Information Sciences 686 (2025) 121316

Contents lists available at ScienceDirect

Information Sciences

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

A dynamic anonymization privacy-preserving model based on hierarchical sequential three-way decisions

Jin Qian [a,c,*], Mingchen Zheng [a], Ying Yu [a], Chuanpeng Zhou [a], Duoqian Miao [b]

[a] School of Information and Software Engineering, 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 Technolgy, Zhenjiang, 212003, Jiangsu, China

ARTICLE INFO

Keywords:
K-anonymity
Differential privacy
Data anonymity
Privacy preservation
Hierarchical sequential three-way decisions

ABSTRACT

Data anonymization is one of the common techniques for ensuring data security and privacy. However, the existing anonymization techniques often suffer lower execution efficiency and unnecessary information loss when dealing with complex data. Therefore, we propose a dynamic anonymity privacy-preserving model based on hierarchical sequential three-way decisions. Specifically, we first divide the data into multiple granularity spaces by attributes and dynamically process the data in the granularity spaces. Then, in a single granularity space, we construct a generalization hierarchy for the data based on the attributes generalization trees and divide it into the positive, negative and boundary regions based on anonymous parameter. Next, we can acquire the positive and boundary regions by generalization and dynamically update the processed data at the next granularity. After that, we suppress the data in the final negative and boundary regions while releasing the positive region. To further improve data availability, we combine the idea of differential privacy by adding noise data to the final boundary region enabling its release and propose an enhanced anonymity model. Finally, we compare our proposed algorithms with other methods on six datasets. Experimental results show that our method effectively reduces processing costs, improves data usability and protects data privacy.

1. Introduction

With the advent of the digital era, a large amount of personal data is being collected, stored, and shared. Leveraging data sharing and analysis has the potential to drive innovation and problem-solving across various fields, including scientific research, healthcare, and finance [1]. Yet, personal data contains inherently sensitive information, and sharing the original data directly without any processing may lead to the leakage of the data owner's private information. Therefore, how to protect the privacy of data owners in data sharing and dissemination has attracted a lot of attention from scholars worldwide [2], and how to achieve a balance between data availability and privacy has gradually become a prominent research topic in the field of privacy protection [3,4].

K-anonymity [5] has been widely used in many fields as an effective method of data anonymization against linking attacks [6]. The model requires that each quasi-identifier value appears at least k times ($k > 1$) when publishing an anonymized dataset, in other words, at least k records must have the same value of the quasi-identifier attribute. Currently, many anonymity models [7–9] are

* Corresponding author at: School of Information and Software Engineering, East China Jiaotong University, Nanchang, 330013, Jiangxi, China.
E-mail addresses: qjqjlqyf@163.com (J. Qian), zhengmc0797@163.com (M. Zheng), yuyingjx@163.com (Y. Yu), chuanpeng0622@163.com (C. Zhou), dqmiao@tongji.edu.cn (D. Miao).

https://doi.org/10.1016/j.ins.2024.121316
Received 20 May 2024; Received in revised form 5 August 2024; Accepted 6 August 2024

Available online 10 August 2024

developed based on k-anonymity. Mehta and Rao [10] presented an enhanced scalable l-diversity method based on the l-diversity model. In the area of data anonymization, researchers strive to strike a balance between enhancing data availability and minimizing information loss [11–13]. Shaham et al. [14] advanced propose a robust framework for the anonymization of spatiotemporal trajectory datasets that enhances privacy and significantly improves dataset utility. Kacha et al. [15] proposed a novel algorithm based on a simple natural-inspired metaheuristic, which effectively solves the limitations of the clustering-based k-anonymity method and improves the data utility. Sopaoglu and Abul [16] developed a k-anonymization method for data streams that additionally protects the sensitivity and enhances classification accuracy. Lan et al. [17] introduced a distributed data processing framework based on the expansion of privacy protection to improve the accuracy and privacy of the algorithm's outputs. Kiran and Shirisha [18] suggested a k-anonymity-based framework for anonymizing categorical data that enhances privacy protection while maintaining data mining model accuracy, effectively addressing the limitations of traditional perturbation methods. Although data anonymization algorithms have been widely researched for data privacy protection, they still need to be improved. For instance, when anonymizing massive data, there exist some security problems [19,20], such as lower efficiency and higher information distortion. On the other hand, existing data anonymization techniques generally adopt the method of unified processing of all data, leading to high processing costs when dealing with a large amount of complex data [21].

As we all know, granular computing [22–24] and three-way decisions (3WD) [25–27] are effective tools to deal with uncertain information following human cognition, which can discover potential knowledge with minimum cost. Granular computing emphasizes understanding and describing the real world from multiple views and levels through the granulation of complex data. By using information granules as the basic units of computation, this approach addresses problems at various coarse and fine granularity levels, selects the most relevant granularity space for decision-making tasks, and thereby reduces complexity while improving accuracy. 3WD mainly divides the universe into three relatively independent regions, and formulates the corresponding processing strategy for each region, making the three types of decisions of accepting, rejecting and delaying respectively. At present, 3WD is widely applied in various fields [28–30]. In recent years, some scholars have proposed many extensions of the 3WD to enhance the classification performance. Yang et al. [31] presented a unified model of sequential three-way decisions and multilevel incremental processing for complex problem solving, which enables efficient decision-making with reduced overall cost. Yao [32] proposed the sequential three-way decisions model. Zhang et al. [33] suggested a new sequential three-way decisions model based on a penalty function. Yang et al. [34] introduced a temporal-spatial composite sequential approach of three-way granular computing. Qian et al. [35] presented a cost-sensitive sequential three-way decision model for the information system with fuzzy decision and achieved better classification performance with lower cost. In order to obtain more generalized decision rules, Qian et al. [36] developed hierarchical sequential three-way decisions by combining sequential three-way decisions and hierarchical rough set model [37] which improved the classification accuracy. Several researchers have combined granular computing and 3WD with privacy preservation to improve the utility of anonymization algorithms. Ye et al. [38] combined data anonymization and rough set theory to propose a top-down optimization algorithm based on hierarchical conditional entropy. Wang et al. [39] utilized fuzzy sets to divide levels for different types of attributes and presented the privacy model for hierarchical data with multi-level sensitivity. Ali et al. [40] designed a privacy enhancement model for IoT based on three-way decisions and differential privacy.

As illustrated before, hierarchical sequential three-way decisions (HS3WD) as an extension model of 3WD has inherent advantages in dealing with complex data. At the same time, the generalization process of attributes can be approximated as the process of upgrading the concept hierarchy tree in the HS3WD model. Qian et al. [41] used a dynamic k-value sequence to anonymize data of different granularities, and proposed a multilevel k-anonymity model based on sequential three-way decisions, which reduces information loss. Therefore, it is necessary and valuable to study the data anonymization model based on HS3WD further to improve the utility of anonymization algorithms.

Based on the above motivations, in this paper, we attempt to combine HS3WD with data anonymization techniques and propose a dynamic k-anonymity model based on HS3WD. Firstly, we construct a hierarchical granularity space from coarse to fine based on the quasi-identifier attribute set, and process the data with sequential three-way decisions in each granularity space with corresponding anonymization strategies. This allows the processing to be dynamically reduced in terms of data, enabling a certain reduction in the processing cost. Then, we combine the idea of differential privacy to propose an enhanced anonymity model, which further improves the data availability. In summary, the contributions of our work are as follows:

(1) We combine HS3WD and k-anonymity to propose a new way of anonymizing data. Different from the way of existing data anonymization techniques deal with the overall data, our model divides the data into multiple granularity spaces based on attribute sets and dynamically processes the data by incrementally adding attributes, resulting in a lower generalized processing cost.
(2) We propose a novel k-anonymity model and its enhancement model based on hierarchical sequential three-way decisions. Different from general binary anonymization, we divide the data into three disjoint parts: the positive region, negative and boundary regions. By incorporating the idea of differential privacy, we add noise data to the boundary region, which enhances data usability while reducing information loss and ensuring security.

The rest of the paper is organized as follows: In Section 2, we briefly introduce the k-anonymity and hierarchical sequential three-way decisions models. Section 3 proposes two specific algorithms under the dynamic anonymity model based on hierarchical sequential three-way decisions, and gives the corresponding examples for illustration. Section 4 gives the related experiments and conclusions. Section 5 summarizes the work of this paper and looks forward to future research directions.

2. Preliminaries

In this section, we will review some basic concepts of k-anonymity and hierarchical sequential three-way decisions. For a detailed description, please refer to Refs. [6,36,42].

2.1. K-anonymity model

In order to defend against linking attacks in data distribution, Sweeney [5] first proposed k-anonymity privacy-preserving model.

Definition 1. (Quasi-Identifier Attributes) Consider a population of entities $U(A_1, A_2, ..., A_n)$ and an external table U_E, for all records $R_i \in U$, if the value combination $R_i(A_j, ..., A_m)$ that contains no identifiers can be uniquely located in U_E, we call the set of attributes $\{A_j, ..., A_m\}$ a quasi-identifier attributes.

Definition 2. (Sensitive-Attributes) Sensitive attributes contain personal sensitive information.

Definition 3. (K-anonymity) Given a population of entities $U(A_1, A_2, ..., A_n)$ and QI be the quasi-identifier associated with it, the anonymized table $U^*(A_1, A_2, ..., A_n)$ is considered to achieve k-anonymity, in which every record is indistinguishable from at least $k-1$ other records.

Definition 4. (Equivalence Class) In a data table $U^*(A_1, A_2, ..., A_n)$ that adheres to k-anonymity, a set of records U^* sharing the same quasi-identifier attribute values is referred to as an equivalence class.

2.2. Hierarchical sequential three-way decisions model (HS3WD)

The model HS3WD [38] is a multi-step decision-making method. The preliminary decision is made based on an attribute set with fewer attributes, and the decision-making for the remaining uncertain objects with more attributes is deferred to the next phase. Therefore, the sequential strategy keeps boosting the accuracy of classification results. In this subsection, we briefly review a general model of hierarchical sequential three-way decisions.

Definition 5. Given a multi-level decision table $S^l = \{U^l, A^l\}$, threshold parameters $(\alpha_t, \beta_t) = \{(\alpha_t^1, \beta_t^1), (\alpha_t^2, \beta_t^2), ..., (\alpha_t^s, \beta_t^s)\}$, an attribute sets A_t, and EC_t is an equivalence relation generated by A_t, then the granular structure is constructed based on A_t, the t-th level of granular structure G_t^l at the l-th conceptual level are defined as:

$$\begin{aligned} G_t &= \{S_t, A_t, EC_t, \alpha_t, \beta_t\}(t = 1, 2, ...n) \\ G_t^l &= \{S_t^l, A_t^l, EC_t^l, \alpha_t^l, \beta_t^l\}(l = 1, 2, ...s) \end{aligned} \tag{1}$$

where G_t represents the t-th level of granular structure, S_t denotes the multi-level decision under the granular structure G_t, and S_t^l represents the multi-level decision table for G_t at the l-th conceptual level.

