基于敏感分级信息熵的匿名方法

石昆正 张攀峰* 董明刚
(桂林理工大学信息科学与工程学院 广西 桂林 541006)

摘 要 针对相似攻击所造成隐私泄露的问题,提出(H,p,k)-匿名模型,通过对敏感属性分级,使等价类中元组不同敏感级别的个数满足设定阈值 H,并设计满足该模型的匿名算法 MAA-SLIE(Micro-aggregation Algorithm based on Sensitive Level Information Entropy)。该算法基于贪心聚类思想,在聚类过程中保证等价类隐私安全指数最大,提高等价类中敏感属性多样性,降低隐私泄露风险,减少信息损失,通过实验验证了算法的合理性和有效性。

关键词 数据匿名 信息熵 微聚集 隐私保护

中图分类号 TP309.2 **文献标志码** A DOI:10.3969/j.issn.1000-386x.2024.05.046

DATA ANONYMITY METHOD BASED ON SENSITIVE HIERARCHICAL INFORMATION ENTROPY

Shi Kunzheng Zhang Panfeng* Dong Minggang
(*School of Information Science and Engineering*, *Guilin University of Technology*, *Guilin 541006*, *Guangxi*, *China*)

Abstract Aiming at the problem of privacy leakages caused by similar attacks, this paper proposes (H,p,k)-anonymous model. By classifying sensitive attributes, the number of tuples with different sensitive level in equivalent classes could meet the set threshold H. An anonymous algorithm MAA-SLIE (micro-aggregation algorithm based on sensitive level information entropy) was designed to satisfy the model. Based on the greedy clustering idea, the algorithm ensured the maximum privacy security index of the equivalence class in the clustering process, improved the diversity of sensitive attributes in the equivalence class, and reduced the risk of privacy leakage and information loss. The rationality and effectiveness of the algorithm were verified through experiments.

Keywords Data anonymity Information entropy Micro-aggregation Privacy protection

0 引 言

在大数据时代背景下,大量的个人数据被收集和发布,例如消费数据、住院数据等。对这些数据进行挖掘与分析,可以促进科学和商业的发展,但数据中包含了大量的个人隐私,会有隐私信息泄露的风险。数据匿名技术可以有效降低个人隐私信息泄露的风险,同时保证数据的可用性。

文献[1]提出了 k-匿名模型,但该模型容易受到同质攻击和背景知识攻击[2]。p-敏感 k-匿名模型[3]和 l-多样性[2]虽然可以避免同质攻击,但会受到相似攻击。t-近邻[4]在一定程度上解决了相似攻击问题,但在发布数据较小的情况下会造成较大的信息损失。文献[5]提出了($\lambda\alpha$,k)-分级匿名模型,根据隐私保护需求的程度不同,将敏感属性划分等级,同时提出一种新层次聚类算法,能够满足不同程序的隐私保护需求,同时有效地减少信息损失。根据敏感属性的分类重要程度受不同准标识属性的影响,文献[6]提出一种基于权重属性熵的分类匿名方法,根据准标识属性权重属性熵的不同,对分类树的划分优化。每种属性对个人隐私影响不同,文献[7]提出一种算法,量化数据中不

收稿日期:2021-01-13。国家自然科学基金项目(61862019);广西自然科学基金项目(2017GXNSFAA198223);广西科技基地和人才专项(2018AD19136);桂林理工大学科研启动基金项目(GLUTQD2017065)。**石昆正**,硕士生,主研领域:数据挖掘,信息安全。**张攀峰**,博士。**董明刚**,博士。

同属性的敏感性,有效保护了个人隐私。文献[8]提出一种贪心聚类匿名方法,有效减少信息损失和运行时间。文献[9]在应对相似攻击问题时,提出了(r,k)-匿名模型并设计了相关算法,降低等价类中敏感属性语义相邻的频率,保证了较高的数据可用性。

在实现数据匿名方法中泛化和隐匿[10]较为常用,但这些方法存在的不足之处有数据可用性差、算法执行时间长等。而微聚集算法[11]相对于泛化和隐匿更简单有效,微聚集算法用等价类的中心值替换元组准标识属性上各个属性的值,很好地解决了语义丢失较多的问题,微聚集算法最开始是用来处理连续数值型数据[11],之后扩展到处理离散型数据[12-13]。文献[14]提出 V-GRAV 算法,改善了欧氏距离受奇异值影响的问题,提高了数据的安全性。文献[15]提出一种微聚集算法,并提出了匿名保护指数,降低隐私泄露风险,但是没有考虑抵抗相似攻击。本文针对匿名模型中,等价类存在高敏感属性值偏多和敏感属性值分布不均的而导致的个体隐私泄露问题,提出(H,p,k)-匿名模型,并设计匿名方法 MAA-SLIE。

1 相关概念

为便于描述,给定一个待发布数据集 $D=(t_1,t_2,\cdots,t_n)$,$t_i(i=1,2,\cdots,n)$为数据集中第 i 个元组。设 V 为数据集 D 上的一个属性类,$t_i[V]$为元组 t_i在属性 V 上的取值。例如,表 1 为某医院病人信息记录表原始数据集。标识属性:可以通过该类别属性直接识别出个体,例如,姓名、身份证号等,在表 1 中被移除。准标识属性:可以通过该类别属性与其他渠道获取的信息或者背景知识进行连接,从而识别出个体的属性,记为 QI,$QI=(QI_1,QI_2,\cdots,QI_d)$,例如表 1 中的 Age 和 ZipCode。敏感属性:数据发布者和个体不愿透露的属性,记为 SA。例如表 1 中的 Disease。

