Expert Systems With Applications 239 (2024) 122343

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Multi-level personalized k-anonymity privacy-preserving model based on sequential three-way decisions

Jin Qian [a,c,*], Haoying Jiang [a], Ying Yu [a], Hui Wang [a], Duoqian Miao [a,b]

[a] School of Software, East China Jiaotong University, Nanchang 330013, Jiangxi, China
[b] Department of Computer Science and Technology, Tongji University, Shanghai 201804, China
[c] School of Computer, Jiangsu University of Science and Technology, Zhenjiang, Jiangsu 212003, China

ARTICLE INFO

Keywords:
Decision-theoretic rough set
Sequential three-way decisions
K-anonymity
Personalized anonymity
Privacy preservation

ABSTRACT

K-anonymity is a widely used privacy-preserving technique which defends against linking attacks by suppression and generalization. The existing k-anonymity algorithms prevent attackers from illegally obtaining private information by constraining at least k records in an equivalence group. However, this unified anonymity method ignores individual differences and leads to a large amount of information loss. To this end, we introduce sequential three-way decisions into k-anonymity, using a dynamic k-value sequence instead of the fixed k-value to achieve personalized k-anonymity. Specifically, we first construct a hierarchical decision table for k-anonymity by attribute generalization trees and sensitive decision values provided with users. Then, we propose a multi-level personalized k-anonymity privacy-preserving model based on sequential three-way decisions, where we anonymize the partitioning granular data with a dynamic k-value sequence, respectively. Furthermore, we present three practical algorithms to implement the proposed model and discuss the differences between them. Finally, the experimental results demonstrate that the proposed model not only provides a more flexible anonymization method to achieve personalized anonymity, but greatly reduces the information loss. This study provides a complete framework for multi-level privacy protection and enriches the application of sequential three-way decisions.

1. Introduction

In era of big data, data sharing has become the mainstream. In order to fully exploit the potential value from big data, the scope of information sharing is now gradually expanding, meanwhile large volumes of data are being released and shared out. However, when data sharing brings convenience to people, it inevitably leads to the problem of privacy leakage. Inappropriate data publishing has led to huge amount of sensitive information being leaked, which endangers the information security of data providers. In recent years, many researches on privacy-preserving technologies have been in full swing (Cao, Wang, Li, Ren, & Lou, 2013; Denham, Pears, & Naeem, 2020; He, Zeadally, Xu, & Huang, 2015; Mehta & Rao, 2022; Wang et al., 2022). How to carry out efficient privacy protection has become a hot topic of research in various fields.

Data anonymization is a basic and widely-adopted privacy protection technology, which ensures the information security while maximizing data availability (Kacha, Zitouni, & Djoudi, 2022; Liang & Samavi, 2020; Mortazavi & Erfani, 2020; Song, Ma, Tian, & Al-Rodhaan, 2019). Since L. Sweeney first proposed the k-anonymity model in 2002 (Sweeney, 2002b), a variety of research results on k-anonymity have appeared. Wong et al. proposed the $(\alpha,\ k)$-anonymity model, which used the parameter "α" to control the frequency of sensitive attribute values to resist the probabilistic attack problem (Wong, Li, Fu, & Wang, 2006). Machanavajjhala et al. proposed the l-diversity model based on the k-anonymity model which can resist homogeneous (Machanavajjhala, Kifer, Gehrke, & Venkitasubramaniam, 2007). In order to resist similarity attacks and skewing problems, Li et al. proposed the t-closeness model (Li, Li, & Venkatasubramanian, 2006). To prevent attribute disclosure, the p-sensitive k-anonymity principle was proposed (Sun, Sun, & Wang, 2011; Truta & Vinay, 2006). For improving the efficiency of k-anonymity, Lin et al. proposed a genetic algorithm-based clustering method for k-anonymity using genetic algorithm (Lin & Wei, 2009). Subsequently, Ye et al. combined rough set theory to implement k-anonymity and improved the data utilization (Ye, Wu, Hu, & Hu, 2013). Considering the risk of disclosure in 1:M data publishing, Gong et al. proposed a novel privacy model called (k, l)-diversity (Gong,

[*] Corresponding author at: School of Software, East China Jiaotong University, Nanchang 330013, Jiangxi, China.
E-mail addresses: qjqjlqyf@163.com (J. Qian), jianghaoyingcx@163.com (H. Jiang), yuyingjx@163.com (Y. Yu), huiwangens@163.com (H. Wang), miaoduoqian@163.com (D. Miao).

https://doi.org/10.1016/j.eswa.2023.122343
Received 17 January 2023; Received in revised form 9 October 2023; Accepted 24 October 2023
Available online 4 November 2023

Luo, Yang, Ni, & Li, 2017). However, while these anonymization models solve most of problems, they all suffer from the same problem: they anonymize all data in a universal approach, and ignore the need for individualized privacy protection.

In fact, real-world datasets are unbalanced and the security requirements for different datasets may differ. To solve this problem, Xiao et al. first proposed the concept of personalized anonymity (Xiao & Tao, 2006). In recent years, a large number of scholars have paid attentions to the personalized anonymity and many research results were subsequently given out. Gedik et al. proposed a flexible privacy personalization framework to support the location k-anonymity, enabling the user to specify the minimum level of anonymity and the maximum temporal and spatial tolerance it desires (Gedik & Liu, 2008). Gao et al. proposed a personalized anonymization model to select the k-anonymized set of trajectories by considering different preference settings for trajectory privacy and data utility ratio in different scenarios (Gao, Ma, Sun, & Li, 2014). Liu et al. proposed a new personalized extended model based on the general $(\alpha,\ k)$-anonymity model, which enables personalized services while effectively protecting privacy (Liu, Xie, & Wang, 2016). Xiong et al. proposed a personalized privacy preservation (PERIO) framework based on game theory and data encryption, which achieves a reasonable balance between crowd-sensing service quality and privacy protection (Xiong et al., 2019). Guo et al. combined and optimized k-anonymity and differential privacy mechanisms to propose a new entropy-based personalized k-anonymity algorithm that improves the security of privacy protection (Guo, Yang, & Wan, 2021). Ren et al. proposed a new personalized (α, β, l, k)-anonymity model of social network, using the parameters α, β, l and k to satisfy the need for personalized privacy protection (Ren & Jiang, 2022). Unfortunately, an obvious weakness of the existing researches is that they only focus on the constraints of attributes and rarely investigate the personalized setting of k-values. For example, for a dataset with both 2-anonymity and 6-anonymity requirements, most existing anonymity models must use 6-anonymity for the entire table to ensure that the data can be safely published. Obviously, this approach is not reasonable and will result in unnecessary information loss. As a result, the study of the personalized settings of k-values is necessary to achieve efficient and cost-effective personalized k-anonymity. In order to further improve the personalized anonymity system, we focus on the personalized setting of k-values in this paper. Fig. 1 depicts the comparison between the traditional anonymity model and our proposed model.

Before implementing the personalized k-anonymity, we first need to separate the dataset for different anonymity requirements, and one key issue is how to divide the data with the lowest cost. The three-way decisions (3WD) (Yao, 2010, 2011, 2018) model provides a new approach to the classified anonymity. As is well-known to all, 3WD is an effective tool to deal with uncertain information, which can discover potential knowledge at minimal cost. In recent years, three-way decisions have been successfully applied in various fields (Hu, 2014; Liang, Pedrycz, Liu, & Hu, 2015; Liang, Xu, Liu, & Wu, 2018; Xu, Zheng, Liu, Yao, & Li, 2022; Yao, 2020; Yu, Zhang, & Wang, 2016; Zhan, Ye, Ding, & Liu, 2022). Sequential three-way decisions model (S3WD) is a commonly used 3WD model, which is closer to the human thinking mode and is recognized as a low-cost and high-efficiency classification decision-making model. As a dynamic multi-stage decision model, S3WD provides a more flexible decision-making mechanism for complex problems and has been well developed and applied in recent years. Qian et al. combined the multi-granulation rough set and sequential three-way decisions, proposed a generalized model of multi-granularity sequential three-way decisions, and enriched the development of multi-granularity three-way decisions (Qian, Liu, Miao, & Yue, 2020). Fang et al. considered the cost of the decision process or decision result, and proposed a granularity-driven sequential three-way decisions model (Fang, Gao, & Yao, 2020). Zhang et al. designed a penalty function to optimize the cost parameters, proposed a new sequential three-way decisions model based on penalty function (S3WDPF) and further improved the classification accuracy (Zhang, Pang, & Wang, 2020). Qian et al. combined sequential three-way decisions and hierarchical rough set model, and proposed a hierarchical sequential three-way decisions model, which can mine hierarchical sequential three-way decision rules under different levels of granularity (Qian, Tang, Yu, Yang, & Gao, 2022). Qian et al. proposed a cost-sensitive sequential three-way decisions model for fuzzy decision information systems and achieved better classification with lower cost (Qian, Zhou, Qian, & Wang, 2022).

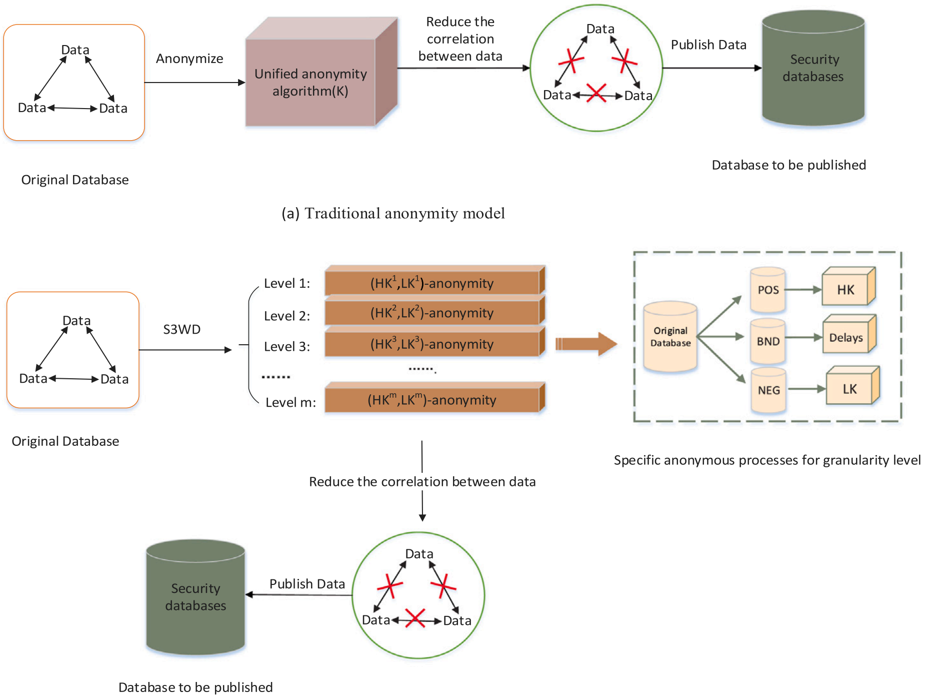
To sum up, existing data anonymization techniques ignore the personalized setting of k-value, which not only ignores the user's personalized needs but also leads to a large amount of unnecessary information loss. Therefore, investigating how to achieve multi-level k-anonymity is necessary to achieve efficient and accurate personalized anonymity. As we all know, the key to achieve multi-level k-anonymity is to efficiently classify the dataset, and S3WD provides a new idea for multi-level classification and anonymity. S3WD is a very cost-effective and efficient classification and decision model, and the progressive multi-stage processing model is well suited to achieve the personalized k-anonymity. However, few people have noticed the advantages of S3WD in dealing with the personalized privacy protection. To this end, we first introduce S3WD into k-anonymity, and present a multi-level personalized k-anonymity privacy-preserving model based on sequential three-way decisions (MKS3WD). In summary, this paper makes the following contributions:

1. We combine the sequential three-way decisions and the k-anonymity to propose a multi-level personalized k-anonymity privacy-preserving model based on sequential three-way decisions (MKS3WD). Different from the traditional k-anonymity with only one fixed k-value, our model uses a dynamic k-value sequence for anonymity and achieves multi-constrained personalized k-anonymity.
2. We present a hierarchical decision table for k-anonymity based on attribute generalization trees and sensitive decision values provided by users, and then propose a sequential three-way decisions model for classification, which is able to divide data with different sensitivities into different granularity structures at the lowest cost
3. We propose three practical algorithms to implement the MKS3WD model, namely, the security downscaling scheme (SD-MKS3WD), the sensitivity extraction scheme (SE-MKS3WD) and the equivalence class extraction scheme (ECE-MKS3WD), which further reduces the information loss while ensuring data security and availability.

The remainder of this paper is organized as follows. In Section 2, we briefly introduce the basic models of k-anonymity, decision-theoretic rough set and sequential three-way decisions. Section 3 presents a hierarchical decision table for k-anonymity and a sequence three-way decisions model for classification, then describe a generalized multi-level personalized k-anonymity privacy-preserving model based on sequential three-way decisions. In Section 4, we propose three practical algorithms to implement our model, some corresponding examples are given for illustration. Section 5 gives the relevant experiments and conclusions. Section 6 summarizes the work of this paper as well as provides an outlook on future research directions.

2. Preliminaries

In this section, we will review some basic concepts of k-anonymity, decision-theoretic rough set and sequential three-way decisions. For a detailed description, please refer to papers (Sweeney, 2002b; Yao & Deng, 2011; Yao & Wong, 1992; Yao, Wong, & Lingras, 1990).


Data
Data
Data
Original Database
Anonymize
Unified anonymity algorithm(K)
Reduce the correlation between data
Data
Data
Data
Publish Data
Security databases
Database to be published
S3WD
Level 1:
Level 2:
Level 3:
……
Level m:
(HK^1,LK^1)-anonymity
(HK^2,LK^2)-anonymity
(HK^3,LK^3)-anonymity
…….
(HK^m,LK^m)-anonymity
Original Database
POS
BND
NEG
HK
Delays
LK
Specific anonymous processes for granularity level
Reduce the correlation between data
Publish Data
Security databases
Database to be published

(a) Traditional anonymity model

(b) Multi-level k-anonymity personalized privacy protection model based on sequential three-way decisions (MKS3WD)

Fig. 1. Comparison of the traditional anonymity model and our proposed model (MKS3WD).