Definition 6. Given a multi-level decision table S^l, a granular structure G_t^l, an equivalence relation EC, an equivalence class $[x]_A$, and a sequence of threshold parameters vectors $(\alpha, \beta) = \{(\alpha^1, \beta^1), (\alpha^2, \beta^2), ..., (\alpha^s, \beta^s)\}$, then the (α^l, β^l)-lower approximation $\underline{apr}_A^l(D_i)$ and the (α^l, β^l)-upper approximation $\overline{apr}_A^l(D_i)$ are defined as:

$$\begin{aligned} \underline{apr}_A^l(D_i) &= \{x | p(D_i|[x]_A) \geq \alpha^l, x \in U_t^l\} \\ \overline{apr}_A^l(D_i) &= \{x | p(D_i|[x]_A) > \beta^l, x \in U_t^l\} \end{aligned} \tag{2}$$

where $U_t^s = U_t$, $U_t^l = \bigcup_{1 \leq i \leq s} \{\overline{apr}_A^{l+1}(D_i) - \underline{apr}_A^{l+1}(D_i)\}$, D_i denotes the equivalence class with including x in the partition $U_t^{l+1}/D = (D_1, D_2, ..., D_s)$.

According to the pair of $< \underline{apr}_A^l(D_i), \overline{apr}_A^l(D_i) >$, for the t-th level granular structure $G = \{G_1, G_2, ..., G_n\}$, we can obtain the three (α_t, β_t)-probabilistic regions as follows:

$$\begin{aligned} &POS_{G_t}(D_i) = \bigcup_{l=1}^{s} \underline{apr}_A^l(D_i); \\ &NEG_{G_t}(D_i) = \bigcup_{l=1}^{s} \left(U_t^l - \overline{apr}_A^l(D_i)\right); \\ &BND_{G_t}(D_i) = U_t - POS_{G_t}(D_i) - NEG_{G_t}(D_i). \end{aligned} \tag{3}$$

Based on the t-level granular structure G_t, the boundary region $BND_{G_t}(D_i)$ can be considered as the universe of the $(t+1)$-level granular structure G_{t+1}.

Table 1
An original information table.

ID		QID			SA
No.	Name	Sex	Age	Zipcode	Disease
1	Alex	F	23	35715	Pneumonia
2	Lily	F	17	35715	Asthma
3	Ethan	M	25	35710	Pneumonia
4	Mia	F	50	35703	Flu
5	Oliver	M	42	35706	HIV
6	Noah	F	46	35706	Cancer
7	Ava	F	68	35724	Flu
8	Emma	M	33	35723	Hepatitis

3. A dynamic anonymity privacy-preserving model based on hierarchical sequential three-way decisions

In this section, we first introduce the multi-hierarchical decision table for k-anonymity, and then, combine the k-anonymity model with the hierarchical sequential three-way decisions model to propose a novel k-anonymity model based on hierarchical sequential three-way decisions (KHS3WD). To further improve the data availability, we incorporate the concept of differential privacy into the KHS3WD model: adding the appropriate amount of noisy data to the portion of the data that delays decision-making. We thereby propose a k-anonymity model based on hierarchical sequential three-way decisions with the introduction of a noise mechanism (KNHS3WD).

3.1. Multi-hierarchical decision table for k-anonymity

In realistic decision-making, the generalization of attributes can naturally form a hierarchy of generalization, which can be represented by an attribute generalization tree. Through the attribute generalization tree, from the perspective of the hierarchical rough set, the generalization process can be described as follows. Low-level (finer) concepts can be replaced with high-level (coarser) concepts, forming a granularity structure from coarse to fine. In order to facilitate the construction of a dynamic anonymization framework based on the HS3WD, in this subsection, we introduce attribute generalization trees as well as multi-hierarchical decision tables for anonymization.

Definition 7. Given an information table $T = \{U, A\}$, $A = \{a_1, a_2, ..., a_n\}$ is a finite nonempty set of attributes and $QID = \{\{a_1^1, a_1^2, ..., a_1^s\}, \{a_2^1, a_2^2, ..., a_2^s\}, ..., \{a_n^1, a_n^2, ..., a_n^s\}\}$, then the attribute generalization tree for attribute a_i can be defined as follows:

$$GA_i = \{a_i^1, a_i^2, \dots, a_i^s\} \tag{4}$$

where a_i^l denotes the values of attribute a_i generalized to the l-th level ($l = 1, 2, ..., s$).

Definition 8. Given an information table $T = \{U, A\}$, and its corresponding the attribute generalization tree, then a multi-hierarchical decision table for k-anonymity under the l-th level of all the attribute generalization tree can be defined as follows:

$$LDT^l = \{U^l, QID^l, D^l\}(l = 1, 2, \dots, s) \tag{5}$$

where s represents the maximum height of the generalization tree, U^l is a non-empty finite set of objects, $QID^l = \{\{a_1^1, a_1^2, ..., a_1^s\}, \{a_2^1, a_2^2, ..., a_2^s\}, ..., \{a_n^1, a_n^2, ..., a_n^s\}\}$ is a set of quasi-identifier attributes and D^l represents the sensitive attribute.

By Definition 8, we can generate s multi-hierarchical decision tables for k-anonymity: $LDT = \{LDT^1, LDT^2, ..., LDT^s\}$ and denote LDT^1 as the original table. Additionally, it should be noted that, in the actual processing, different attributes may be generalized to different levels, resulting in the generation of attributes generalization trees with inconsistent heights. Taking attribute a_i^l as an example, when $l < s$ indicating that the height of the attributes generalization tree of the attribute a_i^l is less than the maximum height of the generalization tree QID, we employ data complementation. This involves supplementing the data in the hierarchical decision table by duplicating the value of the maximum level of generalization for the attributes a_i^l.

Example 1. As shown in Table 1, "NO." and "Name" are identification attributes (ID), while "Sex", "Age" and "Zipcode" act as quasi-identifier attributes (QID); SA represents the sensitive attribute; "Disease" denotes the values of the sensitive attributes. According to Definition 7, we can construct the attribute generalization tree for each QID, as illustrated in Fig. 1. Subsequently, utilizing the generalization trees of attributes, we can derive the multi-hierarchical decision table for k-anonymity. By observing Fig. 1, it's apparent that the generalization tree of attribute "Sex" has two levels, which is less than the level of the QID attribute's generalization tree. Therefore, we adopt data complementation: repeating the value of the highest level of the attribute generalization tree of "Sex", which maintains the consistency of the data as much as possible, so as to obtain a complete multi-hierarchical decision table, as shown in Table 2. □

Table 2
Multi-hierarchical decision table of Table 1.

U	QID									D
	Sex			Age			Zipcode			SA
	a_1^1	a_1^2	a_1^3	a_2^1	a_2^2	a_2^3	a_3^1	a_3^2	a_3^3	disease
u_1	F	*	*	23	[1,30]	[1,70]	35715	3571*	357**	Pneumonia
u_2	F	*	*	17	[1,30]	[1,70]	35715	3571*	357**	Asthma
u_3	M	*	*	25	[1,30]	[1,70]	35710	3571*	357**	Pneumonia
u_4	F	*	*	50	[50,70)	[1,70]	35703	3570*	357**	Flu
u_5	M	*	*	42	(30,50)	[1,70]	35706	3570*	357**	HIV
u_6	F	*	*	46	(30,50)	[1,70]	35706	3570*	357**	Cancer
u_7	F	*	*	68	[50,70)	[1,70]	35724	3572*	357**	Flu
u_8	M	*	*	33	(30,50)	[1,70]	35723	3572*	357**	Hepatitis

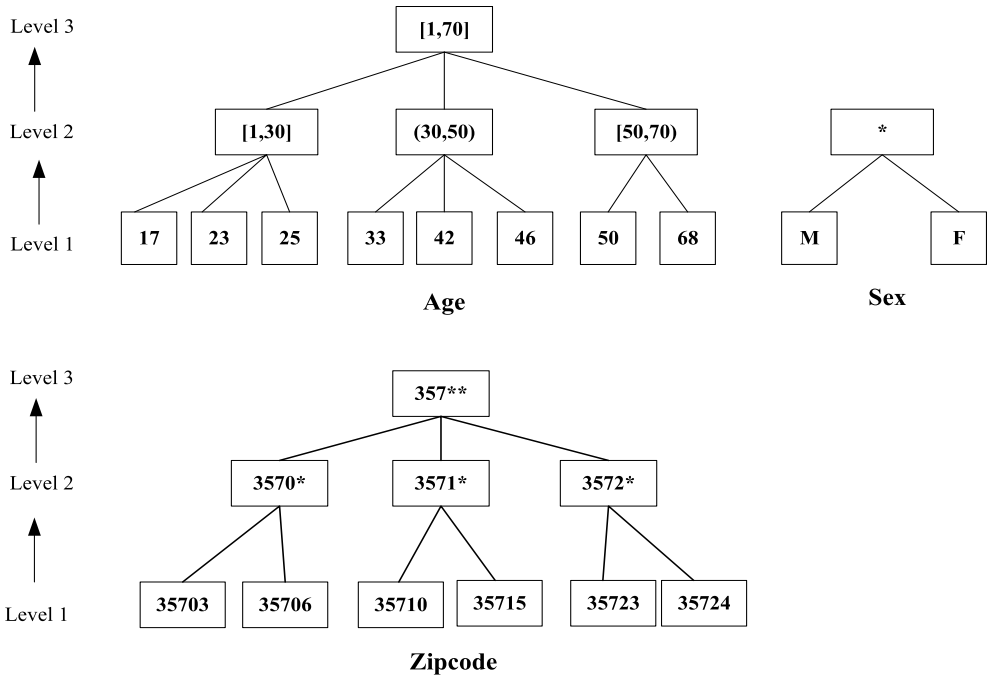

Fig. 1. Attribute generalization tree for each QID.

Table 3
A multi-hierarchical decision table LDT included LDT^1, LDT^2 and LDT^3 for k-anonymity.

LDT^1				LDT^2				LDT^3				
U	a_1^1	a_2^1	a_3^1	U	a_1^2	a_2^2	a_3^2	U	a_1^3	a_2^3	a_3^3	d
u_1	F	23	35715	u_1	*	[1,30]	3571*	u_1	*	[1,70]	357**	PN
u_2	F	17	35715	u_2	*	[1,30]	3571*	u_2	*	[1,70]	357**	AS
u_3	M	25	35710	u_3	*	[1,30]	3571*	u_3	*	[1,70]	357**	PN
u_4	F	50	35703	u_4	*	[50,70)	3570*	u_4	*	[1,70]	357**	Flu
u_5	M	42	35706	u_5	*	(30,50)	3570*	u_5	*	[1,70]	357**	HIV
u_6	F	46	35706	u_6	*	(30,50)	3570*	u_6	*	[1,70]	357**	Cancer
u_7	F	68	35724	u_7	*	[50,70)	3572*	u_7	*	[1,70]	357**	Flu
u_8	M	33	35723	u_8	*	(30,50)	3572*	u_8	*	[1,70]	357**	HEP

From the above definition, we know that multi-hierarchical decision is composed of several hierarchical decision tables at a single level so we can also obtain a multi-hierarchical decision table LDT included LDT^1, LDT^2 and LDT^3, as shown in Table 3. Thus, we can unify the hierarchical treatment of numerical and categorical data according to the constructed multi-hierarchical decision table, which provides the basis for the subsequent construction of the multi-hierarchical sequential dynamic anonymization framework.

