

基于哈夫曼的 k -匿名模型隐私保护数据压缩方案

于玥^{1,2}, 林宪正³, 李卫海^{1,2}, 俞能海^{1,2}

(1. 中国科学技术大学网络空间安全学院, 安徽 合肥 230001; 2. 中国科学院电磁空间信息重点实验室, 安徽 合肥 230001;
3. 华为技术有限公司 2012 实验室, 香港 999077)

摘要: k -匿名模型作为常用的数据匿名技术, 广泛应用于数据发布阶段的隐私保护。随着大数据时代的快速发展, 海量数据的产生给数据存储带来了新的挑战。然而, 存储器的成本较高且存储空间有限, 通过硬件升级来无限制地扩充存储空间并不可行。为此, 使用数据压缩技术可以减少存储成本和通信开销。为减少数据发布阶段使用匿名技术产生的数据的存储空间, 提出了 k -匿名模型隐私保护数据压缩方案。对于 k -匿名模型的原始数据, 按照设定的规则及原始数据同匿名数据之间的预设泛化层次关系计算两者的差值, 并根据差值数据具有的频率特性对差值进行哈夫曼编码压缩。通过存储差值可以间接获得原始数据, 从而减少原始数据的存储空间。对于 k -匿名模型的匿名数据, 根据模型的泛化规则或预设泛化层次关系, 匿名数据通常具有较高的重复性, 且设定的 k 值越大, 匿名数据的泛化程度越高、重复性越强。对匿名数据设计实现哈夫曼编码压缩, 减少匿名数据的存储空间。实验结果表明, 所提方案能够显著降低 k -匿名模型的原始数据及匿名数据的压缩率。在使用的 5 个 k -匿名模型及不同 k 值的设定情况下, 与 Windows 11 的 zip 工具相比, 所提方案的原始数据压缩率和匿名数据压缩率平均降低了 72.2%、64.2%。

关键词: k -匿名模型; 隐私保护; 数据压缩存储; 哈夫曼编码

中图分类号: TP393

文献标志码: A

DOI: 10.11959/j.issn.2096-109x.2023054

Privacy-preserving data compression scheme for k -anonymity model based on Huffman coding

YU Yue^{1,2}, LIN Xianzheng³, LI Weihai^{1,2}, YU Nenghai^{1,2}

1. School of Cyber Technology and Science, University of Science and Technology of China, Hefei 230001, China

2. Key Laboratory of Electro-magnetic Space Information, Chinese Academy of Sciences, Hefei 230001, China

3. 2012 Labs, Huawei Technology Co. Ltd, Hong Kong 999077, China

Abstract: The k -anonymity model is widely used as a data anonymization technique for privacy protection during the data release phase. However, with the advent of the big data era, the generation of vast amounts of data poses

收稿日期: 2022-12-15; 修回日期: 2023-03-24

通信作者: 于玥, yysponge@mail.ustc.edu.cn

基金项目: 国家自然科学基金 (62071446)

Foundation Item: The National Natural Science Foundation of China (62071446)

引用格式: 于玥, 林宪正, 李卫海, 等. 基于哈夫曼的 k -匿名隐私保护数据压缩方案[J]. 网络与信息安全学报, 2023, 9(4): 64-73.

Citation Format: YU Y, LIN X Z, LI W H, et al. Privacy-preserving data compression scheme for k -anonymity model based on Huffman coding [J]. Chinese Journal of Network and Information Security, 2023, 9(4): 64-73.

challenges to data storage. However, it is not feasible to expand the storage space infinitely by hardware upgrade, since the cost of memory is high and the storage space is limited. For this reason, data compression techniques can reduce storage costs and communication overhead. In order to reduce the storage space of the data generated by using anonymization techniques in the data publishing phase, a compression scheme was proposed for the original data and anonymized data of the k -anonymity model. For the original data of the k -anonymity model, the difference between the original data and the anonymized data was calculated according to the set rules and the pre-defined generalization level. Huffman coding compression was applied to the difference data according to frequency characteristics. By storing the difference data, the original data can be obtained indirectly, thus reducing the storage space of the original data. For anonymized data of the k -anonymity model, the anonymized data usually have high repeatability according to the generalization rules of the model or the pre-defined generalization hierarchy relations. The larger the value of k , the more generalized and repeatable the anonymized data becomes. The design of Huffman coding compression was implemented for anonymous data to reduce storage space. The experimental results show that the proposed scheme can significantly reduce the original data and the anonymous data compression rate of the k -anonymity model. Across five models and various k -value settings, the proposed scheme reduces the compression rate of raw and anonymized data by 72.2% and 64.2% on average compared to the Windows 11 zip tool.

