计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2021,30(4):131−138 [doi: 10.15888/j.cnki.csa.007870]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

面向高维数据发布的个性化差分隐私算法①

马苏杭[1,2], 龙士工[1,2], 刘　海[1,2], 彭长根[1,2], 李思雨[1]

[1](贵州大学 计算机科学与技术学院, 贵阳 550025)
[2](贵州大学 贵州省公共大数重点实验室, 贵阳 550025)
通讯作者: 龙士工, E-mail: 2418424924@qq.com

摘　要: 在高维数据隐私发布过程中, 差分隐私预算大小直接影响噪音的添加. 针对不能合理地为多个相对独立的低维属性集合合理分配隐私预算, 进而影响合成发布数据集的安全性和可用性, 提出一种个性化隐私预算分配算法 (PPBA). 引入最大支撑树和属性节点权重值降低差分隐私指数机制挑选属性关系对的候选空间, 提高贝叶斯网络精确度, 提出使用贝叶斯网络中节点动态权重值衡量低维属性集合的敏感性排序. 根据发布数据集安全性和可用性的个性化需求, 个性化设置差分隐私预算分配比值常数q值, 实现对按敏感性排序的低维属性集合个性化分配拉普拉斯噪音. 理论分析和实验结果表明, PPBA 算法相比较于同类算法能够满足高维数据发布安全性和可用性的个性化需求, 同时具有更低的时间复杂度.

关键词: 贝叶斯网络; 差分隐私; 最大支撑树; 动态权重值; 个性化比例分配

引用格式: 马苏杭,龙士工,刘海,彭长根,李思雨.面向高维数据发布的个性化差分隐私算法.计算机系统应用,2021,30(4):131–138. http://www.c-s-a.org.cn/1003-3254/7870.html

Personalized Differential Privacy Algorithm for High-Dimensional Data Publishing

MA Su-Hang[1,2], LONG Shi-Gong[1,2], LIU Hai[1,2], PENG Chang-Gen[1,2], LI Si-Yu[1]

[1](College of Computer Science and Technology, Guizhou University, Guiyang 550025, China)
[2](Guizhou Provincial Key Laboratory of Public Big Data, Guizhou University, Guiyang 550025, China)

Abstract: In the process of privacy preserving high-dimensional data publishing, the size of the differential privacy budget directly affects the addition of noise. The privacy budget cannot be allocated reasonably for independent low-dimensional attribute sets, compromising the security and restricting availability of composite data sets. Then a Personalized Privacy Budget Allocation (PPBA) algorithm is proposed. The maximum support tree and weight values of attribute nodes are introduced to reduce the candidate space of attribute relationship pairs selected by the differential privacy index mechanism and enhance the accuracy of the Bayesian network. The dynamic weight values of nodes in the Bayesian network are set to rank the sensitivity of low-dimensional attribute sets. According to the personalized requirements for security and availability of published data sets, the constant allocation ratio q of differential privacy budgets is customized for the personalized allocation of Laplace noise to the low-dimensional attribute sets sorted by sensitivity. Theoretical analysis and experimental results reveal that the PPBA algorithm can meet the personalized requirements for security and availability of high-dimensional data publishing, with lower time complexity.

Key words: Bayesian network; differential privacy; maximum support tree; dynamic weight value; personalized proportional distribution

① 基金项目: 国家自然科学基金 (62062020, 62002081, U1836205); 贵州省科技计划 (黔科合重大专项字 [2018]3001)
Foundation item: National Natural Science Foundation of China (62062020, 62002081, U1836205); Science and Technology Plan of Guizhou Province ([2018]3001)
收稿时间: 2020-08-18; 修改时间: 2020-09-10; 采用时间: 2020-09-18; csa 在线出版时间: 2021-03-30

1 引 言

随着移动互联网的发展, 数据规模也以前所未有的速度不断增长, 数据属性之间的相互关系变得复杂多样, 高维数据已是一种常见的数据发布类型. 随着数据挖掘和分析技术的发展, 高维数据的发布具有更高的信息价值, 但高维数据中通常包含大量隐私信息, 如果使用不当将造成隐私泄露[1,2]. 为了保证高维数据发布过程中不会泄露隐私信息, 在发布之前使用差分隐私[3,4] 保护技术进行处理. 如果直接对高维数据进行差分隐私处理, 存在添加噪音过多, 数据可用性差等问题. 其中差分隐私预算的分配方式直接影响数据的可用性与安全性关系, 而不同数据机构对于发布数据集安全性和可用性之间的关系需求各不相同, 数据保护级别更高的数据机构更注重数据的安全性; 而主要提供数据进行应用的数据机构则更倾向于数据的可用性.

