*-dimensional sensitive attributes, the *d*-dimensional bucket is denoted as Bucket (S_1, S_2, ..., S_d) ($2 \leq d \leq n$), and each dimension of the multi-sensitive attributes corresponds to a one-dimensional bucket. According to the values of each dimension, the records are mapped to the corresponding bucket.

Definition 5 (Weights clustering). Suppose that there are *n* records in the data table *T*, each data record has *d*-dimensional sensitive attributes. For the sake of simplicity, we ignore the identifier and the quasi-identifier when forming multiple clusters which is denoted as $T_c\{C_1, C_2, ..., C_q\}(|C_q| \geq 1, 1 \leq q \leq n)$. The weights of the sensitive attribute values of each dimension in the same cluster are equal or similar, and the weights of different clusters are quite different.

Definition 6 (Weighted average value). It represents the average of the weights all sensitive attribute in each record.

Definition 7 (Weighted standard value). It represents the degree of deviation between the weight of different sensitive attribute in each record. The greater the value is, the greater the degree of difference. On the contrary, the smaller the value is, the smaller the degree of difference.

Definition 8 (Suppression technology). Some records that cannot be released or some records that do not satisfy privacy-preserving are hidden.

Figure 1 Flow chart of PMSWCD model

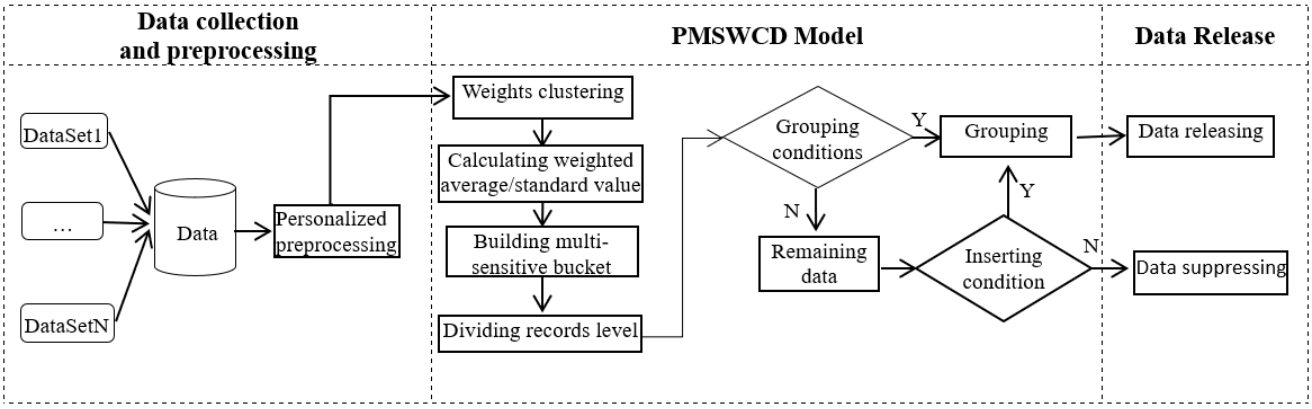


Table 1 Some customer data of a third-party company

Identifier	*Quasi-ID*		*Sensitive Attributes*			
ID	*Zip*	*Age*	*Home Address (weight)*	*Phone Number (weight)*	*Name of Goods (weight)*	*Credibility (weight)*
t_1	0052	29	HeBei**(0.8)	132****7752(0.9)	Food (0.3)	B (0.9)
t_2	0029	31	TianJin**(0.6)	135****6851(0.2)	Clothing (0.3)	A (0.9)
t_3	0031	35	ShanDong**(0.5)	134****5612(0.7)	Phone (0.6)	B (0.9)
t_4	0058	28	HeBei**(0.1)	136****5762(1.0)	Toy (0.3)	B (0.9)
t_5	0062	32	TianJin**(0.5)	136****8623(0.1)	Clothing (0.1)	A (0.9)
t_6	0046	36	BeiJing(0.4)	137****6752(0.4)	Toy (0.1)	A (0.9)
t_7	0039	40	HeBei**(0.2)	139****4231(0.9)	Food (0.2)	B (0.9)
t_8	0075	30	ShanDong**(0.8)	187****1234(0.8)	Phone (0.6)	C (0.9)
t_9	0048	46	BeiJing**(0.6)	152****2564(0.2)	Book (0.7)	A (0.9)
t_{10}	0089	38	TianJin**(0.2)	136****8962(0.8)	Book (0.8)	B (0.9)

4 Personalised privacy-preserving model of multi-sensitive attributes with weights clustering and dividing (PMSWCD)

4.1 The overall framework of data release

In order to improve the validity and security of data releasing, the MSB grouping technology is improved in this paper. The main framework is shown in Figure 1. It is mainly composed of three modules: data collection and preprocessing layer, model layer and data release layer. Specific explanations are as follows.