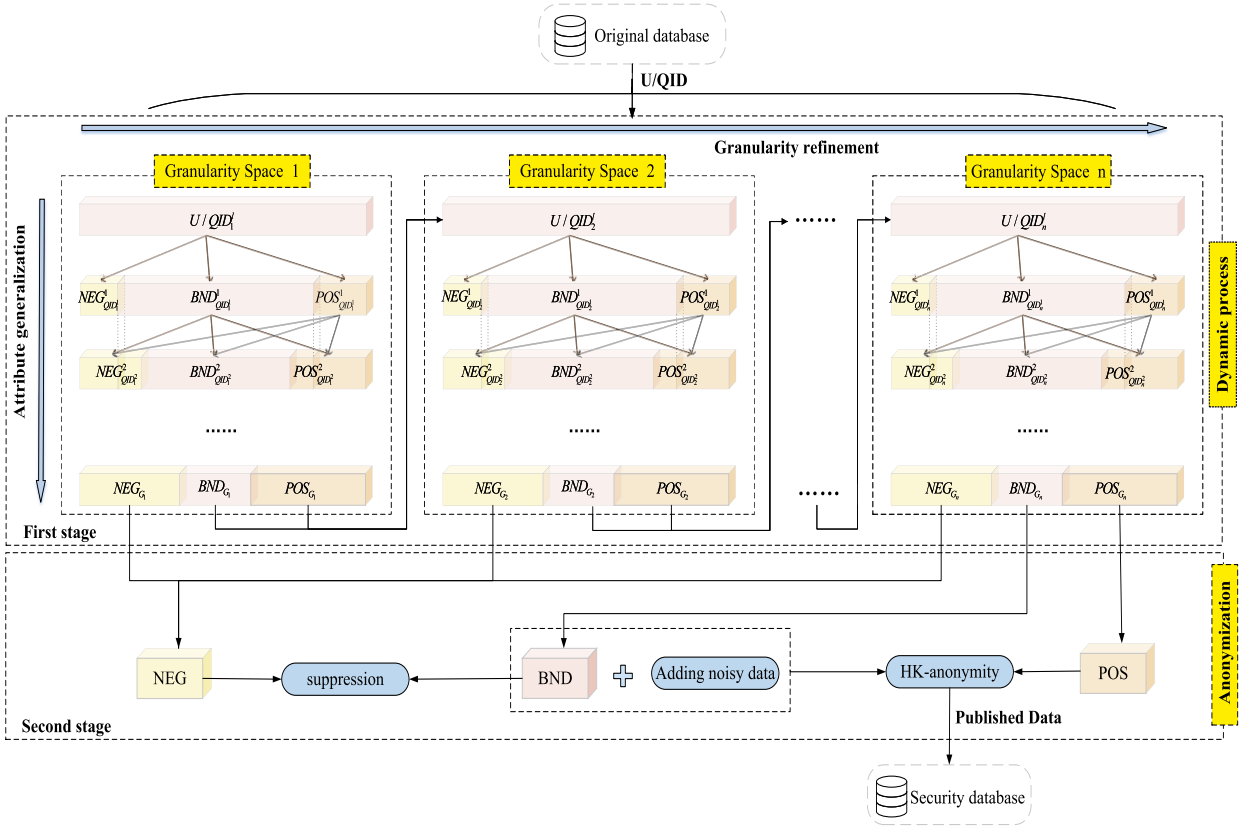

Fig. 2. A dynamic anonymization framework based on hierarchical sequential three-way decisions.

3.2. A novel k-anonymity based on hierarchical sequential three-way decisions (KHS3WD)

This subsection focuses on the processing of the k-anonymity framework based on hierarchical sequential three-way decisions and proposes a k-anonymity algorithm based on HS3WD.

Definition 9. Given a multi-level decision table $LDT_t^l = \{U_t, QID_t^l, D_t^l\}(l = 1,2,...,s, t = 1,2,...,n)$, a sequence of attribute sets $QID_1^l \subset QID_2^l \subset ... \subset QID_n^l \subseteq QID_t^l$, a pair of k-values (HK, LK) and a hierarchical granular structure $G = \{G_1, G_2, ..., G_n\}$ with respect to QID_t^l. For the t-th level of granular structure $G_t = \{LDT_t^l, QID_t^l\}$ at the l-th generalization level, the three regions can be defined as follows:

$$\begin{aligned} POS_{QID_t^l}(f(u)) &= \{u \in U_t^l | f(u) \geq HK\} \\ BND_{QID_t^l}(f(u)) &= \{u \in U_t^l | LK < f(u) < HK\} \\ NEG_{QID_t^l}(f(u)) &= \{u \in U_t^l | f(u) \leq LK\} \end{aligned} \tag{6}$$

where $U_{t+1} = POS_{QID_t}^s(f(u)) \bigcup BND_{QID_t}^s(f(u))$ and $f(u)$ denotes the size of the equivalence group in which the record u is located.

It is crucial to make clear that the objects represented by the boundary region comply with the LK anonymity condition. Although they do not quite satisfy the HK anonymity requirement, these objects are near to it. We choose to adjust the anonymity requirement accordingly, meaning we opt to appropriately lower the anonymity threshold. Consequently, data that meets the LK anonymity requirement is categorized into the boundary region, while data failing to meet the LK anonymity requirement is allocated to the negative region. Then, we take the appropriate actions, such as generalizing these parts of the data upward to the highest level (the s-th level). The data will be suppressed and not be able to proceed on to the next level of granularity if the generalization to s-th level fails to satisfy the LK anonymity requirement.

Current data anonymization techniques typically use a uniform approach for all data. However, this approach is too costly when dealing with large amounts of complex data. HS3WD provides us with a new processing thinking: through the multi-level decision table, the original data is divided into different coarse and fine granularity structures, and processing the data on the coarse granularity can avoid a lot of unnecessary information loss. For example, it may be possible to determine that the data does not satisfy the anonymity requirement by using only two quasi-identifier attributes, whereas it is more difficult to satisfy the anonymity condition

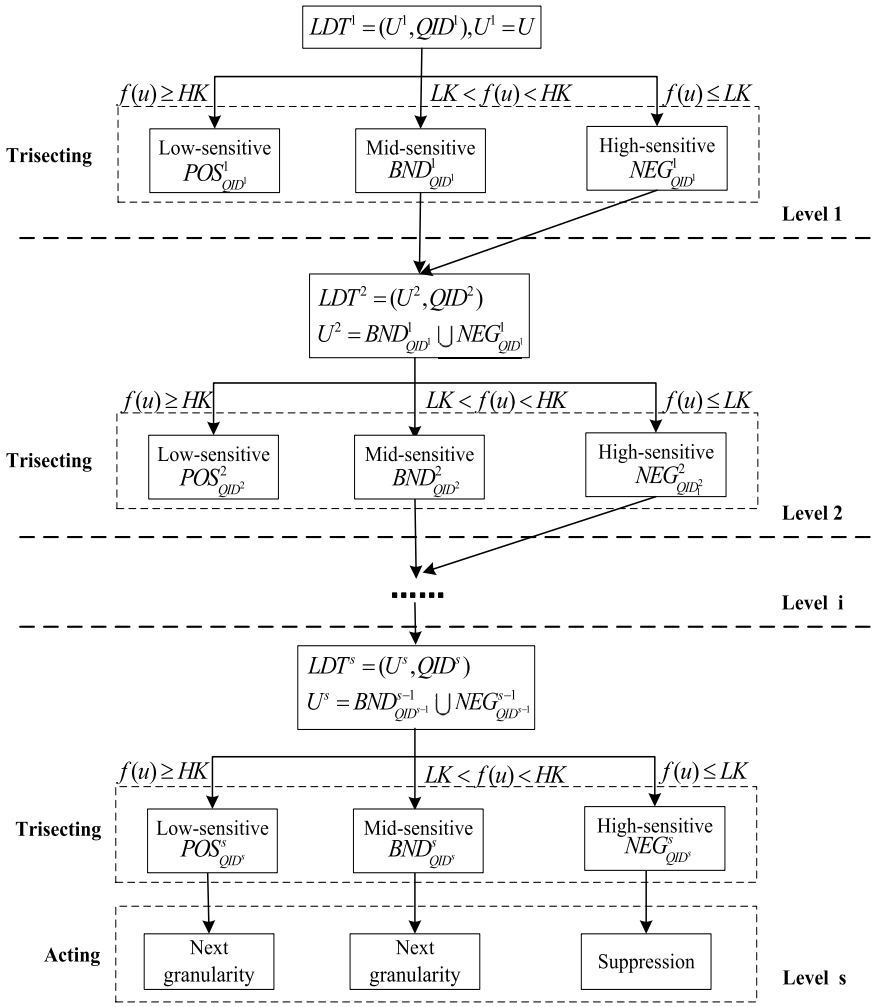

Fig. 3. Attribute generalization process at a single granularity.

using three quasi-identifier attributes. Clearly, the processing cost of attribute generalization for two attributes is significantly less than for three attributes. Therefore, if we can process the data at a coarse granularity structure, it will reduce a lot of unnecessary costs.

To this end, we propose a dynamic anonymization model based on HS3WD shown in Fig. 2. Our dynamic anonymization framework is divided into two main stages:

(1) Dynamic processing stage. In this stage, we divide the original dataset into an n-th level granularity structure space according to the number of quasi-identifier attributes through the HS3WD model. Fig. 3 illustrates the attribute generalization process for a single granularity space. In each granularity structure, we divide the dataset into s-level hierarchies according to Definition 8. At each level, the data is divided into positive, negative, and boundary regions based on the anonymous parameter (HK, LK) according to the division rules outlined in Definition 9. Then, we generalize the negative and boundary regions until they are generalized the s-th level. After generalizing the s-th level, the data still classified in the boundary and positive regions is processed further. This data is treated as the universe for the next granularity level. Meanwhile, the negative region is suppressed and not processed further at the next granularity level. When all the data in the granularity space have been processed, we can calculate the results of the division of the entire data table into the positive, negative, and boundary regions.

(2) Anonymization stage. In this stage, we process the results of the calculations from the previous stage. We anonymize the data in the positive region, suppress the data in the negative region and delay the processing of the data in the boundary region, which can be either suppressed or added with noise before anonymization.

It should be noted that in the dynamic process stage, we suppress the data in the negative region of a certain granularity space, which will make the data to be processed in the next granularity space become less.