Keywords: k -anonymity model, privacy preservation, data compression storage, Huffman coding

0 引言

随着大数据时代的到来以及计算机计算性能的不不断提升, 互联网、教育、医疗等行业存储了海量数据, 数据的共享范围逐渐扩大。在数据发布的过程中, 如果数据拥有方对数据处理不当, 可能会泄露用户的个人隐私信息^[1]。数据匿名技术是重要的隐私保护技术之一, 其通过匿名化技术弱化敏感信息与个体之间的关联性, 并对原始数据中的敏感信息进行保护处理, 从而有效地降低数据发布过程中个人隐私信息披露的风险。 k -匿名模型是最常用的匿名保护数据发布技术, 其基本思想是通过泛化和隐匿原始数据集中的部分属性值形成满足 k -匿名要求的匿名数据集并用于数据发布, 保护隐私的同时在一定程度上保持数据的可用性^[2-3]。

基于 k -匿名的模型广泛应用于公共数据库或微数据发布阶段的隐私保护、位置隐私保护等领域。Wu 等^[4]提出了一个基于反馈的 k -匿名方案, 用于众包数据库的隐私保护。在该方案中, 合成样本被公布给人类工作者, 其结果被用来指导对匿名策略的选择。Kabir 等^[5]提出了一个基于微聚集方法的 k -匿名模型, 用于保护云计算中的微数据。该模型以特定的方式对数据集中的所有记录进行排序, 并提出一个微聚集方法来创建 k -匿名

聚类。Wang 等^[6]提出了一个基于客户端的个性化 k -匿名方案, 实现在网络系统中对自主车辆查询服务的隐私保护。Fei 等^[7]提出了一种基于 k -匿名原则的隐私保护的两级模式方案, 其降低了位置服务隐私保护的成本并提高了用户位置隐私的安全性。Xing 等^[8]提出了一种基于双 k -匿名的位置隐私保护方法, 该方法可以隐藏用户位置和请求信息, 并应用于社会化车联网。

为保证用户隐私数据的安全性, 使用匿名化技术对发布数据进行处理是必要的, 同时海量数据的产生给数据存储带来了压力。然而, 存储器的存储空间有限且成本较高, 无法通过硬件升级来无限扩充空间。在这种情况下, 可以使用数据压缩技术减少存储成本和通信开销^[9]。目前已有对结构化数据进行压缩的相关研究^[10-11], 但没有针对匿名数据及其原始数据进行压缩的研究。 k -匿名模型通常根据预设的泛化层级或一定的泛化规则将原始数据转化为匿名数据, 可以将两者间的关系具体表示为差值, 并根据差值的特性对其进行压缩存储。原始数据可由差值及匿名数据间接获得, 从而减少原始数据存储占用的空间。匿名数据相较于原始数据更具重复性, 因此可以利用匿名数据的重复性对其进行压缩。哈夫曼编码是一种经典的熵编码算法^[12], 它被广泛应用于数据压缩^[13]、图像处理^[14]、数据安全^[15-16]、音频压

缩^[17]等领域。哈夫曼编码主要依据字符的频率进行压缩,适用于 k -匿名模型的数据压缩。

本文的主要工作总结如下。

1) 针对 k -匿名模型的原始数据,提出一种压缩方案,将原始数据与匿名数据之间的关联性具体化为差值。使用哈夫曼编码对差值进行压缩,减少原始数据的存储空间。

2) 针对 k -匿名模型的匿名数据,提出一种压缩方案,由于泛化后的数据具有一定的重复性,且泛化层次越高重复性越强,使用哈夫曼进行压缩,减少匿名数据的存储空间。

3) 实验结果表明,提出的 k -匿名模型原始数据压缩方案和匿名数据压缩方案的压缩率相较于 Windows 11 的 zip 工具,平均降低了 72.2%、64.2%。

1 预备知识

1.1 k -匿名模型

k -匿名模型要求数据集中的每个记录至少与 $k-1$ 条记录无法区分,以抵御链接攻击。 k -匿名模型相关定义如下。

定义 1 原始数据表 T 。原始数据表 T 由显式标识符 EI、准标识符 QI、敏感属性 SA 及非敏感属性 NSA 组成,可以将 T 形式化地表示为 $T = (EI, QI, SA, NSA)$ 。

1) 显式标识符 EI

每条记录的显式标识符都可以唯一标识原始数据集 T 中的个体身份。在数据发布前必须被删除或隐匿。

2) 准标识符 QI

假定一个原始数据表 $T(A_1, A_2, \dots, A_n)$, 其中 A_1 至 A_n 代表组成数据表的属性, T 的准标识符 $QI = \{A_i, \dots, A_j\}$ 由一组属性组成, 其中 $\{A_i, \dots, A_j\} \subseteq \{A_1, A_2, \dots, A_n\}$, 则准标识符为能够以较高概率识别出目标个体的最小属性集合。

3) 敏感属性 SA

需要保护的敏感信息,一般无法预先获知,也无法通过敏感属性唯一确定一个用户记录。

4) 非敏感属性 NSA

非敏感属性即非敏感信息,单一的属性可作为非敏感属性,多个非敏感属性组合有可能成为

准标识符。

定义 2 匿名数据表 T' 。原始数据通过匿名化技术处理之后得到的数据表为匿名数据表 T' 。

定义 3 k -匿名。假定 $T(A_1, A_2, \dots, A_n)$ 是一个表, QI 是表 T 相关的准标识符, 当且仅当在 $T[QI]$ 中出现的每一个记录至少出现 k 次, 则称表 T 满足 k -匿名。

k -匿名模型的实现主要通过匿名化和模型算法。匿名化主要包括泛化和隐匿方法。泛化的思想是通过降低准标识符属性的精度,使得在准标识符属性上值相同的元组个数增加。准标识符属性分为数值型和分类型两种类型,对于数值型属性通常将数值泛化成对应的区间,对于分类型属性则泛化为相对应原属性值更一般的值来取代。泛化操作需要根据属性的泛化层次结构实现,泛化层次结构主要包括域泛化层次(DGH, domain generalization hierarchy)和值泛化层次(VGH, value generalization hierarchy)。图1为数值型属性 Age = {32, 33, 36, 37, 38} 的域泛化层次和值泛化层次。图2为分类型属性 Country = {China, India, Italy, Poland, Cuba, Canada} 的域泛化层次和值泛化层次。隐匿的思想是去除单元值或一个元组的所有值。