表 1 原始数据表

ID	Age	ZipCode	Disease
t_1	21	114235	HIV
t_2	26	114751	HIV
t_3	25	115032	Flu
t_4	48	124151	Diabetes
t_5	32	115012	Cancer
t_6	45	115451	Fever
t_7	35	115313	Hepatitis
t_8	50	113726	Flu

1.1 等价类

等价类 EC(Equivalence-Class)为数据集中若干个元组的集合,且每个元组在准标识属性上的取值相同。

1.2 k-匿名模型

在发布数据集 D^* 中,有 m 个等价类,每个等价类至少有个元组在准标识属性上取值不可区分。即 $D^*=(EC_1,EC_2,\cdots,EC_m)$,$\forall EC_w\in D^*$,$|EC_w|\geqslant k$,$\forall t_i,t_j\in EC_w$,有 $t_i[QI]=t_j[QI]$,其中 $i\neq j,w\in[1,m]$。

在 k-匿名模型的数据集中,识别出个体的概率至多为$\frac{1}{k}$,可以避免个体被识别,但 k-匿名模型不能很好地抵抗同质攻击。

1.3 p-敏感 k-匿名模型

在发布数据集 D^* 中,有 m 个等价类,每个等价类至少有 k 个元组在准标识属性上取值不可区分,且有 $p(p\leqslant k)$个元组在敏感属性上取值不同。即 $D^*=(EC_1,EC_2,\cdots,EC_m)$,$\forall EC_w\in D^*$,$|EC_w|\geqslant k$,$\forall t_i,t_j\in EC_w$,有 $t_i[QI]=t_j[QI]$,且$|t_i[SA]\neq t_j[SA]|\geqslant p$,其中 $i\neq j,w\in[1,m],p\leqslant k$。其中$|t_i[SA]\neq t_j[SA]|$为元组 i 和元组 j 的敏感属性不相同个数。

相比于 k-匿名模型,满足 p-敏感 k-匿名模型的数据集可以避免同质攻击,但会受到相似攻击。例如表 2 为满足 2-敏感 2-匿名的数据集中,通过背景知识即使获得某位病人的年龄为 28 岁,邮编为 114235,也不能确定该病人是患有 HIV 还是 Cancer。但该等价类中的敏感属性为 Cancer 和 HIV,故可得知病人患有严重的疾病,也会造成隐私泄露。

表 2 2-敏感 2-匿名模型数据表

EC	Age	ZipCode	Disease
1	[26~32]	11****	HIV
	[26~32]	11****	Cancer
2	[21~25]	11****	HIV
	[21~25]	11****	Flu
3	[48~50]	1*****	Diabetes
	[48~50]	1*****	Flu
4	[35~45]	115***	Hepatitis
	[35~45]	115***	Fever

1.4 敏感级别

将敏感属性取值按敏感程度可以划分为不同敏感级别,例如“疾病”作为敏感属性,其取值可以分为

不同敏感级别,高敏感级别有“Cancer”“HIV”,中敏感级别有“Diabetes”“Hepatitis”,低敏感级别有“Flu”“Fever”。

定义 1(敏感级别信息熵)　设某个等价类中包含 n 个元组,这 n 个元组中有 $u(u\leqslant n)$ 个不同的敏感属性,其敏感级别可分为 $l_1,l_2,\cdots,l_s(s\leqslant u)$。$p_{l_i}$表示 l_i敏感级别的所有元组在等价类 n 个元组中的占比,则定义敏感属性 l_i的敏感级别信息熵 $H(l_i)$为:

$$H(l_i)=-p_{l_i}\mathrm{lb}p_{l_i} \tag{1}$$

定义 2(等价类敏感级别信息熵)　设 $H(l_i)$为等价类 EC 中敏感级别为 l_i的敏感级别信息熵,则定义等价类敏感级别信息熵 $HEC(EC)$为:

$$HEC(EC)=\sum_{i=1}^{s}w_i\cdot H(l_i),w_i\in(0,1) \tag{2}$$

式中:w_i为 l_i敏感级别的权重,根据不同敏感级别取不同值。

1.5　属性距离

为降低数据匿名的信息损失,不同类型属性之间的距离采用不同的计算方法,本文将准标识属性分为连续数值型属性、等级型属性、分类型属性和布尔型属性。

1.5.1　连续数值型属性

在数据集 D 中有不同的连续数值型属性,例如“年龄”和“考试成绩”的取值区间和计量单位都有所不同,这些差异会影响元组间距离的计算,可通过归一化来减小这种影响:

$$t[\dot{QI}_n]_i=\frac{t[QI_n]_i-X(QI_n)_{\min}}{X(QI_n)_{\max}-X(QI_n)_{\min}} \tag{3}$$

式中:$X(QI_n)_{\min}$和 $X(QI_n)_{\max}$分别为连续数值型属性 QI_n值域区间的最小值和最大值。$t[QI_n]_i$为元组 t_i在属性 QI_n上的取值,$t[\dot{QI}_n]_i$是 $t[QI_n]_i$归一化后的值。