To prove the utility of our proposed framework, we propose a novel k-anonymity based on hierarchical sequential three-way decisions, which is handled as shown in Algorithm 1. More specifically, we first divide all quasi-identifier attributes in the dataset to obtain a sequence of attribute sets $QID^l_1 \subset QID^l_2 \subset \ldots \subset QID^l_n \subseteq QID^l_l$ as input to the algorithm and construct the granularity space

$G_t(t = 1, 2, ..., n)$ based on the different attribute sets. Lines 2 to 5 denote that, in each granularity space, the data in the multi-level decision table LDT_t^l is divided into m equivalence groups $\{C_1, C_2, ...C_m\}$ using the l-th level quasi-identifier attributes according to Definition 9. In lines 6 to 17, we evaluate whether the equivalence group satisfies the anonymity requirement based on its size. If it satisfies, it is classified as the positive region and used as the processing data in the subsequent granularity space. If it does not, it is classified as either the boundary region or negative region, and data classified as the boundary and negative regions are generalized upwards to the $l + 1$ generalization level, where they are reclassified into the three regions until they are generalized to the s-th level. After the s levels of generalization have been processed, the data of the negative region will be suppressed and will not be processed in the next granularity space. In contrast, the data of the positive and boundary regions will progress to the next granularity space, where they will undergo further subdivision into three regions. The time complexity of the Algorithm 1 can be calculated from the above analysis as $O(n \cdot s)$.

Algorithm 1: KHS3WD: a novel k-anonymity algorithm based on HS3WD.

Input: (1) Original datasets, U; total attribute nums, n; k-value pair (HK, LK);
(2) A multi-level decision table $LDT = \{LDT_1^l, LDT_2^l, ...LDT_n^l\}(t = 1, 2, ..., n)$.
(3) A sequence of attribute sets $QID_1^l \subset QID_2^l ... QID_n^l \subseteq QID_t^l$.
Output: Security database Q.

1 Initialize $U_1^l = U$, $Q = \varnothing$.
2 **for** $t \leftarrow 1$ **to** n **do**
3 $Dealdata = U_t^l$;
4 **for** $l \leftarrow 1$ **to** s **do**
5 $Dealdata/QID_t^l = \{C_1, C_2, ..., C_m\}$ according to Definition 9;
6 **if** $|C_l| \geq HK$ **then**
7 $POS^l_{QID_t^l}(C) = \{u \in U_t^l || C_l| \geq HK\}$;
8 **else if** $LK < |C_l| < HK$
9 $BND^l_{QID_t^l}(C) = \{u \in U_t^l | LK < |C_l| < HK\}$;
10 **else**
11 $NEG^l_{QID_t^l}(C) = \{u \in U_t^l || C_l| \leq LK\}$
12 $Dealdata -= POS^l_{QID_t^l}(C)$;
13 **if** $l \leq s$ **then**
14 $l = l + 1$; **turn to** line 10;
15 **else**
16 break;
17 **end**
18 $NEG_{G_t} = NEG^s_{QID_t^s}(C)$;
19 $POS_{G_t} = POS_G \bigcup POS^l_{QID_t^l}(C)$ and $BND_{G_t} = U - POS_{G_t} - NEG_{G_t}$;
20 **if** $t + 1 \leq n$ **then Computed**
21 $U^l_{t+1} = POS_{G_t}(C) \bigcup BND_{G_t}(C)$, $t = t + 1$; **turn to** line 8;
22 **else** break;
23 **end**
24 $Q = POS_{G_n}(C)$;
25 **Output** Q;

3.3. An enhanced anonymity model based on KHS3WD

In model KHS3WD, we divided the original dataset into three regions and suppressed the data in the boundary region. We recognize that the data in the boundary region closely adheres to anonymity requirements, and suppressing this part of the data will lead to information loss. To solve this problem, in this subsection, we enhance KHS3WD and propose a k-anonymity model based on hierarchical sequential three-way decisions with a noise mechanism (KNHS3WD) as shown in Algorithm 2. This model reduces information loss by adding an appropriate amount of noise data to the boundary region, ensuring it meets anonymization requirements.

The main flowchart of the algorithm is shown in Fig. 4. We reprocess the data in the boundary region after generalization to the s-th level so that it satisfies the anonymity requirement, thus enhancing data usability. In particular, we input a k-value pair and the amount of noise data N. The granularity structure and equivalence groups division are consistent with KNHS3WD. We use the idea of three-way decisions to classify the equivalence groups. Next, we manipulate the classification results accordingly. The negative region indicates that if it still does not satisfy the anonymity requirement after generalization, then the anonymity requirement must not be satisfied either in the next granularity space, we suppress this data. We store the data meeting the requirement at each level in each granularity space in the secure dataset Q for release. Therefore, based on the above analysis, it is easy to see that the time complexity of the proposed Algorithm 2 is also $O(n \cdot s)$.

To make it easier to understand, we illustrate with a specific example, as shown in Example 2.

Example 2. Given a data table $U = \{u_1, u_2, ..., u_{19}\}$ which has three attributes $\{A_1, A_2, A_3\}$, $A_i = \{a_i^1, a_i^2, a_i^3\}$, $QID_1 = \{A_1, A_2\}$, $QID_2 = \{A_1, A_2, A_3\}$, $QID_1^3 = \{\{a_1^1, a_2^1\}, \{a_1^2, a_2^2\}, \{a_1^3, a_2^3\}\}$, $QID_2^3 = \{\{a_1^1, a_2^1, a_3^1\}, \{a_1^2, a_2^2, a_3^2\}, \{a_1^3, a_2^3, a_3^3\}\}$ and a pair of anonymous parameters $(HK, LK) = (6, 3)$, we can conclude the following:

Algorithm 2: KNHS3WD: K-anonymity model based on HS3WD with the introduction of a noise mechanism.

Input: (1) Original datasets, U; total attribute numbers, n; k-value pair (HK, LK);
(2) A multi-level decision table $LDT = \{LDT_1^l, LDT_2^l, ...LDT_n^l\}$, $(t = 1, 2, ..., n)$;
(3) A sequence of attribute sets $QID_1^l \subset QID_2^l \subset \ldots \subset QID_n^l \subseteq QID_t^l$; $G_t = \{LDT_t^l, QID_t^l, EC_t^l\}$; Noisy data N.
Output: Security database Q.

1 Initialize $U_1^l = U$, $Q = \varnothing$.
2 **for** $t \leftarrow 1$ ***to*** n **do**
3 $Dealdata = U_t^l$;
4 **for** $l \leftarrow 1$ ***to*** s **do**
5 $Dealdata/QID_t^l = \{C_1, C_2, ..., C_m\}$ according to Definition 9;
6 **if** $|C_l| \geq HK$
7 $POS^l_{QID_t^l}(C) = \{u \in U_t^l || C_l| \geq HK\}$;
8 **else if** $LK < |C_l| < HK$
9 $BND^l_{QID_t^l}(C) = \{u \in U_t^l | LK < |C_l| < HK\}$;
10 **if** $t = n$ **and** $l = s$
11 $N+ = (HK - |C_l|)$;
12 **else**
13 $NEG^l_{QID_t^l}(C) = \{u \in U_t^l || C_l| \leq LK\}$
14 $Dealdata- = POS^l_{QID_t^l}(C)$;
15 **if** $l \leq s$ **then**
16 $l = l + 1$;**turn to** line 10;
17 **else**
18 break;
19 **end**
20 $NEG_{G_t} = NEG^s_{QID_t^s}(C)$;
21 $POS_{G_t} = POS_G \bigcup POS^l_{QID_t^l}(C)$;
22 **if** $t = n$ **and** $l = s$
23 $BND_{G_t} = BNG_{G_t} \bigcup noisydata$;
24 **else**
25 $BND_{G_t} = U - POS_{G_t} - NEG_{G_t}$;
26 **if** $t + 1 \leq n$ **then Computed**
27 $U^l_{t+1} = POS_{G_t}(C) \bigcup BND_{G_t}(C)$, $t = t + 1$; **turn to** line 8;
28 **else** break;
29 **end**
30 $Q = POS_{G_n}(C) \bigcup BND_{G_n}(C)$;
31 **Output** Q;

(1) We divide the original data table into two granularity spaces for processing.

① For the first level of the first granularity space G_1, $Dealdata = U_1^l = U$. Consider the division of equivalence groups: $Dealdata/QID_1^1 = \{\{5, 3, 14, 11, 15, 16, 8\}, \{10, 9, 7, 4, 2\}, \{12, 13\}, \{6, 1\}, \{17\}, \{18\}, \{19\}\}$, we can calculate the following three regions:

$$POS^1_{QID_1^1}(C) = \{5, 3, 14, 11, 15, 16, 8\},$$
$$BND^1_{QID_1^1}(C) = \{10, 9, 7, 4, 2\},$$
$$NEG^1_{QID_1^1}(C) = \{\{12, 13\}, \{6, 1\}, \{17\}, \{18\}, \{19\}\}.$$

② For the second level of the first granularity space G_1, $Dealdata = BND^1_{QID_1^1}(C) \cup NEG^1_{QID_1^1}(C)$. Consider the division of equivalence groups: $Dealdata/QID_1^2 = \{\{10, 9, 7, 4, 2, 19\}, \{12, 13, 1, 6\}, \{17\}, \{18\}\}$, we can calculate the following three regions:

$$POS^2_{QID_1^2}(C) = \{10, 9, 7, 4, 2, 19\},$$
$$BND^2_{QID_1^2}(C) = \{12, 13, 6, 1\},$$
$$NEG^2_{QID_1^2}(C) = \{\{17\}, \{18\}\}.$$