4.1.1 Data collection and preprocessing

Collection of data through enterprise survey and web crawlers. Through enterprise investigation, we can understand the storage and release of data information of courier services companies. Crawler technology can collect data from web pages. The collected information is built into the original database. Table 1 are some customer data of a third-party company. For convenience, we preprocessed the data. Preprocessing involves setting weights for sensitive attributes. As follows: Set a weight of 0~1 for each sensitive attribute according to the user's requirements for privacy protection of each sensitive attribute. Table 1 contains four sensitive attributes, such as home address, telephone number, commodity name and credibility. In this paper, the first two sensitive attributes in Table 1 are taken as examples for analysis, and the weight $w(0 \le w \le 1)$ of the sensitive attribute value is given respectively, that is, the importance of the sensitive attribute value in this record. For example, the weight of the home address of the record t_1 is 0.8, and the weight of the contact phone attribute is 0.9, and so on.

4.1.2 PMSWCD model

Firstly, cluster the data records according to the weight of sensitive attributes to form multiple clusters, and calculate the weighted average value and weighted standard value of the records in the cluster. Then the weighted multi-dimensional buckets are built for multi-sensitive attributes, and the records in the data table are mapped to the corresponding multi-dimensional buckets. Finally, we select important records by the hierarchical division method to satisfy the users' different needs.

4.1.3 Data release

Use the maximum weight preference algorithm, we select the data record in the multi-dimensional bucket to build a group satisfying l-diversity. Then make the first judgment on the data that does not meet the group's requirements. Finally, the quasi-identifiers of each group are anonymised. Hide the remaining data that cannot be inserted into groups and publish anonymous tables.

Our paper mainly improves the model layer and solves the following problems. (1) Users can set the weight of sensitive attributes according to their own needs, which can meet the needs of personalised privacy-preserving. (2) Our core idea is clustering and dividing. Clustering data sets according to the weight of sensitive attributes before grouping not only reduces the times of calculating weighted average values and weighted standard values, but also facilitates the comparison and selection of data records during grouping, so as to improve algorithm efficiency and reduce CPU running time. (3) We further process the clustered data and the remaining data to reduce data suppressing.

4.2 Personalised multi-sensitive attributes with weights clustering and dividing model

4.2.1 Clustering

According to the weight set by each sensitive attribute in data set T, all records are clustered to form multiple clusters, denoted as data set Tc. After clustering, the weight of the sensitive attribute values of the same cluster is similar or the same, while the weight of the attribute values of different clusters is greatly different, and the intersection between any two clusters is an empty set. The records are clustered to form 5 clusters, C_1 $\{t_1, t_8\}$, C_2 $\{t_2, t_5, t_9\}$, C_3 $\{t_3\}$, C_4 $\{t_4, t_7, t_{10}\}$, C_5 $\{t_6\}$. It is found that there is often an internal relationship between data records, especially when the amount of data is large, some data records have similar values on a quasi-identifier. For this reason, the quasi-identifier attributes with many common characteristics can be grouped into one class through clustering, and the data set can be divided into several subsets accordingly, so as to reduce the degree of data concealment and the generalisation rate of quasi-identifier, and improve the availability of data. The purpose of weight clustering at this point can bring benefits to the following steps, which not only reduces the number of times of calculating weighted average and weighted standard average, but also facilitates the comparison and selection of data records during grouping, thus improving algorithm efficiency and reducing CPU running time.

4.2.2 Calculating the weighted average and weighted standard value for each record

The Weight Average Value ($WAve^n$) is denoted as

$$WAve^n = \frac{1}{d}\sum_{i=1}^{d} S_i^n \tag{1}$$

In equation (1), $WAve^n$ represents the average value of the weight of each sensitive attribute in each record, d represents the number of sensitive attributes in the NTH record, and S_i^n represents the weight of the i-dimensional sensitive attributes in the nth record. The greater the value of $WAve^n$, the more important the data for this record.