目前已有的面向高维数据发布的差分隐私算法有概率图模型[5–7]、阈值过滤技术[8] 以及投影技术[9], 这些技术通过维度转换达到降维效果, 减少噪音添加对数据可用性的影响. 降维效果的好坏直接影响数据的可用性, 而阈值过滤技术和投影技术忽略了高维属性之间普遍存在依赖关系, 采用直接截断的降维方法, 大大降低了数据的可用性. 文献 [5–7] 利用指数机制[3,10] 挑选属性关系对, 受候选空间大小和隐私预算分配方式的影响, 空间越大挑选的属性关系对越不准确. 同时, 单一的隐私预算分配方式为敏感性不同的属性数据分配相同的隐私预算, 导致隐私预算无法根据数据可用性与安全性的个性化需求合理分配, 存在隐私浪费的问题.

基于在高维数据发布过程中, 数据安全性与可用性受降维算法效果和隐私预算分配方式的影响, 为满足发布数据集安全性与可用性的个性化需求, 本文提出个性化隐私预算分配 (Personnalized Privacy Budget Allocation, PPBA) 算法, 主要内容如下.

(1) 对基于概率图模型的贝叶斯网络算法进行优化, 引入最大支撑树和最大权重值, 减少指数机制挑选属性关系对的搜索空间, 避免敌手进行多次查询对比分析, 泄露隐私信息. 提高数据可用性和安全性.

(2) 依据动态权重值确定贝叶斯网络中低维属性集合敏感性由大到小的排序. 受文献 [11–13] 启发, 根据不同用户数据可用性与安全性需要, 个性化设置隐私预算分配比值常数q,为不同敏感性的属性集合合理分配差分隐私 (Laplace[10]) 噪声.

(3) 理论证明所提出的 PPBA 算法满足ε-差分隐私, 并在真实数据集上进行性能评估. 实验结果表明能够满足数据可用性与安全性个性化需求, 同时降低了时间复杂度.

2 相关工作

数据独立发布算法和数据相关发布算法是主要的 2 类面向高维数据发布的差分隐私算法. 独立发布算法的典型代表是 PriView[14],该算法假设所有属性都是相互独立的,这在真实数据集中是不存在的, 且缺少正式的推理机制. 而 PrivBayes 算法[5]、加权贝叶斯网络算法[6]、联合树算法[7] 是典型的数据相关发布算法.

PrivBayes 算法利用指数机制挑选属性关系对形成贝叶斯网络, 对联合分布概率进行推理, 存在候选空间较大, 数据可用性和安全性得不到保障的问题. 文献 [6] 对贝叶斯网络进行优化, 利用最大权重值提高贝叶斯网络推理的准确性, 但仍然存在挑选属性关系对候选空间较大的问题. 文献 [7] 通过指数机制构造 Markov 网, 引入高通滤波技术缩减指数机制搜索空间. 并结合相应的后置技术对 Markov 网分割来获得完全团图, 生成满足差分隐私的联合树, 利用联合树中各个团后置处理之后的联合分布表合成最终的高维数据. 文献 [5–7] 在高维数据相关发布得到广泛的应用, 但在面对不同数据机构对于数据安全性与可用性的个性化需求, 缺少个性化的隐私预算分配策略.

针对不同数据类型关于隐私预算分配问题, 为了兼顾数据安全性与可用性的效率, 文献 [11] 以差分隐私保护结合主流决策树分类方法, 提出等差分配隐私预算的方式, 改善决策树的分类准确率. 文献 [12] 针对树索引结构提出等差数列分配和等比数列分配两种方式. 避免对树的某一层分配过小, 数据可用性过低; 分配过大, 不能对这层数据提供足够安全保障的问题.

3 基础知识

本节内容主要对面向高维数据发布的个性化差分隐私算法所使用的贝叶斯网络、差分隐私概念进行说明.

3.1 贝叶斯网络

文章在论述过程中涉及较多数学符号, 为了更好地对下文相关内容进行解释, 给出相关符号定义, 如表 1 所示.

表 1 符号定义表

符号	描述
D	原始高维数据集D
n	数据集D的元组个数
Ar	数据集D中的属性集合
d	属性集合Ar中的属性个数
N	贝叶斯网络
$\Pr[Ar]$	高维数据集D的原始分布
$\Pr_N[Ar]$	根据贝叶斯网络推理原始数据集的近似分布
K	贝叶斯网络的最大父节点数, 即贝叶斯网络的度值
WV	属性节点的权重值
DWV	属性节点的动态权重值
CM	属性值的多样性
$P(X,M)$	属性节点与父节点集合的联合概率
M	父节点集合
$dom(X)$	属性变量的域

定义 1. 贝叶斯网络. 贝叶斯网络N为一个有向无环图, N中每一个节点代表高维数据集 D 中一个字段属性, 如果N中两个属性节点之间存在着直接依赖关系, 则两个属性字段节点之间用一条弧 (或有向边) 直接相连. 贝叶斯网络N使用 (属性字段节点, 属性字段节点的父节点集合) 对来表示.