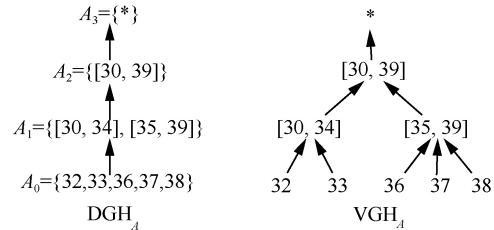


图1 属性 Age 的域泛化层次和值泛化层次
Figure 1 The generalization process of domain and value of age attribute

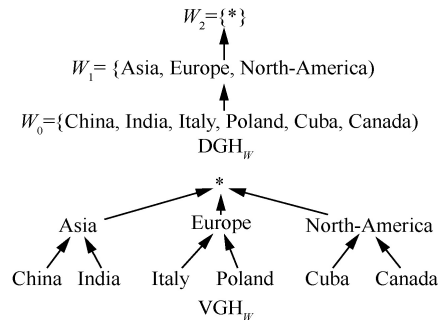


图2 Country 属性的域泛化层次和值泛化层次
Figure 2 The generalization process of domain and value of country attribute

k -匿名模型的泛化算法从泛化约束的角度可以分为全域泛化算法和局域泛化算法。经典的全域泛化算法包括 Datafly 算法^[18]、Incognito 算法^[19]等，主要的局域泛化算法包括 Top-down 算法^[20]、Mondrian^[21]、Basic Mondrian^[22]等。

1.2 哈夫曼编码

哈夫曼编码是一种可变长编码，通常用于无损数据压缩，其主要目的是为出现频率较高的字符分配长度较短的码字，从而最大程度地减少存储空间。哈夫曼的编码过程由两部分组成。首先，统计原始数据中各字符出现的频率，根据字符频率生成对应的哈夫曼树以及编码表。其次，扫描原始数据的每个字符并查询编码表，使用对应的码字替换字符。在解码的过程中，解码器读取压缩文件以确定编码表中对应的字符并进行替换。下面使用一个示例来解释编码过程。初始字符及其字符频率如图 3 所示。

BCAADDDCCACACAC				
字符	B	C	A	D
频率	1	6	5	3

图 3 初始字符及其字符频率

Figure 3 Initial string and its character frequency

按照字符出现的频率进行升序排列，并生成对应的哈夫曼树，如图 4 所示。生成哈夫曼树的算法如下，假设共有 n 个字符。

- 1) 创建 n 个节点并为每个字符分配一个节点，节点的权值字符的频率，将其作为有 n 棵树的森林；
- 2) 在森林中选择两个根节点的权值最小的树进行合并，作为一棵新树的左、右子树，且新树的根节点权值为其左、右子树根节点权值之和。
- 3) 从森林中删除选取的两棵树，并将新树加入森林。
- 4) 重复步骤 2)、步骤 3)，直到森林中只剩一棵树，该树即所求的哈夫曼树。

根据哈夫曼树可以获得对应的编码表。在图 4 中，左子树的前缀为 0，右子树的前缀为 1，由此可以获得每个叶节点字符的编码值。编码表由每个字符及其对应的码字组成，除此之外还需存

储每个码字的长度。表 1 是根据图 4 中的 Huffman 树获得的编码表。

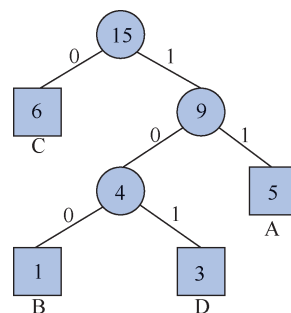


图 4 根据字符频率生成的哈夫曼树

Figure 4 Huffman tree generated based on symbol frequency

表 1 哈夫曼编码表
Table 1 Huffman coding table

字符	码字	码字长度
A	11	2
B	100	3
C	0	1
D	101	3

2 方案设计

本文针对 k -匿名模型原始数据与匿名数据之间存在的关联性，提出了 k -匿名模型的原始数据压缩方案。该方案计算原始数据同匿名数据之间的差值，由于差值不同的数据的频率相差较大，数据具有一定重复性，故使用哈夫曼编码对其进行压缩存储。通过存储差值可以间接获得原始数据，并减少原始数据的存储空间。对于匿名数据，随着模型的 k 值变大，其泛化层级也会变高，使得匿名化后的数据具有一定的重复性。针对这一特点，本文提出了 k -匿名模型的匿名数据压缩方案，通过哈夫曼进行压缩以减少其存储空间。由于 k -匿名模型的泛化操作对象为准标识符属性，故本文只讨论对于准标识符属性的压缩。

2.1 k -匿名模型的原始数据压缩

通过设定规则将原始数据与匿名数据之间的关系进行具体化表示，即计算其差值。根据差值的统计特性对其进行压缩存储。

给定一个原始数据表 T ， T 的准标识符为 $QI = \{D_1, \dots, D_n, C_1, \dots, C_m\}$ ， D_i 和 C_j 分别为数值型属性和分类型属性，其中 $1 \leq i \leq n$ ， $1 \leq j \leq m$ ，

n 和 m 分别为 QI 包含的数值型属性和分类型属性的个数。经过匿名化处理后得到的匿名数据表为 T' 。对于准标识符中的数值型属性 D_i ，假定其原始数据的值为 x ，在经过泛化处理 x 存在 3 种情况。