The Weight Standard Value ($WSve^n$) is denoted as

$$WAve^n = \frac{1}{d}\sum_{i=1}^{d}\left(S_i^n - WAve^n\right)^2 \tag{2}$$

In equation (2), d represents the number of sensitive attributes of the NTH record, S_i^n represents the weight of the i-dimensional sensitive attributes of the nth record, and $WAve^n$ represents the average value of the weight of each sensitive attribute in each record. $WSve^n$ represents the

difference degree between the weights of sensitive attributes in each record. The greater the value, the greater the difference degree. On the contrary, the smaller the number, the smaller the difference.

4.2.3 Building multi-dimensional bucket

According to the values of sensitive attributes S_1 and S_3, the data in Table 1 is mapped, and the resulting multi-dimensional bucket grouping is shown in Table 2. Each cell represents a bucket. The blank space indicates that the bucket contains no data. t_1~t_{10} are the IDs of the data in Table 1.

Table 2 2-*d* bucket groups

	Group 1	*Group 2*	*Group 3*	*Group 4*	*Group 5*
HeBei**	$\{t_1, t_7\}$				$\{t_4\}$
TianJin**		$\{t_2, t_5\}$		$\{t_{10}\}$	
ShanDong**			$\{t_3, t_8\}$		
BeiJing**				$\{t_9\}$	$\{t_6\}$

4.2.4 Dividing record level

According to the weighted average ($WAve^n$), the importance level of records is defined, and then the priority selection strategy based on the maximum weight is used to construct groups that meet *l*-diversity. The greater $WAve^n$ the more important the record is, otherwise the opposite is true. All our records are divided into levels A, B and C. If the range of $WAve^n \in [0.6,1]$ is in the data set, classify the data as A, 'A' means important. If the range of $WAve^n \in [0.4,0.6]$ is in the data set, classify the data as level B, 'B' means general. If the range of $WAve^n \in [0,0.4]$ is in the data set, classify the data as level C, 'C' is not important. The purpose of the record level table is to judge the importance of records. Important records are released firstly.

4.2.5 Release data

In order to release important data as much as possible, while satisfying the privacy requirements of users, we use the maximum weight priority selection strategy to select the record for group which satisfy multi-sensitive attributes *l*-diversity, which ensure that important data are released.

The steps of the maximum weight priority selection strategy are as follows:

Step 1: If there exists the cluster of $WAve^n \in [0.6, 1]$ in data set, according to the record levels these data are classified as A level. Successively select the record with larger weighted average and smaller weighted standard value to form group until all the records in the cluster are traversal completed. Otherwise, go to Step 2.

Step 2: If there exists the cluster of $WAve^n \in [0.4, 0.6)$ in data set, according to the record levels these data are classified as B level. Successively select the record with larger weighted average and smaller weighted standard value to form group until all the records in the cluster are traversal completed. Otherwise, go to Step 3.

Step 3: If it exists the cluster of $WAve^n \in [0, 0.4)$ in data set, according to the record levels these data are classified as C level. In order to guarantee the release of sensitive attributes with higher weights in the records, select the record with larger weighted average and larger weighted standard value form group until all the records in the cluster are traversal completed.

Group records to be published. Then the published data is divided into two data tables QIT and ST. QIT is a quasi-identifier attribute table, which contains records with quasi-identifier attributes and group numbers; ST is a sensitive attribute table, which contains records with sensitive attributes and group numbers.

4.3 Personalised privacy-preserving model of multi-sensitive attributes with weights clustering and dividing (PMSWCD)

In this paper, we propose PMSWCD algorithm to release data, and we use algorithm to denote them. In this algorithm, we bring in clustering, division and generalisation methods, and then select important records to release.

Algorithm PMSWCD

Input: Data table *T*, diversity parameter *l*.

Output: Quasi-identifier attribute table QIT, Sensitive attribute table ST.