2.1. K-anonymity model

K-anonymity is an important anonymity model, which defends against link attacks by suppression and generalization (Sweeney, 2002a, 2002b). Here are some brief introductions of k-anonymity.

Definition 1. T is an original information table, according to the characteristics of attributes, the attributes of T can be divided into the following four categories.

- **Identification Attributes (ID):**
  The attributes which can clearly identify the identity of the objects.
- **Quasi-Identifier Attributes (QID):**
  Also known as semi-sensitive attributes, the union of several quasi-identifier attributes can determine the identity of the objects with a higher probability.
- **Sensitive Attributes (SA):**
  Attributes that contain sensitive information about the objects.
- **Other Attributes (OA):**
  Attributes that are different from the above three and do not require special treatment.

Definition 2 (*Equivalence Class and Equivalence Group*)**.** The objects with the same quasi-identifier attribute values are equivalence classes to each other. The sets of identical equivalence classes are equivalence groups.

Definition 3 (*K-anonymity Sweeney, 2002b*)**.** Given an information table $T = (U, At)$, where $U = \{x_1, x_2, \dots, x_n\}$ is a finite non-empty set of objects, $At = \{a_1, a_2, \dots, a_m\}$ is a finite nonempty set of attributes. QID denotes the quasi-identifier attribute in T, $QID \subseteq At$. $T[QID]$ denotes the set of all quasi-identifier attribute value sequences in T. T is said to satisfy k-anonymity if and only if each sequence of values in $T[QID]$ appears with at least k occurrences in $T[QID]$.

2.2. Decision-theoretic rough set models (DTRS)

Decision-theoretic rough set (DTRS) (Yao & Wong, 1992; Yao et al., 1990) is a probabilistic extension of algebraic rough set model. By introducing Bayesian risk decision theory, DTRS uses tolerance threshold pairs (α, β) to partition the data $(0 \le \beta \le \alpha \le 1)$.

Definition 4. Given a set of states $\Omega = \{O, \neg O\}$ indicating that the object x is in O and not in O, respectively. The set of action is given by $A = \{a_P, a_N, a_B\}$, which represents the three different decisions for the object x. These three decisions deciding $x \in$ POS(O), $x \in$ BND(O) and $x \in$ NEG(O), where $POS(O)$, $BND(O)$ and $NEG(O)$ represent the probabilistic positive, boundary and negative regions, respectively. When object x belongs to O, $\{\lambda_{PP}, \lambda_{BP}, \lambda_{NP}\}$ denotes the loss incurred for making decision of $\{a_P, a_N, a_B\}$. Similarly, $\{\lambda_{PN}, \lambda_{BN}, \lambda_{NN}\}$ denotes the loss incurred for making corresponding decision when object x belongs to $\neg O$. For the object x, the expected losses associated with taking different decisions can be expressed as (Yao & Wong, 1992; Yao et al., 1990):

$$\begin{aligned} R(a_P|[x]) &= \lambda_{PP} P(O|[x]) + \lambda_{PN} P(\neg O|[x]), \\ R(a_B|[x]) &= \lambda_{BP} P(O|[x]) + \lambda_{BN} P(\neg O|[x]), \\ R(a_N|[x]) &= \lambda_{NP} P(O|[x]) + \lambda_{NN} P(\neg O|[x]). \end{aligned} \tag{1}$$

According to Bayesian decision-theoretic framework, when $0 \le \lambda_{PP} \le \lambda_{BP} \le \lambda_{NP}$ and $0 \le \lambda_{NN} \le \lambda_{BN} \le \lambda_{PN}$, the minimum-risk decision rules are given as follows:

(P1) if $P(O|[x]) \ge \alpha$ and $P(O|[x]) \ge \gamma$, decide $x \in POS(O)$,
(B1) if $P(O|[x]) \le \alpha$ and $P(O|[x]) \ge \beta$, decide $x \in BND(O)$,
(N1) if $P(O|[x]) \le \beta$ and $P(O|[x]) \le \gamma$, decide $x \in NEG(O)$.

where α, β and γ are defined as:

$$\begin{aligned} \alpha &= \frac{(\lambda_{PN} - \lambda_{BN})}{(\lambda_{PN} - \lambda_{BN}) + (\lambda_{BP} - \lambda_{PP})}, \\ \beta &= \frac{(\lambda_{BN} - \lambda_{NN})}{(\lambda_{BN} - \lambda_{NN}) + (\lambda_{NP} - \lambda_{BP})}, \\ \gamma &= \frac{(\lambda_{PN} - \lambda_{NN})}{(\lambda_{PN} - \lambda_{NN}) + (\lambda_{NP} - \lambda_{PP})}. \end{aligned} \tag{2}$$

When $0 \leq \beta \leq \gamma \leq \alpha \leq 1$, we can obtain the decision rules:

(P2) if $P(O|[x]) \geq \alpha$, decide $x \in POS(O)$,
(B2) if $\beta < P(O|[x]) < \alpha$, decide $x \in BND(O)$,
(N2) if $P(O|[x]) \leq \beta$, decide $x \in NEG(O)$.

Thus, given two parameters α and β for a decision class D_i with respect to an equivalence relation EC, the probabilistic lower and upper approximations can be defined by:

$$\begin{aligned} \underline{apr}_{EC}^{(\alpha,\ \beta)}(D_i) &= \{x \in U | P(D_i|[x]_{EC}) \geq \alpha\}, \\ \overline{apr}_{EC}^{(\alpha,\ \beta)}(D_i) &= \{x \in U | P(D_i|[x]_{EC}) > \beta\}. \end{aligned} \tag{3}$$

Similarly, the probabilistic positive, boundary and negative regions are given by:

$$\begin{aligned} POS_{EC}^{(\alpha,\ \beta)}(D_i) &= \underline{apr}_{EC}^{(\alpha,\ \beta)}(D_i) \\ &= \{x \in U | P(D_i|[x]_{EC}) \geq \alpha\}; \end{aligned} \tag{4}$$

$$\begin{aligned} BND_{EC}^{(\alpha,\ \beta)}(D_i) &= \overline{apr}_{EC}^{(\alpha,\ \beta)}(D_i) - \underline{apr}_{EC}^{(\alpha,\ \beta)}(D_i) \\ &= \{x \in U |\ \beta < P(D_i|[x]_{EC}) < \alpha\}; \end{aligned} \tag{5}$$

$$\begin{aligned} NEG_{EC}^{(\alpha,\ \beta)}(D_i) &= U - \underline{apr}_{EC}^{(\alpha,\ \beta)}(D_i) \cup \overline{apr}_{EC}^{(\alpha,\ \beta)}(D_i) \\ &= \{x \in U | P(D_i|[x]_{EC}) \leq \beta\}. \end{aligned} \tag{6}$$

where $P(D_i|[x]_{EC})$ denotes the conditional probability, $P(D_i|[x]_{EC}) = \frac{[x]_{EC} \cap D_i}{[x]_{EC}}$.

In conclusion, DTRS is a low-cost classification decision model, which divides the parameters of the probabilistic positive, boundary and negative regions by calculating the cost (risk) concept. DTRS further improves the Pawlak rough set model and is the basis for three-way decisions.

2.3. Sequential three-way decisions (S3WD)

As we all know, sequential three-way decisions model (S3WD) is a progressive multi-stage decision model (Yao, 2013; Yao & Deng, 2011). In S3WD, we can select appropriate threshold pairs $(\alpha,\ \beta)^l$ depending on the DTRS, and then we divide the data into different granularity spaces to make appropriate decisions for each space separately. With the granularity calculation, S3WD can obtain the most accurate results with the best cost effectiveness. In this subsection, we briefly review a general model of sequential three-way decisions.

Definition 5. For an information table S $= (U, At = C \bigcup D, V, f)$, given a decision class D_i^l, a dynamic threshold parameter sequence $(\alpha,\ \beta)^l = \{(\alpha^1,\ \beta^1), (\alpha^2,\ \beta^2), \ldots, (\alpha^l,\ \beta^l))$. For an equivalence relation EC, the $(\alpha^l,\ \beta^l)$-lower approximation $\underline{apr}_{EC}^{(\alpha^l,\ \beta^l)}$ and the $(\alpha^l,\ \beta^l)$-upper approximation $\overline{apr}_{EC}^{(\alpha^l,\ \beta^l)}$ are defined by

$$\begin{aligned} \underline{apr}_{EC}^{(\alpha^l,\ \beta^l)}(D_i^l) &= \{x \in U^l | P(D_i^l|[x]_{EC}) \geq \alpha^l\}, \\ \overline{apr}_{EC}^{(\alpha^l,\ \beta^l)}(D_i^l) &= \{x \in U^l | P(D_i^l|[x]_{EC}) > \beta^l\}. \end{aligned} \tag{7}$$

where $U^1 = U$, $U^{l+1} = BND_{EC}^{(\alpha^l, \beta^l)}(D_i^l) = \overline{apr}_{EC}^{(\alpha^l,\ \beta^l)}(D_i^l) - \underline{apr}_{EC}^{(\alpha^l,\ \beta^l)}(D_i^l)$, D_i^l represents the equivalence class including x in the partition U^l/D_i^l, and $[x]_{EC}$ represents the equivalence class including x in the partition U^l/EC.

The pair $< \underline{apr}_{EC}^{(\alpha^l,\ \beta^l)}, \overline{apr}_{EC}^{(\alpha^l,\ \beta^l)} >$ is called the l^{th}-level lower and upper approximations induced by EC with respect to D_i^l in U^l. Therefore, we can obtain the three probabilistic regions as follows

$$\begin{aligned} POS_{EC}^{(\alpha^l, \beta^l)}(D_i^l) &= \underline{apr}_{EC}^{(\alpha^l,\ \beta^l)}(D_i^l) \\ &= \{x \in U^l | P(D_i^l|[x]_{EC}) \geq \alpha^l\}; \end{aligned} \tag{8}$$

$$\begin{aligned} BND_{EC}^{(\alpha^l, \beta^l)}(D_i^l) &= \overline{apr}_{EC}^{(\alpha^l,\ \beta^l)}(D_i^l) - \underline{apr}_{EC}^{(\alpha^l,\ \beta^l)}(D_i^l) \\ &= \{x \in U^l |\ \beta^l < P(D_i^l|[x]_{EC}) < \alpha^l\}; \end{aligned} \tag{9}$$

$$\begin{aligned} NEG_{EC}^{(\alpha^l, \beta^l)}(D_i^l) &= U^l - POS_{EC}^{(\alpha^l, \beta^l)}(D_i^l) \cup BND_{EC}^{(\alpha^l, \beta^l)}(D_i^l) \\ &= \{x \in U^l | P(D_i^l|[x]_{EC}) \leq \beta^l\}. \end{aligned} \tag{10}$$

Table 1
An original information table.

ID		*QID*			*SA*	
NO.	Name	Sex	Age	Unit	GPA(d)	f
1	Mary	F	20	CT1	3.6	0.8
2	Kelly	F	21	IS2	4.0	0.5
3	Tome	M	24	ME1	4.3	0.3
4	Mango	M	15	BE1	3.4	0.9
5	Lisa	F	16	BS1	4.0	0.5
6	Alice	F	18	BS2	4.5	0.1
7	Peter	M	22	AM2	3.0	0.2
8	White	F	22	CT3	3.8	0.9
9	Lili	M	25	AM2	2.8	0.7
10	Amy	F	19	BE3	4.4	0.6

3. Multi-level personalized k-anonymity privacy-preserving model based on sequential three-way decisions

In this section, we first propose a hierarchical decision table for k-anonymity. Then, in order to introduce sequential three-way decisions into k-anonymity, we present a sequential three-way decision model for classification. Finally, we propose a generalized multi-level personalized k-anonymity privacy-preserving model by combining sequential three-way decisions model for classification and k-anonymity model.

3.1. Hierarchical decision table for k-anonymity

K-anonymity is a widely-adopted privacy protection model, in which suppression and generalization are two major operations. In the real-world applications, the generalization of attribute can form a complete attribute generalization tree and have obvious hierarchical relations. Thus, to facilitate the construction of multi-level personalized k-anonymity privacy-preserving model, in this subsection, we define attribute generalization tree and attribute generalization forest, and then define a hierarchical decision table for k-anonymity using the attribute generalization forest and the personalized sensitive decision values.

Definition 6. Given an information table $T = \{U, At\}$, where $U = \{x_1, x_2, \ldots, x_n\}$ is a finite non-empty set of objects, $At = \{a_1, a_2, \ldots, a_m\}$ is a finite nonempty set of attributes. Suppose attribute a_i has c attribute generalization levels, then $GT_i = \{a_i^1, a_i^2, \ldots, a_i^c\}$ is an attribute generalization tree with respect to a_i, where a_i^l represents the set of values for attribute a_i generalized to the lth level $(l = 1, 2, \ldots, c)$. $GT = \bigcup_{i=1}^m GT_i$ is the attribute generalization forest of T.

Definition 7. Let $HDT = \{U, QID, D = (d, f)\}$ be a hierarchical decision table for k-anonymity, where $U = \{x_1, x_2, \ldots, x_n\}$ is a finite non-empty set of objects, $QID = \{\{a_1^1, a_1^2, \ldots, a_1^c\}, \{a_2^1, a_2^2, \ldots, a_2^c\}, \ldots, \{a_m^1, a_m^2, \ldots, a_m^c\}\}$ is a finite non-empty set of quasi-identifier attributes, c is the maximum height of the attribute generalization trees for QID, D denotes the sensitive attribute, d denotes the sensitive attribute value of D, and f denotes the personalized sensitive decision value of D, $f \in [0, 1]$.

It is worth noting that different attributes may generate attribute generalization trees with inconsistent heights in the real-world problems. Thus, if there exists $|H(a_i)| < c$, where $|H(a_i)|$ is the height of the attribute generalization tree for attribute a_i, and c is the maximal height of the generalization trees for QID, we supplement the data in the hierarchical decision table by repeating the values of the maximum generalization level for a_i.