通过挑选属性间的依赖关系, 实现高维数据的维度转换, 构建贝叶斯网络进行联合分布的推理. 通过例子解释说明, 高维数据集属性集合为Ar_1, 有 A、B、C、D 共 4 个属性, 未进行维度转换形成贝叶斯网络时, 其联合分布的计算如下式所示:

$$\Pr[Ar_1]=\Pr[A]*\Pr[B|A]*\Pr[C|A,B]*\Pr[D|A,B,C] \tag{1}$$

若在属性依赖关系的挑选中使用最大父节点个数值即度值为 2 的贝叶斯网络算法对该数据集进行处理, 形成如图 1 所示 4 个属性字段节点构成的 2 度贝叶斯网络图.

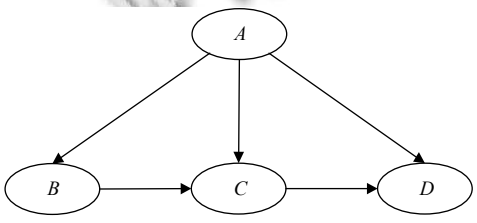

图 1 2 度贝叶斯网络

则该贝叶斯网络用 4 个相对独立的低维属性集合$(A,\emptyset)$,$(B,\{A\})$, $(C,\{A,B\})$,$(D,\{A,C\})$, 来表示, 其中联合分布$\Pr_N[Ar_1]$的计算如式 (2) 所示.

$$\Pr_N[Ar_1]=\Pr[A]*\Pr[B|A]*\Pr[C|A,B]*\Pr[D|A,C] \tag{2}$$

未进行维度转化处理之前该数据集属性之间存在 6 种属性关系, 当使用 2 度贝叶斯网络算法之后降低到 5 种属性关系. $\Pr_N[Ar_1]$相比$\Pr[Ar_1]$在数据量较多的情况下具有更低的计算复杂度, 为多个相对独立的低维属性集合加入更少的噪声.

3.2 差分隐私

差分隐私保护技术通过向原始数据集添加满足差分隐私的噪音生成邻近数据集, 使得原始数据集与邻近数据集在查询输出中具有概率不可区分性.

定义 2. ε-差分隐私[10]. 对于任意两个相邻数据集D_1和D_2, 它们之间相差最多为一条记录, 若一个随机函数 A 满足ε-差分隐私保护, $Range(A)$表示随机函数A的取值范围, 则对于所有的$S\subseteq Range(A)$有:

$$\Pr[A(D_1)\in S]\le e^{\varepsilon}\Pr[A(D_2)\in S] \tag{3}$$

其中, $\Pr[E]$表示事件 E 的披露风险, ε为隐私预算参数, 代表了差分隐私保护水平, 其值越小, 不可区分性越大, 隐私保护级别越高.

定义 3. 敏感度[10]. 敏感度是由函数本身决定的, 不同函数具有不同的敏感度, 敏感度过低会使发布数据集的安全性得不到保障, 敏感度过高则使发布数据集的发布结果实用性降低.

给定 F 是将一个数据集映射到一个固定大小实数向量的函数, 那么函数 F 的敏感度为:

$$S(F)=\max\|F(D_1)-F(D_2)\|_1 \tag{4}$$

其中, D_1和D_2为任意两个邻近数据集, 二者仅相差一个数据元组.

为了在给定的隐私预算内, 将全部隐私预算合理分配到多个相对独立的低维属性集合中, 使整个数据发布过程中满足差分隐私, 可以利用差分隐私的序列组合性质.

性质 1. 差分隐私序列组合性[11]. 给定数据集D,相互独立的差分隐私随机算法$A_1,A_2,\cdots,A_i$分别满足ε_i-差分隐私, 其中$1\le i\le d$, 则序列组合$\{A_1,A_2,\cdots,A_i\}$满足ε-差分隐私, 其中$\varepsilon=\sum_{i=1}^{d}\varepsilon_i$.

定义 4. 互信息函数. 1948 年香农提出信息熵[14] 的概念, 属性之间互信息 I 的大小代表属性之间的关联程度. 高维数据集 D 属性节点 X 与 Y 之间的互信息

I 如式 (5) 所示.

$$I(X:Y)=\sum_{x\in dom(X)}\sum_{y\in dom(Y)}P(X=x,Y=y)*\log\frac{P(X=x,Y=y)}{P(X=x)P(Y=y)} \tag{5}$$

其中, 满足差分隐私的噪音机制主要有指数机制、Laplace 机制.

