·安全技术·　　文章编号：1000-3428(2019)04-0124-06　　文献标志码：A　　中图分类号：TP309.2

基于贝叶斯网络与语义树的隐私数据发布方法

郝志峰[1,2]，王日宇[1]，蔡瑞初[1]，温　雯[1]

（1. 广东工业大学 计算机学院，广州 510006；2. 佛山科学技术学院 数学与大数据学院，广东 佛山 528000）

摘　要：为在隐私预算相同的条件下提高发布数据的可用性，在 PrivBayes 的基础上，提出一种改进的隐私数据发布方法 PrivBayes_Hierarchical。基于贝叶斯网络隐私数据发布方法的思想，引入语义树对含有层次关系的数据属性进行抽象，使用贝叶斯网络描述数据属性之间的依赖关系。利用格雷码减少随机噪声对数据精度的影响，并对贝叶斯网络结构学习方法进行优化，以减少不必要的隐私预算消耗，提高数据可用性。实验结果表明，该方法在公开数据集下可以获得比 PrivBayes 更高的数据精度，从而提升隐私数据集的可用性。

关键词：差分隐私；数据发布；贝叶斯网络；数据分析；隐私保护

中文引用格式：郝志峰，王日宇，蔡瑞初，等. 基于贝叶斯网络与语义树的隐私数据发布方法[J]. 计算机工程，2019，45(4)：124-129.

英文引用格式：HAO Zhifeng，WANG Riyu，CAI Ruichu，et al. Privacy data publishing method based on Bayesian network and semantic tree[J]. Computer Engineering，2019，45(4)：124-129.

Privacy Data Publishing Method Based on Bayesian Network and Semantic Tree

HAO Zhifeng[1,2]，WANG Riyu[1]，CAI Ruichu[1]，WEN Wen[1]

（1. School of Computers，Guangdong University of Technology，Guangzhou 510006，China；
2. School of Mathematics and Big Data，Foshan University，Foshan，Guangdong 528000，China）

【Abstract】In order to improve the availability of published data under the same privacy budget，an improved privacy data publishing method PrivBayes_Hierarchical is proposed based on PrivBayes. Based on the idea of Bayesian network privacy data publishing method，semantic tree is introduced to abstract the data attributes with hierarchical relationships，and Bayesian network are used to describe the dependencies between data attributes. Using Graycode to reduce the impact of random noise on the data accuracy and to optimize the Bayesian network structure learning method to reduce unnecessary privacy budget consumption and improve the availability of data. Experimental results show that this method can obtain higher data precision than PrivBayes for public data and improve the availability of private data sets.

【Key words】differential privacy；data publishing；Bayesian network；data analysis；privacy preserving

DOI：10.19678/j.issn.1000-3428.0050368

0　概述

近年来，随着互联网与信息技术的快速发展，各大企事业单位、商业公司以及公共事业平台等都积累了大量的各种类型的数据。合理并有效地利用这些数据，可以加快社会生产效率，提升商业收益。比如，互联网电商平台可以利用积累下来的商品数据和用户购买历史记录进行个性化推荐，不仅可以增强用户体验，还能提高平台收益。但是，在合理利用已有数据提升业务的同时，还应保证隐私数据不被泄露。由于很多数据属于个人隐私或商业机密，如医院积累的病人数据，若被不法分子获取，会对病人和医院造成不可估量的损失。因此，保护数据隐私不被窃取逐渐成为一个研究热点。

在早期的工作中，保护数据隐私的方法有：基于数据失真的发布技术[1-3]，基于数据加密的发布技术以及基于条件的限制发布技术[4-6]。数据失真通过对数据进行泛化或抑制实现，但对攻击者背景知识

基金项目：广东省自然科学基金(2014A030306004，2014A030308008)；广东省科技计划项目(2015B010108006，2015B010131015)；广东特支计划(2015TQ01X140)；广州市珠江科技新星(201610010101)；广州市科技计划项目(201604016075)。