Example 1. Table 1 is an original information table. "No." and "Name" are the identification attributes (ID); "Sex", "Age" and "Unit" are the quasi-identifier attributes (QID); "SA" is the sensitive attribute; "GPA" denotes the sensitive attribute value of SA; "f" denotes

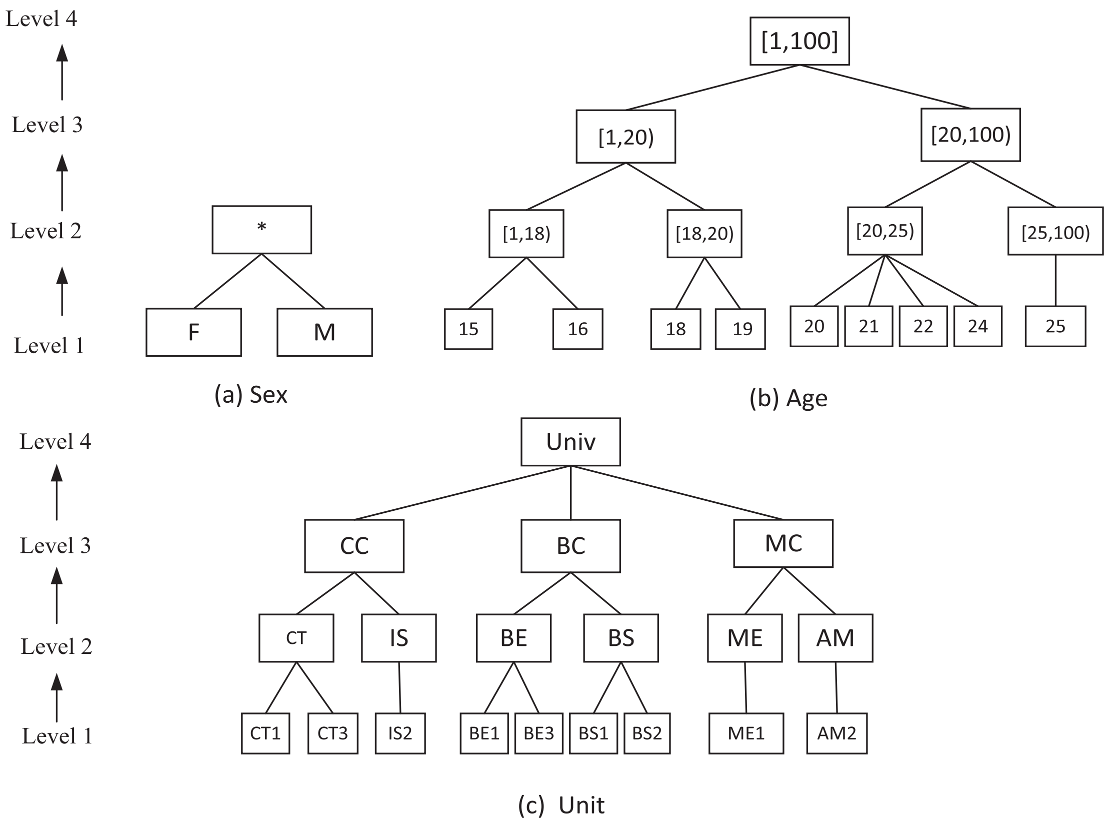


Fig. 2. The attribute generalization forest of Table 1.

Table 2
The hierarchical decision table for k-anonymity of Table 1.

U	QID												D	
	Sex				Age				Unit				SA	
	a_1^1	a_1^2	a_1^3	a_1^4	a_2^1	a_2^2	a_2^3	a_2^4	a_3^1	a_3^2	a_3^3	a_3^4	GPA(d)	f
x_1	F	*	*	*	20	[20, 25)	[20, 100)	[1, 100]	CT1	CT	CC	Univ	3.6	0.8
x_2	F	*	*	*	21	[20, 25)	[20, 100)	[1, 100]	IS2	IS	CC	Univ	4.0	0.5
x_3	M	*	*	*	24	[20, 25)	[20, 100)	[1, 100]	ME1	ME	MC	Univ	4.3	0.3
x_4	M	*	*	*	15	[1, 18)	[1, 20)	[1, 100]	BE1	BE	BC	Univ	3.4	0.9
x_5	F	*	*	*	16	[1, 18)	[1, 20)	[1, 100]	BS1	BS	BC	Univ	4.0	0.5
x_6	F	*	*	*	18	[18, 20)	[1, 20)	[1, 100]	BS2	BS	BC	Univ	4.5	0.1
x_7	M	*	*	*	22	[20, 25)	[20, 100)	[1, 100]	AM2	AM	MC	Univ	3.0	0.2
x_8	F	*	*	*	22	[20, 25)	[20, 100)	[1, 100]	CT3	CT	CC	Univ	3.8	0.9
x_9	M	*	*	*	25	[25, 100)	[20, 100)	[1, 100]	AM2	AM	MC	Univ	2.8	0.7
x_{10}	F	*	*	*	19	[18, 20)	[1, 20)	[1, 100]	BE3	BE	BC	Univ	4.4	0.6

the personalized sensitive decision value of SA. From Definition 6, we can obtain the attribute generalization forest of QID, which are shown in Fig. 2. Then, we can generate a hierarchical decision table for k-anonymity by the attribute generalization forest. To do this, we first obtain that $QID = \{\{a_1^1, a_1^2\}, \{a_2^1, a_2^2, a_2^3, a_2^4\}, \{a_3^1, a_3^2, a_3^3, a_3^4\}\}$ by observing Fig. 2. According to the previous analysis, $c = 4$ and $|H(a_1)| = 2 < 4$, we can supplement the attribute "Sex" by repeating the value of a_1^2. Finally, we can obtain the final set of quasi-identified attributes $QID = \{\{a_1^1, a_1^2, a_1^3, a_1^4\}, \{a_2^1, a_2^2, a_2^3, a_2^4\}, \{a_3^1, a_3^2, a_3^3, a_3^4\}\}$, where $a_1^2 = a_1^3 = a_1^4$. Table 2 illustrates the hierarchical decision table for k-anonymity of Table 1.

In the end, by constructing a hierarchical decision table for k-anonymity, we can perform the unified hierarchical process for both numerical and category attributes. At the same time, we also consider the personalized sensitive values and use them as the sensitive decision values. All these works provide the basis for the construction of multi-level personalized k-anonymity model.

3.2. Sequential three-way decisions model for classification

In order to classify the data with different security requirements into the appropriate granular spaces in a reasonable manner, in what follows, we combine the hierarchical decision table for k-anonymity and the sequential three-way decisions to propose a sequential three-way decisions model for classification.

Definition 8. Given a hierarchical decision table $HDT = \{U, QID, D = (d, f)\}$, a dynamic threshold parameter sequence $(\alpha, \beta)^l = \{(\alpha^1, \beta^1), (\alpha^2, \beta^2), \ldots, (\alpha^l, \beta^l)\}$, EC_t is an equivalence relation induced by $QID_t (t = 1, 2, \ldots, l)$, then the tth level of granular structure GS_t is defined as

$$GS_t = \{HDT_t, QID_t, EC_t, \alpha^t, \beta^t\} \tag{11}$$

where HDT_t denotes a hierarchical decision table for k-anonymity under the tth level granular structure GS_t, QID_t denotes the quasi-identifier attributes under GS_t.

Definition 9. For a hierarchical decision table $HDT = \{U, QID, D = (d, f)\}$, where $U = \{x_1, x_2, \ldots, x_n\}$ is a finite non-empty set of objects, $QID = \{a_1, a_2, \ldots, a_m\}$ is a finite nonempty set of quasi-identifier attributes, where D denotes the sensitive attribute, d denotes the sensitive attribute values of D, and f denotes the personalized sensitive decision values of D. Given a dynamic threshold parameter sequence $(\alpha, \beta)^l = \{(\alpha^1, \beta^1), (\alpha^2, \beta^2), \ldots, (\alpha^l, \beta^l)\}$. For a multilevel granular structure $GS = \{GS_1, GS_2, \ldots, GS_l\}$, the (α^l, β^l)-lower approximation $\underline{apr}_{GS_l}^{(\alpha^l, \beta^l)}$ and the (α^l, β^l)-upper approximation $\overline{apr}_{GS_l}^{(\alpha^l, \beta^l)}$ are defined by

$$\begin{aligned} \underline{apr}_{GS_l}^{(\alpha^l, \beta^l)}(D) &= \{x \in U^l \mid f_x(D) \geq \alpha^l\}, \\ \overline{apr}_{GS_l}^{(\alpha^l, \beta^l)}(D) &= \{x \in U^l \mid f_x(D) > \beta^l\}. \end{aligned} \tag{12}$$

where $U^1 = U$, $U^{l+1} = BND^{(\alpha^l,\beta^l)}_{GS_l}(D) = \overline{apr}^{(\alpha^l,\beta^l)}_{GS_l}(D) - \underline{apr}^{(\alpha^l,\beta^l)}_{GS_l}(D)$, and $f_x(D)$ represents the sensitive decision value of object x corresponding to the sensitive attribute D.

The pair $< \underline{apr}^{(\alpha^l,\beta^l)}_{GS_l}, \overline{apr}^{(\alpha^l,\beta^l)}_{GS_l} >$ is called the lth-level lower and upper approximations induced by GS_l with respect to D in U^l. Thus, we can obtain the positive, boundary and negative regions $POS^{(\alpha^l,\beta^l)}_{GS_l}(D)$, $BND^{(\alpha^l,\beta^l)}_{GS_l}(D)$ and $NEG^{(\alpha^l,\beta^l)}_{GS_l}(D)$ as follows

$$\begin{aligned} POS^{(\alpha^l,\beta^l)}_{GS_l}(D) &= \underline{apr}^{(\alpha^l,\beta^l)}_{GS_l}(D) \\ &= \{x \in U^l | f_x(D) \geq \alpha^l\}; \end{aligned} \tag{13}$$

$$\begin{aligned} BND^{(\alpha^l,\beta^l)}_{GS_l}(D) &= \overline{apr}^{(\alpha^l,\beta^l)}_{GS_l}(D) - \underline{apr}^{(\alpha^l,\beta^l)}_{GS_l}(D) \\ &= \{x \in U^l | \beta^l < f_x(D) < \alpha^l\}; \end{aligned} \tag{14}$$

$$\begin{aligned} NEG^{(\alpha^l,\beta^l)}_{GS_l}(D) &= U^l - POS^{(\alpha^l,\beta^l)}_{GS_l}(D) \cup BND^{(\alpha^l,\beta^l)}_{GS_l}(D) \\ &= \{x \in U^l | f_x(D) \leq \beta^l\}. \end{aligned} \tag{15}$$

It is worth mentioning that since the sequential three-way decisions is a cost-effective way of decision making, the threshold pair $(\alpha,\beta)^l$ satisfies the decision-theoretic rough set and can achieve the highest accuracy at the lowest cost. Furthermore, our proposed sequential three-way decision model for classification only adjusts the decision values, and still inherit the advantages of low-cost data classification of sequential three-way decisions. Example 2 illustrates the application of sequential three-way decision model for classification with a specific case.

Example 2. Given a set of objects $U = \{x_1, x_2, \ldots, x_{17}\}$, a decision class D, and sensitive decision values $f_U(D)$ = {0.41, 0.89, 0.54, 0.94, 0.42, 0.48, 0.61, 0.17, 0.56, 0.70, 0.53, 0.85, 0.87, 0.80, 0.62, 0.68, 0.65}. Let $(\alpha,\ \beta)^3$ = {(0.8 ,0.2), (0.6, 0.5), (0.55, 0.54)}, based on Definition 9, we can conclude the following

(a) For the first level of granular, $U^1 = U$, $(\alpha^1,\ \beta^1) = (0.8, 0.2)$, we can compute
$POS^{(\alpha^1,\beta^1)}_{GS_1}(D) = \{4, 2, 13, 12, 14\}$;
$BND^{(\alpha^1,\beta^1)}_{GS_1}(D) = \{10, 16, 17, 15, 7, 9, 3, 11, 6, 5, 1\}$;
$NEG^{(\alpha^1,\beta^1)}_{GS_1}(D) = \{8\}$;

(b) For the second level of granular, $U^2 = BND^{(\alpha^1,\beta^1)}_{GS_1}$(D), $(\alpha^2,\ \beta^2) = (0.6, 0.5)$, we can compute
$POS^{(\alpha^2,\beta^2)}_{GS_2}(D) = \{10, 16, 17, 15, 7\}$;
$BND^{(\alpha^2,\beta^2)}_{GS_2}(D) = \{9, 3, 11\}$;
$NEG^{(\alpha^2,\beta^2)}_{GS_2}(D) = \{6, 5, 1\}$;

(c) For the last level of granular, $U^3 = BND^{(\alpha^2,\beta^2)}_{GS_2}$(D), $(\alpha^3,\ \beta^3) = (0.55, 0.54)$, we can compute
$POS^{(\alpha^3,\beta^3)}_{GS_3}(D) = \{9\}$;
$BND^{(\alpha^3,\beta^3)}_{GS_3}(D) = \emptyset$;
$NEG^{(\alpha^3,\beta^3)}_{GS_3}(D) = \{3, 11\}$.

However, since the risk functions λ may not be consistent in different application scenarios, the threshold sequences $(\alpha^l,\ \beta^l)$ used may not be the same. In this paper, since we cannot consider all scenarios, the threshold sequence $(\alpha^l,\ \beta^l)$ used in our experiments may also not be applicable to all application scenarios and can only represent some of them, and we obtain better results, which reflects the usability of the proposed model. Fortunately, this problem can be well solved in real applications. This best thresholds can be solved in the practical applications by calculating accurate risk assessment values given by experts, which is beyond the scope of this paper.

3.3. A generalized multi-level personalized k-anonymity privacy-preserving model based on sequential three-way decisions (MKS3WD)

In this subsection, we define a multi-level k-anonymity and then combine the sequential three-way decisions model for classification and the k-anonymity model to propose a generalized multi-level personalized k-anonymity privacy-preserving model based on sequential three-way decisions (MKS3WD).