设元组 t_i和元组 t_j在连续数值型属性 QI_n上的取值经过归一化之后分别为 $t[\dot{QI}_n]_i$和 $t[\dot{QI}_n]_j$,则 t_i和 t_j在连续属性 QI_n上的距离为:

$$d_l(t_i,t_j)=|(t[\dot{QI}_n]_i-t[\dot{QI}_n]_j)| \tag{4}$$

设 EC 为数据集 D 的一个等价类,则等价类中所有元组在连续数值型属性 QI_n上的中心表示为:

$$C(EC)_n=\frac{1}{n}\sum_{i=1}^{n}t[QI_n]_i \tag{5}$$

式中:n 表示等价类 EC 中元组的个数。

1.5.2　等级型属性

在准标识属性 QI 取值的区间内有一定的等级差的关系或者有前后顺序关系的属性称为等级型属性 QI_g。例如“学历”的取值为“小学”“初中”“高中”“大专”“本科”。其中“小学”与“初中”之间的距离比“小学”与“本科”之间的距离小。为度量等级型属性取值之间的距离对其从高阶级到低阶级进行编号,则元组 t_i和元组 t_j在等级型属性 QI_g上的距离为:

$$d_g(t_i,t_j)=\frac{|t[QI_g]_i-t[QI_g]_j|}{|B(QI_g)|} \tag{6}$$

式中:$|t[QI_g]_i-t[QI_g]_j|$为元组 t_i和元组 t_j在等级型属性上两个编号值之差的绝对值,$|B(QI_g)|$为等级型属性 QI_g上按顺序排列最大的编号值与最小的编号值之差的绝对值,例如“小学”与“大专”之间的距离 $d_g(\text{小学},\text{大专})=\frac{|0-3|}{|4|}$。

定义 3(中心均值频率数)　求得一组数的算术平均值 $\overline{K}$,以 $\overline{K}$ 为中点计算取值区间 $\overline{K}$ 两侧值频率 $\overleftarrow{K}_g$和 $\overrightarrow{K}_g$,比较 $\overleftarrow{K}_g$和 $\overrightarrow{K}_g$大小,若 $\overleftarrow{K}_g<\overrightarrow{K}_g$,则 $C(EC)_g=\lceil\overline{K}\rceil$;若 $\overleftarrow{K}_g>\overrightarrow{K}_g$,则 $C(EC)_g=\lfloor\overline{K}\rfloor$;若 $\overleftarrow{K}_g=\overrightarrow{K}_g$,则 $C(EC)_g=\mathrm{round}(\overline{K})$。本文使用中心均值频率数定义等级型属性的中心,既能较多地保留了等级型属性的语义,又能减少聚类带来的信息损失。

$$C(EC)_g=\begin{cases}\lceil\overline{K}\rceil & \overleftarrow{K}_g<\overrightarrow{K}_g\\ \lfloor\overline{K}\rfloor & \overleftarrow{K}_g>\overrightarrow{K}_g\\ \mathrm{round}(\overline{K}) & \overleftarrow{K}_g=\overrightarrow{K}_g\end{cases} \tag{7}$$

$$\overline{K}=\frac{t[QI_g]_1+t[QI_g]_2+\cdots+t[QI_g]_n}{|EC|} \tag{8}$$

$$\overleftarrow{K}_g=\frac{\sum_{i=1}^{n}\mathrm{down\_count}}{|EC|} \tag{9}$$

$$\overrightarrow{K}_g=\frac{\sum_{i=1}^{n}\mathrm{up\_count}}{|EC|} \tag{10}$$

式中:$t[QI_g]$为元组在等级型属性 QI_g上的取值,$|EC|$为等价类中元组的个数。式(8)计算等价类中元组在等级型属性上的均值。式(9)和式(10)分别计算等价类中元组取值小于均值和大于均值的比率。

$$\mathrm{down\_count}=\begin{cases}1 & t[QI_g]_i<\overline{K}\\ 0 & \text{其他}\end{cases}\quad i=1,2,\cdots,n \tag{11}$$

$$\mathrm{up\_count}=\begin{cases}1 & t[QI_g]_i>\overline{K}\\ 0 & \text{其他}\end{cases}\quad i=1,2,\cdots,n \tag{12}$$

$$\mathrm{round}(\overline{K})=\left[x+\frac{1}{2}\right] \tag{13}$$

设 EC 为数据集 D 的一个等价类,则等价类中所

有元组在等级型属性 QI_g 上的中心用中心均值频率数 $C(EC)_g$ 定义。

1.5.3 分类型属性

在分类型属性 QI_c 的取值中,构建分类树向上寻找父节点,通过计算节点间的距离来确定分类型属性值之间的距离。例如分类型属性“国籍“构建分类树如图 1 所示。

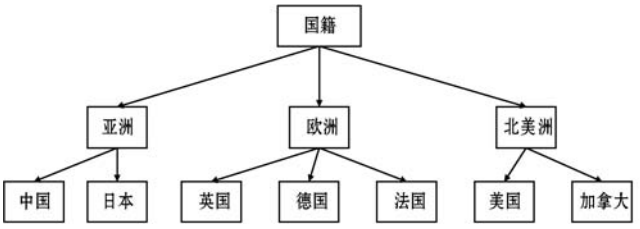


图 1 “国籍”属性分类树

设 $root_c$ 为分类型属性 QI_c 分类树根节点路径长度,$u(t_i,t_j)$ 为元组 t_i 和 t_j 到其最近邻公共祖先节点的距离。分类型属性距离为:

$$d_c(t_i,t_j)=\frac{u(t_i,t_j)}{root_c} \tag{14}$$

分类型属性 QI_c 的中心为等价类 EC 内所有元组两两之间距离和的最小值的元组在 QI_c 上的取值。设 EC 为数据集 D 的一个等价类,则等价类中所有元组在分类型属性 QI_c 上的中心为:

$$C(EC)_c=\underset{t_i,t_j\in EC,i\neq j}{\operatorname{argmin}}\sum d_c(t[QI_c]_i,t[QI_c]_j) \tag{15}$$