1) x 被泛化到对应泛化层次的区间 $[a, b]$ ，其中 $a \leq x \leq b$ 。

2) x 被泛化到最高泛化层次，最高泛化层次为字符 $*$ 。

3) x 的泛化层次为第一个层次，即泛化前后 x 的值不变。

针对数值型属性 D_i 计算其差值 σ ，若 x 被泛化到区间 $[a, b]$ ，则 $\sigma = x - a$ 。若 x 被泛化为 $*$ ，统计原始数据中 D_i 的不同数值出现的频率，按频率对数值进行降序排序，并建立关于 $*$ 的字典 $Z_{D_i(*)}$ 。假定 D_i 共有 d 个不同的数值，将排序后的数值设定为字典的键，字典的值依次分配为 0 至 d 。在这种情况下， σ 为字典 $Z_{D_i(*)}$ 中键 x 对应的值。若泛化前后 x 的值不变，则 $\sigma = 0$ 。

对于准标识符中的分类型属性 C_j ，假定其原始数据的值为 y ，在经过泛化处理 y 存在两种情况。

1) y 被泛化到对应泛化层级的值 y' ；

2) y 的泛化层次为第一个层次，即泛化前后 y 的值不变。

针对分类型属性 C_j 计算其差值 σ ，若 y 被泛化到对应泛化层级的值 y' ， y' 通常对应于原始数据表 T 中多个不同值。假定 y' 对应于 T 中 c 个不同的值 y_k ，其中 $0 < k \leq c$ 。根据原始数据表 T 与匿名数据表 T' 之间的泛化层级的信息，建立 y' 的字典 $Z_{C_j(y')}$ 。统计 y_k 的频率，按频率对其进行降序排序。将排序后的值设定为字典的键，字典的值依次分配为 0 至 c 。 σ 为字典 $Z_{C_j(y')}$ 中键 y_k 对应的值。若泛化前后 y 的值不变，则 $\sigma = 0$ 。

通过匿名数据表 T' 、差值 σ 、字典 $Z_{D_i(*)}$ 和 $Z_{C_j(y')}$ 可以计算得出原始数据表 T ，存储差值及字典可以减少存储原始数据表所占用的存储空间。由于在对 σ 进行设置的过程中，对泛化为 $*$ 的数值型以及泛化为 y' 的分类型的原始数据按频率进行排序，故 σ 为 0 等较小的值出现的频率较

高，其数据具有一定的重复性。统计 σ 不同值的个数和频率，并使用哈夫曼编码进行压缩可以有效减少其存储空间。对原始数据的压缩过程如算法 1 所示。

算法 1 原始数据压缩

输入 原始数据表 T ，匿名数据表 T' ，准标识符 $QI = \{D_1, \dots, D_n, C_1, \dots, C_m\}$

输出 Huffman 压缩结果及编码表 R_o ，字典 Z

1) $R_o \leftarrow 0$

2) for $D_i \in QI$ do

3) for $x \in D_i$ do

4) if x 被泛化为 $[a, b]$ then

5) $\sigma = x - a$

6) end if

7) if x 被泛化为 $*$ then

8) $\sigma = Z_{D_i(*)}.get(x)$

9) end if

10) if x 泛化前后不变 then

11) $\sigma = 0$

12) end if

13) end for

14) $R_o \leftarrow R_o + \text{Huffman}(\sigma)$

15) $Z \leftarrow Z + Z_{D_i(*)}$

16) end for

17) for $C_j \in QI$ do

18) for $y \in C_j$ do

19) if y 被泛化为 y' then

20) $\sigma = Z_{C_j(y')}.get(y)$

21) end if

22) if y 泛化前后不变 then

23) $\sigma = 0$

24) end if

25) end for

26) $R_o \leftarrow R_o + \text{Huffman}(\sigma)$

27) $Z \leftarrow Z + Z_{C_j(y')}$

28) end for

29) return R_o, Z

2.2 k -匿名模型的匿名数据压缩

在匿名化操作的过程中，匿名数据表 T' 由原始数据表 T 经过泛化算法运算获得，通常根据泛

化程度的需要对 k 值进行设定。由于泛化算法会根据预设的泛化层级将不同的数据泛化为同一数据, 泛化后 T' 中的数据会具有一定的重复性, 且 k 值设定的数值越大, T' 的泛化程度越高, 数据的重复性越大, 利用这一特点对匿名数据进行压缩。

给定一个匿名数据表 T' , T' 的准标识符为 $QI' = \{A_1, A_2, \dots, A_n\}$, 其中 n 为 QI' 包含的属性的个数。对于 T' 中的 QI' 属性数据, 其具有重复性, 适用于哈夫曼编码进行压缩。分别统计 QI' 中每个属性的不同值的个数和频率, 使用哈夫曼进行压缩。特别地, 当泛化程度非常高时, 泛化后的数据只具有一个值 (最高泛化层级 *), 则直接记录该值及其对应的数量。对匿名数据的压缩过程如算法 2 所示。