作者简介：郝志峰(1968—)，男，教授，主研方向为信息安全、机器学习、人工智能；王日宇(通信作者)，硕士研究生；蔡瑞初，教授；温　雯，副教授。

收稿日期：2018-01-31　　**修回日期：**2018-03-13　　**E-mail：**zfhao@gdut.edu.com

有较强的假设。数据加密通过使用密码学对数据进行保护,但在通常情况下加密算法实现效率比较低,数据可用性也较差。限制发布根据数据特性和保护目标有选择性地发布局部敏感数据。文献[7-9]提出针对统计数据库隐私泄露问题的差分隐私技术。该技术不仅有严格的数学定义,还可以在允许攻击者拥有最大背景知识的前提下,向原有数据集插入少量的噪声,既能保证隐私数据的安全性,也能保证较高的数据可用性[10]。

针对隐私数据查询及发布问题,目前主要分为交互式与非交互式 2 种场景[11-12]。交互式场景是指通过开放数据查询接口,返回经过噪声处理之后的统计结果,用户不能直接得到整个数据集。非交互式场景指直接发布一个经过隐私保护技术处理之后的"净化"数据集,用户可以在数据集上自定义各种操作,不会受到交互式场景下查询接口的限制。

本文主要研究如何改进编码方式以减少随机噪声带来的数据偏差,将数据属性的语义层次关系引入到基于贝叶斯网络的隐私数据发布方法 PrivBayes[13] 中,提出一种语义树数据发布方法 PrivBayes_Hierarchical。

1　相关工作

在隐私保护研究的内容中,主要考虑 2 个问题:

1)数据的安全性,确保敏感数据不会被窃取和泄露。

2)数据的可用性,尽量减少噪声带来的误差,使其在后续分析工作中保持较高的精度。

数据集统计特征的发布方法,如直方图发布[14],只能描述数据集的部分特征,在使用场景上存在很多的局限性。基于匿名技术对敏感数据进行泛化和压缩的 k-anonymity 模型[5] 的改进以及扩展模型,如 l-diversity[6]、t-closeness[15] 和(α, k)-anonymity[16]等,相继被科研人员提出。这些传统隐私保护模型都是根据攻击方式的不同,对攻击者拥有的背景知识做了一定的假设,因此,模型的应用场景受到很多制约。并且这些隐私保护模型没有提供数学方法来论证其算法的隐私保护水平,也就无法对其进行严格的定量分析与评估。

差分隐私保护模型很好地克服了传统隐私保护方法的主要缺陷,不再对攻击者拥有的背景知识做任何假设,并使用严格的数学推理作为理论基础,对隐私保护模型进行量化,提供可靠的分析论证方法。基于数据失真的差分隐私保护技术通过向原始数据添加噪声来保护敏感信息,维持了原始数据统计特征的稳定性[17-18]。

1.1　差分隐私保护模型

若 D 为需要发布的敏感数据集,差分隐私保护模型使用随机算法 A 对 D 进行计算来产生输出数据集,且输出数据集不能泄露任何一条特定记录。

定义 1(ε-差分隐私[7]) 对于任意 2 个具有相同属性结构的数据集 D_1 和 D_2,D_1 和 D_2 之间只相差一条记录(邻近数据集)。设有随机算法 A,该算法所有可能的输出集合为 O,在数据集 D_1 和 D_2 的任意输出结果为 $S(S\subseteq O)$,若随机算法 A 满足:

$$\Pr[A(D_1)=S]\leqslant e^{\varepsilon}\times\Pr[A(D_2)=S] \tag{1}$$

其中,$\Pr[\cdot]$表示事件发生的概率,参数 ε 为隐私保护预算,则算法 A 即为满足 ε 差分隐私的随机算法。

从定义 1 可以看出,参数 ε 越小,输出结果相同的概率就越大,则隐私保护程度也就越高。若参数 $\varepsilon=0$,则对邻近数据集会有 2 个概率分布相同的结果。根据不同使用场景下对隐私保护程度的要求,ε 可以取不同的值,如 0.05、0.2、1.0 等。

差分隐私保护模型需要对原始数据插入适量的噪声,有很多噪声机制可以达到所需要的效果。对于连续性数据,通常使用 Laplace 机制,而对于离散型数据,则使用指数机制。