1.5.4 布尔型属性

布尔型属性的取值之间既没有等级关系也没有分类关系,则称该属性为布尔型属性 QI_b。例如“性别”为一个布尔型属性,其取值有“男性”“女性”。布尔型属性距离为:

$$d_b=\begin{cases}0 & (t[QI_b]_i=t[QI_b]_j)\\ 1 & (t[QI_b]_i\neq t[QI_b]_j)\end{cases} \tag{16}$$

式中:$t[QI_b]_i$ 为元组 i 在布尔属性 QI_b 上的取值。

设 EC 为数据集 D 的一个等价类,则布尔型属性 QI_b 的中心 $C(EC)_b$ 为等价类中元组取值出现频率最高的值:

$$C(EC)_b=\underset{t_i\in EC}{\operatorname{argmax}}f(t[QI_b]_i) \tag{17}$$

式中:$f(t[QI_b]_i)$ 为元组 t_i 在布尔型属性 QI_b 上取值在等价类中所出现的频率。

1.6 信息损失

本文基于元组间距离来衡量信息损失,元组间距离越小则等价类中元组相似度越高,数据的可用性越高。

1.6.1 元组间距离

设 $t_i,t_j\in D$,则元组 t_i 和元组 t_j 之间的距离为元组在各个准标识属性上距离之和为:

$$d_t(t_i,t_j)=\sum_{x=1}^{|QI|}|t[QI_x]_i-t[QI_x]_j| \tag{18}$$

1.6.2 元组数据匿名化的信息损失

设 C 为等价类的中心,则等价类内元组被数据匿名化后产生的信息损失为 t 与 C 之间的距离为:

$$d_t(t,C)=\sum_{x=1}^{|QI|}|t[QI_x]-C[QI_x]| \tag{19}$$

1.6.3 等价类数据匿名化的信息损失

设等价类 EC 内元组数据匿名化产生的信息损失为 $d_t(t,C)$,则等价类数据匿名化所产生的信息损失记为 $InfoLoss(EC)$,即:

$$InfoLoss(EC)=\sum_{t\in EC}d_t(t,C) \tag{20}$$

为提高等价类中敏感属性的多样性,降低信息损失,本文提出隐私安全指数的概念,定义如下:

定义 4(隐私安全指数) 等价类 EC 的敏感级别信息熵为 $HEC(EC)$,等价类 EC 数据匿名化的信息损失为 $InfoLoss(EC)$,则等价类 EC 的隐私安全指数为:

$$PSI(EC)=w_1\cdot HEC(EC)+w_2\cdot\frac{1}{InfoLoss(EC)} \tag{21}$$

式中:$w_1+w_2=1$,通过不同的 w 值来侧重匿名数据集的隐私保护性还是数据可用性。在聚类过程中,隐私安全指数越大,等价类敏感级别信息熵越大,数据的安全性越高,数据匿名化的信息损失越小。

为评估数据匿名化的效果,本文提出熵隐私保护度和信息损失率的概念,定义如下:

定义 5(熵隐私保护度) 设 $R=\{EC_1,EC_2,\cdots,EC_n\}$ 为数据匿名前聚类处理的等价类集合,$EC_i(i=1,2,\cdots,n)$ 为 R 中任意一个等价类,则熵隐私保护度表示为:

$$EPP(R)=\frac{\sum_{EC_i\in R}HEC(EC_i)}{|R|} \tag{22}$$

式中:$|R|$ 为集合 R 中的等价类个数。

定义 6(信息损失率) 设 $R=\{EC_1,EC_2,\cdots,EC_n\}$ 为数据匿名前聚类处理的等价类集合,$EC_i(i=1,2,\cdots,n)$ 为 R 中任意一个等价类,则信息损失率为:

$$R_{IL}=\frac{\sum_{i=1}^{n}InfoLoss(EC_i)}{|D|} \tag{23}$$

式中:$|D|$ 为数据集元组的个数。

2　敏感级别信息熵的匿名模型

为避免相似攻击,提高数据可用性,本文引入信息熵,并在 p-敏感 k-匿名模型基础上加以改进。

定义 7((H,p,k)-匿名模型)　在发布数据集 D^* 中,有 m 个等价类,每个等价类至少有 k 个元组在准标识属性上取值不可区分,有 $p(p\leqslant k)$ 个元组在敏感属性上取值不同,且每个等价类中元组不同敏感级别的个数不少于 H。即 $D^*=(EC_1,EC_2,\cdots,EC_m)$,$\forall EC_w\in D^*$,$|EC_w|\geqslant k$,$\forall t_i,t_j\in EC_w$,有 $t_i[QI]=t_j[QI]$,且 $|t_i[SA]\neq t_j[SA]|\geqslant p$,$|EC_w[SL]|\geqslant H$,其中 $i\neq j,w\in[1,m],p\leqslant k$。其中 $|EC_w[SL]|$ 为等价类中元组不同敏感级别的个数。