Definition 10 (*Multi-Level k-anonymity*)**.** Given a hierarchical decision table $HDT = \{U, QID, D = (d, f)\}$, a dynamic k-value sequence $(HK, LK)^l = \{(HK^1,\ LK^1),\ (HK^2,\ LK^2),\ \cdots,\ (HK^l,\ LK^l)\}$, a dynamic threshold parameter sequence $(\alpha,\ \beta)^l = \{(\alpha^1,\ \beta^1), (\alpha^2,\ \beta^2), \ldots, (\alpha^l,\ \beta^l)\}$, for a multilevel granular structure $GS = \{GS_1, GS_2, \ldots, GS_l\}$, if $POS^{(\alpha^t,\beta^t)}_{GS_t}(D)$ satisfies HK^t and $NEG^{(\alpha^t,\beta^t)}_{GS_t}(D)$ satisfies $LK^t(t = 1, 2, \ldots, l)$, then T satisfies multi-level k-anonymity.

As we all know, the existing k-anonymity models use a fixed parameter k to anonymize the entire table, which is not suitable for handling unbalanced data in the real world. To solve this drawback, we focus on how to personalize data anonymization with different parameters k for different security requirements in this paper. To do this, we combine the sequential three-way decisions model for classification and the k-anonymity model to propose a multi-level personalized k-anonymity privacy-preserving model, using a dynamic k-value sequence $(HK, LK)^l$ to implement the multi-level k-anonymity. According to Definition 10, the dataset satisfies multi-level k-anonymity when and only when all the granular datasets satisfy the corresponding anonymous requirements. Fig. 3 shows our model.

Specifically, our proposed model is implemented in two major stages. In the first stage, we divide the dataset into *l*-level granular structures by the sequential three-way decisions model for classification, and then at each granularity level, we divide the data into three regions by the sensitive decision values. High and low sensitive data are divided into the positive region and negative region, respectively, while the data with the ambiguous sensitivity will form the boundary region. In the second stage, we anonymize the positive and negative regions with high requirements HK^l and low requirements LK^l, respectively. For the objects in the boundary regions, we delay anonymizing them and process them at the next level for the finer anonymization. As the level increases, the degree of anonymity decreases in the positive regions and increases in the negative regions, namely, $HK^1 > HK^2 > \cdots > HK^l > LK^l > \cdots LK^2 > LK^1$.

However, this hierarchical anonymization approach inevitably raises a new question: how do we handle the remaining data in the upper levels that cannot be anonymized? To address this problem, we present three practical algorithms to implement our model using secure downscaling scheme (SD-MKS3WD), sensitivity extraction scheme (SE-MKS3WD) and equivalence class extraction scheme (ECE-MKS3WD). In what follows, we will illustrate these three algorithms in details.

4. Three practical algorithms of multi-level personalized k-anonymity privacy-preserving model based on sequential three-way decisions

In this section, we mainly investigate how to handle data that cannot be anonymized in the upper levels. It is well known that k-anonymity defends against linking attacks by ensuring that there are at least k records in the equivalence group to interfere with the attacker's judgment. Thus, when the number of records in the equivalence group is less than k or the remaining records cannot satisfy k-anonymity even generalized to the root node, the remaining data will not be anonymized successfully, and the number of such unanonymizable data will increase with the diversity of k-values.

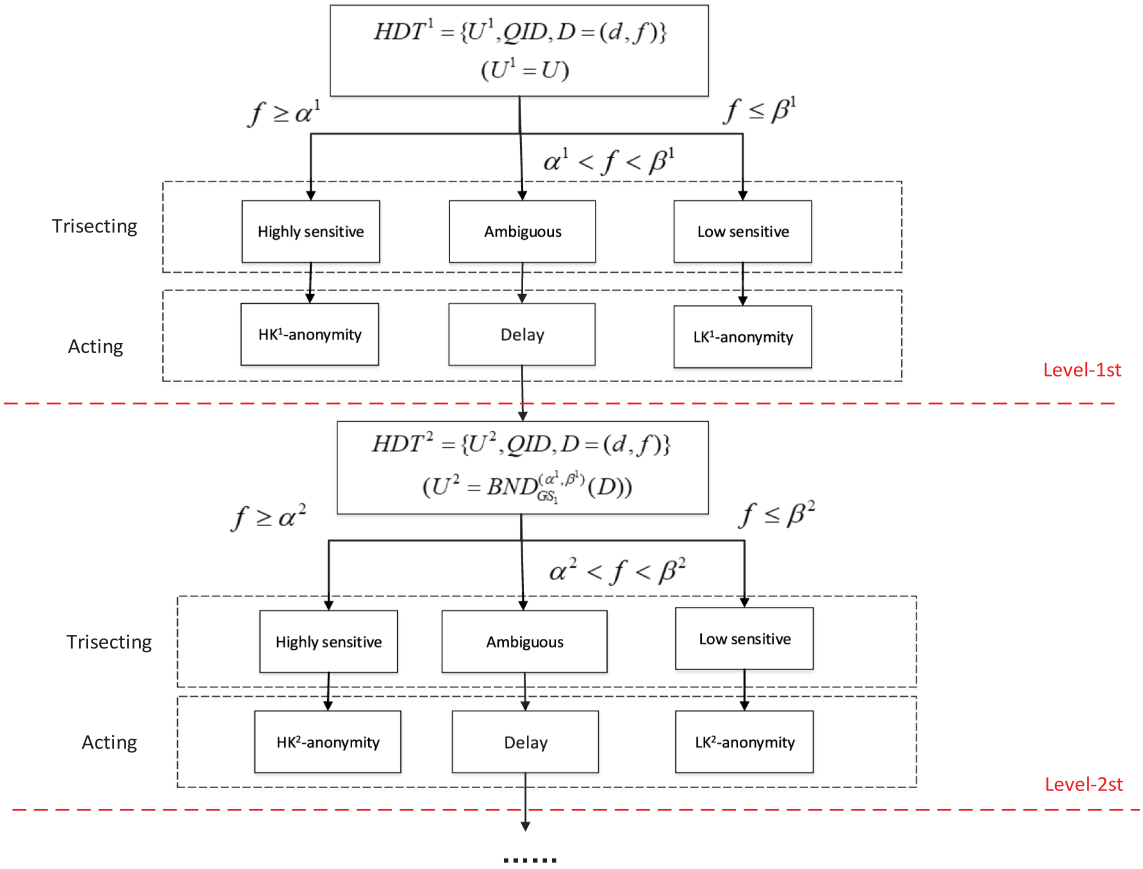
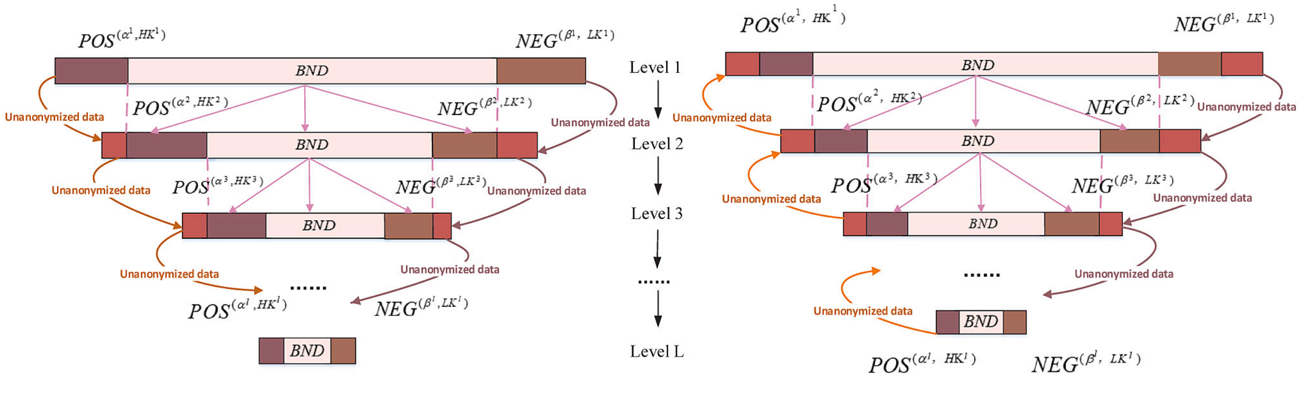


Fig. 3. A generalized multi-level personalized k-anonymity privacy-preserving model based on sequential three-way decisions (MKS3WD).



(a) SD-MKS3WD

(b) SE-MKS3WD and ECE-MKS3WD

Fig. 4. Three practical algorithms of multi-level personalized k-anonymity privacy-preserving model based on sequential three-way decisions.

To solve this drawback, we propose three practical algorithms to implement our model. SD-MKS3WD downscales the data that cannot be anonymized from the upper level to the lower level to process the remaining data, SE-MKS3WD processes the remaining data by extracting the data with the closest sensitivity from the lower levels, and ECE-MKS3WD directly extracts the equivalence of the remaining data from the lower levels to achieve the anonymization of remaining data. These two data extraction algorithms (SE-MKS3WD and ECE-MKS3WD) use different strategies to extract data from the lower levels, respectively. Fig. 4 describes the core ideas of these three algorithms.

4.1. Multi-level personalized k-anonymity privacy-preserving model based on sequential three-way decisions using secure downscaling (SD-MKS3WD)

The security downscaling scheme is the simplest, where we merge the data that cannot be anonymized in the positive and negative regions with the objects in the boundary region, and then use them as a new theoretical domain for the next level. In this way, we can link different granular structures so that the data that cannot be anonymized in the upper granular structure can be fully utilized, and the data availability is improved. Fig. 4(a) shows the model of SD-MKS3WD.

Specifically, as shown in Algorithm 1, we first sort the dataset with the sensitive decision value f, then use the sequential three-way decisions model for classification to divide them into different granular structures GS_l. Next, we set up the appropriate k-value pairs (HK^l, LK^l) for every granular structure and anonymize them separately. For the data that cannot be anonymized at GS_t $(t = 1, 2, \dots, l-1)$, we merge them with the objects at GS_{t+1} for (HK^{t+1}, LK^{t+1})-anonymity and so on. Moreover, it is important to note that for the implementation of k-anonymity in our algorithm, we adopt the means of suppression and generalization. Specifically, we build the hierarchical decision table for k-anonymity, please refer to Section 3.1 for details. First, we take the quasi-identified attributes of level 1 in the hierarchical decision table to divide the data into equivalence classes, and if the divided equivalence groups satisfy k-anonymity, they are imported into the secure dataset Q and are not involved in the subsequent operations. On the contrary, the data that does not satisfy the anonymity requirement is generalized in the next step by taking the values of the attributes at level 2 in the hierarchical decision table to divide the equivalence classes again, and so on until the anonymization is completed or the quasi-identified attribute reaches the maximum generalization level. If there are data that still do not satisfy the anonymity requirement even if the quasi-identified attribute is generalized to the top level, we suppress it, i.e., delete the whole record. Therefore, on the basis of the above analysis, this is easy to see that the time complexity of the SD-MKS3WD is $O(m \cdot l)$.

To make it easier to understand, we also provide a concise explanation with a specific example as shown in Example 3. To simplify the description, in the following we replace $POS^{(\alpha^t,\beta^t)}_{GS_t}(D)$, $BND^{(\alpha^t,\beta^t)}_{GS_t}(D)$, $NEG^{(\alpha^t,\beta^t)}_{GS_t}(D)$ with $POS_{GS_t}(D)$, $BND_{GS_t}(D)$, $NEG_{GS_t}(D)$, respectively.

Example 3 (*Continued with Example 2*). Consider a partition of equivalence classes $U/QID = \{\{4, 2, 13, 12, 14, 7\}, \{10, 16, 17, 15, 9\}, \{3, 11, 5\}, \{6, 1\}, \{8\}\}$, a dynamic anonymous parameters sequence $(HK, LK)^3 = \{(6,1), (5,2), (4,3)\}$, we can conclude the following

(a) For the first level of granular, $U^1 = U$, $(\alpha^1, \beta^1) = (0.8, 0.2)$, $(HK^1, LK^1) = (6, 1)$, we can compute
$POS_{GS_1}(D) = \{4, 2, 13, 12, 14\}$;
$BND_{GS_1}(D) = \{10, 16, 17, 15, 7, 9, 3, 11, 6, 5, 1\}$;
$NEG_{GS_1}(D) = \{8\}$;
The data that cannot be anonymized in the positive and negative region are $\{4, 2, 13, 12, 14\}$ and $\varnothing$, respectively.

(b) For the second level of granular, $U^2 = \{4, 2, 13, 12, 14\} \cup BND_{GS_1}(D) = \{4, 2, 13, 12, 14, 10, 16, 17, 15, 7, 9, 3, 11, 6, 5, 1\}$, $(\alpha^2, \beta^2) = (0.6, 0.5)$, $(HK^2, LK^2) = (5, 2)$, we can compute
$POS_{GS_2}(D) = \{4, 2, 13, 12, 14, 10, 16, 17, 15, 7\}$;
$BND_{GS_2}(D) = \{9, 3, 11\}$;
$NEG_{GS_2}(D) = \{6, 5, 1\}$;
Since $\{4, 2, 13, 12, 14, 7\}$ satisfies 5-anonymity and $\{6, 1\}$ satisfies 2-anonymity, the data that cannot be anonymized in the positive and negative region are $\{10, 16, 17, 15\}$ and $\{5\}$, respectively.

(c) For the last level of granular, $U^3 = BND_{GS_2}(D) \cup \{10, 16, 17, 15\} \cup \{5\} = \{10, 16, 17, 15, 9, 3, 11, 5\}$, $(\alpha^3, \beta^3) = (0.55, 0.54)$, $(HK^3, LK^3) = (4, 3)$, we can compute
$POS_{GS_3}(D) = \{10, 16, 17, 15, 9\}$;
$BND_{GS_3}(D) = \varnothing$;
$NEG_{GS_3}(D) = \{3, 11, 5\}$;
Since $\{10, 16, 17, 15, 9\}$ satisfies 4-anonymity and $\{3, 11, 5\}$ satisfies 3-anonymity, all data satisfy the anonymity requirements.
Output securely publishable dataset $Q = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11}, x_{12}, x_{13}, x_{14}, x_{15}, x_{16}, x_{17}\}$, the algorithm terminates.