定义 2(敏感度[12]) 现有函数 $f:D\rightarrow R^d$,该函数将数据集 D 映射到维度为 d 的实数向量,函数 f 的敏感度定义为:

$$S(F)=\max_{D_1,D_2}\|f(D_1)-f(D_2)\|_1 \tag{2}$$

其中,D_1 和 D_2 为邻近数据,$\|f(D_1)-f(D_2)\|_1$ 为 $f(D_1)$ 与 $f(D_2)$ 的 1 阶范数距离。

从定义 2 可以看出,敏感度衡量了函数 f 在任意 2 个邻近数据集上的最大输出变化量。

定义 3(Laplace 机制[12]) 现有数据集 D,任意查询函数 $f:D\rightarrow R^d$,若算法 A 满足:

$$A(D)=f(D)+N \tag{3}$$

其中,$N\sim Lap(S(f)/\varepsilon)$,则随机算法 A 即为满足 ε-差分隐私的随机算法。

从定义 3 可以看出,该机制通过向真实数据插入服从 Laplace 概率分布的噪声实现差分隐私保护效果,加入的噪声量与敏感度 $S(f)$ 成正比,与隐私保护预算 ε 成反比。

定义 4(指数机制[19]) 现有数据集 D,任意查询函数 $f:D\rightarrow R$,打分函数 $q:(D\times R)\rightarrow O$。其中,打分函数 q 将查询函数 f 的值域 R 映射到实数集 O,f 的查询结果 $r(r\in R)$ 越接近真实值,获得分数 $o(o\in O)$ 越高。若算法 A 以正比于 $\exp\left(\frac{\varepsilon q(D,r)}{2S(q(D,r))}\right)$ 的

概率从值域 R 中选择并输出 r，则随机算法 A 即为满足 ε-差分隐私的随机算法。

从定义 4 可以看出，r 获得的分数越高，被选中作为最终输出的概率就会越大。与 Laplace 机制稍有不同，指数机制下使用的敏感度基于打分函数 q 来定义：

$$S(q)=\max_{D_1,D_2,r}\left|q(D_1,r)-q(D_2,r)\right| \tag{4}$$

1.2 贝叶斯网络

设任意数据集 D 的属性字段集合为 A，且集合大小为 b，即数据集 D 含有 b 个不同的属性字段。贝叶斯网络 N 是一个有向无环图，图的顶点代表数据集中的属性字段，顶点之间的有向边（弧）表示属性字段之间的依赖关系，可用来描述属性字段之间的概率分布关系。图 1 所示为一个定义在 5 个属性字段上的贝叶斯网络。

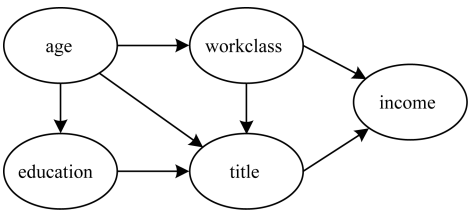

图 1 含有 5 个顶点的贝叶斯网络

定义 5（顶点的入度）在贝叶斯网络 N 中，以某顶点为弧头，终止于该顶点的弧的数目称为顶点的入度。

定义 6（顶点的出度）在贝叶斯网络中，以某顶点为弧尾，起始于该顶点的弧的数目称为顶点的出度。

对于贝叶斯网络结构的学习算法有 SGS 算法[20]、随机 K2 算法[21]和基于最大信息系数的贪婪算法[22]等。研究人员将差分隐私模型和贝叶斯网络相结合，提出了很多数据发布算法，如文献[13]利用贝叶斯网络构建属性字段的依赖关系，结合数据采样来生成“净化”数据集的 PrivBayes 算法；文献[23]通过分析贝叶斯网络节点的权重关系，对其改造后的加权贝叶斯隐私数据发布算法。这些数据发布算法都利用贝叶斯网络模型作为“净化”数据集产生的依据，但在面临稀疏数据集以及实数域上的连续属性字段时，这些算法并没有把这些因素导致的可用性下降考虑在内，从而在实际应用中会有一定的局限性。

2 基于语义树的数据发布算法

从定义 4 可以看出，在应用差分隐私保护框架时，对于离散型数据应使用指数机制，而指数机制的关键是构造一个合适的打分函数。理想的打分函数应满足在保护隐私的情况下，使查询结果与真实值偏差尽可能小。