命题 1. 基于互信息函数的指数机制. 指数机制[10]主要用于处理输出结果为非数值型结果. 在维度转换过程中, 属性节点的关联程度作为指数机制挑选属性关系对的依据, 打分函数为属性间的互信息函数 I, 其中 $\Delta I(X:Y)$ 为互信息函数 I 的敏感度, 以正比于 $\exp\left(\frac{\varepsilon I(X:Y)}{2\Delta I(X:Y)}\right)$ 的概率挑选出具有最大依赖关系的维度属性, 组成多个满足 ε 差分隐私的相对独立的低维属性集合. 其中文献 [5] 中给出了维度转换过程中互信息敏感度的计算方法, 见式 (6); 由于在指数机制挑选过程中, 除挑选属性关系对外无其它隐私消耗, 由差分隐私组合性质[11], 该过程满足对应 ε-差分隐私.

$$\Delta(I(X:Y))=\frac{2}{n}\log\frac{n+1}{2}+\frac{n-1}{n}\log\frac{n+1}{n-1} \tag{6}$$

命题 2. 基于联合分布的拉普拉斯机制. 拉普拉斯机制[11]通过 Laplace 分布产生噪声扰动真实值达到差分隐私保护. 在贝叶斯网络中对多个相对独立的低维属性集合, 计算其联合分布 P. $P^*=P+Z$ 为向其联合分布概率中添加拉普拉斯噪音 Z, 其中 Δf 为联合分布函数敏感度, $Z\sim Lap(\Delta f/\varepsilon)$ 为服从尺度参数 $\Delta f/\varepsilon$, 方差为 $2\Delta f^2/\varepsilon^2$ 的 Laplace 分布. 由于在该过程中除为联合分布添加拉普拉斯噪音外无其它隐私消耗, 由差分隐私组合性质[11]满足对应 ε 值的差分隐私.

4 PPBA 算法

4.1 最大支撑树

本节对最大支撑树的定义和构建过程进行解释说明, 通过最大支撑树限制指数机制挑选属性关系对的候选空间.

命题 3. 最大支撑树. 利用高维数据属性之间的互信息得到的一种树状网络结构, 通过依次计算两两属性间的互信息, 只保留与该属性具有最大互信息的属性之间的无向边, 完成最大支撑树的建立. 根据最大支撑树减少挑选属性关系对的候选空间, 确定贝叶斯网络度值 K.

算法 1. 最大支撑树

输入: Data D
输出: VT
1. Initialize: $T=\emptyset$, $VT=\emptyset$;
2. ①for i=1 to d
for j= 1 to d and $j\neq i$
Compute $I(X_i,X_j)$, add $I(X_i,X_j)$ to T
②Select Max $I(X_i,X_j)$, add (X_i,X_j) to VT;
3. Return VT;

根据算法 1 输出的 V_T 集合, 其中 VT 集合用于存储最大支撑树的无向边 (X_i,X_j), 以图 1 为例将图中有向边转化为无向边, 由连接关系可知 A、B、C、D 四个属性节点无向边个数分别为 3、3、2、2 其中最大值为 3, 则选取 K 值为 3.

4.2 个性化比例分配

本节内容主要对个性化比例分配方法所涉及的敏感性排序和比例分配的计算过程进行解释.

(1) 依据动态权重值对低维属性集合进行敏感性排序

在文献 [6] 中分别给出了 CM、WV、DWV 值的计算方法, 根据文献 [6] 中对属性节点动态权重值的定义, 动态权重值可以很好地代表属性节点在贝叶斯网络中的重要性, 重要性越高, 对于贝叶斯网络精确度和数据集的可用性影响越大, 该属性值隐私泄露对数据集的安全性影响越大. 故选取动态权重值作为敏感性的衡量依据.

假设图 1 中各属性 CM 值如表 2 中所示, 则由文献 [6] 的计算方法, 对图 1 中 4 个属性权重值计算结果如表 2 所示.

表 2 属性权重值计算结果表

i	X_i	M_i	CM	WV	DWV
1	A	$\emptyset$	15	0.3333	0.5555
2	B	$\{A\}$	10	0.2222	0.1556
3	C	$\{A、B\}$	12	0.2667	0.1668
4	D	$\{B、C\}$	8	0.1778	−0.1222

根据动态权重值大小进行排序, 则属性节点的敏感性排序为 A、C、B、D.

(2) 个性化比例分配计算

高维数据集经贝叶斯网络处理之后, 将数据集划分为 d 个相对独立的低维属性集合, 依据属性节点的动态权重值对低维属性集合进行敏感性由大到小排序, 根据隐私预算分配策略将总的隐私预算合理分配

到每个低维属性集合. 通过个性化设置分配比值常数$q(q>1)$, 从敏感性最高的低维属性集合起, 使该节点低维属性集合与前一个敏感性更高的低维属性集合分配的隐私预算大小比值为常数q $(q>1)$, 从而将隐私预算ε划分为$\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_d$分别分配至$d$个低维属性集合.