表 3 为满足(2,2,2)-匿名模型的数据集。相对 p-敏感 k-匿名模型,(H,p,k)-匿名模型限制每个等价类的敏感级别信息熵,避免相似攻击导致的敏感属性泄露。

表 3　(2,2,2)-匿名模型数据表

EC	Age	ZipCode	Disease
1	[21,25]	11 ****	HIV
	[21,25]	11 ****	Flu
2	[26,35]	11 ****	HIV
	[26,35]	11 ****	Hepatitis
3	[32,45]	115 ***	Cancer
	[32,45]	115 ***	Fever
4	[48,50]	1 *****	Diabetes
	[48,50]	1 *****	Flu

3　MAA-SLIE 算法

3.1　算法思想

本节提出一种基于敏感级别信息熵的微聚集算法(Micro-aggregation Algorithm based on Sensitive Level Information Entropy,MAA-SLIE),其基本思想为:利用贪心和聚类思想划分,保证等价类隐私安全指数最大,等价类的敏感级别信息熵也就最大,提高了等价类中敏感属性的多样性的同时又能减少信息损失。在聚类过程中,使等价类至少有 k 个元组在准标识属性上取值不可区分,有 $p(p\leqslant k)$ 个元组在敏感属性上取值不同,且每个等价类中元组不同敏感级别的个数不少于 H,保证(H,p,k)-匿名模型的实现,算法伪代码描述如下:

算法 1　MAA-SLIE

输入: dataset D, Quasi-identifier QI, Sensitive attribute SA, Sensitive attribute level SAL, (H, p, k)-anonymous model parameters H, p, k(1 < p≤k)。

输出: Anonymous data table D * 。

1: R = ∅;

2: if the number of different sensitive attribute values in D < p or the number of tuples in D < k;

3: return;

4: find central tuple t_{center} of D, D = D − t_{center};

5: while the number of different sensitive attribute values in D≥p and the number of tuples in D≥k do

6: find the tuple t_s furthest from t_{center} in D;

7: generate class $G_i=\{t_s\}$, i = 1,2,⋯,n, $t_{center}=t_s$;

8: while the number of tuples in G_i < k do

9: while the number of different sensitive attribute of tuples in G_i < p do

10:　if t[SA] = t_j[SA], (t ∈ D, ∀ t_j ∈ G_i) then / * D 中元组的 SA 与 G_i 中元组的 SA 都相同 * /

11:　D = D∪G_i, break_mark = Ture;

12:　　break

13:　t′ = Argmax$_{(t\in D)\wedge(\forall t_j\in G,t[SA]\neq t_j[SA])\wedge((G\cup t)[SL]\geqslant H)}$ (PSI(G∪t)); / * 从 D 中找与 G 中敏感属性不同的元组,该元组使 G 的 PSI 最大且满足(G∪t)[SL]≥H * /

14:　G_i = G_i∪t′, D = D − t′;

15:　end while

16: if break_mark = Ture then

17:　break

18:　t″ = Argmax$_{(t\in D)}$(PSI(G_i∪t));

19:　G_i = G_i∪t″, D = D − t″;

20:　end while

21: R = R∪G_i, del G_i, i + = 1;

22: end while

23: while remaining tuples in D do

24:　　randomly select tuples t from D, D = D − t;

25:　　G_i = Argmax$_{(G_i\in Q)}$(PSI(G_i∪t))

26:　R = R − G_i, G_i = G_i∪t;

27:　　R = R∪G_i;

28: end while

29: process each class in R in turn, replace the attribute value of each tuple in the class on the quasi-identity attribute with the central attribute value of the class to obtain the anonymous data set D * ;

30: Get cluster result D *

end

3.2　算法描述

步骤 2　如果 D 中不同敏感属性的元组个数小于

p,或 D 中元组个数小于 k 时,算法返回。

步骤 5 当 D 满足敏感属性取值不同的元组个数大于 p,且元组个数大于 k 时,重复执行步骤 6 – 步骤 22。

步骤 10 如果 D 中没有满足与 G 中元组敏感属性值不同的元组,G 中剩余元组加入 D,跳出循环。

步骤 13 从 D 中找到一个元组 t,其中 t 的敏感属性值与 G 类中所有的元组敏感属性值都不相同,与 G 合并后的等价类的隐私安全指数最大,且等价类敏感级别信息熵大于 lbH。

步骤 23 – 步骤 28 把 D 中剩余元组加入到合并后隐私安全指数最大的类。

步骤 29 – 步骤 30 进行数据匿名化,用等价类中各个准标识属性的中心替代元组在其上的值,得到满足 (H,p,k)-匿名模型的匿名数据集 D′。

3.3 算法分析

3.3.1 算法的正确性分析

在生成等价类过程中,当 D 中敏感属性取值不同的元组个数大于 p,且元组个数大于 k 时,MAA-SLIE 算法最终会得到满足 (H,p,k)-匿名模型的匿名数据集。从原始数据集中找到使等价类隐私安全指数最大的元组,根据式(21),隐私安全指数越大,隐私泄露风险越低,隐私抵抗效果越好,并且等价类内元组相似性越高,数据匿名的信息损失越低。