③ For the third level of the first granularity space G_1, $Dealdata = BND^2_{QID_1^2}(C) \cup NEG^2_{QID_1^2}(C)$. Consider the division of equivalence groups: $Dealdata/QID_1^3 = \{\{12, 13, 1, 6, 17\}, \{18\}\}$, we can calculate the following three regions:

$$POS^3_{QID_1^3}(C) = \varnothing,$$

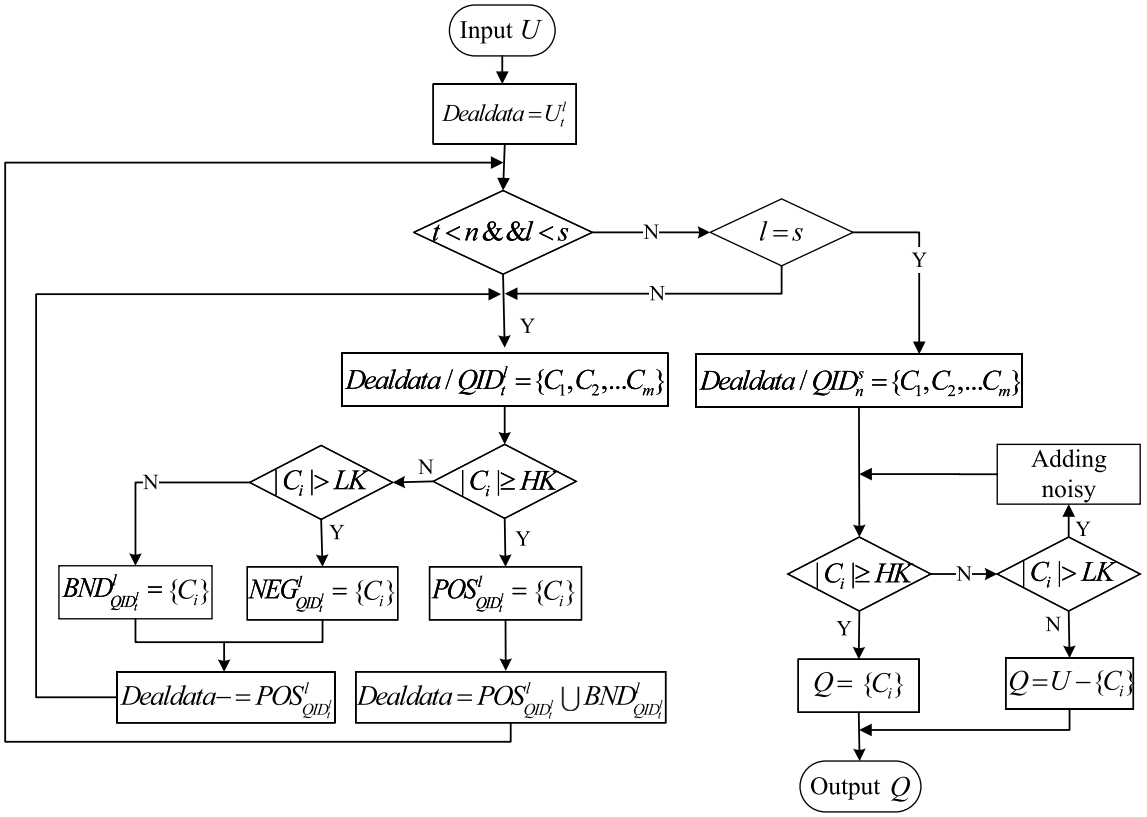

Fig. 4. The main flowchart of KNHS3WD algorithm.

$$BND^3_{QID^3_1}(C) = \{12, 13, 6, 1, 17\},$$

$$NEG^3_{QID^3_1}(C) = \{18\}.$$

The negative region is $NEG_{G_1} = NEG^3_{QID^3_1}(C)$, the positive region of this generalization process to the end is $POS_{G_1} = POS^1_{QID^1_1}(C) \bigcup POS^2_{QID^2_1}(C) \bigcup POS^3_{QID^3_1}(C)$ and the boundary region is $BND_{G_1} = U - POS_{G_1} - NEG_{G_1}$, the negative region is highly sensitive data, which are suppressed, so we get the data that needs to be deleted in the first granularity space as $\{18\}$.

(2) The data to be processed at the next granularity is updated as $U_2^l = POS_{G_1} \bigcup BND_{G_1}$.

① For the first level of the second granularity space G_2, $Dealdata = U_2^l$. Consider the division of equivalence groups: $Dealdata/QID_2^1 = \{\{5, 3, 14, 11, 15, 16\}, \{10, 9, 7, 4\}, \{2, 12, 13\}, \{6, 1\}, \{8\}, \{17\}, \{19\}\}$, we can calculate the following three regions:

$$POS^1_{QID^1_2}(C) = \{5, 3, 14, 11, 15, 16\},$$

$$BND^1_{QID^1_2}(C) = \{10, 9, 7, 4\},$$

$$NEG^1_{QID^1_2}(C) = \{\{2, 12, 13\}, \{6, 1\}, \{8\}, \{17\}, \{19\}\}.$$

② For the second level of the second granularity space G_2, $Dealdata = BND^1_{QID^1_2}(C) \cup NEG^1_{QID^1_2}(C)$. Consider the division of equivalence groups: $Dealdata/QID_2^2 = \{\{10, 9, 7, 4, 19\}, \{2, 12, 13\}, \{6, 1\}, \{8\}, \{17\}\}$, we can calculate the following three regions:

$$POS^2_{QID^2_2}(C) = \varnothing,$$

$$BND^2_{QID^2_2}(C) = \{10, 9, 7, 4, 19\},$$

$$NEG^2_{QID^2_2}(C) = \{\{2,12,13\},\{6,1\}\},\{8\},\{17\}\}.$$

③ For the third level of the second granularity space G_2, $Dealdata = BND^2_{QID^2_2}(C) \cup NEG^2_{QID^2_2}(C)$. Consider the division of equivalence groups: $Dealdata/QID^3_2 = \{\{10,9,7,4,19,17\},\{2,12,13,6,1\},\{8\}\}$, we can calculate the following three regions:

$$POS^3_{QID^3_2}(C) = \{10,9,7,4,19,17\},$$

$$BND^3_{QID^3_2}(C) = \{2,12,13,6,1\},$$

$$NEG^3_{QID^3_2}(C) = \{8\}.$$

Similarly, we can get the second granularity space for the data that needs to be deleted as $\{8\}$. For the boundary region, we add noisy data u_{20} to it, which is the same as any record in the boundary region, here we only consider the number of data to be added without considering the specific content of the added data, so the boundary region is $BND_{G_2} = BND^3_{QID^3_2}(C)' = \{2,12,13,6,1,20\}$. In addition, we can separately calculate the positive region as $POS_{G_2} = \{\{5,3,14,11,15,16\},\{10,9,7,4,19,17\}$ and the negative region as $NEG_{G_2} = \{8\}$.

After all the granularity spaces are processed, we anonymize the classification results, for both the positive region data $POS_{G_3}(C)$ and the boundary region data $BND_{G_3}(C)$ satisfy 6-anonymity, so the final output is a securely published dataset $Q = \{u_1,u_2,u_3,u_4,u_5,u_6,u_7,u_9,u_{10},u_{11},u_{12},u_{13},u_{14},u_{15},u_{16},u_{17},u_{19},u_{20}\}$. □

3.4. Algorithm analysis and discussion

3.4.1. Security analysis

In this paper, we combine k-anonymity and HS3WD to propose a dynamic anonymization model. Essentially, the privacy-preserving effect of the proposed algorithm is based on k-anonymity, which satisfies the classified anonymous principle in Definition 9 with high security. The specific security analysis is illustrated below.

Theorem 1. *For any hierarchical granular structure G^l_t with respect to $QID^l_t(l = 1,2,...,s,t = 1,2,...,n)$, where QID^l_t represents the set of quasi-identifier attributes of the l-th level of generalization hierarchy containing t attributes, the risk of privacy leakage in $POS_{QID^l_t}$ is $\frac{1}{HK}$, and the risk of privacy leakage for the dataset is $\frac{1}{HK}$.*

Proof. The dynamic anonymization algorithms proposed in this paper first use attribute generalization trees to construct multi-level decision tables for anonymization, and during the anonymization process, the objects that satisfy the HK anonymity requirement are divided into the positive region, which means that there exists at least $HK-1$ equivalence classes for each record in the positive region, thus reducing the correct identification rate of the adversary. In addition, the algorithm reprocesses the data in the boundary region to make such objects also satisfy the HK anonymization requirement by adding noisy data, when a small amount of noise can be tolerated. Both attribute generalization and adding noise can effectively cut off the correlation between quasi-identifier attributes, defend against linking attacks by adversaries, and prevent information leakage. Referring to Definition 3, it's evident that the risk of privacy leakage for a dataset satisfying HK-anonymity is $\frac{1}{HK}$, therefore, for our algorithms, the risk of privacy leakage for the positive region as well as the securely published dataset is $\frac{1}{HK}$. □

3.4.2. Algorithm discussion

This paper introduces hierarchical sequential three-way decisions into k-anonymity, then proposes a novel dynamic anonymity privacy-preserving model based on hierarchical sequential three-way decisions, providing a more effective solution for the needs of the big data era. However, there are still some problems that need to be further discussed.

(1) On the one hand, the algorithms proposed in this paper are an extended version based on k-anonymity, which means that the security of the algorithm depends on the security of k-anonymity. Currently, the k-anonymity algorithm still has some shortcomings, such as the inability to defend against homogeneity attacks and background knowledge attacks. In order to address this problem, we can learn about the latest data anonymization techniques, and explore whether they can be applied to the framework introduced in this paper to further improve the security.

(2) On the other hand, these algorithms construct different granularity spaces by sequentially adding attributes one by one, whether the order of adding attributes affects the algorithm's effect, namely, which attribute is preferred to be added for the best algorithmic processing, this problem can be combined with the optimal attribute selection by decision-maker, and the specific solution is also the focus of our research work after that.

Table 4
Description of the datasets.

No.	Datasets	$\|U\|$	$\|QID\|$
1	Adult	48842	8
2	MAGIC Gamma Telescope	19020	10
3	Agaricus-lepiota	8124	16
4	Abalone	4177	8
5	Wine Quality (red)	1599	11
6	Obesity	2111	16

Table 5
Setting of parameters.

	Dataset1	Dataset2	Dataset3	Dataset4	Dataset5	Dataset6
CASE1	(10,6)	(10,6)	(6,4)	(10,6)	(10,6)	(6,4)
CASE2	(30,20)	(30,20)	(10,6)	(30,20)	(20,10)	(10,6)
CASE3	(60,45)	(40,30)	(14,10)	(40,30)	(30,20)	(14,10)
CASE4	(80,65)	(60,45)	(20,15)	(90,70)	(40,30)	(20,15)
CASE5	(100,80)	(100,80)	(25,20)	(100,80)	(80,65)	(30,20)

4. Experiment results and analysis

The primary goal of the experiments is to assess the performance of our proposed approach in terms of data availability, privacy, and processing efficiency. To ensure an accurate evaluation, we compare our methods KHS3WD and KNHS3WD with the traditional k-anonymity (KA) and the k-anonymity algorithm based on noise mechanism and 3WD (KN3WD). We implement these algorithms in Java and run these experiments on a personal computer with Microsoft Windows 10, AMD Ryzen 7 5800H with Radeon Graphics 3.20 GHz; 16.0 GB (RAM) memory. The software is IntelliJ IDEA Community Edition 2023.3.2.

4.1. Datasets

The six datasets we used are all publicly accessible datasets in the UCI Machine Learning Repository, as illustrated in Table 4. Since there are different anonymization requirements for different datasets, we set different parameters (HK, LK) for each of the six datasets as shown in Table 5. Before starting our experiments, we need to preprocess the selected datasets. We delete the records that have too many missing values and employ Rosetta software (http://www.lcb.uu.se/tools/rosetta/) to transform the continuous data into discrete values. Furthermore, we stratify the experimental data by constructing the attribute generalization trees for the quasi-identifier attributes (QID) according to the general social cognition. Finally, we supplement the data information with insufficient attribute generalization hierarchies using the data complementation method.

4.2. Cost metric

During the process of anonymization, we will inevitably lose some original information, which in turn leads to information loss. At the same time, data generalization reduces the accuracy of attribute values on quasi-identifiers, all of which have a direct impact on data usability. Thus, information loss is an important metric for measuring the performance of anonymization algorithms. Additionally, we evaluate the algorithms using three more metrics: generalized processing cost (GPC), information suppression rate (ISR), and information distortion rate (IDR). GPC measures the overall processing cost, while ISR indicates the amount of lost information during anonymization. We also consider IDR because adding noisy data may lead to distortion of the data and affect its utility.