由图 1 中属性节点的低维属性集合敏感性由大到小的排序为 A、C、B、D. 总隐私预算ε大小, 根据需要设置的比值常数为q $(q\geq 1)$.

由等比数列性质式 (7)、式 (8):

$$\varepsilon=\sum_{i=1}^{d}\varepsilon_i=\frac{\varepsilon_1\left(1-q^d\right)}{1-q};q>1 \tag{7}$$

$$\varepsilon=\sum_{i=1}^{d}\varepsilon_1 d;q=1 \tag{8}$$

得:

$$\varepsilon_1=\frac{\varepsilon(1-q)}{1-q^d};q>1 \tag{9}$$

$$\varepsilon_1=\frac{\varepsilon}{d};q=1 \tag{10}$$

$$\varepsilon_i=\varepsilon_1 q^{d-1};q\geq 1 \tag{11}$$

取ε=0.5 时, 分别设q值为 1、1.1、1.3, 则 A、B、C、D 各属性节点分配的ε值由式 (9), 式 (10) 计算结果如表 3 所示.

表 3 ε分配表

q	A	C	B	D
1	0.1667	0.1667	0.1667	0.1667
1.1	0.1077	0.1185	0.1303	0.1433
1.3	0.0808	0.1050	0.1366	0.1775

由以上分析和表 3 可知, 当给定总的隐私预算和低维属性集合按敏感性由高到低的排序, 用户只需调整q值, 就可以改变隐私预算的分配方式. 当$q=1$时, 每个低维属性集合分配的隐私预算相同, 即均匀分配隐私预算. 当$q>1$时, 按低维属性集合排序, 每个集合分配的隐私预算以q倍增加, 随着q值的增加, 越重要的低维属性集合分配的隐私预算越小, 对应的保护强度越高, 数据的可用性则相应降低. 不难理解只要稍微改变q值, 就可以改变隐私预算分配方式.

4.3 PPBA 算法实现

本节描述 PPBA 算法的具体实现细节如算法 2.

算法 2. PPBA 算法

输入: D、K、q、ε
输出: N、D^*
1. Initialize: $N=\emptyset$,$V=\emptyset$;
2. Select X_1; add X_1 to V;add $(X_1,\emptyset)$ to N;
3. ① for i=2 to d
② Initialize $\Omega=\emptyset$;
③ for 每一个属性字段$X\in Ar/V$, 并且
④ add (X,M) to Ω
⑤ end for
⑥ 从Ω中选择使$\exp(\frac{\varepsilon_i I(X_i,M_i)}{2\Delta I(X_i,M_i)})$最大的$(X_i,M_i)$; add$(X_i,M_i)$ to N;add X_i to V;
⑦ end for
4. Return N;
5. 依据N, 计算低维属性集合属性节点的DWV值;
6. 根据DWV值, 将低维属性集合敏感性由大到小排序, 计算为每个集合分配的ε_i值
7. ① for i=1 to d do
② Add $\lambda_i=\frac{\Delta f}{\varepsilon_i}$ to $P(X_i|M_i)$;
③ return $P*(X_i,M_i)$;
④ end for
8. Return D^*

PPBA 算法主要分为两个部分, 1–4 步为算法第一部分, 实现满足ε/2-差分隐私的贝叶斯网络. 由最大支撑树确定贝叶斯网络的度值K, 第 2 步选择具有最大权重值的属性节点作为贝叶斯网络的首节点. 第 3 步以互信息函数为满足ε/2-差分隐私指数机制的打分函数, 从属性字段集合中选择 d–1 个低维属性集合对加入贝叶斯网络N, 其中V用于存储属性节点, $\binom{V}{K}$表示V的所有子集元素个数为$\min(K,|V|)$. 第 4 步返回满足差分隐私的贝叶斯网络N.

算法第 2 部分, 合成满足ε-差分隐私的发布数据集. 5–7 步根据数据可用性和安全性需求设置q值, 为每个属性集合分配满足ε/2-差分隐私 Laplace 机制的隐私预算. 为属性节点X_i的条件分布$P(X_i|M_i)$加入服从 Laplace 分布的噪音, 得到$P*(X_i|M_i)$. 第 8 步根据$P*(X_i|M_i)$形成原始数据集的近似联合分布, 抽样合成满足ε-差分隐私的合成发布数据集D^*.

4.4 满足差分隐私证明

证明. 在 PPBA 算法中, 根据命题 1 和命题 2 在指数机制挑选属性关系对和对条件分布添加拉普拉斯噪音的过程中由差分隐私序列组合性质[11] 分别满足ε/2-差分隐私保护, 其它行为不会产生额外的隐私预算. 根据差分隐私组合性质中的序列组合性[11], 证得 PPBA 算法满足ε-差分隐私.