文献[13]提出的 PrivBayes 敏感数据发布算法利用贝叶斯网络来表示属性字段之间的关系，并提出使用 F 函数作为打分函数，用来替代互信息函数，避免了由于互信息函数敏感度过高导致的插入噪声过大，从而影响数据可用性的问题。但 F 函数要求所有属性变量是二元的，故对于多类别离散属性和连续属性，需要对其做一定的编码变换才可使用。

2.1 编码方式

在现有的编码方法中，最常用的是二进制编码。具体来说，对于具有 n 个不同取值的离散型变量 X，把 X 转换为 $\lceil \mathrm{lb}(n) \rceil$ 个二进制变量 $X_1, X_2, \cdots, X_{\lceil \mathrm{lb}(n) \rceil}$，每个二进制变量 X_i 对应 X 取值二进制时表示的第 i 位。对于连续性变量 Y，首先对其进行离散化，将整个值域划分成 b 个小区间，然后使用与离散型变量 X 类似的方法，将 Y 转换为 $\lceil \mathrm{lb}(b) \rceil$ 个二进制变量。图 2 和图 3 分别为离散型变量和连续型变量的二进制编码。

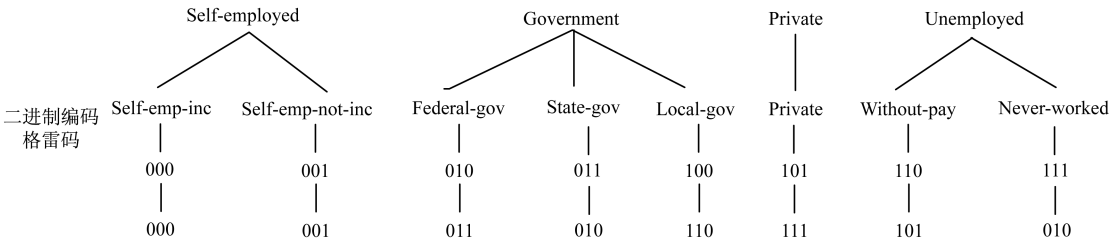

图 2 离散型变量“workclass”的编码

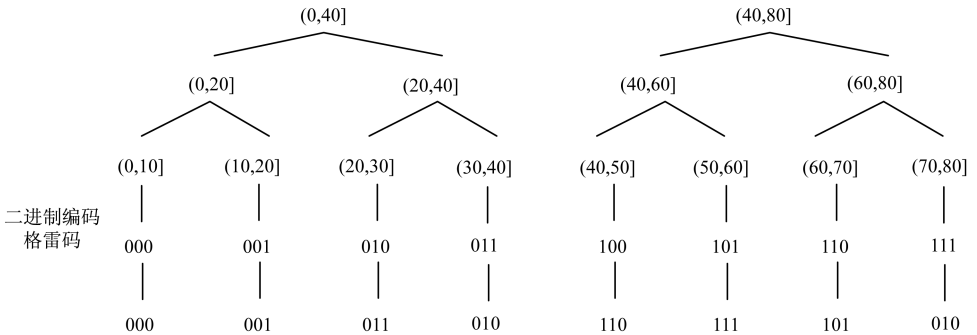

图 3 连续型变量“age”的编码

以图 3 为例,它给出了连续型变量“age”的二进制编码。首先将“age”的值域(0,80]平均切分成 8 个小区间,然后将每个小区间用 3 个二进制位进行编码,如用 000 表示第 1 个区间(0,10],用 001 表示第 2 个区间(10,20],以此类推。将所有属性字段进行二进制编码,原始数据集 D 则转换成了新的数据集 D_b,D_b 再经过 PrivBayes 隐私数据发布算法处理生成“净化”数据集 D_b^*,最后用同样的规则对 D_b^* 中所有属性字段进行二进制解码即可得到最终待发布的隐私数据集。

二进制编码方式为贝叶斯网络的构造带来了灵活性。例如,当且仅当连续属性“age”大于等于 60 时,二类属性“is retired”取值为真。PrivBayes 在构造贝叶斯网络阶段,只需要使用“age”二进制编码后属性的前 2 位即可。因为前 2 位已经将“age”离散化为 4 个区间(0,20]、(20,40]、(40,60]、(60,80],对于保留“is retired”和“age”的相关关系已经足够。