Definition 10. (Generalized Processing Cost, GPC) Given a multilevel decision table $MT^l = \{U^l, QID_t^l, D^l\}(l = 1, 2, ..., s, t = 1, 2, ..., n)$, N noisy data and n quasi-identifier attributes in MT^l have s attribute levels, the generalization processing cost in the process of getting the security data is defined as follows:

$$GPC = \frac{\sum_{t=1}^{n}\sum_{l=1}^{s}(t * l * |M_t^l|) + m * s * |M_m^s| + n * s * N}{n * s * |MT^l|} \tag{7}$$

where $|M_m^s|$ denotes the number of suppressed records at m attributes and $|M_t^l|$ represents the number of records that satisfy the anonymity requirement for t attributes generalized to the l-th level.

It is worth stating that our proposed algorithms take a moderate amount of the added noisy data. However, in the comparison experiments, if the data that does not satisfy the anonymity requirement is only suppressed without subsequent processing, then $N = 0$ in Eq. (7).

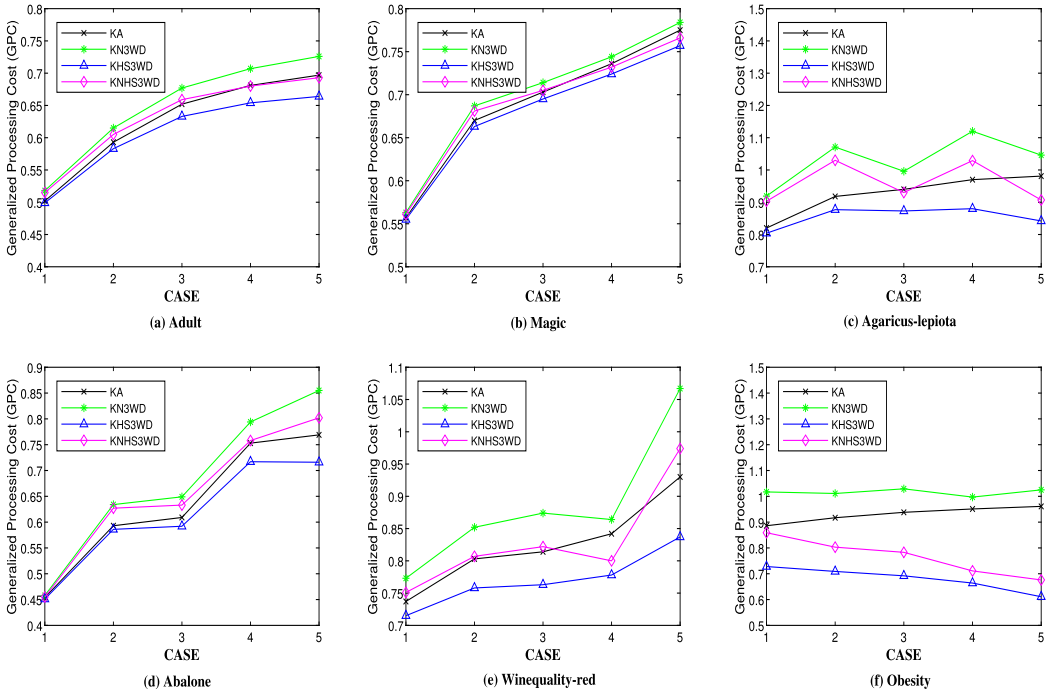

Fig. 5. Comparison of the generalized processing cost of four algorithms at the different (HK, LK).

Definition 11. (Information Suppression Rate, ISR) Given an original data table MT and an anonymized table MT', the information suppression rate is defined as follows:

$$ISR = \frac{|MT| - |MT'|}{|MT|} \tag{8}$$

where $|MT|$ indicates the total number of records in the original data table and MT' represents the total number of records in the secure data table to be released.

Definition 12. (Information Distortion Rate, IDR) Given an original data table MT, N is the added noisy data, then the information distortion rate is defined as follows:

$$IDR = \frac{|N|}{|MT| + |N|} \tag{9}$$

where $|MT|$ indicates the total number of records in the original data table and $|N|$ denotes the number of added noisy data.

Definition 13. (Information Loss Rate, ILR) The sum of GPC, ISR and IDR as the total information loss rate is defined as:

$$ILR = GPC + ISR + IDR \tag{10}$$

Note that ILR takes into account both generalized processing cost GPC, suppression rate ISR, and distortion rate IDR for a more comprehensive evaluation capability. Based on the above analysis, it is easy to see that the smaller the ILR is, the smaller the information loss is and the higher the data availability becomes.

4.3. Comparison of the generalized processing cost at the different (HK, LK)

In this subsection, we compare the generalization cost of the traditional k-anonymity (KA), the k-anonymity model based on three-way decisions and differential privacy (KN3WD), and our proposed two algorithms (KHS3WD, KNHS3WD) on six datasets, and analyze the effect of different cases, namely, different pairs of k-values (HK, LK), on the generalization cost. To facilitate the experiments, we use different CASES to denote different k-value pairs (HK, LK) as shown in Table 5. KHS3WD is an improved algorithm for K anonymization based on HS3WD, which does not involve the addition of noisy data, whereas KNHS3WD combines the idea of differential privacy with the addition of noisy data based on the KHS3WD algorithm. Fig. 5 shows the experimental results on six datasets.

We can conclude the following by observing Fig. 5:

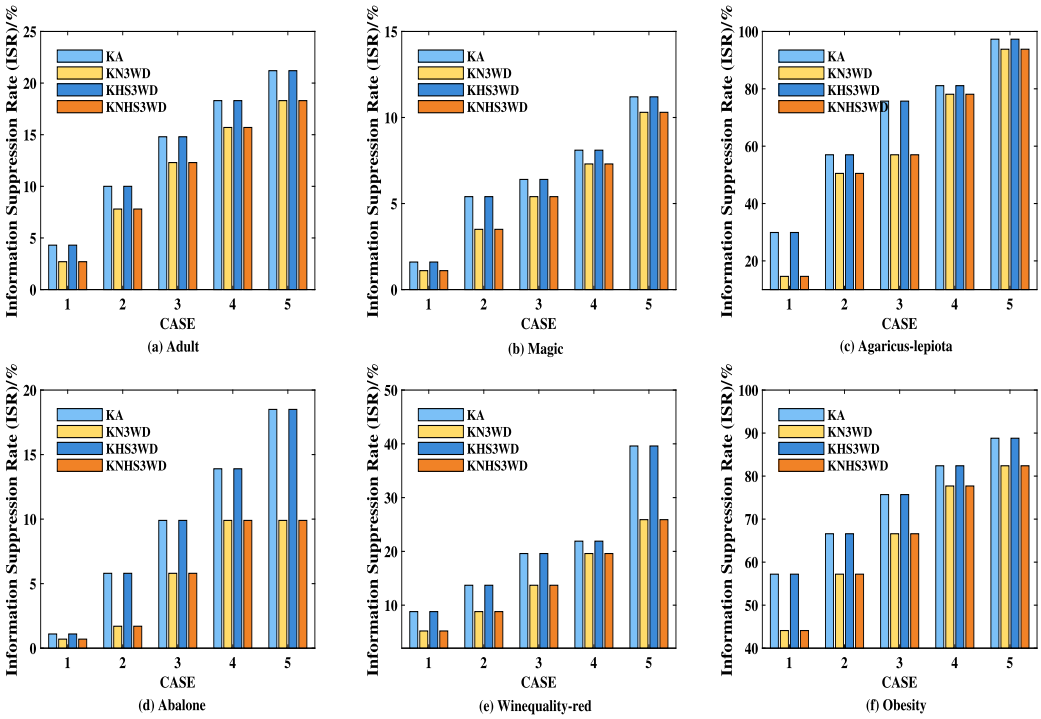

Fig. 6. Comparison of the information suppression rate of four algorithms at the different (HK, LK).

(1) For all datasets, without adding noise, our proposed algorithm KHS3WD has a lower generalization processing cost than the traditional k-anonymity algorithm.

(2) We improve the data availability by adding noise, from the experimental results we can see that also in the case of adding noise, our proposed algorithm KNH3WD has a lower generalization processing cost than KN3WD, which is enough to prove the superiority of our algorithm.

(3) Moreover, it should be noted that KN3WD has a higher processing cost than the traditional k-anonymity algorithm because KN3WD adds noise and also needs to process the noisy data compared to the traditional k-anonymity. Thus it introduces a certain generalized processing cost, but the availability of the data improves in this way, as can be seen in Fig. 6, and as we will explain in more detail later on.

4.4. Comparison of the information suppression rate at the different (HK, LK)

In the previous subsection, we mentioned that the purpose of adding noise is to reduce the information suppression rate and thus improve the usability of the data. In what follows, we compare the information suppression rate of the four algorithms under different scenarios, and Fig. 6 illustrates the experimental results for the six datasets.

From Fig. 6, one can notice that as the value of k-value pairs goes up, there is a rise in the rate of suppression. This is because the higher the k-value pair, the stricter the anonymity requirements are. Consequently, more data needs to be removed to ensure anonymity, leading to an increase in the rate of information suppression. When a uniform processing approach is taken to the data, KN3WD shows a decrease in the information suppression rate compared to the traditional k-anonymity algorithm. This shows that the processing mechanism of adding noise is effective in improving the availability of data. When the processing method proposed in this paper is used, KNHS3WD also shows a decrease in the information suppression rate compared to the KHS3WD algorithm. In individual situations, adding noise significantly reduces the data suppression rate.

By comparing the information suppression rates of the four algorithms with the same k-value pairs, one can find that the information suppression rate of KA is the same as that of KHS3WD. This means that the amount of data to be suppressed is the same no matter what kind of processing we use on the data, so the information suppression rate of the algorithms is also the same. Combined with Fig. 5, we can find that KA and KHS3WD proposed in this paper have the same information suppression rate, but the generalized processing cost of KHS3WD is lower than that of KA. Therefore, we can conclude that different data processing methods do not affect the amount of suppressed data and have lower generalized processing cost, which ensures the correctness of the algorithm's anonymous processing. From this point of view, we can also show the effectiveness and correctness of the algorithm proposed in this paper.