01: Personalised setting weight according to user's needs
02: Weights clustering;
03: Calculating the weighted average value and the weighted standard value for each record and save them in the record;
04: Building multi-sensitive bucket and dividing records level;
05: **while**
06: **S**elect the record t_i by the maximum weight first selection strategy;
07: **if** (the record t_i satisfies *l*-diversity)
08: Insert t_i into the group and delete t_i from the bucket;
09: **else**
10: create a new *l*-diversity group and delete t_i from the bucket;
11: **end while** (no record is selected)
12: **for each** the records which do not satisfy *l*-diversity group
13: **S**elect the record t_i by the maximum weight first selection strategy;
14: **if (** a group G_i still satisfy *l*-diversity after adding the record t_i)
15: Insert t_i into the group G_i ;
16 : **else**
17 : Insert t_i into the remaining dataset;
18 : **end for**
19: Suppression all the remaining records;
20: Generalising the quasi-identifier attributes of data in all groups;
21: return QIT and ST;

Taking the data in Table 1 as an example, $l = 3$ is taken to explain the execution process of the above algorithm. The dataset T_c after weight clustering contains 5 clusters: $C_1\{t_1, t_8\}$, $C_2\{t_2, t_5, t_9\}$, $C_3\{C_3\}$, $C_4\{t_4, t_7, t_{10}\}$, $C_5\{t_6\}$, and the corresponding D-dimensional bucket ($D = 2$) is shown in Table 2. According to the maximum weight priority selection strategy, we first select record t_1 of cluster C_1 into group G_1, as $G_1\{t_1\}$, while deleting t_1 in bucket; then select record t_8, and t_1 and t_8 satisfy multi-sensitive attributes l-diversity, so we insert t_8 into the group G_1, as $G_1\{t_1, t_8\}$, while deleting t_8 in bucket; the next step is to select record t_3 in cluster C_3. Due to multi-sensitive attributes l-diversity rule, t_3 cannot be inserted into the group G_1, so a new group G_2 is built, and denoted as $G_2\{t_3\}$, and deleting t_3 in bucket; then select record t_4 in cluster C_4, and t_3 and t_4 satisfy multi-sensitive attributes l-diversity, so we insert t_4 into the group G_2, as $G_2\{t_3, t_4\}$, and deleting t_4 in bucket. Then we can get the groups $G_1\{t_1, t_8, t_{10}\}$, $G_2\{t_2, t_3, t_4\}$, $G_3\{t_5, t_7, t_9\}$, and the remaining record is t_6. Once again, determine whether the remaining data can be inserted into the group. At this time, the remaining record t_6 can be inserted into the group G_1, as $G_1\{t_1, t_6, t_8, t_{10}\}$. Therefore, the final groups are $G_1\{t_1, t_6, t_8, t_{10}\}$, $G_2\{t_2, t_3, t_4\}$, $G_3\{t_5, t_7, t_9\}$. The final released data is shown in Table 3 and Table 4.

Table 3 QIT of algorithm

Zip	*Age*	*Group ID*
[0050–0090]	[28–38]	1
[0020–0060]	[28–40]	2
[0020–0060]	[28–40]	2
[0020–0060]	[28–40]	2
[0030–0070]	[30–40]	3
[0050–0090]	[28–38]	1
[0030–0070]	[30–40]	3
[0050–0090]	[28–38]	1
[0030–0070]	[30–40]	3
[0050–0090]	[28–38]	1

Table 4 ST of algorithm

Group ID	*Home Address*	*Phone Number*
1	HeBei**	132****7752
	TianJin**	137****6752
	TianJin**	187****1234
	BeiJing**	136****8962
2	TianJin**	135****6851
	ShanDong**	134****5612
	HeBei**	136****5762
3	TianJin**	136****8623
	HeBei**	139****4231
	BeiJing**	152****2564

5 Experiment and analysis

5.1 Experimental environment and data

5.1.1 Experimental environment

Processor Intel(R) Core (TM)i5-5200U CPU, memory 4GB, operating system Windows 10(×64), MATLAB and Java are used as the main test language.

5.1.2 Experimental data

We collected 7000 pieces of customer information of logistics company as experimental data. The description of partial data is shown in Table 1.

5.2 Evaluation criteria

5.2.1 Release ratio of important data

If the weighted average value of the record belongs to the range of [0.4, 1], it is defined as 'important data', the release ratio (*Relratio*) of important record is defined as equation (3).

$$Relratio = \frac{number'\left(WAve \in [0.4,1]\right)}{number\left(WAve \in [0.4,1]\right)} \tag{3}$$

In equation (3), $number'(WAve \in [0.4,1])$ represents the number of important records in the published data, and $number(WAve \in [0.4,1])$ represents the number of important records in the initial data. The larger the *Relratio* is, the more important data records are released, and the more it can meet the personalised needs of specific users.