Algorithm 1: Secure downscaling scheme (SD-MKS3WD).

```
input : An universal set of object, $U$, a dynamic threshold
        sequence, $(\alpha,\ \beta)^l = \{(\alpha^1, \beta^1), (\alpha^2, \beta^2), \cdots, (\alpha^l,\ \beta^l)\}$, the
        maximum attribute level, $m$, a dynamic k-value
        sequence, $(HK, LK)^l = \{(HK^1, LK^1), (HK^2, LK^2)$
        $, \cdots, (HK^l, LK^l)\}$.
output: Secure Database $Q$.
1  Import sensitive decision values and sort them from highest to
     lowest;
2  $t = 1, U^t = U, Q = \varnothing$ ;
3  for $t \leftarrow 1$ to $l$ do
4      if $U^t = \varnothing$ or $t > l$, then break;
5      end if;
6      Compute $POS_{GS_t}(D) = \{x \in U^t | f_x(D) \geq \alpha^t\}$,
         $BND_{GS_t}(D) = \{x \in U^t |\ \beta^t < f_x(D) < \alpha^t\}$ and
         $NEG_{GS_t}(D) = \{x \in U^t | f_x(D) \leq \beta^t\}$;
7      $c = 1$;
8      for $c \leftarrow 1$ to $m$ do
9          if $POS_{GS_t}(D) = \varnothing$ or $NEG_{GS_t}(D) = \varnothing$ then
10         break;
11         end if;
12         $UnsafeData = \varnothing$ ;
13         check whether $POS'_{GS_t}(D)$ satisfies $HK^t$-anonymity or
             $NEG'_{GS_t}(D)$ satisfies $LK^t$-anonymity;
14         if satisfies then
15         $Q+ = POS'_{GS_t}(D)$ or $Q+ = NEG'_{GS_t}(D)$ ;
16         else
17         $UnsafeData+ = POS'_{GS_t}(D)$ or
             $UnsafeData+ = NEG'_{GS_t}(D)$ ;
18         end if;
19         end check;
20         if $UnsafeData \neq \varnothing$ then
21         $c = c + 1, POS_{GS_t}(D) = UnsafeData$ or
             $NEG_{GS_t}(D) = UnsafeData$; turn to line 8;
22         else
23         break;
24         end if;
25     end
26     $U^{t+1} = BND_{G_t}(D)$+$UnsafeData$;
27     $t = t + 1$; turn to line 3;
28 end
29 Output $Q$;
```

However, since the degree of anonymity decreases level by level for positive regions $(HK^{t+1} < HK^t)$, SD-MKS3WD may cause some data with high anonymity requirements to reduce their requirements for anonymity, which affects the security of the data. To solve this problem, we treat the remaining data in the positive and negative regions differently and propose two new algorithms: sensitivity extraction scheme (SE-MKS3WD) and equivalence class extraction scheme (ECE-MKS3WD).

Specifically, we first use the three-way decisions to divide the dataset into three regions, and the positive and negative regions are divided into two separate parts with different anonymization approaches. Due to the highly sensitive data are concentrated in the positive region, we adopt a stricter anonymization approach for the positive region to prevent privacy leakage: we process the highly sensitive records first, and only after ensuring that all highly sensitive records are successfully anonymized, the next level of processing is performed $(HK^t > HK^{t+1})$. However, for the negative region, we process the low-sensitive data first, and the low-sensitive records that cannot be anonymized will enter the high-sensitive level to assist the high anonymization implementation $(LK^t < LK^{t+1})$. Fig. 4(b) depicts the two different strategies in the positive and negative regions. Obviously, compared to the traditional hierarchical anonymity, our model is more flexible and requires less costly cost.

In what follows, we will introduce these two algorithms (SE-MKS3WD and ECE-MKS3WD) in details.

4.2. Multi-level personalized k-anonymity privacy-preserving model based on sequential three-way decisions using sensitivity extraction (SE-MKS3WD)

In this subsection, we improve the approach for handling positive regions in SD-MKS3 W to achieve strict anonymity, and prioritize the anonymization of data with high anonymization requirements in the upper levels. For data that cannot be anonymized in the upper level, we extract the data with the closest sensitivity to the remaining data from the lower levels until all remaining data are successfully anonymized.

As shown in Algorithm 2, we sort the dataset in order of the sensitive decision values from highest to lowest in line 1. From line 2 to 30, we use the sequential three-way decisions for classification to divide the dataset into different granular structures and anonymize each granular structure individually. To do this, in line 6, we divide the positive, negative and boundary regions for every granularity level by the sensitive decision values. From line 7 to 22, we determine whether the data in the positive region satisfy HK-anonymity, and directly store the data which satisfies HK-anonymity into the secure database Q. During this process, we still adopt the k-anonymity approach from Algorithm 1. Finally, from line 23 to 27, for the remaining data, we extract the data with the closest sensitivity to them from the lower levels. Due to the sorting in the line 1, we only need to extract the first data from the next level. From line 28 to 30, we start the anonymization process for the next level only when all the remaining data in the upper level are successfully anonymized. In line 32, we export the final dataset that can be safely published. In summary, it is easy to observe that the time complexity of Algorithm 2 is $O(m \cdot l^2)$.

Example 4 illustrates the SE-MKS3WD using the same case in Example 2. Note that Algorithm 2 only describes how the sensitivity extraction scheme handles the data in the positive region, and the negative region is handled in the same way as in Algorithm 1 and will not be repeated.

Algorithm 2: Sensitivity extraction scheme (SE-MKS3WD).

```
input : An universal set of object, $U$; a dynamic threshold
        sequence, $(\alpha,\ \beta)^l = \{(\alpha^1,\beta^1),(\alpha^2,\beta^2),\cdots,(\alpha^l,\ \beta^l))$; the
        maximum attribute level, $m$; a dynamic k-value
        sequence, $(HK, LK)^l = \{(HK^1, LK^1), (HK^2, LK^2), \cdots,$
        $(HK^l, LK^l)\}$.
output: Secure Database $Q$.
1  Import sensitive decision values and sort them from highest to
    lowest;
2  $t = 1, U^t = U, Q = \varnothing$ ;
3  for $t \leftarrow 1$ to $l$ do
4      if $U^t = \varnothing$ or $t > l$, then break;
5      end if;
6      Compute $POS_{GS_t}(D) = \{x \in U^t | f_x(D) \geq \alpha^t\}$,
        $BND_{GS_t}(D) = \{x \in U^t |\ \beta^t < f_x(D) < \alpha^t\}$ and
        $NEG_{GS_t}(D) = \{x \in U^t | f_x(D) \leq \beta^t\}$;
7      $c = 1$;
8      for $c \leftarrow 1$ to $m$ do
9          if $POS_{GS_t}(D) = \varnothing$ then break;
10         end if;
11         $UnsafeData = \varnothing$ ;
12         check whether $POS'_{GS_t}(D)$ satisfies $HK^t$-anonymity;
13         if satisfies then
14         $Q += POS'_{GS_t}(D)$;
15         else
16         $UnsafeData += POS'_{GS_t}(D)$;
17         end if; end check;
18         if $UnsafeData \neq \varnothing$ then
19         $c = c + 1, POS_{GS_t}(D) = UnsafeData$; turn to line 8;
20         else break;
21         end if;
22     end
23     if $UnsafeData \neq \varnothing$ and $t + 1 \leq l$ then Compute
        $U^{t+1} = BND_{GS_t}(D)$,
        $POS_{GS_{t+1}}(D) = \{x \in U^{t+1} | f_x(D) \geq \alpha^{t+1}\}$;
24     if $POS_{G_{t+1}}(D) \neq \varnothing$ then
25     Extract the first record $POS_{GS^1_{t+1}}(D)$ from $POS_{GS_{t+1}}(D)$,
        $POS_{GS_t}(D) = UnsafeData + POS_{GS^1_{t+1}}(D)$,
        $BND_{GS_t}(D) = BND_{GS_t}(D) - POS_{GS^1_{t+1}}(D)$; turn to line 7;
26     else $t = t + 1$, turn to line 23;
27     end if;
28     else
29     $U^{t+1} = BND_{GS_t}(D)$, $t = t + 1$; turn to line 3;
30     end if;
31 end
32 Output $Q$;
```

Example 4 (*Continued with Example 2*)**.** Consider a partition of equivalence classes U/QID = {{4, 2, 13, 12, 14, 7}, {10, 16, 17, 15, 9}, {3, 11, 5}, {6, 1}, {8}}, a dynamic k-value sequence $(HK, LK)^3$ = {(6, 1), (5, 2), (4, 3)}, we can conclude the following

(a) For the first level of granular, $U^1 = U$, $(\alpha^1,\ \beta^1) = (0.8, 0.2)$, $(HK^1, LK^1) = (6, 1)$, we can compute
$POS_{GS_1}(D) = \{4, 2, 13, 12, 14\}$;
$BND_{GS_1}(D) = \{10, 16, 17, 15, 7, 9, 3, 11, 6, 5, 1\}$;
$NEG_{GS_1}(D) = \{8\}$;
The data that cannot be anonymized in the positive and negative region are $\{4, 2, 13, 12, 14\}$ and $\varnothing$, respectively. Thus, in order to handle the remaining data in the positive region, we can extract the data with the closest sensitivity from the next levels until anonymization is completed. According to Algorithm 2, we can compute $POS_{GS_2}(D) = \{10, 16, 17, 15, 7\}$. Since we sorted the data by sensitive attribute values in advance, we only need to extract the records in $POS_{GS_2}(D)$ one by one sequentially until all the remaining data have been successfully anonymized. $POS^{i}_{GS_1}(D)$ denotes the dataset after the ith extraction in the positive region. The following is the exact extraction process.
$POS^{1}_{GS_1}(D)$ = {4, 2, 13, 12, 14, 10};
$POS^{2}_{GS_1}(D)$ = {4, 2, 13, 12, 14, 10, 16};
$POS^{3}_{GS_1}(D)$ = {4, 2, 13, 12, 14, 10, 16, 17};
$POS^{4}_{GS_1}(D)$ = {4, 2, 13, 12, 14, 10, 16, 17, 15};
$POS^{5}_{GS_1}(D)$ = {4, 2, 13, 12, 14, 10, 16, 17, 15, 7}.
Obviously, after extracting the data five times, {4, 2, 13, 12, 14, 7} satisfies 6-anonymity and the data in the positive region that cannot be anonymized is $\{10, 16, 17, 15\}$. At this time, $POS_{GS_2}(D) = \varnothing$, we can continue to compute $POS_{GS_3}(D) = \{9\}$, make the next data extraction:
$POS^{6}_{GS_1}(D)$ = {10, 16, 17, 15, 9};
Since the data of next levels is empty ($POS_{GS_2}(D) = \varnothing$ and $POS_{GS_3}(D) = \varnothing$), the anonymity of the first level finishes.

(b) For the second level of granular, $U^2 = \{3, 11, 6, 5, 1\}$, $(\alpha^2,\ \beta^2)$ = (0.6, 0.5), $(HK^2, LK^2) = (5, 2)$, we can compute
$POS_{GS_2}(D) = \varnothing$;
$BND_{GS_2}(D)$ = {3, 11};
$NEG_{GS_2}(D)$ = {6, 5, 1}.
Since $\{6, 1\}$ satisfies 2-anonymity, the data which cannot be anonymized in the positive and negative region are $\varnothing$ and $\{5\}$, respectively.

(c) For the last level of granular, $U^3 = \{3, 11, 5\}$, $(\alpha^3,\ \beta^3) = (0.55, 0.54)$, $(HK^3, LK^3) = (4,\ 3)$, we can compute
$POS_{GS_3}(D) = \varnothing$;
$BND_{GS_3}(D) = \varnothing$;
$NEG_{GS_3}(D) = \{5,\ 3,\ 11\}$;
Since $\{5,\ 3,\ 11\}$ satisfies 3-anonymity, all data satisfy the anonymity requirements.
Output securely publishable dataset $Q = \{x_1,\ x_2,\ x_3,\ x_4,\ x_5,\ x_6,\ x_7,\ x_8,\ x_{11},\ x_{12},\ x_{13},\ x_{14}\}$, the algorithm terminates.

4.3. Multi-level personalized k-anonymity privacy-preserving model based on sequential three-way decisions using equivalence class extraction (ECE-MKS3WD)

By observing Example 3, it is not difficult to find that directly extracting the data with the closest sensitivity may result in invalid extracted data (For example, the dataset $\{x_{10}, x_{16}, x_{17}, x_{15}\}$ in Example 3). To address this problem, in this subsection, we construct another algorithm (ECE-MKS3WD) to process the remaining data in the upper levels by extracting the equivalence classes from the lower levels, as shown in Algorithm 3. Different from SE-MKS3WD, ECE-MKS3WD extracts only the equivalence classes of the remaining data from the lower levels to avoid extracting invalid data. Thus, Algorithm 3 has the same time complexity as Algorithm 2 as $\mathrm{O}(m \cdot l^2)$.

In order to compare the advantages of different schemes for the same dataset, Example 5 still uses the same case from Example 2 to describe the ECE-MKS3WD. Note that same as Algorithm 2, Algorithm 3 only describes how to handle data in the positive region, and the negative region is handled in the same way as Algorithm 1, and will not be repeated.

Example 5 (*Continued with Example 2*)**.** Consider a partition of equivalence classes $U/QID = \{\{4,\ 2,\ 13,\ 12,\ 14,\ 7\}, \{10,\ 16,\ 17,\ 15,\ 9\}, \{3,\ 11,\ 5\}, \{6,\ 1\}, \{8\}\}$, a dynamic k-value sequence $(HK, LK)^3 = \{(6,\ 1), (5,\ 2), (4,\ 3)\}$, and we can conclude the following

(a) For the first level of granular, $U^1 = U$, $(\alpha^1,\ \beta^1) = (0.8, 0.2)$, $(HK^1, LK^1) = (6, 1)$, we can compute
$POS_{GS_1}(D) = \{4, 2, 13, 12, 14\}$;
$BND_{GS_1}(D) = \{10, 16, 17, 15, 7, 9, 3, 11, 6, 5, 1\}$;
$NEG_{GS_1}(D) = \{8\}$;
The data that cannot be anonymized in the positive and negative region are $\{4, 2, 13, 12, 14\}$ and $\varnothing$, respectively. To avoid extracting invalid data, different from Example 4, we extract the equivalent class of the remaining data from the next levels directly, and we can compute $POS_{GS_2}(D) = \{10, 16, 17, 15, 7\}$, Since $\{4, 2, 13, 12, 14, 7\}$ is an equivalence group, we extract $\{7\}$ from $POS_{GS_2}(D)$, we can compute
$POS^1_{GS_1}(D) = \{4,\ 2,\ 13,\ 12,\ 14,\ 7\}$;
Since $POS^1_{GS_1}(D)$ satisfies 6-anonymity, the first level of anonymity is completed.