算法 2 匿名数据压缩

输入 匿名数据表 T' , 准标识符 $QI' = \{A_1, A_2, \dots, A_n\}$

输出 Huffman 压缩结果及编码表 R_a

- 1) $R_a \leftarrow 0$
- 2) for $A_i \in QI'$ do
- 3) if A_i 只包含一个值 then
- 4) 将该值及其对应的数量存储为 u
- 5) $R_a \leftarrow R_a + u$
- 6) end if
- 7) if A_i 包含多个值 then
- 8) $R_a \leftarrow R_a + \text{Huffman}(A_i)$
- 9) end if
- 10) end for
- 11) return R_a

3 实验结果与分析

3.1 数据集及实验环境

本文实验使用的数据集来源于 UC Irvine Machine Repository 的 Adult 数据集^[23], 该数据集是研究 k -匿名最常采用的数据源之一。通过对数据集进行完整性和适用性检测, 最终选用 Adult 数据集中的 adult.data 文件作为实验的数据样本, 其大小为 3.79 MB, 包含 15 个属性和 32 561 条数据。删除含有空值和非法值的元组后, 共有 30 162 条数据。选择数据集中的 9 个属性作为准标识符的属性集, 表 2 中描述了准标识符属性的数据结构。

表 2 数据结构示意
Table 2 Schematic of data structure

编号	属性	类别	不同值个数	可泛化层级
1	Age	数值型	72	5
2	Education	分类型	16	4
3	Marital-status	分类型	7	3
4	Native-country	分类型	41	3
5	Occupation	分类型	14	3
6	Race	分类型	5	2
7	Gender	分类型	2	2
8	Salary-class	分类型	2	2
9	Workclass	分类型	7	3

各属性的泛化层次定义如下。

Age, 泛化层次为 5 层。第 1 个层次: 原始数据; 第 2 个层次: 5 年间隔; 第 3 个层次: 10 年间隔; 第 4 个层次: 20 年间隔; 第 5 个层次: *。

Education, 泛化层次为 4 层。第 1 个层次: 原始数据; 第 2 个层次: Undergraduate、Graduate、High School、Primary School、Professional Education; 第 3 个层次: Higher Education、Primary Education、Secondary Education; 第 4 个层次: *。

Marital-status, 泛化层次为 3 层。第 1 个层次: 原始数据; 第 2 个层次: spouse present、spouse not present; 第 3 个层次: *。

Native-country, 泛化层次为 3 层。第 1 个层次: 原始数据; 第 2 个层次: North America、Asia、Europe、Africa、South America; 第 3 个层次: *。

Occupation, 泛化层次为 3 层。第 1 个层次: 原始数据; 第 2 个层次: Technical、Non-technical、Other; 第 3 个层次: *。

Race、Gender、Salary-class, 泛化层次均为 2 层。第 1 个层次: 原始数据; 第 2 个层次: *。

Workclass, 泛化层次为 3 层。第 1 个层次: 原始数据; 第 2 个层次: Government、Non-Government、Unemployed; 第 3 个层次: *。

对于模拟实验, 使用的操作系统为 Windows 11; 硬件环境为 11th Gen Intel Core i5-1135G7, 4 核 2.4 GHz 主频, 内存为 16 GB; 编译环境 Python 3.8.6。

3.2 原始数据压缩实验

为验证所提出的 k -匿名模型的原始数据压缩

方案的有效性, 本文采用 5 个经典的 k -匿名模型进行对比实验, 包括两个全域泛化算法模型 Incognito^[19]、Datafly^[18], 以及 3 个局域泛化算法模型 Top-down^[20]、Mondrian^[21]、Basic Mondrian^[22]。实验设置模型的 k 值分别为 5、10、50、100, 对比本文方法以及 Windows 11 的 zip 工具的压缩效果。

全域泛化模型原始数据的压缩率对比如图 5 所示, 图中 Huf 代表本文方案。对于 Incognito 模型, 当 k 值分别为 5、10、50、100 时, 本文方案的压缩率相较于 Windows 11 的 zip 压缩率分别降低了 76.7%、74.7%、70.3%、68.6%; 对于 Datafly 模型, 其压缩率分别降低了 70.9%、68.7%、68.7%、68.7%。从图中可以看出, 当 k 值增大时压缩率存在小幅上涨, 这是由于泛化层次变高, 导致匿名数据的信息损失度增大, 在计算匿名数据同原始数据的差值时字典变大。相较于低 k 值, 高 k 值的数据频率有所降低。特别地, 对于 Datafly 模型, 由于其模型本身对数据的泛化程度较高、信息损失度较大。当 k 值为 10 时, Datafly 模型中部分准标识符属性已达到最大泛化层级, 故继续增大 k 值其压缩率保持不变。

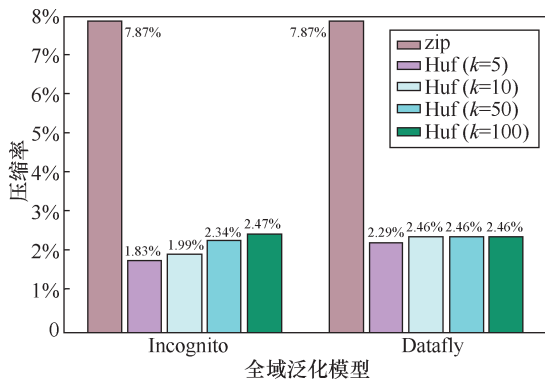


图 5 全域泛化模型原始数据的压缩率对比
Figure 5 Comparison of raw data compression rates of full-domain generalization models