3.3.2 算法的复杂性分析

设 D 中元组个数为 n,准标识属性个数为 d,匿名数据划分 m 个等价类。算法在步骤 2 和步骤 3 判断 D 中元组个数与敏感属性值个数时,遍历一次 D,时间复杂度为 $O(n)$。在步骤 4 中,求 D 中心元组,遍历一次 D,时间复杂度为 $O(n)$。在步骤 5 – 步骤 22 中,每得到一个新的等价类 G 时,最多遍历 k 次 D,然后计算准标识属性上的距离,生成一个等价类的时间复杂度为 $O(dkn)$,共生成 m 个等价类,时间复杂度为 $O(dkmn)$,步骤 22 结束后,D 中元组个数至多还剩余 $n-mk$ 个。因此,算法步骤 23 – 步骤 28 的循环次数为 $n-mk$,每次循环遍历 R 一遍,时间复杂度为 $O(dm)$,故步骤 23 – 步骤 28的时间复杂度为 $O(dm(n-mk))$。步骤 29 生成结果集,把所有元组准标识属性的值替换为所在等价类的中心,时间复杂度为 $O(dn)$。总体执行过程中,MAA-SLIE 算法的总时间复杂度为 $O(dkmn)$。

4 实验及结果分析

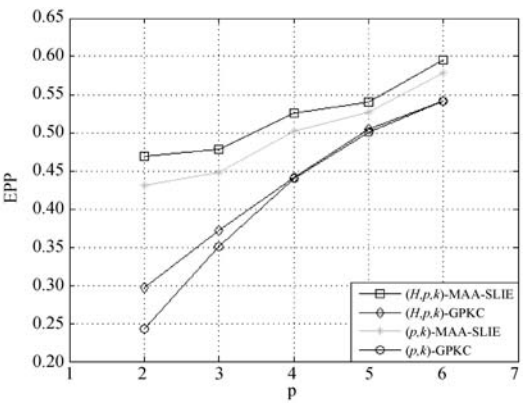
4.1 实验数据集与参数

本节验证 MAA-SLIE 算法性能,选择与 GPKC (GreedyPKClustering)算法[16]((H,p,k)-GPKC)、p-敏感 k-匿名模型在 MAA-SLIE((p,k)-MAA-SLIE)和 p-敏感 k-匿名模型在 GPKC 算法((p,k)-GPKC)进行对比,实验从隐私保护率,信息损失和执行时间三个方面进行对比分析。本实验选用 UCI 机器学习数据库中的 Adult 数据集作为实验数据集,去除有缺失值的数据,得到 45 222 个数据记录,包含有 15 个属性,本文选取其中 3 000 个数据记录作为实验数据集进行实验。实验选取其中 8 个属性:Age、Gender、Race、Education、Marital Status、Native Country、Work Class、Occupation。其中 Age 为连续数值型属性,Education 为等级型属性,Marital Status、Native Country、Work Class 为分类型属性,Gender、Race 为布尔型属性,Occupation 为敏感属性。实验环境为英特尔 Core i5-8300H @ 2.30 GHz CPU,8 GB RAM,操作系统为 Microsoft Windows 10,算法均在 IDLE(Python 3.7 64-bit)下实现。

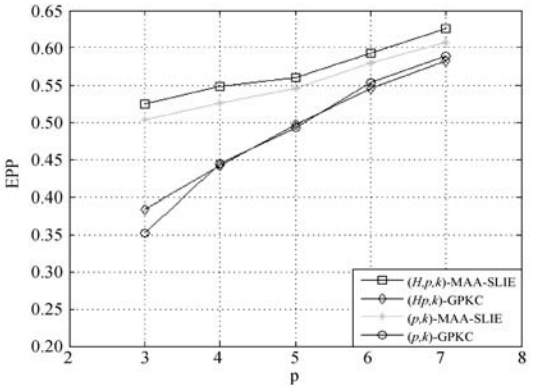
本文将 Adult 数据集中准标识属性的权重均设为 1,将敏感级别分为 5 个级别,在计算等价类敏感级别信息熵时,式(2)中的加权参数 w 由低敏感级别到高敏感级别分别取 0.1、0.2、0.4、0.6、0.8,高敏感级别属性的权重较高,低敏感级别属性的权重较低,这样既能减少信息损失又能对高敏感级别的属性更好保护。隐私安全指数中的 w_1、w_2 取 0.5。选取不同的 H、k 和 p 值,并对每一组不同取值重复进行 10 次实验。

4.2 隐私泄露风险分析

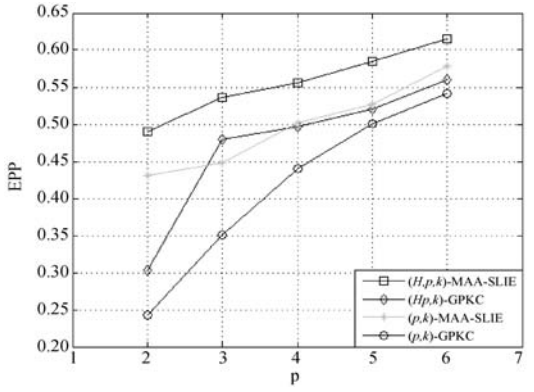
如图 2 所示,当 $H=2$,k 分别为 8 和 12,随着 p 的增加,等级类中不同敏感属性值增多,EPP 也随之增加,故数据安全性提高。当 $k=8$,H 分别为 2 和 3 时,H 越大熵隐私保护度也越大,因为等价类中包含了更多不同敏感级别的元组。在同等条件下算法 MAA-SLIE 比算法 GPKC 的熵隐私保护度要高,因为在聚类的过程中算法 MAA-SLIE 要求聚类时隐私安全指数最大,即等价类的敏感属性的敏感级别信息熵最大,敏感属性值最丰富,数据的安全性最高。