(b) For the second level of granular, $U^2 = \{10, 16, 17, 15, 9,\ 3, 11, 6, 5, 1\}$, $(\alpha^2,\ \beta^2) = (0.6,\ 0.5)$, $(HK^2, LK^2) = (5, 2)$, we can compute
$POS_{GS_2}(D) = \{10,\ 16,\ 17,\ 15\}$;
$BND_{GS_2}(D) = \{9,\ 3,\ 11\}$;
$NEG_{GS_2}(D) = \{6,\ 5,\ 1\}$;
Since $\{6,\ 1\}$ satisfies 2-anonymity, the data which cannot be anonymized in the positive and negative region are $\{10,\ 16,\ 17,\ 15\}$ and $\{5\}$, respectively. And we can compute $POS_{GS_3}(D) = \{9\}$, Since $\{10, 16, 17, 15, 9\}$ is an equivalence group, we extract $\{9\}$ from $POS_{GS_3}(D)$, we can compute
$POS^1_{GS_2}(D) = \{10,\ 16,\ 17,\ 15,\ 9\}$;
Since $POS^1_{GS_2}(D)$ satisfies 5-anonymity, the second level of anonymity is completed.

Algorithm 3: Equivalence class extraction scheme (ECE-MKS3WD).

```
input : An universal set of object, $U$; a dynamic threshold
        sequence, $(\alpha,\beta)^l = \{(\alpha^1,\ \beta^1), (\alpha^2,\ \beta^2), \cdots, (\alpha^l,\ \beta^l)$; the
        maximum attribute level, $m$; and a dynamic k-value
        sequence, $(HK, LK)^l = \{(HK^1, LK^1), (HK^2, LK^2), \cdots,$
        $(HK^l, LK^l)\}$.
output: Secure Database $Q$.
1  Import sensitive decision values and sort them from highest to
     lowest;
2  $t = 1, U^t = U, Q = \varnothing$ ;
3  for $t \leftarrow 1$ to $l$ do
4      if $U^t = \varnothing$ or $t > l$, then break;
5      end if;
6      Compute $POS_{GS_t}(D) = \{x \in U^t | f_x(D) \geq \alpha^t\}$,
         $BND_{GS_t}(D) = \{x \in U^t |\ \beta^t < f_x(D) < \alpha^t\}$ and
         $NEG_{GS_t}(D) = \{x \in U^t | f_x(D) \leq \beta^t\}$;
7      $c = 1$;
8      for $c \leftarrow 1$ to $m$ do
9          if $POS_{GS_t}(D) = \varnothing$ then break;
10         end if;
11         $UnsafeData = \varnothing$ ;
12         check whether $POS'_{GS_t}(D)$ satisfies $HK^t$-anonymity;
13         if satisfies then
14         $Q+ = POS'_{GS_t}(D)$;
15         else
16         $UnsafeData+ = POS'_{GS_t}(D)$;
17         end if;
18         end check;
19         if $UnsafeData \neq \varnothing$ then
20         $c = c + 1, POS_{GS_t}(D) = UnsafeData$; turn to line 8;
21         else break;
22         end if;
23     end
24     if $UnsafeData \neq \varnothing$ and $t + 1 \leq l$ then Compute
         $U^{t+1} = BND_{GS_t}(D)$,
         $POS_{GS_{t+1}}(D) = \{x \in U^{t+1} | f_x(D) \geq \alpha^{t+1}\}$;
25     if $POS_{G_{t+1}}(D) \neq \varnothing$ then
26     Extract an equivalence class $EC_{GS_{t+1}}$ of $UnsafeData$ from
         $POS_{GS_{t+1}}(D)$, $POS_{GS_t}(D) = UnsafeData + EC_{GS_{t+1}}$,
         $BND_{GS_t}(D) = BND_{GS_t}(D) - EC_{GS_{t+1}}$; turn to line 7;
27     else $t = t + 1$, turn to line 24;
28     end if;
29     else
30     $U^{t+1} = BND_{GS_t}(D)$, $t = t + 1$; turn to line 3;
31     end if;
32 end
33 Output $Q$;
```

(c) For the last level of granular, $U^3 = \{3, 11, 5\}$, $(\alpha^3,\ \beta^3) = (0.55, 0.54)$, $(HK^3, LK^3) = (4,\ 3)$, we can compute
$POS_{GS_3}(D) = \varnothing$;
$BND_{GS_3}(D) = \varnothing$;
$NEG_{GS_3}(D) = \{5,\ 3,\ 11\}$.
Since $NEG_{GS_3}(D)$ satisfies 3-anonymity, all data satisfy the anonymity requirements.
Output securely publishable dataset $Q = \{x_1,\ x_2,\ x_3,\ x_4,\ x_5,\ x_6,\ x_7,\ x_8,\ x_9,\ x_{10},\ x_{11},\ x_{12},\ x_{13},\ x_{14},\ x_{15},\ x_{16},\ x_{17}\}$, the algorithm terminates.

By observing Example 5, it is clear that the ECE-MKS3WD further improves the data availability compared to the SE-MKS3WD. Moreover, it is worth noting that in all three algorithms, for data that cannot be anonymized even after the algorithm ends, we judge them as unsuitable for publication and discard them directly.

4.4. Algorithm analysis and discussion

(1) Security Analysis

In this paper, we combine the k-anonymity model and sequential three-way decisions to propose a multi-level personalized k-anonymity privacy-preserving model based on sequential three-way decisions. Therefore, fundamentally, the privacy-preserving effect of the proposed methods is based on the k-anonymity. According to the previous analysis, the proposed models all satisfy the multi-level k-anonymity principle in Definition 10, and have high security. The specific security analysis is provided as follows.

Theorem 1. *In the proposed methods of this paper, for any* t $(t = 1, 2, \ldots, l)$*, where* l *represents the total level of division, the risk of privacy leakage in the positive probability region* $(POS_{GS_t}^{(\alpha^t,\beta^t)}(D))$ *is* $\frac{1}{HK^t}$*, and the risk of privacy leakage in the negative probability region* $(NEG_{GS_t}^{(\alpha^t,\beta^t)}(D))$ *is* $\frac{1}{LK^t}$.

Proof. In this paper, three practical algorithms (SD-MKS3WD, SE-MKS3WD and ECE-MKS3WD) are proposed to realize the proposed model. According to the previous analysis, these three algorithms adopt the same data anonymization principle, and all of them satisfy the multi-level k-anonymity principle.

Specifically, each of these algorithms first classifies data with different sensitivities into different levels of granularity structure by the sequential three-way decisions model for classification, and then makes the data in the different granularity structures satisfy the personalized k-anonymity, respectively. According to Definition 10, for any t $(t = 1, 2, \ldots, l)$, where l represents the total level of division, the proposed algorithms ensure that the probabilistic positive regions all satisfy HK^t-anonymity and the probabilistic negative regions all satisfy LK^t-anonymity. Meanwhile, according to Definition 3, it is not difficult to conclude that the privacy leakage risk of a dataset that satisfies k-anonymity is $\frac{1}{k}$. Thus, for our algorithms, the risk of privacy leakage in the positive probability region $(POS_{GS_t}^{(\alpha^t,\beta^t)}(D))$ is $\frac{1}{HK^t}$, and the risk of privacy leakage in the negative probability region $(NEG_{GS_t}^{(\alpha^t,\beta^t)}(D))$ is $\frac{1}{LK^t}$.

(2) Algorithm Discussion

According to the previous analysis, this paper introduces the sequential three-way decisions into the k-anonymity, and then proposes a multi-level personalized k-anonymity privacy protection model based on sequential three-way decisions. The proposed methods make the data anonymization more flexible and provide a new idea for the realization of multilevel privacy protection, but there are still some vulnerabilities and problems that need further exploration and research.

First of all, the proposed model is an extended model based on k-anonymity, therefore, the security of the proposed algorithms depends on the security of k-anonymity. In fact, k-anonymity has been proved to be vulnerable to many attacks at present, thus, the security of the proposed model needs to be improved. For this problem, we can learn about the latest data anonymization techniques and then explore whether they can be applied to the proposed algorithmic framework to improve the security, which is also a focus of our next work. In addition, "how to set the appropriate dynamic k-value sequence" is also a problem that needs to be solved. As known, personalized anonymity is an uncertainty problem, and the personalization sensitivity value directly affects the degree of anonymity that the data should attain. However, the personalization sensitive value generally depends on the personal preference of the data provider, which is ambiguous and unstable. Therefore, how to set the appropriate k-value sequence is also an uncertain problem with strong subjectivity. In order to solve this problem, we may try to combine various aspects such as the subjective needs of data providers and the environmental needs of anonymized data distribution, and propose a comprehensive metric function to provide a theoretical basis for the setting of dynamic k-value sequences. The specific solution is also what we need to focus on afterwards.

Table 3
Description of the datasets.

No.	Dataset	$\|U\|$	$\|QID\|$
1	Glass Identification	214	9
2	Statlog	690	14
3	Winequality-red	1599	11
4	Abalone	4177	8
5	Agaricus-lepiota	8124	16
6	MAGIC Gamma Telescope	19 020	10

5. Experimental results and analysis

In this section, some experiments are conducted to evaluate the utility of the proposed algorithms. We implement these experiments on a personal computer with Intel(R) Core (TM) i5-1135G7 @ 2.40 GHz 2.42 GHz; 16.0 GB (RAM) memory. The software is Eclipse IDE 2022-03.

5.1. Data sets

The datasets we used are all publicly available datasets from the UCI Machine Learning Repository, which are described in Table 3. Before starting our experiments, we pre-processed the selected datasets. We remove the records with missing values and employ Rosetta software (http://www.lcb.uu.se/tools/rosetta/) to transform the continuous data into the discrete values Since the original datasets do not have the personalized sensitive values, we randomly inserted a personalized sensitive value in the range [0, 1] for each record in order to facilitate the subsequent experiments. In addition, we constructed the attribute generalization trees for the quasi-identifier attributes (QID) based on the general social cognition, and then stratified the experimental data.

5.2. Cost metric

During the anonymization process, we inevitably lose some original information and generate some information loss, which directly affect the usability of the data. Therefore, information loss is an important metric to measure the performance of a privacy-preserving algorithm. In this paper, considering the k-anonymity approach used in the proposed algorithms, we define two evaluation functions to measure the information loss, namely, information suppression rate (ISR) and information generalization rate (IGR).

Definition 11 (*Information Suppression Rate, ISR*)**.** Given an original information table T, T' is a dataset that can be safely published after anonymization, the information suppression rate of T is defined as follows

$$ISR = \frac{|T| - |T'|}{|T|} \tag{16}$$

where $|T|$ and $|T'|$ denote the total number of records in the original information table T and the securely published dataset T', respectively.

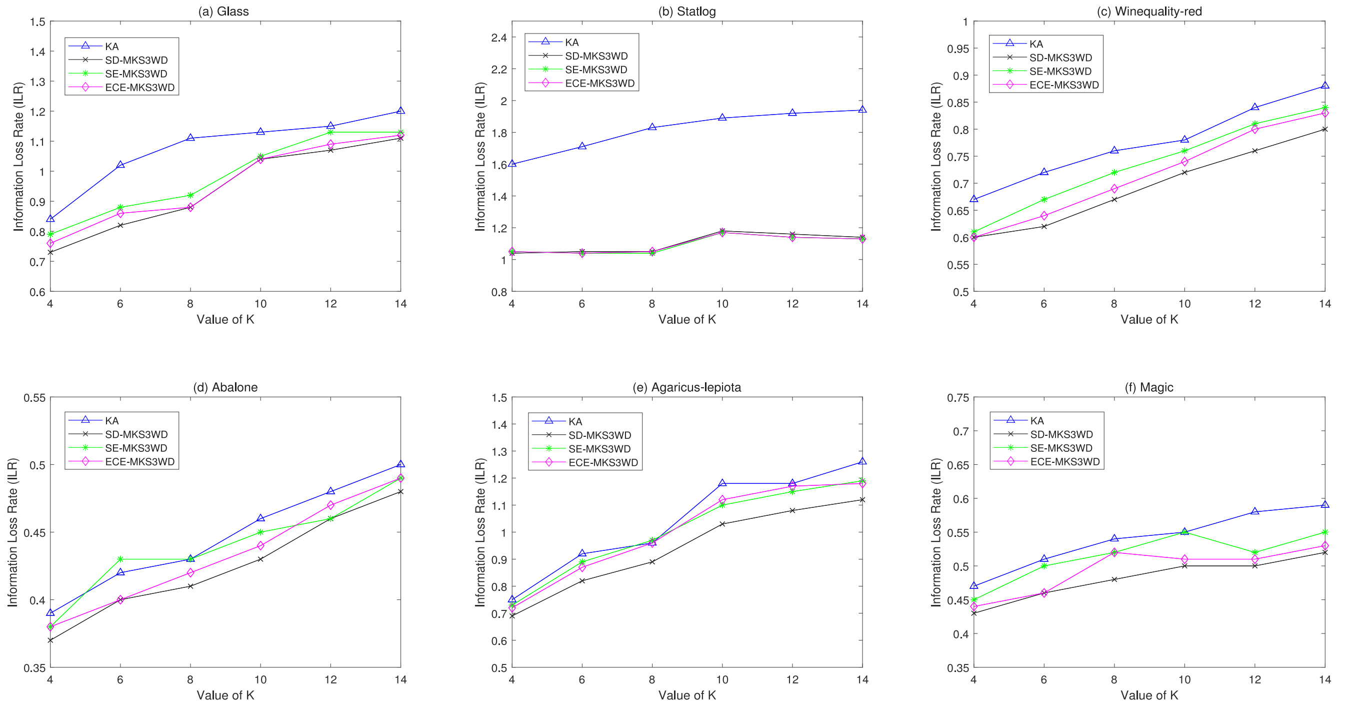


Fig. 5. Comparison of the information losses of KA, SD-MKS3WD, SE-MKS3WD, ECE-MKS3WD at different k-values.