局域泛化模型原始数据的压缩率对比如图 6 所示, 图中 Huf 代表本文方案。对于 Top-down 模型, 当 k 值分别为 5、10、50、100 时, 本文方案的压缩率相较于 Windows 11 的 zip 压缩率分别降低了 75.3%、72.4%、67.1%、65.4%; 对于 Mondrian 模型, 其压缩率分别降低了 78.7%、77.4%、71.9%、71.5%; 对于 Basic Mondrian 模

型, 其压缩率分别降低了 76.4%、74.2%、72.7%、72.3%。

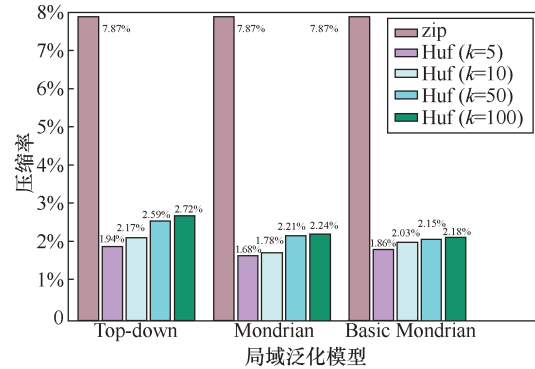


图 6 局域泛化模型原始数据压缩率对比
Figure 6 Comparison of raw data compression rates for local generalization models

对于不同 k 值的全域泛化模型 Incognito、Datafly, 本文方案的压缩率平均降低了 72.6%、69.3%。对于局域泛化模型 Top-down、Mondrian、Basic Mondrian, 本文方案的压缩率平均降低了 70.1%、74.9%、73.9%, 证明本文提出的针对 k -匿名模型的原始数据压缩方案优于 Windows 11 的 zip 压缩工具。

表 3 罗列了 5 个模型及其最大泛化程度对应的 k 值, 即当表 2 中的 9 个准标识符属性均达到最大泛化层次时, 5 个模型对应的 k 值。其中对于 Datafly 模型, $k=10$ 时除 Gender 属性和 Education 属性外其余准标识符属性均达到最大泛化层次, 继续增大 k 值会使用隐匿方法去除不满足 k -匿名条件的记录, 导致数据表失真严重从而使数据表失去可用性, 故 Datafly 模型的 k 值取 10。

表 3 模型及其最大泛化程度对应的 k 值
Table 3 The model and its maximum generalization degree corresponding to the value of k

模型	k 值
Incognito	1 800
Datafly	10
Top-down	2 700
Mondrian	14 100
Basic Mondrian	15 000

Windows 11 的 zip 压缩工具对原始数据的压缩率为 7.87%, 当全部 9 个准标识符属性均达到最大泛化层次时, 本文提出的原始数据压缩方案

的压缩率为 3.46%，即无论 k 值如何增大，压缩率最大为 3.46%，相较于 Windows 11 的 zip 压缩工具，本文方案仍有更优的压缩率。需要注意的是，表 3 列出的 k 值是为了说明达到最大泛化程度时模型的 k 值大小，而实际应用中通常不会将 k 值设置过大，这是因为 k 值过大会导致数据失去可用性。

表 4 列出了 k 值为 5、10、50 和 100 时，本文提出的原始数据压缩方案和 Windows 11 的 zip 压缩工具的压缩时间对比，从表中可以看出，本文所提的压缩方案需要花费的时间更少、开销更小。

表 4 原始数据压缩时间对比
Table 4 Original data compression time comparison

模型	压缩时间/ms			
	$k=5$	$k=10$	$k=50$	$k=100$
Incognito	33.4	33.6	34.9	35.4
Datafly	34.5	35.2	35.2	35.2
Top-down	33.6	34.4	35.7	36.4
Mondrian	32.9	33.2	34.5	34.7
Basic_Mondrian	33.5	34.1	34.3	34.4
zip	43.5	43.5	43.5	43.5

3.3 匿名数据压缩实验

匿名数据压缩实验的模型设置同 3.2 节原始数据压缩实验。表 5 列出了全域泛化模型匿名数据压缩率对比（即相较于 zip，本文方案的压缩率降低的百分比）。从表中可知，对于不同 k 值的 Incognito 及 Datafly 模型，本文方案的压缩率相较于 zip 平均降低了 60.6%、72%。

表 5 全域泛化模型匿名数据压缩率对比
Table 5 Comparison of anonymized data compression rates for the full domain generalization models

模型及 k 值设置	zip 压缩率	本文方案压缩率	压缩率对比
Incognito ($k=5$)	7.77%	2.87%	63.1%
Incognito ($k=10$)	7.54%	2.82%	62.6%
Incognito ($k=50$)	6.75%	2.74%	59.4%
Incognito ($k=100$)	6.08%	2.59%	57.4%
Datafly ($k=5$)	3.42%	1.06%	69%
Datafly ($k=10$)	3.18%	0.86%	73%
Datafly ($k=50$)	3.18%	0.86%	73%
Datafly ($k=100$)	3.18%	0.86%	73%

表 6 列出了局域泛化模型匿名数据的压缩率

对比（即相较于 zip，本文方案的压缩率降低的百分比）。从表中可知，对于不同 k 值的 Top-down、Mondrian、Basic Mondrian 模型，本文所提方案的压缩率相较于 zip 平均降低了 61.5%、68.6%、58.2%，证明本文提出的针对 k -匿名模型的匿名数据压缩方案优于 Windows 11 的 zip 压缩工具。

表 6 局域泛化模型匿名数据压缩率对比
Table 6 Comparison of anonymized data compression rates for the local generalization models