5.2.2 Additional information loss

After the initial grouping, in order to reduce the occultation rate, on the premise of satisfying l-diversity, the remaining data were divided into the existing groups. So, in the final group, some sensitive record values may repeat that $l < |G|$ (G the packet data record number), it is to a certain extent, increased the risk of privacy, therefore introduced additional information loss degree to measure the risk.

In data set T, if there exist multi-sensitive attribute l-diversity group $G\{G_1, G_2, ..., G_m\}$, $|G_i| \geq l (1 \leq i \leq m)$, where m represents the number of groups, then the additional information is defined as equation (4).

$$AddInfo = \frac{\sum_{1 \leq i \leq m} |Gi - l|}{ml} \tag{4}$$

5.2.3 Suppression ratio

The suppression ratio represents the proportion of the data records processed by occult in all the data records in the data table. The suppression ratio is defined as equation (5).

$$Suppratio = \frac{|T_s|}{T} \tag{5}$$

In equation (5), $|T_s|$ represents the number of suppression records, and $|T|$ represents the total amount of data in the data table. The smaller the suppression rate is, the better the released data will be. Ideally, the value is 0.

5.3 Experimental analysis

The experiment will analyse the performance of the proposed algorithm from four aspects: release ratio of important data, additional information loss, suppression ratio, and execution time. with various data size $|T|$ ($K = 10^3$), various diversity parameters l, and various the number of sensitive attribute d, we use our PMSWCD algorithm and WMBF algorithm to privacy-preserving of the data on the data set, and then calculate the values of each evaluation index, and the results are shown in the following chart.

5.3.1 Analysis for release ration of important data

Figure 2(a–c) gives the release ratio of important data in two algorithms under different parameters.

Figure 2(a) shows the release ratio of important data under different data size $|T|$, when l is 3 and d is 3. It can be seen that with the increase of data size $|T|$, the release ratio of important data of the two algorithms has a corresponding change, but kept above 0.80. In data size $|T|$ under the condition of the same value, our PMSWCD algorithm of release ratio of important data higher than WMBF algorithm, more can meet the personalised needs of specific users.

From Figure 2(b), it can be seen that with the increase of diversity parameter l, the release ratio of important data of the two algorithms shows an obvious downward trend. Our PMSWCD algorithm also shows a corresponding downward trend, but it is still close to 1.0. In diversity parameter l under the condition of the same value, the release ratio of important data our PMSWCD algorithm is significantly higher than WMBF algorithm, and the released data is of higher value to users and can better meet the personalised needs of users. The reason is that in the WMBF algorithm, with the increase of the diversity parameter l, there are more and more l-diversity groupings, fewer and fewer records in the bucket, and the randomness of records being selected increases. Therefore, it is possible to ignore important data records and give priority to non-important data records.

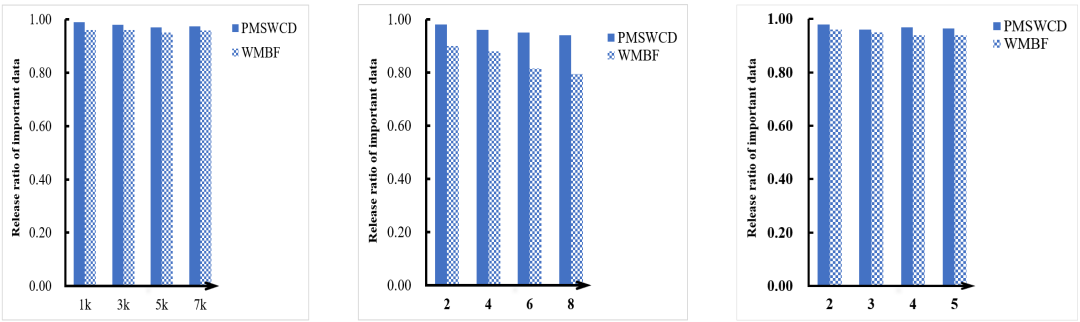
From Figure 2(c), it can be seen that with the increase of number d, the release ratio of important data of the two algorithms has a corresponding change, but kept above 0.90. In number d under the condition of the same value, our PMSWCD algorithm of release ratio of important data is higher than WMBF algorithm.

In conclusion, it can be seen that the release ratio of important data of our PMSWCD algorithm is always higher than that of algorithm WMBF algorithm, which can reach above 95%, and it is less affected by parameter changes.