5 实验与分析

根据实验测试结果, 对比分析 PPBA 算法、加权 PrivBayes 算法、PrivBayes 算法的数据可用性、数据安全性与可用性之间个性化平衡需求的实验以及算法时间性能 3 个方面.

5.1 实验环境

实验中, 采用美国 UCI (University of California, Irvine) 所提供的机器学习库中的成人数据集, 该数据集由美国人口普查数据组成, 共计 32 561 个元组. 在该数据集中一共选取了 10 个属性字段: Age, Workclass, Educatio, Maritalstatus, Race, Occupation, Relationship, Sex, Native, Country, Income. 在实验之前将数据集划分为测试数据集和训练数据集, 并对数据集做删除缺省值, 属性离散化等数据预处理操作.

实验中所使用的软硬件参数如下:

(1) 操作系统: Windows10;

(2) 硬件参数: IntelCoreTM I5, 2.4 GHz CPU, 8 GB DDR 内存;

(3) 编译环境及工具: Python3.6, Pycharm.

5.2 贝叶斯网络精确度分析

贝叶斯网络与原始数据的拟合度直接影响发布数据的可用性. 在贝叶斯网络结构学习中使用 K2[15] 算法中的评分函数确定网络结构的好坏, 本实验选择 K2Score 函数分别对 3 个算法生成的贝叶斯网络进行评分, 评分越高, 贝叶斯网络与原始数据拟合度越高. 其中由于 K2 函数公式特性计算网络评分值均为负值. 实验分别选取 1000、5000、10000、15000、20000、25000、30000 大小数据集对比 3 个算法生成的贝叶斯网络的精确度, 结果如图 2 所示.

从图 2 可以看出随着数据集不断增大, PPBA 算法生成的贝叶斯网络的精确性高于 PrivBayes 算法, 原因是随着数据集不断增大, 属性维度之间的依赖关系越来越复杂, 相较于加权 PrivBayes 算法和 PrivBayes 算法, PPBA 算法利用最大支撑树, 将指数机制属性关系对的挑选空间控制在较优的范围, 提高贝叶斯网络的精确度, 在数据集不断增大, 属性关系越来越复杂的情况下, 优势更为明显.

5.3 个性化分配隐私预算下数据可用性与数据安全性分析

PPBA 算法将实验数据集低维属性集合按敏感性由大到小排序, 取 q 值大小分别为 1.0、1.2、1.3、1.5、1.6、1.8、2.0. 观察取不同 q 值下, 将$\varepsilon = 0.5$的隐私预算分配给低维属性集合, 结果如图 3 所示. 图 3 横坐标为按敏感性由大到小进行排序的低维属性集合的属性节点, 1 为敏感性最高的低维属性集合的节点, 以此类推. 从图 3 看出, 在 q 值为 1.0 时各属性集合分配均等的隐私预算. 随着 q 值不断增大, 越敏感的属性集合分配的隐私预算越小, 对其隐私保护强度越大, 反之, 敏感性越小属性分配的隐私预算越大, 隐私保护强度越小. 从而实现隐私预算合理分配.

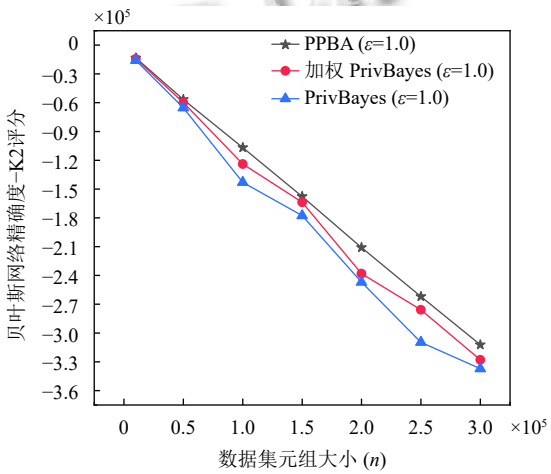

图 2 贝叶斯网络精确度对比图

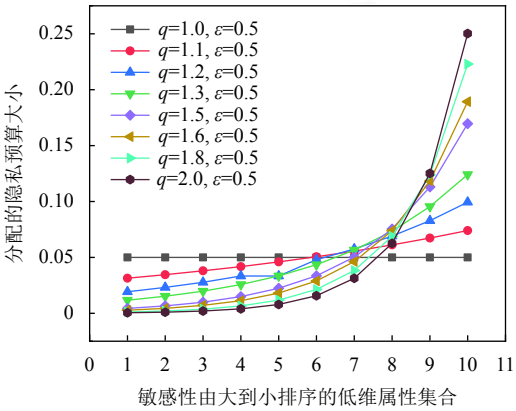

图 3 敏感性排序下为属性集合分配的隐私预算

发布数据集所需的可用性与安全性之间的个性化平衡是衡量隐私预算分配优劣极重要指标. 选取训练数据集大小分别为 1000、5000、10000、15000、20000、25 000、30 000 的数据, 使用加权 PrivBayes (ε= 1.0) 算法, PrivBayes (ε= 1.0) 算法, 以及q取值 1.0、1.1、1.2、1.3、1.5 下的 PPBA (ε= 1.0) 算法生成满足ε-差分隐私的合成发布数据集. 使用以上算法生成的合成