(a) $H=2$, $k=8$



(b) $H=2$, $k=12$

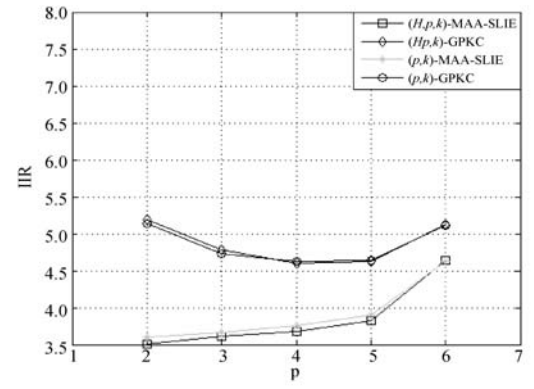


(c) $H=3$, $k=8$

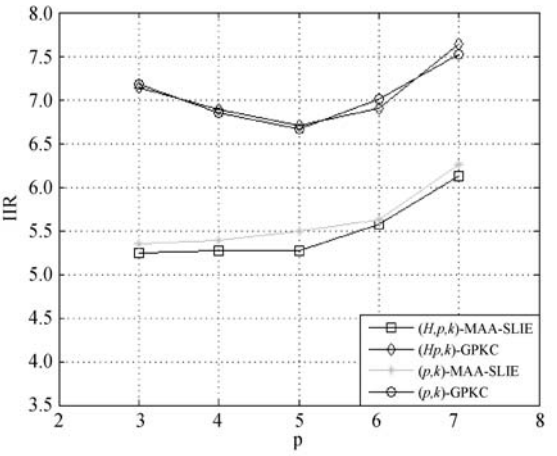
图 2　H、p 值变化下的熵隐私保护度

4.3　信息损失分析

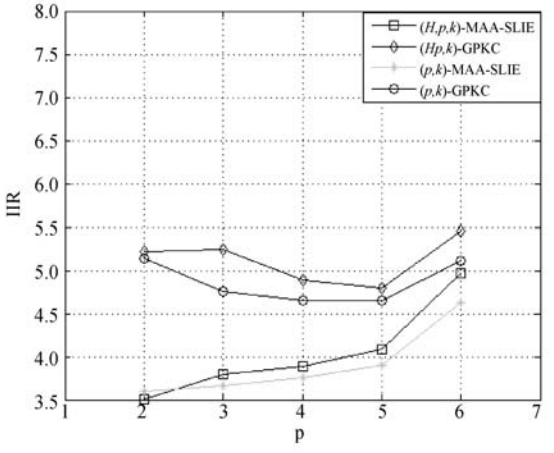
如图 3 所示,当 $H=2$,k 分别为 8 和 12 时,随着 p 的增长,信息损失率也随之增加,因为随着等价类需要加入更多不同敏感属性值的元组,排除部分相似度较高但敏感属性值相同的元组,信息损失因此增加。在相同 p 值的情况下,k 值越高信息损失率越高,因为等价类中元组越多,元组间距离因此增加。当 $k=8$,H 分别为 2 和 3 时,H 越大信息损失率也较大,因为要求等价类有更多不同的敏感级别的元组,因此会筛选掉相同敏感级别的元组。对比 MAA-SLIE 与 GPKC,MAA-SLIE 的信息损失率比 GPKC 低。因为 MAA-SLIE 要求聚类时隐私安全指数最大,故元组间的距离最小,信息损失量最少。



(a) $H=2$, $k=8$



(b) $H=2$, $k=12$



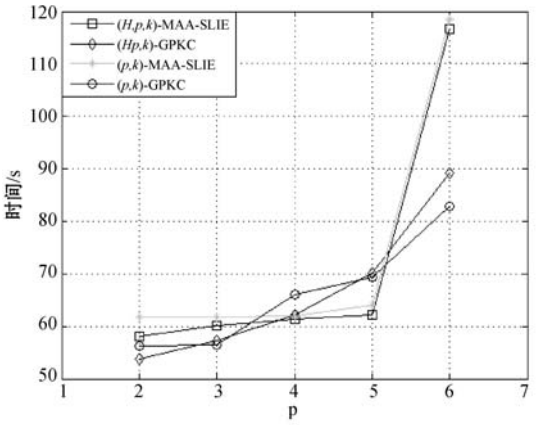
(c) $H=3$, $k=8$

图 3　H、p 值变化下的信息损失率

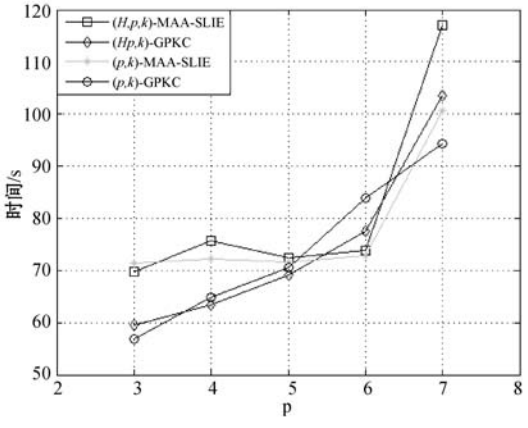
由图 2 和图 3 可知,同样 k 值的情况下,随 p 值的增加数据的熵隐私保护度和信息损失率都增加,因此数据匿名的安全保护程度与信息损失率是彼此矛盾的。