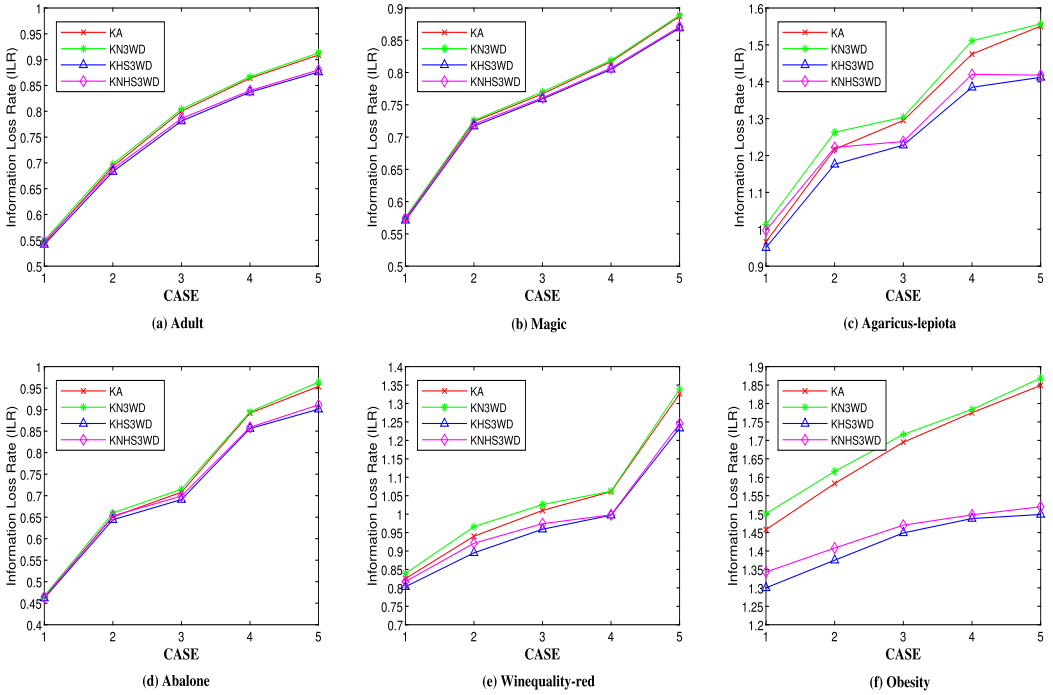

Fig. 7. Comparison of the information loss rate of four algorithms at the different (HK, LK).

4.5. Comparison of the information loss rate at the different (HK, LK)

To further investigate the performance of our proposed algorithms, we introduce the information loss rate as a synthetically evaluated metric. In this subsection, we explore the effect of different k-value pairs on the information loss and compare the variation of information loss produced by the four algorithms in different cases on six datasets. The experimental results are shown in Fig. 7.

By observing the experimental results, one can easily see that as the k-value increases, the information loss roughly shows a gradual increase, however, the information loss of the KNHS3WD algorithm decreases slightly from CASE4 to CASE5 on the Agaricus-lepiota dataset. This is due to the fact that the anonymity requirements become stricter as the k-value pairs increase, resulting in KNHS3WD deleting more data in the CASE5 (14,10). Although there's an increase in the suppression rate, the reduction in the generalized processing costs and information distortion outweighs this rise. Based on Definition 13, a slight decrease in the final information loss rate can be calculated.

On the other hand, observing the distances between the lines in Fig. 7, one can find that the information loss rate of all four algorithms increases as the value of k increases in all datasets. However, the information loss rates of both our proposed KHS3WD and KNHS3WD algorithms are smaller than those of the traditional k-anonymity algorithm and KN3WD algorithm, which suggests that our proposed algorithms are able to reduce the information suppression rate while still maintaining a low information loss rate. In other words, our proposed algorithm improves data availability while preserving data privacy and has a low generalized processing cost.

4.6. Comparison of the information loss rate at different numbers of attributes

To clarify the applicability of our proposed algorithms, we also compare the effect of these four algorithms on information loss with different numbers of attributes. Experimental results are shown in Fig. 8.

From Fig. 8 we can see that as the number of attributes increases, the information loss rate is gradually increasing. This is because as the number of attributes increases the conditions for judging equivalence classes become more stringent, leading to greater information loss. Of course, there are exceptions. One can see Fig. 8(f) that the number of attributes in the Obesity dataset from 14 to 16, the information loss of our proposed algorithms KHS3WD and KNHS3WD decreases, which is because the information suppression rate is almost the same at 14 attributes as at 16 attributes. In other words, there is no effect of the 16th attribute on the information suppression rate, which is most probably be due to the characteristics of the dataset itself, which was able to determine the security data and suppression data with 14 attributes. In this situation, our proposed algorithms KHS3WD and KNHS3WD are more superior. These algorithms process the dataset at the 14th attribute, reducing the cost of generalizing records at the 16th attribute and decreasing information loss.

Additionally, by observing the distance between the four fold lines, we observe a widening gap between the KA and KHS3WD lines, as well as between the KN3WD and KNHS3WD lines, which shows that, as the number of attributes increases, our proposed

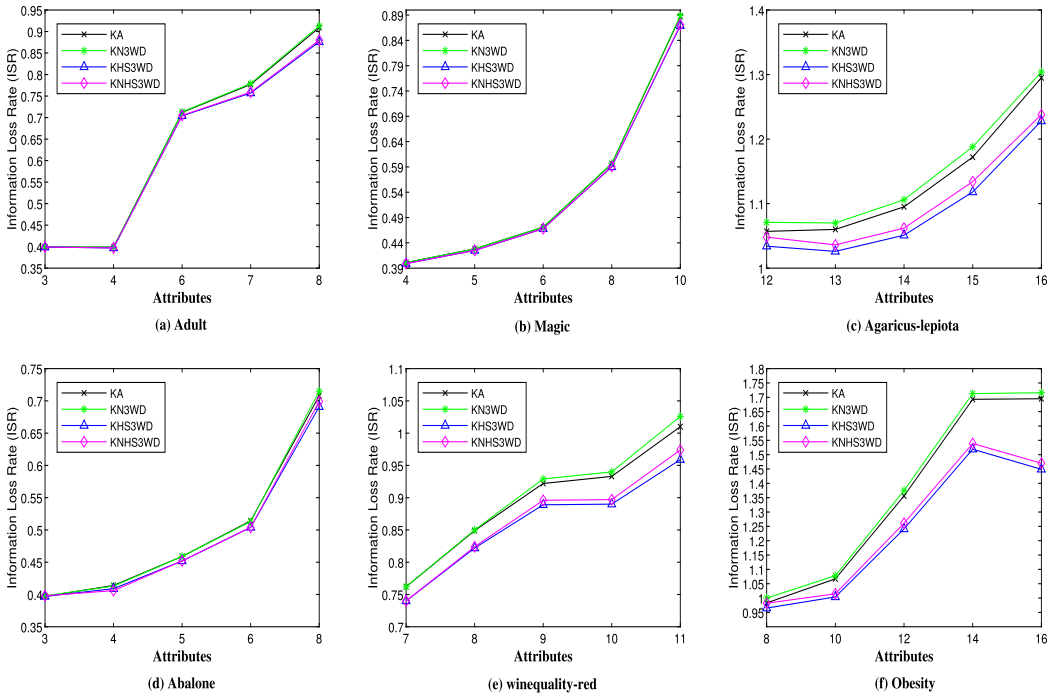

Fig. 8. Comparison of the information loss rate of four algorithms at different number of attributes.

Table 6
Confusion Matrix.

	Predicted positive	Predicted negative
Actual positive	True Positive (TP)	False Negative (FN)
Actual negative	False Positive (FP)	True Negative (TN)

algorithms cause less information loss and have more obvious advantages. Furthermore, from Figs. 8(c), (e) and (f), it becomes evident that our proposed algorithm is particularly well-suited for multi-attribute datasets.

4.7. Comparison of the utility of the proposed algorithms at different (HK, LK)

Finally, to validate the utility of our proposed algorithms, we use F-Measure to comprehensively evaluate the performance of our methods and compare it with those of algorithms KA and KN3WD on six datasets.

When anonymizing data, four different results may be produced. The possible outcomes are defined in terms of a confusion matrix as shown in Table 6. The four possible outcomes are specifically defined:

- True Positive (TP): Data requiring anonymization will be anonymized;
- True Negatives (TN): Data not requiring anonymization remains unaltered;
- False Positive (FP): Data not requiring anonymization will also be anonymized;
- False Negatives (FN): Data requiring anonymization remains unaltered.

Using these four results, the following equations for Precision and Recall can be obtained:

$$Precision = \frac{TP}{TP + FP} \tag{11}$$

$$Recall = \frac{TP}{TP + FN} \tag{12}$$

F-Measure is a metric defined as the weighted harmonic mean of Precision and Recall. It tends to agree with the smaller of the two values, the higher the F-measure value, the higher both Precision and Recall are, reflecting a more comprehensive evaluation of the algorithm's performance. It is computed by the following formula:

$$F = \frac{2 \times Precision \times \mathrm{Re}\,call}{Precision + \mathrm{Re}\,call} \tag{13}$$

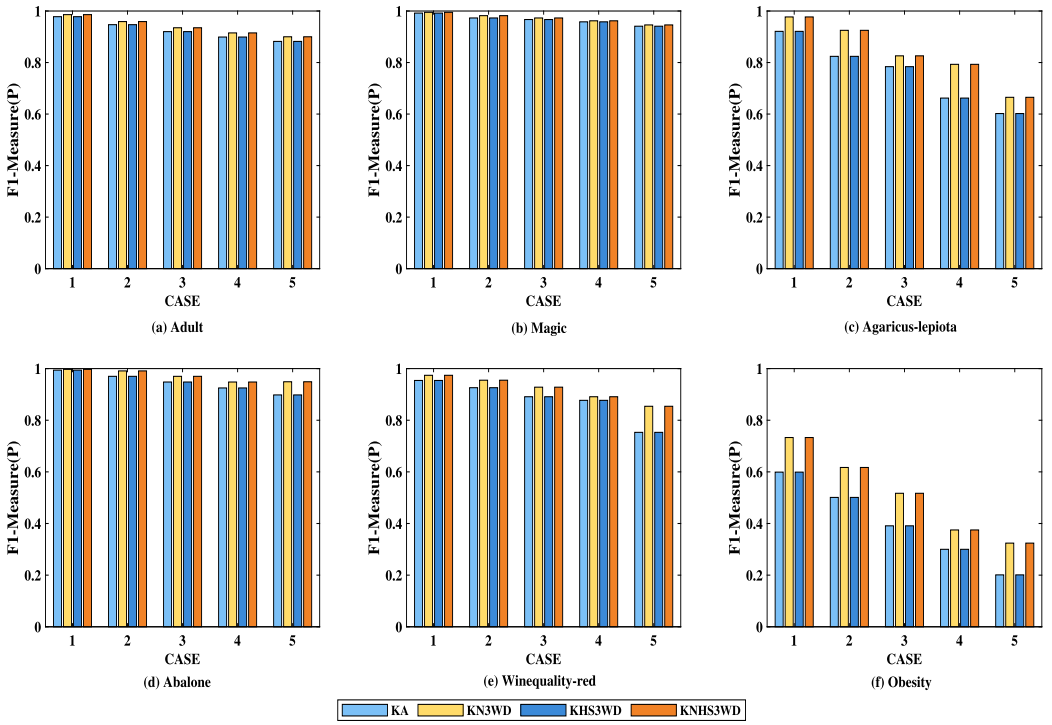

Fig. 9. Comparison of the utility of four algorithms at different (HK, LK).