发布数据集训练 SVM 分类模型, 利用 SVM 分类模型[16]对测试数据集进行测试. 选取训练得到的 SVM 模型分类器对测试数据集中“Sex”属性进行分类. SVM 分类的结果以及q值分别选取 1.0、1.1、1.3、1.5 时通过 Laplace 方差计算隐私损失所得的隐私保护强度结果分别如图 4、图 5 所示. 从图 4 看出q值逐渐增大, 在数据集不大的情况下, 会出现 PPBA 算法 SVM 准确率低于加权 PrivBayes 算法和 PrivBayes 算法的现象, 但随着数据集的不断增大, PPBA 算法的分类准确率均高于加权 PrivBayes 算法和 PrivBayes 算法, 更进一步的说明 PPBA 算法更适用于高维数据集的情况下. 从图 5 看出q值越大, 隐私保护强度越高. 结合图 4、图 5, 根据用户对发布数据集安全性与可用性的需求, 当用户数据集元组大于 15000 的情况下, 对 SVM 分类准确率要求为 80% 与 82% 之间, 但同时要求隐私保护强度不低于 0.001% 与 0.002% 之间, 根据图 4, q取值 1.2 可以达到数据可用性与安全性的最优平衡需求. 当用户对隐私要求保护强度为 0.007% 与 0.008% 之间, 数据可用性需求为 79% 到 80% 之间, 结合图 4, 图 5, 可个性化设置q取值为 1.5. 从而证明 PPBA 算法可以根据用户需要满足数据可用性与隐私保护强度之间个性化选择的平衡.

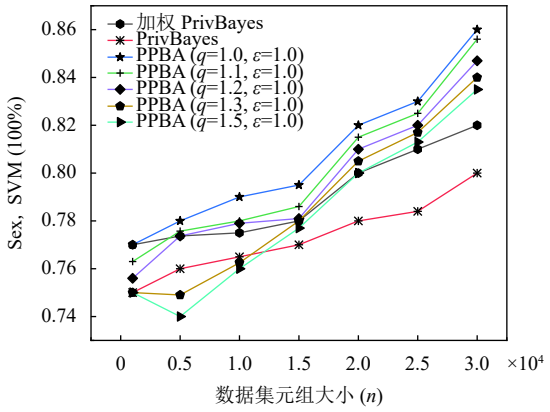

图 4 Sex 属性下 SVM 分类准确率

5.4 时间性能对比分析

在实验中, 将 PPBA 隐私保护算法 (ε= 1.0, q=1.0)、加权 PrivBayes 隐私保护算法 (ε= 1.0) 和 PrivBayes 隐私保护算法 (ε= 1.0) 在合成发布数据集过程中, 按照训练数据集由小到大进行运行时间对比分析. 由于加权 PrivBayes 隐私保护算法、PrivBayes 隐私保护算法随机生成贝叶斯网络, 运行时间具有不确定性, 实验选择每个数据集下运行 10 次取平均值的方式衡量时间性能. 对比分析结果如图 6 所示, PPBA 算法运行时间相对 PrivBayes 算法、加权 PrivBayes 算法时间更短, 究其原因 PPBA 算法利用属性节点权重值确定首节点, 最大支撑树确定最大父节点个数 K 值, 减少属性关系候选空间, 避免 K 值过大, 内存资源的浪费, 具有更优的时间性能. 但由于实验计算机性能有限, 数据预处理工作量大等问题, 整体耗时较长, 实验结果有待改进.