Obviously, the lower the information suppression rate, the closer the final published data will be to the original information table, and the higher the data availability and information utility.

Definition 12 (*Information Generalization Rate, IGR*). Given a hierarchical decision table $T = \{U, QID, D = (d, f)\}$, all quasi-identifier attributes in T have m attribute levels. Suppose the anonymized table T' has $|S_i|$ rows at the ith attribute level ($i = 1, 2, \ldots, m$), then the information generalization rate from the anonymity of T is defined as follows

$$IGR = \frac{\sum_{i=1}^{m}(i \times |S_i|)}{m \times |T'|} \tag{17}$$

where $|T'|$ denotes the total number of records in the table T', m is the maximum number of attribute levels, and $|S_i|$ represents the number of attributes generalized to the ith level.

It can be seen that with the increase of the information generalization rate, the availability of data decreases, and when all the attributes are generalized to the highest level, the information generalization rate equal to 1. In order to combine the measures of information suppression rate (ISR) and information generalization rate (IGR) into a single framework, we define another measure called information loss rate (ILR) as follows.

Definition 13 (*Information Loss Rate, ILR*). The sum of ISR and IGR as the final information loss rate is defined as follows

$$ILR = ISR + IGR = \frac{|T| - |T'|}{|T|} + \frac{\sum_{i=1}^{m}(i \times |S_i|)}{m \times |T'|} \tag{18}$$

The measure of ILR considers both the suppression rate ISR and the generalization rate IGR, which has a more comprehensive evaluation capability. According to the above analysis, it is easy to observe that a smaller value of ILR indicates lower information loss and higher data availability.

Table 4
Setting of parameters.

No.	K	$(HK, LK)^l$
1	4	{(4, 1), (3, 2)}
2	6	{(6, 1), (5, 2), (4, 3)}
3	8	{(8, 1), (7, 2), (6, 3), (5, 4)}
4	10	{(10, 1), (9, 2), (8, 3), (7, 4), (6, 5)}
5	12	{(12, 1), (11, 2), (10, 3), (9, 4), (8, 5), (7, 6)}
6	14	{(14, 1), (13, 2), (12, 3), (11, 4), (10, 5), (9, 6), (8, 7)}

5.3. Comparison of the information losses at different k-values

In this subsection, we compare the information loss of the traditional k-anonymity (KA) and our proposed three algorithms (SD-MKS3WD, SE-MKS3WD, ECE-MKS3WD) on six datasets and analyze the effect of different k-values on information loss. To facilitate the experiment, we use the dynamic k-value sequence $(HK, LK)^l = \{(k, 1), (k - 1, 2), \ldots, (k - 1 + l, l)\}$ as the anonymous parameter sequence, where k denotes the anonymous parameters used in the traditional k-anonymity and l denotes the total granularity level of the sequential three-way decisions, as detailed in Table 4. Fig. 5 shows the experimental results on the six datasets.

From the differences between the broken lines in Fig. 5, we can conclude as follows:

- As the value of k increases, the information loss tends to increase. This is because the larger the value of k, the stricter the anonymity requirement, and the enhanced security will inevitably lead to greater information loss.
- For all datasets, our proposed three algorithms produce lower information loss than the traditional k-anonymity, which is sufficient to demonstrate the superiority of our algorithms.
- Among these three proposed algorithms, both SE-MKS3WD and ECE-MKS3WD generate higher information loss than SD-MKS3WD. This is because although extracting data can ensure absolute security, it inevitably raises the anonymity requirements to achieve the high level of

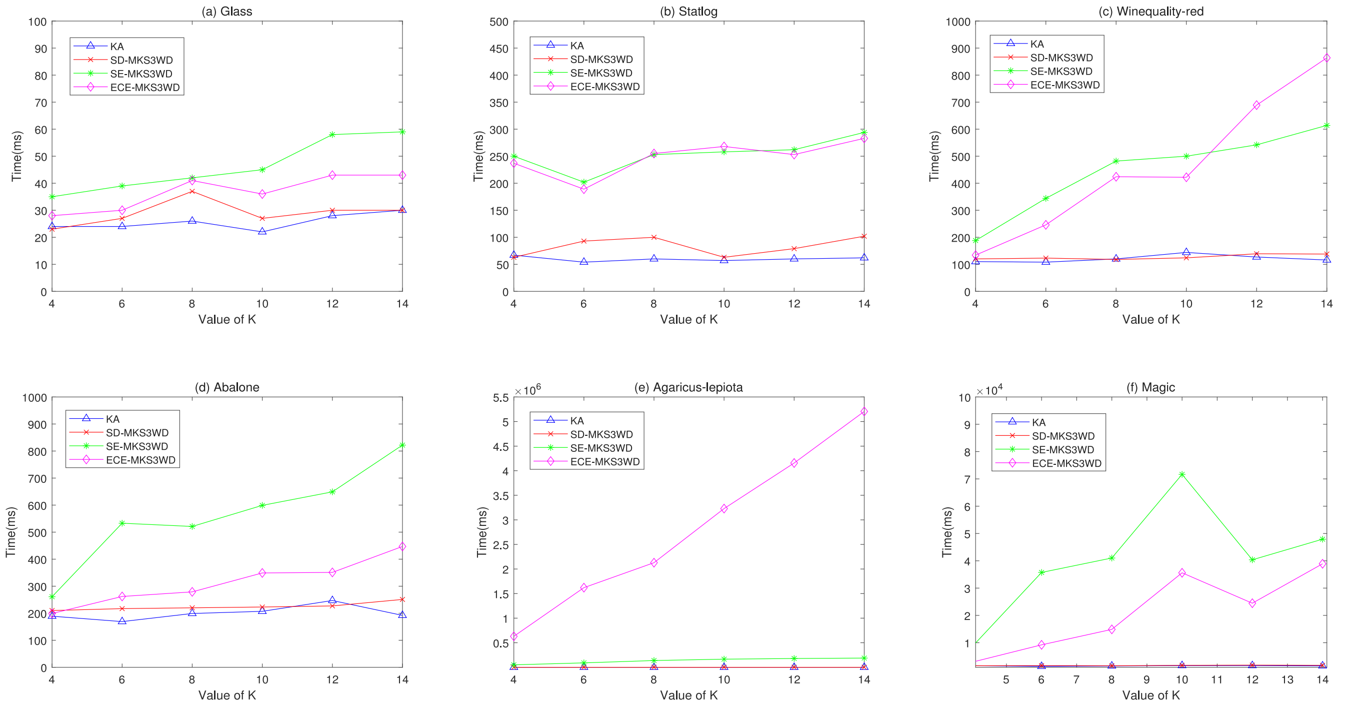


Fig. 6. Comparison of the runtimes of KA, SD-MKS3WD, SE-MKS3WD, ECE-MKS3WD at different k-values.

anonymity, leading to extra information loss, which is well illustrated in Example 4.

- From Fig. 5, we can easily observe that ECE-MKS3WD produces lower information loss than SE-MKS3WD for most datasets. This also demonstrates that the approach of extracting equivalence classes is superior to the approach of directly extracting the data with the closest sensitivity in most situations.

5.4. Comparison of the runtimes at different k-values

To further investigate the performance of our proposed algorithms, we also compare the running time of these four algorithms at different k-values. Fig. 6 shows the experimental results on the six datasets.

From Fig. 6, we can easily see that our algorithms generally take more time than the traditional k-anonymity, this is because we additionally consider the personalized anonymity. However, the experimental results show that among these three algorithms, the running time of SD-MKS3WD is less than those of SE-MKS3WD and ECE-MKS3WD, and is even close to the traditional k-anonymity (KA). This suggests that, in terms of the running time, SD-MKS3WD is the best choice for personalized k-anonymity when the security requirements are less stringent. For example, when the security level is divided into 10 levels, the user's security tolerance for "obesity" is 3–6. In other words, level 6 is the optimal security level, but level 3 is also tolerable for the user if level 6 is not guaranteed. In such case, the security degradation will not affect the user's anonymity request excessively, and the SD-MKS3WD is certainly suitable to realize personalized anonymity with high efficiency and low information loss.

Furthermore, the comparison of SE-MKS3WD and ECE-MKS3WD in running time indicates that ECE-MKS3WD outperforms SE-MKS3WD in most cases, which is because extracting equivalence classes can avoid extracting invalid data. However, for datasets with a larger number of quasi-identified attributes, the running time of ECE-MKS3WD is more than that of SE-MKS3WD because judging the equivalence classes may consume more extra time, such as "Agaricus-lepiota", which is shown in Fig. 6(e).

5.5. Comparison of the information losses under different granularity levels

Finally, we investigate the effects of different granularity levels on the information loss, and compare the variation of information loss generated by the proposed three algorithms under different granularity levels on six datasets. Fig. 7 depicts the experimental results, where the horizontal coordinates represent the three anonymization algorithms corresponding to the proposed models, the vertical coordinates indicate the information loss under different granularity levels, and the anonymization parameter k is equal to 8.

By observing the experimental results, it is easy to see that with the increase of the granularity level, the information loss generated by anonymity shows a trend of first decreasing and then increasing. This is because as the granularity level increases, the data division becomes more detailed and the anonymous diversity increases can avoid a lot of unnecessary information loss. Thus, we can also conclude that a reasonable increase in the number of sequential levels can reduce the information loss.

However, the re-increase of information loss also illustrates that dividing excessive granular structures will lead to extra information loss. This is because the excessive personalized anonymity can increase the difficulty of anonymity and reduce the availability of data, resulting in new information loss. Therefore, when we perform hierarchical anonymization, designing an appropriate level of anonymity is essential to reduce unnecessary information loss.

5.6. Comparison of the utility of the proposed algorithm to anonymize data at different k-values

To further validate the utility of the proposed algorithm to anonymize data, we use F1-Measure to evaluate the utility of KA, SD-MKS3WD, SE-MKS3WD and ECE-MKS3WD to anonymize data on six experimental datasets.

First, we define True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN), as follows:

- True Positive (TP): Data that needs to be anonymized is anonymized;

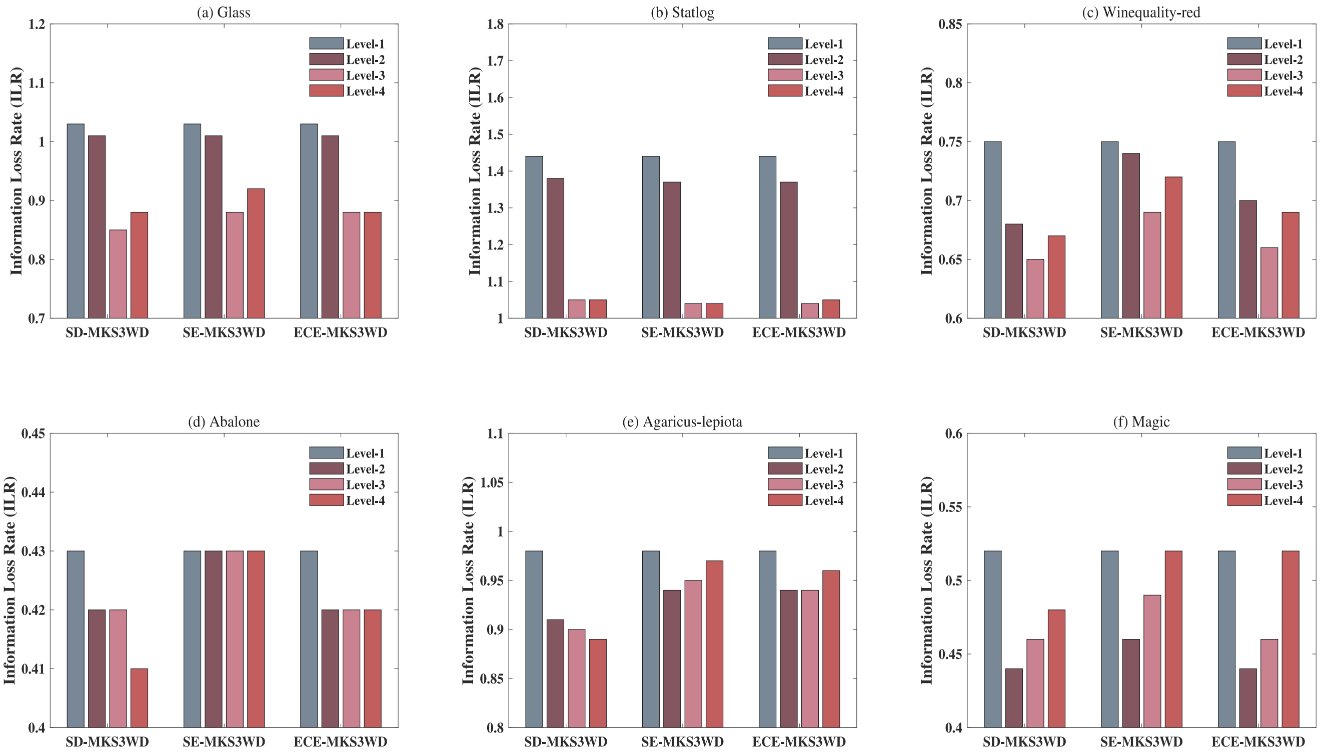


Fig. 7. Comparison of the information losses of SD-MKS3WD, SE-MKS3WD, ECE-MKS3WD under different granularity levels (k = 8).

- True Negative (TN): Data that does not need to be anonymized is not anonymized;
- False Positive (FP): Data that does not need to be anonymized is anonymized;
- False Negative (FN): Data that needs to be anonymized is not anonymized.

Then with the above definitions, Precision and Recall are defined as follows:

$$Precision = \frac{TP}{TP + FP} \tag{19}$$

$$Recall = \frac{TP}{TP + FN} \tag{20}$$

It is easy to observe that the Precision calculates the probability of actual positive samples among all samples predicted to be positive, and the Recall calculates the probability of actual positive samples among all samples that should be positive. Thus, in order to balance Precision and Recall, we use F1-Measure to evaluate the model performance, defined as follows:

$$F = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{21}$$

The value of F1-Measure ranges from 0 to 1, with 1 representing the best performance and 0 representing the worst performance. When F is higher, the model performance is better. The experimental results are shown in Fig. 8.