F-Measure takes values between 0 and 1, with 1 and 0 representing the best and the worst performance, respectively. When the value of F-Measure is higher, the model performance is better. The experimental results are shown in Fig. 9.

From Fig. 9, one can observe a decreasing trend in the F-Measure values across the six experimental datasets as the k-value pairs increase. This indicates a declining utility of the algorithm for anonymizing the data. The reason behind this trend is that with increasing k-value pairs, the requirement for data anonymization becomes more stringent, making it more challenging to anonymize the data effectively. However, the F-Measure values of our proposed algorithm remain very close to 1 for most datasets, suggesting its effectiveness. Additionally, in Fig. 9(f), although the effectiveness of our proposed algorithm is not as strong as in other datasets, its utility remains comparable to that of the comparison algorithms KA and KN3WD.

In conclusion, from the analysis of the above experimental results, the anonymization utility of the proposed algorithms in this paper has performed excellently on most of the datasets, which confirms that the proposed algorithm has a good performance, although some datasets are generally anonymized, but this is mainly due to the characteristics of the datasets themselves, and perhaps these datasets may not be suitable for privacy preservation through data anonymization.

5. Conclusions

In this paper, we propose a novel anonymization algorithm (KHS3WD) that incrementally processes the original dataset attribute by attribute. In this algorithm, we partition the original dataset into different granularity levels based on the number of attributes, utilizing the concept of hierarchical sequential three-way decisions. At each granularity level, we further divide the sequential levels of attribute generalization, conducting sequential processing through attribute generalization and classifying the data. For the data that does not currently meet the anonymity requirements, we generalize their attributes and proceed to the next sequential level until the sequential process is completed, then move to the next granularity level. This gradual reduction of the dataset between granularities helps to decrease the processing cost for subsequent data. To further improve data availability, we introduce the KNHS3WD algorithm by combining the concept of differential privacy. Experimental results demonstrate the effectiveness and usability of our proposed model. In addition, our proposed model theoretically can be extended using other data anonymization techniques.

In future work, we investigate whether the order of attribute addition will have some influence on the experimental results, and the optimal attributes can be added sequentially through expert decision-making or voting mechanism to further reduce the information loss and improve the utility of the algorithm. In addition, we can also apply the framework proposed in this paper to other current anonymization techniques in the field of data anonymization to study their specific anonymization results, so as to seek a balance between data privacy and usability.

CRediT authorship contribution statement

Jin Qian: Writing – review & editing, Supervision, Methodology, Conceptualization. **Mingchen Zheng:** Writing – original draft, Validation, Software, Methodology, Conceptualization. **Ying Yu:** Writing – review & editing, Validation, Software. **Chuanpeng Zhou:** 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.

Acknowledgements

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 under Grant Nos. 20232ACB202013.

References

[1] K. Salehzadeh Niksirat, L. Velykoivanenko, N. Zufferey, M. Cherubini, K. Huguenin, M. Humbert, Wearable activity trackers: a survey on utility, privacy, and security, ACM Comput. Surv. 56 (7) (2024) 1–40.
[2] P. Wang, Y. Lei, Y. Ying, D. Zhou, Differentially private stochastic gradient descent with low-noise, Neurocomputing 585 (2024) 127557.
[3] A. Yazdinejad, A. Dehghantanha, H. Karimipour, G. Srivastava, R.M. Parizi, A robust privacy-preserving federated learning model against model poisoning attacks, IEEE Trans. Inf. Forensics Secur. 19 (2024) 6693–6708.
[4] B. Denham, R. Pears, M.A. Naeem, Enhancing random projection with independent and cumulative additive noise for privacy-preserving data stream mining, Expert Syst. Appl. 152 (2020) 113380.
[5] L. Sweeney, k-anonymity: a model for protecting privacy, Int. J. Uncertain. Fuzziness Knowl.-Based Syst. 10 (05) (2002) 557–570.
[6] B.C. Fung, K. Wang, R. Chen, P.S. Yu, Privacy-preserving data publishing: a survey of recent developments, ACM Comput. Surv. 42 (4) (2010) 1–53.
[7] A. Machanavajjhala, D. Kifer, J. Gehrke, M. Venkitasubramaniam, l-diversity: privacy beyond k-anonymity, acm transactions on knowledge discovery from data, ACM Trans. Knowl. Discov. Data 1 (1) (2007) 3–es.
[8] R. Wong, J. Li, A. Fu, K. Wang, (α, k)-anonymous data publishing, J. Intell. Inf. Syst. 33 (2009) 209–234.
[9] N. Li, T. Li, S. Venkatasubramanian, Closeness: a new privacy measure for data publishing, IEEE Trans. Knowl. Data Eng. 22 (7) (2009) 943–956.
[10] B.B. Mehta, U.P. Rao, Improved l-diversity: scalable anonymization approach for privacy preserving big data publishing, J. King Saud Univ, Comput. Inf. Sci. 34 (4) (2022) 1423–1430.
[11] W. Zheng, Z. Wang, T. Lv, Y. Ma, C. Jia, K-anonymity algorithm based on improved clustering, in: Algorithms and Architectures for Parallel Processing: 18th International Conference, ICA3PP 2018, Guangzhou, China, November 15-17, 2018, Proceedings, Part II 18, Springer, 2018, pp. 462–476.
[12] W. Mahanan, W.A. Chaovalitwongse, J. Natwichai, Data privacy preservation algorithm with k-anonymity, World Wide Web 24 (5) (2021) 1551–1561.
[13] Y. Liang, R. Samavi, Optimization-based k-anonymity algorithms, Comput. Secur. 93 (2020) 101753.
[14] S. Shaham, M. Ding, B. Liu, S. Dang, Z. Lin, J. Li, Privacy preserving location data publishing: a machine learning approach, IEEE Trans. Knowl. Data Eng. 33 (9) (2020) 3270–3283.
[15] L. Kacha, A. Zitouni, M. Djoudi, Kab: a new k-anonymity approach based on black hole algorithm, J. King Saud Univ, Comput. Inf. Sci. 34 (7) (2022) 4075–4088.
[16] U. Sopaoglu, O. Abul, Classification utility aware data stream anonymization, Appl. Soft Comput. 110 (2021) 107743.
[17] Q. Lan, B. Song, Y. Li, G. Li, Distributed differentially private ranking aggregation, IEEE Trans. Comput. Soc. Syst. 11 (1) (2022) 503–513.
[18] A. Kiran, N. Shirisha, K-anonymization approach for privacy preservation using data perturbation techniques in data mining, Mater. Today Proc. 64 (2022) 578–584.
[19] J. He, J. Du, N. Zhu, Research on k-anonymity algorithm for personalized quasi-identifier attributes, Netinfo Secur. 8 (2020) 19–26.
[20] M. Cunha, R. Mendes, J.P. Vilela, A survey of privacy-preserving mechanisms for heterogeneous data types, Comput. Sci. Rev. 41 (2021) 100403.
[21] J. Qian, M. Zheng, C. Zhou, C. Liu, Y. Xiaodong, Recent advancement in multi-granulation three-way decisions, J. Data Acquis. Proces. Shu Ju Cai Ji Yu Chu Li 39 (2) (2024) 361–375.
[22] L.A. Zadeh, Fuzzy sets and information granularity, in: Fuzzy Sets, Fuzzy Logic, and Fuzzy Systems: Selected Papers by Lotfi a Zadeh, World Scientific, 1996, pp. 433–448.
[23] W. Pedrycz, A. Skowron, V. Kreinovich, Handbook of Granular Computing, John Wiley & Sons, 2008.
[24] A. Bargiela, W. Pedrycz, Granular computing, in: Handbook on Computer Learning and Intelligence: Volume 2: Deep Learning, Intelligent Control and Evolutionary Computation, World Scientific, 2022, pp. 97–132.
[25] D. Liu, X. Yang, T. Li, Three-way decisions: beyond rough sets and granular computing, Int. J. Mach. Learn. Cybern. 11 (2020) 989–1002.
[26] X. Yang, Y. Zhang, H. Fujita, D. Liu, T. Li, Local temporal-spatial multi-granularity learning for sequential three-way granular computing, Inf. Sci. 541 (2020) 75–97.
[27] Y. Yao, Three-way granular computing, rough sets, and formal concept analysis, Int. J. Approx. Reason. 116 (2020) 106–125.
[28] C. Jiang, Y. Yao, Effectiveness measures in movement-based three-way decisions, Knowl.-Based Syst. 160 (2018) 136–143.
[29] X. Yue, Y. Chen, B. Yuan, Y. Lv, Three-way image classification with evidential deep convolutional neural networks, Cogn. Comput. 14 (6) (2022) 2074–2086.
[30] J. Qian, D. Wang, Y. Yu, X. Yang, S. Gao, E3wd: a three-way decision model based on ensemble learning, Inf. Sci. 667 (2024) 120487.
[31] X. Yang, T. Li, H. Fujita, D. Liu, Y. Yao, A unified model of sequential three-way decisions and multilevel incremental processing, Knowl.-Based Syst. 134 (2017) 172–188.
[32] Y. Yao, Three-way decision and granular computing, Int. J. Approx. Reason. 103 (2018) 107–123.
[33] Q. Zhang, G. Pang, G. Wang, A novel sequential three-way decisions model based on penalty function, Knowl.-Based Syst. 192 (2020) 105350.

[34] X. Yang, T. Li, D. Liu, H. Fujita, A temporal-spatial composite sequential approach of three-way granular computing, Inf. Sci. 486 (2019) 171–189.
[35] W. Qian, Y. Zhou, J. Qian, Y. Wang, Cost-sensitive sequential three-way decision for information system with fuzzy decision, Int. J. Approx. Reason. 149 (2022) 85–103.
[36] J. Qian, D. Tang, Y. Yu, X. Yang, S. Gao, Hierarchical sequential three-way decision model, Int. J. Approx. Reason. 140 (2022) 156–172.
[37] Q. Feng, D. Miao, Y. Cheng, Hierarchical decision rules mining, Expert Syst. Appl. 37 (3) (2010) 2081–2091.
[38] M. Ye, X. Wu, X. Hu, D. Hu, Anonymizing classification data using rough set theory, Knowl.-Based Syst. 43 (2013) 82–94.
[39] J. Wang, G. Cai, C. Liu, J. Wu, X. Li, A multi-level privacy-preserving approach to hierarchical data based on fuzzy set theory, Symmetry 10 (8) (2018) 333.
[40] W. Ali, M. Nauman, N. Azam, A privacy enhancing model for Internet of things using three-way decisions and differential privacy, Comput. Electr. Eng. 100 (2022) 107894.
[41] J. Qian, H. Jiang, Y. Yu, H. Wang, D. Miao, Multi-level personalized k-anonymity privacy-preserving model based on sequential three-way decisions, Expert Syst. Appl. 239 (2024) 122343.
[42] J. Qian, C. Hong, Y. Yu, C. Liu, D. Miao, Generalized multigranulation sequential three-way decision models for hierarchical classification, Inf. Sci. 616 (2022) 66–87.