格雷码是二进制编码的一种变体。在一组数的编码中,若任意 2 个相邻的代码只有一位二进制数不同,另外最大数与最小数之间也仅一位数不同,即“首尾相连”,则称这种编码为格雷码。如图 2 和图 3所示,也分别给出了离散型变量和连续型变量的格雷编码,与二进制编码相比,仅是这些代码的顺序发生了变化。正是由于雷格码连续的代码中仅有一位二进制不同,因此将其应用在 PrivBayes 中,会对插入的噪声有更好的鲁棒性。

以图 3 所示的“age”属性为例,现有一个编码为 010 的数据记录,PrivBayes 产生的噪声改变了该代码的第 1 位或第 3 位,即噪音数据为 110 或 011。若采用二进制编码,对应真实数据所在区间为(20,30],噪音数据所在区间为(60,70]或(30,40];若采用格雷码,对应真实数据所在区间为(30,40],噪音数据所在区间为(40,50]或(20,30]。由此可以看出,采用格雷码可以使噪音数据具有更大的概率邻接真实数据,且具有更好的稳定性。

2.2　属性取值的层次关系

根据定义 3 和定义 4,差分隐私框架对原始数据加入的噪声量与查询(打分)函数的敏感度成正比,与隐私预算 ε 的大小成反比。

对于原始数据来说,加入的噪声会对其原始分布进行“抹平”。加入的噪声越大,“抹平”程度越高。具体来说,对于连续属性,噪声量越大,其分布越会趋于均匀分布;对于离散属性,噪声量越大,其取值分布越会趋于等概率随机抽样。以一个取值集合为{足球,排球,篮球,网球}的离散型属性为例,根据差分隐私框架的指数机制,不同 ε 大小的抽样概率示例如表 1 所示。

表 1　指数机制应用示例

项目	可用性	概率		
		$\varepsilon=0$	$\varepsilon=0.1$	$\varepsilon=1$
足球	30	0.25	0.424	0.924
排球	25	0.25	0.330	0.075
篮球	8	0.25	0.141	1.5E-05
网球	2	0.25	0.105	7.7E-07

在安全性较高的情形下,隐私预算往往会比较小,如何有效控制 ε 就成为重点。在真实数据集中,如图 2 和图 3,特别是连续型属性,可以根据其语义构成具有层次结构的树。树的叶子节点表示为所有真实取值,上层节点可以依据语义表示为对下一层节点的泛化。因此,上层节点本身拥有一定的噪音量,合理利用该特性可达到隐私需要和精度需求的平衡。

为利用这种语义树的层次关系,本文使用新的父节点集合搜索算法 MaximalParentSets 来构造贝叶斯网络,在此基础上进行 PrivBayes 后续的噪声干扰和数据抽样操作。MaximalParentSets 算法描述如下:

算法 1　MaximalParentSets 算法

输入　候选属性集合,属性个数 k

输出　父节点集合

```
if V = ∅ then return {∅};
pick an arbitrary attribute X from V;
initialize S = ∅, U = ∅
for i = 0 to height(X) - 1 do {
foreach Z ∈ MaximalParentSets(V\{X}, k - 1) do{
    if Z ∈ U then continue;
    add Z to U; add Z∪{X^(i)} to S;
}
}
foreach Z ∈ MaximalParentSets(V\{X}, k) do {
if Z ∈ U then continue;
add Z to S;
}
return S;
```

MaximalParentSets 会递归地计算贝叶斯网络单个节点所有可能的父节点集的集合,并满足 k 度贝叶斯网络的定义,即每个父节点集合的元素数量不大于 k。与 PrivBayes 不同,本文将父节点的语义层级对孩子节点的影响考虑在内,从而平衡数据的可用性和安全性。经过 MaximalParentSets 算法处理后的集合内,每个节点包含各自的层次信息,通过这种隐含的层次信息来减少由于隐私预算 ε 不足而导致插入噪声过大带来的误差。

3 实验结果与分析

本文使用 UCI 公开机器学习数据集 Adult 进行实验评估,该数据集由美国 1994 年人口普查的 45 222 条个人信息数据组成,包含 15 个属性字段,其中有 5 个字段属于连续型属性。