模型及 k 值设置	zip 压缩率	本文方案压缩率	压缩率对比
Top-down ($k=5$)	7.70%	2.91%	62.2%
Top-down ($k=10$)	7.51%	2.84%	62.2%
Top-down ($k=50$)	6.94%	2.67%	61.5%
Top-down ($k=100$)	6.66%	2.65%	60.2%
Mondrian ($k=5$)	7.26%	2.58%	64.5%
Mondrian ($k=10$)	6.46%	2.19%	66.1%
Mondrian ($k=50$)	4.34%	1.28%	70.5%
Mondrian ($k=100$)	3.36%	0.90%	73.2%
Basic Mondrian ($k=5$)	7.87%	2.96%	62.4%
Basic Mondrian ($k=10$)	7.57%	2.92%	61.4%
Basic Mondrian ($k=50$)	6.10%	2.81%	53.9%
Basic Mondrian ($k=100$)	4.95%	2.22%	55.2%

表 7 和表 8 分别列出了全域泛化模型、局域泛化模型的匿名数据压缩时间对比。从表中可以看出，对于不同算法及不同 k 值设置，本文提出的匿名数据压缩方案的压缩时间均少于 Windows 11 的 zip 压缩工具，需要的计算开销更小。

表 7 全域泛化模型的匿名数据压缩时间对比
Table 7 Comparison of anonymized data compression time for the full domain generalization models

模型及 k 值设置	压缩时间/ms	
	zip	本文方案
Incognito ($k=5$)	43.6	28.7
Incognito ($k=10$)	42.9	27.5
Incognito ($k=50$)	42.1	26.9
Incognito ($k=100$)	41.5	26.2
Datafly ($k=5$)	32.5	18.2
Datafly ($k=10$)	31.4	16.6
Datafly ($k=50$)	31.4	16.6
Datafly ($k=100$)	31.4	16.6

表8 局域泛化模型的匿名数据压缩时间对比
Table 8 Comparison of anonymized data compression time for the local generalization models

模型及 k 值设置	压缩时间/ms	
	zip	本文方案
Top-down ($k=5$)	41.5	27.4
Top-down ($k=10$)	40.8	26.8
Top-down ($k=50$)	39.4	25.7
Top-down ($k=100$)	38.9	25.5
Mondrian ($k=5$)	40.8	24.7
Mondrian ($k=10$)	38.4	19.3
Mondrian ($k=50$)	34.5	17.8
Mondrian ($k=100$)	32.3	17.5
Basic Mondrian ($k=5$)	42.5	28.1
Basic Mondrian ($k=10$)	41.1	27.6
Basic Mondrian ($k=50$)	37.9	26.3
Basic Mondrian ($k=100$)	35.8	22.7

4 讨论

现阶段还没有针对 k -匿名脱敏前后数据压缩方案的设计, 基于数据发布阶段 k -匿名技术的广泛使用及其产生的大量数据表, 本文针对 k -匿名隐私保护数据进行压缩方案的设计。方案设计主要基于 k -匿名模型本身具有的泛化层级及任意一个记录都至少有 k 个记录与之相同的特点, 结合现有的熵编码算法(即 Huffman 编码)进行压缩。

对于原始数据的压缩, 本文方案根据泛化层级计算原始数据表和匿名数据表之间的差值, 其中差值占用的存储空间少于原始数据表。故本文所提原始数据压缩方案相较使用现有压缩方法对原始数据直接进行压缩, 其压缩率有所下降。根据本文设定的计算差值的规则, 差值具有一定的概率分布特性, 适用于 Huffman 编码, 故使用 Huffman 编码进行压缩可以进一步降低压缩率。对于匿名数据的压缩, 满足 k -匿名的数据表包含的任意一个记录都有至少 k 个记录与之相同, 即脱敏后的数据具有大量的重复, 故使用 Huffman 编码对每个准标识符属性的不同属性值进行压缩相较使用现有压缩方法对匿名数据直接进行压缩会有更好的效果。实验结果证明, 本文所提压缩方案优于 Windows 11 的 zip 工具。

本文的压缩方案基于数据的统计特性, 而费

诺编码、香农编码、Huffman 编码均考虑了数据的统计特性, 即为出现频率高的字符分配较短码字, 为出现频率低的字符分配较长码字。然而, 费诺编码比较适合于对分组概率相等或接近的信源编码, 只有在数据的概率分布呈现特定形式的条件下才能达到最佳码的性能; 香农编码有系统的、唯一的编码方法, 但在很多情况下编码效率不是很高, 且使用香农编码方法的编码结果其平均码长不是最短的, 即不是紧致码。相较于费诺编码和香农编码, Huffman 编码对信源的统计特性没有特殊要求, 编码效率比较高, 对编码设备的要求也比较简单, 因此, Huffman 编码的综合性能优于费诺编码和香农编码, 故本文使用 Huffman 编码进行压缩。

5 结束语

本文分别针对 k -匿名模型的原始数据及匿名数据提出压缩方案, 利用原始数据及匿名数据之间的关系计算其差值, 并根据差值的统计特性进行哈夫曼压缩, 减少原始数据存储空间。对于匿名数据, 由于泛化后的数据具有高度的重复性, 利用这一特点设计实现针对匿名数据的哈夫曼压缩方案。实验结果表明, 与 Windows 11 的 zip 工具相比, 本文提出的原始数据压缩方案的压缩率约降低了 72.2%, 匿名数据压缩方案的压缩率约降低了 64.2%, 且本文所提方案对于 k -匿名模型具有普适性。未来可针对大数据应用场景下的其他隐私保护技术产生的数据的特点, 研究其压缩方案。