4.4　执行时间分析

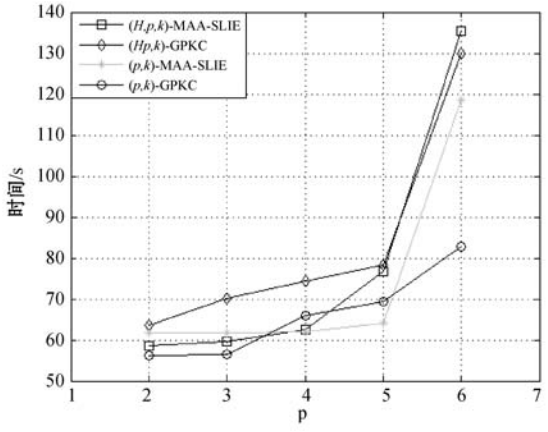
图 4 为在 $H=2$,k 值分别为 8 和 12 时两种算法运行时间比较,k 值固定随着 p 值增加,聚类过程中需要比对的元组变少,聚类过程中时间变短,但是会有更多剩余的元组不满足等价类不同敏感属性值的要求,在剩余元组处理阶段需要更多次的计算,因此执行时间增加。在 $k=8$,H 分别为 2 和 3 时,H 越大算法执行时间越长。等价类中不同敏感级别元组更多,更多的元组在聚类筛选时被排除,因此会增加剩余元组处理阶段的时间。MAA-SLIE 比 GPKC 在多数情况下会多花一些时间,因为在聚类同样大小的等价类,MAA-SLIE 不仅需要考虑元组间的距离还需计算等价类敏感属性的敏感级别信息熵,但两个算法的执行时间相差不大。



(a) $H=2$, $k=8$



(b) $H=2$, $k=12$



(c) $H=3$, $k=8$

图 4 H、p 值变化下的执行时间

5 结 语

本文为避免相似攻击导致的隐私泄露问题提出了一种基于等价类敏感级别信息熵的(H,p,k)-匿名模型,并设计匿名方法 MAA-SLIE。该方法在聚类过程中通过贪心法保证隐私安全指数最大,从而等价类中敏感属性取值多样性提高,降低隐私泄露风险通过微聚集算法减少数据概化的信息损失。通过多组实验结果的比较和分析,验证了该算法的有效性和合理性。本文只考虑单一敏感属性,后续研究方向是多敏感属性的数据匿名。

参 考 文 献

[1] Latanya S. k-ANONYMITY: A model for protecting privacy [J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems,2002,10(5):557 - 570.

[2] Machanavajjhala A, Gehrke J, Kifer D, et al. L-diversity: Privacy beyond k-anonymity[C]//22nd International Conference on Data Engineering (ICDE'06). IEEE,2006.

[3] Campan A, Truta T M, Cooper N. P-sensitive k-anonymity with generalization constraints[J]. Transactions on Data Privacy, 2010,3(2):65 - 89.

[4] Li N, Li T, Venkatasubramanian S. t-Closeness: Privacy beyond k-anonymity and l-diversity[C]//Proceedings of IEEE 23rd International Conference on Data Engineering. IEEE, 2007.

[5] 桂琼,程小辉. 基于聚类的分级匿名方法[J]. 计算机应用,2013,33(2):412 - 416.

[6] 廖军,蒋朝惠,郭春,等. 一种基于权重属性熵的分类匿名算法[J]. 计算机科学,2017,44(7):42 - 46.

[7] Majeed A, Lee S. Attribute susceptibility and entropy based data anonymization to improve users community privacy and utility in publishing data[J]. Applied Intelligence, 2020, 50(2):2555 - 2574.

[8] 姜火文,曾国荪,马海英. 面向表数据发布隐私保护的贪心聚类匿名方法[J]. 软件学报,2017,28(2):341 - 351.

[9] 桂琼,吕永军,程小辉. 基于敏感信息邻近抵抗的匿名方法[J]. 计算机工程,2020,46(12):142 - 149.

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

[11] Domingo-Ferrer J, Mateo-Sanz J M. Practical data-oriented microaggregation for statistical disclosure control[J]. IEEE Transactions on Knowledge & Data Engineering, 2002, 14(1):189 - 201.

[12] Domingo-Ferrer J, Torra V. Ordinal, continuous and heterogeneous k-anonymity through microaggregation [J]. Data Mining & Knowledge Discovery, 2005, 11(2):195 - 212.

[13] Torra V. Microaggregation for categorical variables: A median based approach[C]//CASC Project Final Conference. DBLP,2004.

[14] 张岐山,郑丽君. 基于灰关联分析的 V-MDAV 算法研究[J]. 计算机应用研究,2020,37(1):107 - 111.

[15] 杨静,王超,张健沛. 基于敏感属性熵的微聚集算法[J]. 电子学报,2014,42(7):1327 - 1337.

[16] Campan A, Truta T M, Miller J, et al. A clustering approach for achieving data privacy[C]//International Conference on Data Mining. DBLP, 2009.