5.3.2 Analysis for additional information loss

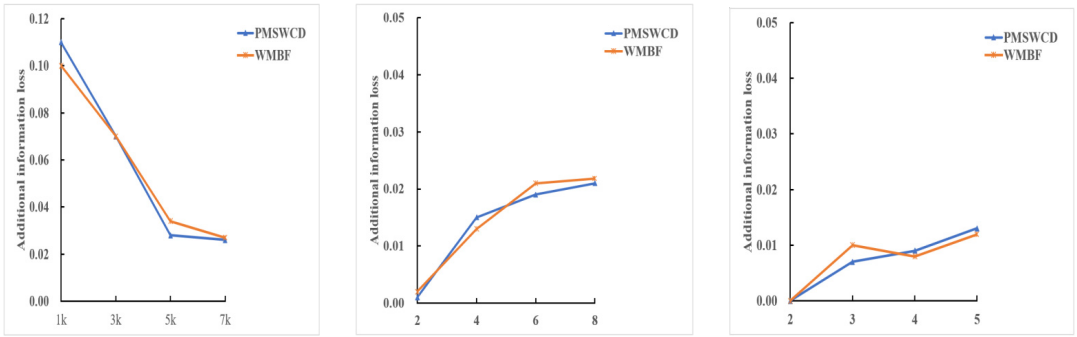
Figure 3(a–c) shows the additional information loss under different parameters in two algorithms.

Figure 2 Release ratio of important data under different parameters between our PMSWCD and WMBF algorithm



(a) Data size $|T|$ (l=3, d=3) (b) Diversity parameter l ($|T|$=6 k, d=3) (c) Number d ($|T|$=6 k, l=3)

Figure 3 Additional information loss under different parameters between our PMSWCD and WMBF algorithm



(a) Data size $|T|$ (l=3, d=3) (b) Diversity parameter l ($|T|$=6 k, d=3) (c) Number d ($|T|$=6 k, l=3)

From Figure 3(a) it can be seen that with the increase of data size $|T|$, the additional information loss of the two algorithms is reduced and less than 0.12. In data size $|T|$ under the condition of the same value, our PMSWCD algorithm of additional information loss lower than WMBF algorithm.

From Figure 3(b), it can be seen that with the increase of diversity parameter l, the additional information loss of the two algorithms shows an obvious rising trend. Our PMSWCD algorithm also shows a corresponding rising trend, and then it becomes stable. In diversity parameter l under the condition of the same value, the additional information loss of our PMSWCD algorithm is significantly lower than that of WMBF algorithm. The reasons are as follows: with the increase of l, the number of groups obtained in the first group becomes less and less, that is, the remaining ungrouped data becomes more and more. Therefore, the number of data size in the second group also becomes more and more, and the additional information loss becomes more and more.

From Figure 3(c) it can be seen that with the increase of number d, the additional information loss of the two algorithms has a corresponding change, but kept below 0.02. In number d under the condition of the same value, our PMSWCD algorithm of additional information loss lower than WMBF.

In conclusion, we can see that: (1) The additional information loss decreases with the increase of data size and is below 0.12. (2) The additional information loss increases with the increase of diversity parameter l. (3) The additional information loss generated in the two algorithms is small, and means that the algorithm is close to optimal.

5.3.3 *Analysis for suppression ratio*

Figure 4(a–b) shows the suppression ratio under different parameters in two algorithms.