参考文献:

- [1] 方滨兴, 贾焰, 李爱平, 等. 大数据隐私保护技术综述[J]. 大数据, 2016, 2(1): 1-18.
FANG B X, JIA Y, LI A P, et al. An overview of big data privacy protection technologies[J]. Big Data, 2016, 2(1): 1-18.
- [2] SAMARATI P, SWEENEY L. Protecting privacy when disclosing information: k -anonymity and its enforcement through generalization and suppression[C]//Proc of the IEEE Symposium on Research in Security and Privacy. 1998: 384-393
- [3] SWEENEY L. K -anonymity: a model for protecting privacy[J]. International Journal of Uncertainty Fuzziness and Knowledge Based Systems, 2002, 10(5): 557-570.
- [4] WU S, WANG X L, WANG S, et al. K -anonymity for crowdsourcing database[J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(9): 2207-2221.
- [5] KABIR H, MAHMOOD A, WANG H, et al. Microaggregation

- sorting framework for k -anonymity statistical disclosure control in cloud computing[J]. IEEE Transactions on Cloud Computing, 2015, 8(2): 408-417.
- [6] WANG J B, CAI Z P, YU J G. Achieving personalized k -anonymity-based content privacy for autonomous vehicles in CPS[J]. IEEE Transactions on Industrial Informatics, 2020, 16(6): 4242-4251.
- [7] FEI F, LI S, DAI H P, et al. A k -anonymity based schema for location privacy preservation[J]. IEEE Transactions on Sustainable Computing, 2019, 4(2): 156-167.
- [8] XING L, JIA X F, GAO J P, et al. A location privacy protection algorithm based on double k -anonymity in the social internet of vehicles[J]. IEEE Communications Letters, 2021, 25(10): 3199-3203.
- [9] JAYASANKAR U, THIRUMAL V, PONNURANGAM D. A survey on data compression techniques: from the perspective of data quality, coding schemes, data type and applications[J]. Journal of King Saud University-Computer and Information Sciences, 2021, 33(2): 119-140.
- [10] LAI H, XU M, XU J, et al. Evaluating data storage structures of MapReduce[C]//2013 8th International Conference on Computer Science & Education. 2013: 1041-1045.
- [11] DING X, TIAN B, LI Y. A scheme of structured data compression and query on Hadoop platform[C]//2015 Third International Conference on Digital Information, Networking, and Wireless Communications (DINWC). 2015: 160-164.
- [12] HUFFMAN D A. A method for the construction of minimum-redundancy codes[J]. Proceedings of the IRE, 1952, 40(9): 1098-1101.
- [13] PAI Y T, CHENG F C, LU S P, et al. Sub-trees modification of Huffman coding for stuffing bits reduction and efficient NRZI data transmission[J]. IEEE Transactions on Broadcasting, 2012, 58(2): 221-227.
- [14] ALZAHIR S, BORICI A. An innovative lossless compression method for discrete-color images[J]. IEEE Transactions on Image Processing, 2014, 24(1): 44-56.
- [15] YIN Z, XIANG Y, ZHANG X. Reversible data hiding in encrypted images based on multi-MSB prediction and Huffman coding[J]. IEEE Transactions on Multimedia, 2019, 22(4): 874-884.
- [16] YI X, YANG K, ZHAO X, et al. AHCM: adaptive Huffman code mapping for audio steganography based on psychoacoustic model[J]. IEEE Transactions on Information Forensics and Security, 2019, 14(8): 2217-2231.
- [17] FIRMANSAL L, SETIAWAN E B. Data audio compression lossless FLAC format to lossy audio MP3 format with Huffman shift coding algorithm[C]//2016 4th International Conference on Information and Communication Technology (ICoICT). 2016: 1-5.
- [18] SWEENEY L. Computational disclosure control: primer on data privacy protection[D]. Massachusetts Institute of Technology, 2001: 67-82.
- [19] LEFEVRE K, DEWITT D J, RAMAKRISHNAN R. Incognito: efficient full-domain k -anonymity[C]//Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data. 2005: 49-60.
- [20] FUNG B C M, WANG K, Y P S. Top-down specialization for information and privacy preservation[C]//21st International Conference on Data Engineering (ICDE'05). 2005: 205-216.
- [21] LEFEVRE K, DEWITT D J, RAMAKRISHNAN R. Mondrian multidimensional k -anonymity[C]//22nd International Conference on Data Engineering (ICDE'06). 2006: 25-35.
- [22] LEFEVRE K, DEWITT D J, RAMAKRISHNAN R. Workload-aware anonymization[C]//Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2006: 277-286.
- [23] ASUNCION A, NEWMAN D. UCI machine learning repository[R]. 2007.

[作者简介]



于玥(1998-),女,黑龙江哈尔滨人,中国科学技术大学硕士生,主要研究方向为数据压缩编码及其安全应用。



林宪正(1981-),男,台湾人,主要研究方向为信道编码算法、代数算法。



李卫海(1975-),男,辽宁大连人,中国科学技术大学副教授,主要研究方向为多媒体内容安全、数据安全。



俞能海(1964-),男,安徽无为,人,中国科学技术大学教授、博士生导师,主要研究方向为视觉计算与人工智能安全、信息隐藏与数据安全、数字媒体内容安全。