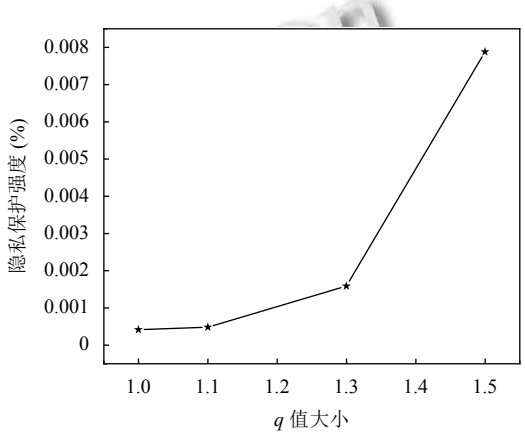

图 5 不同q值下隐私保护强度

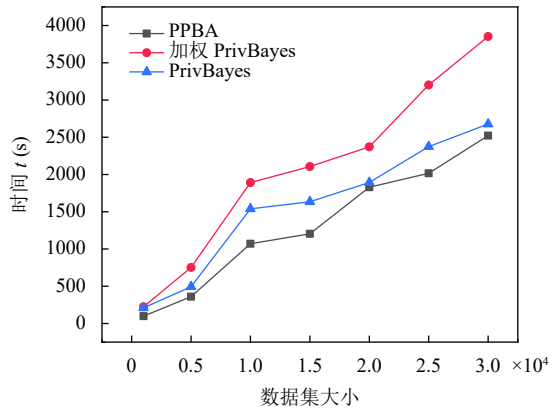

图 6 时间性能对比图

6 总结与展望

面向高维数据隐私发布, 不同数据发布用户对于数据安全性和可用性的个性化需求, 本文提出个性化差分隐私预算分配算法 (PPBA), 通过最大权重值和最大支撑树, 降低属性关系对的挑选空间, 构建更优的贝叶斯网络, 按照高维数据隐私保护强度和数据可用性间的平衡需要, 个性化设置比例常数q值,依据集合的敏感性排序, 为低维属性集合分配合理的隐私预算, 合成

发布满足差分隐私数据集. 通过实验验证 PPBA 算法形成的贝叶斯网络更优, 具有更低的时间复杂度, 且满足根据用户需求, 个性化实现隐私预算分配. 接下来的研究工作会围绕整个算法过程中差分隐私预算分配策略再利用, 延长隐私预算使用周期, 提高发布数据的可用性等问题进行研究.

参考文献

1 Palanisamy B, Li C, Krishnamurthy P. Group privacy-aware disclosure of association graph data. Proceedings of 2017 IEEE International Conference on Big Data. Boston, MA, USA. 2017. 1043–1052.

2 Zhou J, Dong XL, Cao ZF. Research advances on privacy preserving in recommender systems. Computer Research and Development, 2019, 56(10): 2033–2048.

3 Dwork C. Differential privacy: A survey of results. Proceedings of the 5th International Conference on Theory and Applications of Models of Computation. Xi'an, China. 2008. 1–19.

4 Xin BZ, Yang W, Geng YY, *et al*. Private FL-GAN: Differential privacy synthetic data generation based on federated learning. Proceedings of 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain. 2020. 2927–2931.

5 Zhang J, Cormode G, Procopiuc CM, *et al*. PrivBayes: Private data release via Bayesian networks. Proceedings of 2014 ACM SIGMOD International Conference on Management of Data. Snowbird, UT, USA. 2014. 1423–1434.

6 王良, 王伟平, 孟丹. 基于加权贝叶斯网络的隐私数据发布方法. 计算机研究与发展, 2016, 53(10): 2342–2352.

7 张啸剑, 陈莉, 金凯忠, 等. 基于联合树的隐私高维数据发布方法. 计算机研究与发展, 2018, 55(12): 2794–2809. [doi: 10.7544/issn1000-1239.2018.20170756]

8 Wang D, Xu JH. Differentially private high dimensional sparse covariance matrix estimation. arXiv: 1901.06413, 2019.

9 Xu CG, Ren J, Zhang YX, *et al*. DPPro: Differentially private high-dimensional data release via random projection. IEEE Transactions on Information Forensics and Security, 2017, 12(12): 3081–3093. [doi: 10.1109/TIFS.2017.2737966]

10 李效光, 李晖, 李凤华, 等. 差分隐私综述. 信息安全学报, 2018, 3(5): 92–104.

11 尚涛, 赵铮, 舒王伟, 等. 基于等差隐私预算分配的大数据决策树算法. 工程科学与技术, 2019, 51(2): 130–136.

12 汪小寒, 韩慧慧, 张泽培, 等. 树索引数据差分隐私预算分配方法. 计算机应用, 2018, 38(7): 1960–1966.

13 陈旋, 刘健, 冯新淇, 等. 基于朴素贝叶斯的差分隐私合成数据集发布算法. 计算机科学, 2015, 42(1): 236–238. [doi: 10.11896/j.issn.1002-137X.2015.01.052]

14 李燕. 基于香农熵和互信息的主题优化方法的研究 [硕士学位论文]. 大连: 大连海事大学, 2017.

15 李淑智, 刘弹, 张熠卓, 等. 贝叶斯网络结构评分函数研究. 2010 年全国模式识别学术会议 (CCPR2010) 论文集. 重庆, 中国. 2010.

16 Juan ROS, Kim J. Photovoltaic cell defect detection model based-on extracted electroluminescence images using SVM classifier. Proceedings of 2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC). Fukuoka, Japan. 2020. 578–582.