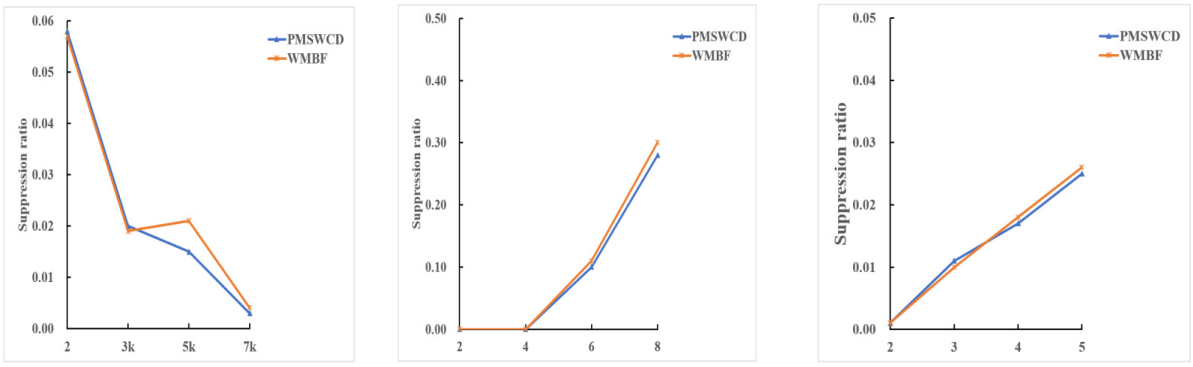
From Figure 4(a), it can be seen that with the increase of data size $|T|$, the suppression ratio of the two algorithms decreases with the increase of data size $|T|$, and when the data size $|T| = 7$ k, suppression ratio close to zero. In data size $|T|$ under the condition of the same value, our PMSWCD algorithm of suppression ratio lower than WMBF algorithm, so has better performance. The reason is that the greater the number of data size, the better the diversification of sensitive attribute values of the record, and then the effect of the grouping becomes better, and the number of suppressed data also gradually decreases.

From Figure 4(b) it can be seen that the suppression ratio of the two algorithms increases with the increase of l. In diversity parameter l under the condition of the same value, the suppression ratio of the PMSWCD algorithm is significantly lower than that of WMBF. The reason is that with the increase of the diversity parameter l, the smaller the number of packets obtained in the process of grouping, the more data to be hidden, so the higher the occult rate. Since WMBF algorithm does not regroup the ungrouped data after the initial grouping, it directly hides it, so the hiding rate is higher.

From Figure 4(c) it can be seen that with the increase of number d, the suppression ratio of the two algorithms has a corresponding change, which increases with the increase of number d; in number d the condition of the same value, the PMSWCD algorithm of suppression ratio of important data lower than WMBF algorithm, so the released data has higher value.

In conclusion from Figure 4: (1) The suppression ratio decreases as the data size $|T|$ increasing, and it will close to 0 when data size $|T| = 7$ k. The reason is that the greater the number of data size, the better the diversification of sensitive attribute values of the record, and then the effect of the grouping becomes better, and the number of suppressed data also gradually decreases. (2) With the increase of l and d, the suppression ratio of the two algorithms continues to increase. Although the proposed algorithm does not reduce the suppression ratio to a great extent, the suppression ratio of the proposed algorithm is always lower than that of WMBF algorithm.

Figure 4 Suppression ratio under different parameters between our PMSWCD and WMBF algorithm



(a) Data size $|T|$ (l=3, d=3) (b) Diversity parameter l ($|T|$=6k, d=3) (c) Number d ($|T|$=6 k, l=3)

6 Conclusion

In this paper, we introduce data release for multi-sensitive attributes in logistics, and analyse the personalised privacy-preserving problem of multi-sensitive attributes values. Based on the idea of multi-sensitive bucket, we propose a personalised privacy-preserving model of Multi-Sensitive attributes with Weights Clustering and Dividing (PMSWCD) to satisfy the requirements of users. We adopted the clustering and dividing method to release data. Then we compared the PMSWCD with WMBF algorithm. The experimental results show that the additional information loss and suppression ratio of two algorithms have a little difference, but the release rate of the important data in our PMSWCD algorithm is above 95%, and the execution time is lower. Therefore, the published data in our PMSWCD algorithm have high availability and achieved the effect of personalised privacy-preserving.

Acknowledgement

This work is partially supported by Humanities and social sciences research project of the Ministry of Education (20YJAZH046), Graduate course construction project of Bistu (2020YKJ17), Higher education research project of Bistu (2020GJZD02). It is recommended by 4th EAI International Conference on Smart Grid and Internet of Things.

References

Gunawan, D. and Mambo, M. (2018) 'Set-valued data anonymization maintaining data utility and data property', *IMCOM'18: The 12th International Conference on Ubiquitous Information Management and Communication*, pp.1–8.

Hasan, A.S.M.T., Jiang, Q.S., Luo, J. et al. (2016) 'An effective value swapping method for privacy preserving data publishing', *Security and Communication Networks*, pp.9–16.

Jiang, H.W., Zeng, G.S. and Ma, H.Y. (2017) 'Greedy clustering-anonymity method for privacy preservation of table data-publishing', *Journal of Software*, pp.341–351.