By Fig. 8, we can get the following conclusions:

- In the six experimental datasets, as the k-value increases, the values of F1-Measure shows a tendency to become smaller, and the utility of algorithms to anonymize data is reduced. This is because, as the k-value increases, it becomes more difficult to anonymize the data.
- From Fig. 8, one can notice that the F1-measures of proposed algorithms are lower than k-anonymity (KA) on some datasets. This is because the proposed algorithms may fail to anonymize a few data successfully in order to achieve accurate personalized anonymization, however, F1-Measure only considers the data utility from whether the data is successfully anonymized or not. In addition, for most datasets, the difference between k-anonymity and the proposed algorithms is not very large or even almost close, which also reflects that the proposed algorithm's data utility is still not far from k-anonymity despite considering personalized anonymity.
- Among all the datasets, SD-MKS3WD has the highest anonymization utility for the proposed algorithms, which is due to the fact that its anonymization requirements are not as strict as those of SE-MKS3WD and ECE-MKS3WD. In addition, ECE-MKS3WD has higher anonymization utility than SE-MKS3WD in some datasets, but some are opposite. This is because the efficiency of data extraction in both algorithms depends on the characteristics of the original dataset, which confirms our previous analysis.
- Finally, it is easy to find that the anonymization utility of the proposed algorithm is excellent on most of the datasets, which confirms that the proposed model has a good performance. While there also exist some that performs average on the dataset, which is caused by the characteristics of the dataset, and maybe these datasets are not suitable for privacy protection with data anonymization.

6. Conclusions

In this paper, we first combine sequential three-way decisions and k-anonymity to propose a multi-level personalized k-anonymity privacy-preserving model. Within this framework, we divide the datasets with different security requirements into different granular spaces by the sequential three-way decisions model for classification, and then use a dynamic k-value sequence to personalize the anonymity of each granularity space separately. Finally, the experimental results show that the proposed model is effective and available. It is worth mentioning that our model is also scalable and can theoretically be applied to most of the current state-of-the-art anonymization models with better anonymization results yet to be obtained, and we will further investigate those in the future.

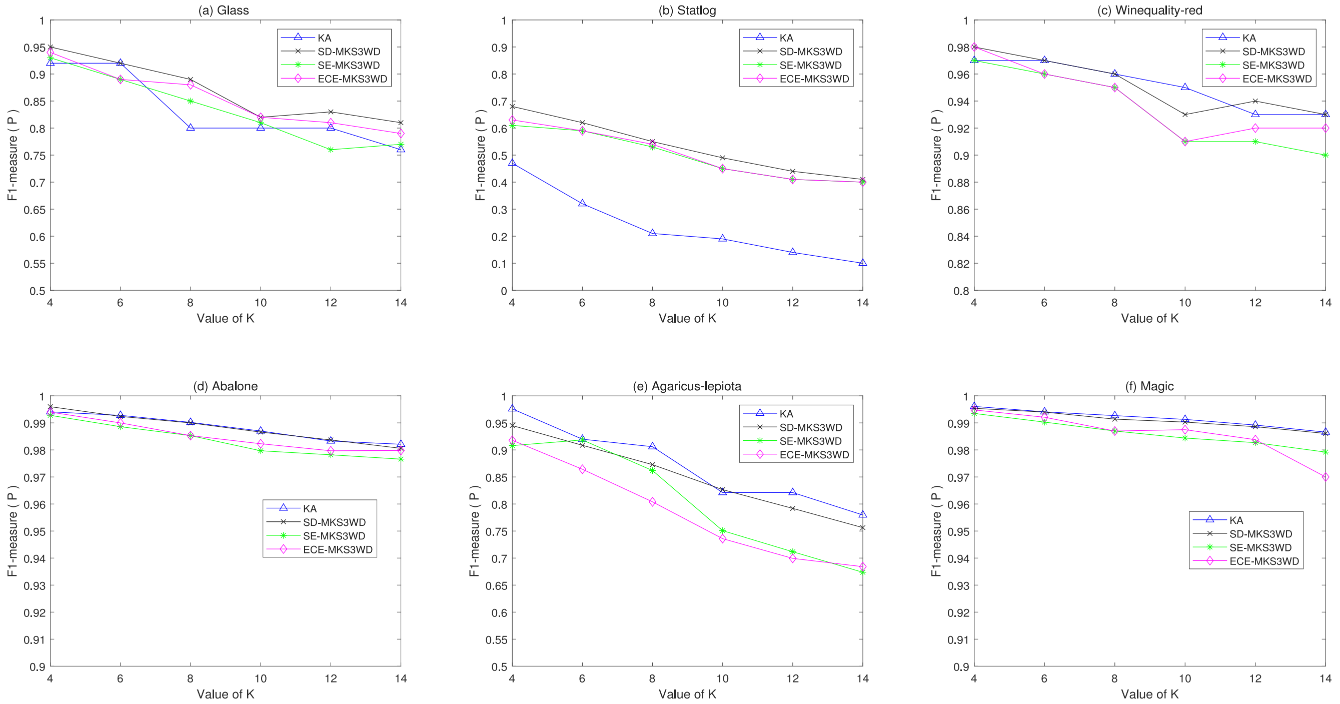


Fig. 8. Comparison of the utility of KA, SD-MKS3WD, SE-MKS3WD, ECE-MKS3WD to anonymize data at different k-values.

CRediT authorship contribution statement

Jin Qian: Conceptualization, Methodology, Writing – review & editing, Supervision. **Haoying Jiang:** Conceptualization, Methodology, Writing – original draft, Software, Validation. **Ying Yu:** Writing – review & editing, Software, Validation. **Hui Wang:** Writing – review & editing. **Duoqian Miao:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

The research is supported by the National Natural Science Foundation of China under Grant Nos. 62066014, 62163016, 61976158, Double thousand plan of Jiangxi Province of China, Jiangxi Province Natural Science Foundation of China under Grant Nos. 20232ACB202013, 20212ACB202001, 20224BAB212014, 20232ABC03A04.

References

Cao, N., Wang, C., Li, M., Ren, K., & Lou, W. (2013). Privacy-preserving multi-keyword ranked search over encrypted cloud data. *Ieee Transactions on Parallel and Distributed Systems*, *25*(1), 222–233.

Denham, B., Pears, R., & Naeem, M. A. (2020). Enhancing random projection with independent and cumulative additive noise for privacy-preserving data stream mining. *Expert Systems with Applications*, *152*, Article 113380.

Fang, Y., Gao, C., & Yao, Y. (2020). Granularity-driven sequential three-way decisions: A cost-sensitive approach to classification. *Information Sciences*, *507*, 644–664.

Gao, S., Ma, J., Sun, C., & Li, X. (2014). Balancing trajectory privacy and data utility using a personalized anonymization model. *Journal of Network and Computer Applications*, *38*, 125–134.

Gedik, B., & Liu, L. (2008). Protecting location privacy with personalized k-anonymity: Architecture and algorithms. *Ieee Transactions on Mobile Computing*, *7*(1), 1–18.

Gong, Q., Luo, J., Yang, M., Ni, W., & Li, X.-B. (2017). Anonymizing 1:M microdata with high utility. *Knowledge-Based Systems*, *115*, 15–26.

Guo, J., Yang, M., & Wan, B. (2021). A practical privacy-preserving publishing mechanism based on personalized k-anonymity and temporal differential privacy for wearable IoT applications. *Symmetry-Basel*, *13*(6), 1043.

He, D., Zeadally, S., Xu, B., & Huang, X. (2015). An efficient identity-based conditional privacy-preserving authentication scheme for vehicular Ad Hoc networks. *Ieee Transactions on Information Forensics and Security*, *10*(12), 2681–2691.

Hu, B. Q. (2014). Three-way decisions space and three-way decisions. *Information Sciences*, *281*, 21–52.

Kacha, L., Zitouni, A., & Djoudi, M. (2022). KAB: A new k-anonymity approach based on black hole algorithm. *Journal of King Saud University-Computer and Information Sciences*, *34*(7), 4075–4088.

Li, N., Li, T., & Venkatasubramanian, S. (2006). T-closeness: Privacy beyond k-anonymity and l-diversity. In *2007 IEEE 23rd international conference on data engineering* (pp. 106–115). IEEE.

Liang, D., Pedrycz, W., Liu, D., & Hu, P. (2015). Three-way decisions based on decision-theoretic rough sets under linguistic assessment with the aid of group decision making. *Applied Soft Computing*, *29*, 256–269.

Liang, Y., & Samavi, R. (2020). Optimization-based k-anonymity algorithms. *Computers & Security*, *93*, Article 101753.

Liang, D., Xu, Z., Liu, D., & Wu, Y. (2018). Method for three-way decisions using ideal TOPSIS solutions at pythagorean fuzzy information. *Information Sciences*, *435*, 282–295.

Lin, J.-L., & Wei, M.-C. (2009). Genetic algorithm-based clustering approach for k-anonymization. *Expert Systems with Applications*, *36*(6), 9784–9792.

Liu, X., Xie, Q., & Wang, L. (2016). Personalized extended (α, k)-anonymity model for privacy-preserving data publishing. *Concurrency and Computation-Practice & Experience*, *29*(6), 3886.

Machanavajjhala, A., Kifer, D., Gehrke, J., & Venkitasubramaniam, M. (2007). L-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *1*(1), 3.

Mehta, B. B., & Rao, U. P. (2022). Improved l-diversity: Scalable anonymization approach for privacy preserving big data publishing. *Journal of King Saud University-Computer and Information Sciences*, *34*(4), 1423–1430.

Mortazavi, R., & Erfani, S. (2020). GRAM: An efficient (k, l) graph anonymization method. *Expert Systems with Applications*, *153*, Article 113454.

Qian, J., Liu, C., Miao, D., & Yue, X. (2020). Sequential three-way decisions via multi-granularity. *Information Sciences*, *507*, 606–629.

Qian, J., Tang, D., Yu, Y., Yang, X., & Gao, S. (2022). Hierarchical sequential three-way decision model. *International Journal of Approximate Reasoning*, *140*, 156–172.

Qian, W., Zhou, Y., Qian, J., & Wang, Y. (2022). Cost-sensitive sequential three-way decision for information system with fuzzy decision. *International Journal of Approximate Reasoning*, *149*, 85–103.

Ren, X., & Jiang, D. (2022). A personalized (α, β, l, k)-anonymity model of social network for protecting privacy. *Wireless Communications & Mobile Computing*, *2022*, 11.

Song, F., Ma, T., Tian, Y., & Al-Rodhaan, M. (2019). A new method of privacy protection: Random k-anonymous. *Ieee Access*, *7*, 75434–75445.

Sun, X., Sun, L., & Wang, H. (2011). Extended k-anonymity models against sensitive attribute disclosure. *Computer Communications*, *34*(4), 526–535.

Sweeney, L. (2002a). Achieving k-anonymity privacy protection using generalization and suppression. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *10*(5), 571–588.

Sweeney, L. (2002b). K-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *10*(05), 557–570.

Truta, T. M., & Vinay, B. (2006). Privacy protection: p-sensitive k-anonymity property. In *22nd International conference on data engineering workshops* (p. 94). IEEE.

Wang, Z., Xu, Y., Yan, Y., Zhang, Y., Rao, Z., & Ouyang, X. (2022). Privacy-preserving indoor localization based on inner product encryption in a cloud environment. *Knowledge-Based Systems*, *239*, Article 108005.

Wong, R. C.-W., Li, J., Fu, A. W.-C., & Wang, K. (2006). (α, K)-anonymity: An enhanced k-anonymity model for privacy preserving data publishing. In *Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 754–759).

Xiao, X., & Tao, Y. (2006). Personalized privacy preservation. In *Proceedings of the 2006 ACM SIGMOD international conference on management of data* (pp. 229–240).

Xiong, J., Ma, R., Chen, L., Tian, Y., Li, Q., Liu, X., et al. (2019). A personalized privacy protection framework for mobile crowdsensing in IIoT. *Ieee Transactions on Industrial Informatics*, *16*(6), 4231–4241.

Xu, Y., Zheng, Z., Liu, X., Yao, A., & Li, X. (2022). Three-way decisions based service migration strategy in mobile edge computing. *Information Sciences*, *609*, 533–547.

Yao, Y. (2010). Three-way decisions with probabilistic rough sets. *Information Sciences*, *180*(3), 341–353.

Yao, Y. (2011). The superiority of three-way decisions in probabilistic rough set models. *Information Sciences*, *181*(6), 1080–1096.

Yao, Y. (2013). Granular computing and sequential three-way decisions. In *Rough sets and knowledge technology: 8th International conference, RSKT 2013, Halifax, NS, Canada, October 11-14, 2013, Proceedings 8* (pp. 16–27). Springer.

Yao, Y. (2018). Three-way decision and granular computing. *International Journal of Approximate Reasoning*, *103*, 107–123.

Yao, Y. (2020). Three-way granular computing, rough sets, and formal concept analysis. *International Journal of Approximate Reasoning*, *116*, 106–125.

Yao, Y., & Deng, X. (2011). Sequential three-way decisions with probabilistic rough sets. In *Proceedings of the 10th IEEE international conference on cognitive informatics and cognitive computing* (pp. 120–125). IEEE.

Yao, Y., & Wong, S. (1992). A decision theoretic framework for approximating concepts. *International Journal of Man-Machine Studies*, *37*(6), 793–809.

Yao, Y., Wong, S., & Lingras, P. (1990). A decision-theoretic rough set model. *Methodologies for Intelligent Systems*, *5*, 17–24.

Ye, M., Wu, X., Hu, X., & Hu, D. (2013). Anonymizing classification data using rough set theory. *Knowledge-Based Systems*, *43*, 82–94.

Yu, H., Zhang, C., & Wang, G. (2016). A tree-based incremental overlapping clustering method using the three-way decision theory. *Knowledge-Based Systems*, *91*, 189–203.

Zhan, J., Ye, J., Ding, W., & Liu, P. (2022). A novel three-way decision model based on utility theory in incomplete fuzzy decision systems. *Ieee Transactions on Fuzzy Systems*, *30*(7), 2210–2226.

Zhang, Q., Pang, G., & Wang, G. (2020). A novel sequential three-way decisions model based on penalty function. *Knowledge-Based Systems*, *192*, Article 105350.