Kanwal, T., Shaukat, S.A.A.S. et al. (2019) 'Privacy-preserving model and generalization correlation attacks for 1:M data with multiple sensitive attributes', *Inf. Sci*., pp.238–256.

Li, J., Bai, Z.H., Yu, R.Y. et al. (2018) 'Mobile location privacy protection algorithm based on PSO optimization', *Acta computer Sinica*, Vol. 41, No. 5, pp.1037–1051.

Li, L.X., Ding, Y.S. and Wang, J.Y. (2017) 'Differential privacy data protection method based on clustering', *CyberC 2017: 2017 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery*, Nanjing, China, pp.11–16.

Li, N.H., Li, T.C. and Venkatasubramanian, S. (2007) 'T-closeness: privacy beyond k-anonymity and l-diversity', *IEEE 2007: International Conference on Data Engineering*, pp.106–115.

Lu, Q.W. et al. (2017) 'Personalized privacy-preserving trajectory data publishing', *Chinese Journal of Electronics*, Vol. 26, No. 2, pp.285–291.

Machanavajjhala, A., Gehrke, J., Kiefer, D. and Venkitasubramaniam, M. (2007) 'l-diversity: privacy beyond k-anonymity', *ICDE 2007: ACM Transactions on Knowledge Discovery from Data*, pp.24–36.

Pramanik, M.I., Lau, R.Y.K. and Zhang, W. (2016) 'K-anonymity through the enhanced clustering method', *IEEE 2016: International Conference on e-Business Engineering (ICEBE)*, pp.85–91.

Raju, N.V.S.L., Seetaramanath, M.N. and Rao, P.S. (2019) 'A novel dynamic KCi-Slice publishing prototype for retaining privacy and utility of multiple sensitive attributes', *Int. J. Inf. Technol Comput. Sci.*, pp.18–32.

Reddy, S.R.P., Raju, K.V. and Kumari, V.V. (2018) 'A novel approach for personalized privacy preserving data publishing with multiple sensitive attributes', *Int. J. Eng. Technol*., pp.197–206.

Sun, X.X., Wang, H., Li, J.Y. et al. (2009) 'Achieving P-sensitive K-anonymity via anatomy', *ICEBE 2009: Proceedings of the 2009 IEEE International Conference on e-Business Engineering*, pp.199–205.

Sweeney, L. (2002) 'k-anonymity: a model for protecting privacy', *International Journal of Uncertainty Fuzziness and Knowledge-Based Systems*, Vol. 10, No. 5, pp.557–570.

Truta, T.M. and Vinay, B. (2006) 'Privacy protection: p-sensitive k-anonymity property', *ICDE 2006: Proceeding of the 22th International Conference on Data Engineering*, pp.94–103.

Wang, H.Y. and Lu, J.X. (2019) 'Personalized privacy protection method for group recommendation', *Journal of Communications*, Vol. 40, No. 9, pp.106–115.

Wong, R.C., Li, J. et al. (2006) '(α, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing', *ACM 2006: Proceedings of the 12th international conference on Knowledge discovery and data mining*, pp.754–759.

Xiao, Y.L. and Li, H.Q. (2021) 'Complete l-diversity grouping algorithm for multiple sensitive attributes and its applications', *Communications and Computer Sciences*, Vol. E104A, No. 7, pp.984–990.

Yang, X.C., Wang, Y.Z., Wang, B. et al. (2008) 'Privacy preserving approaches for multiple sensitive attributes in data publishing', *Chinese Journal of Computers*, Vol. 31, No. 4, pp.574–587.

Ye, Y., Wang, L., Han, J. and Qin, S. (2017) 'An anonymization method combining anatomy and permutation for protecting privacy in microdata with multiple sensitive attributes', *ICMLC 2017: International Conference on Machine Learning and Cybernetics*, Ningbo, China, pp.1–13.

Zhou, C.L., Chen, Y.H., Tian, H. et al. (2020) 'Network k nearest neighbor query method for protecting location privacy and query content privacy', *Acta software Sinica*, Vol. 31, No. 2, pp.229–250.

Zhu, H., Tian, S., Xie, M. et al. (2014) 'Preserving privacy for sensitive values of individuals in data publishing based on a new additive noise approach', *IEEE 2014: Proceeding of the 3rd International Conference on Computer Communication and Networks*, pp.1–6.