实验使用支持向量机(Support Vector Machine,SVM)算法对 Adult 数据集进行分类任务以评估隐私保护算法的数据精度,目标变量为 salary 属性。实验使用原始数据的 80% 作为训练集,20% 作为测试集。在训练集上分别使用 PrivBayes 和改进后的 PrivBayes_Hierarchical 算法进行数据隐私保护操作得到仿真数据集,然后使用 SVM 算法对仿真数据集分别进行训练得到隐私保护算法对应的 SVM 模型,最后将训练出来的 SVM 模型应用到测试集上,记录其相应的分类正确率。为减少误差,实验使用交叉验证并将平均分类正确率作为数据精度的评估标准。

实验结果如图 4 所示,横轴代表隐私预算大小,纵轴表示 SVM 分类准确率。从图 4 可以看出,在隐私预算较小时,利用语义树的层次结构并结合格雷码的 PrivBayes_Hierarchical 方法,可以显著提高数据可用性,获得比 PrivBayes 方法更理想的结果。

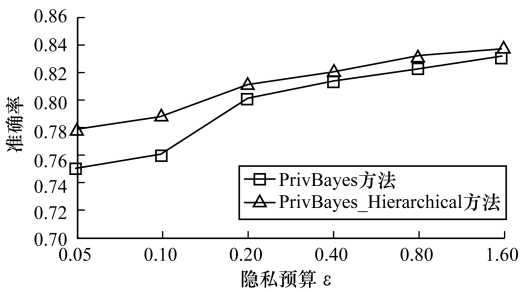

图 4 2 种方法的 SVM 分类准确率

对于数据集中连续型属性的编码处理,本文有 2 种区间切分方法,一种是根据区间大小均匀划分(Range),另一种是根据数据分布均匀划分(Distribute)。例如,将一个值域为[0,100)的连续型属性划分成10 个小区间,第1 种方法可划分为[0,10),[10,20),[20,30),…,[90,100)。第 2 种由于需要考虑数据在区间内的分布情况,划分时应尽量保证落入每个小区间的数据个数相等,结果可能为[0,5),[5,20),[20,50),…,[90,100)。理论上,在切分个数相同的情况下,隐私预算越高,2 种切分方法产生的结果越相近。

2 种区间切分方法实验对比结果如图 5 所示,横轴代表隐私预算大小,纵轴表示编码后的数据与原始数据的距离。由图 5 可以看出,按照数据分布均匀划分区间的方式在隐私预算紧张的情况下,数据可用性高于根据区间大小均匀划分的方法。

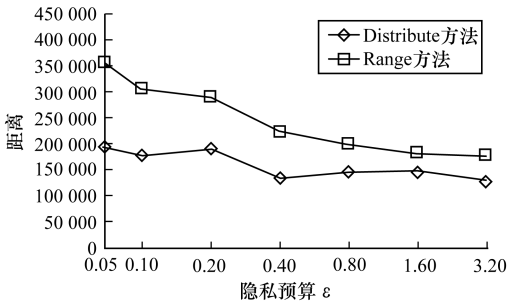

图 5 不同区间划分方法距离差异对比

4 结束语

本文基于贝叶斯网络数据发布算法的思想,对 PrivBayes 方法进行扩展,提出一种改进的隐私数据发布方法 PrivBayes_Hierarchical。利用具有层次语义关系的属性特征,减少隐私保护后的数据误差。同时,改进数据编码方式,提高数据集的可用性。实验结果表明,该方法可以在有限的隐私预算条件下,获得更高的数据精度,并提升数据集的可用性。下一步将对该方法的数据集应用范围进行扩展,如针对包含空值较多的稀疏数据集,考虑空值对贝叶斯网络构造的影响,以及在空值较多的情形下算法是否具有鲁棒性,以保证隐私数据集的可用性。

参考文献

[1] MURALIDHAR K, SARATHY R. Security of random data perturbation methods[J]. ACM Transactions on Database Systems, 1999, 24(4): 487-493.

[2] KARGUPTA H, DATTA S, WANG Q, et al. On the privacy preserving properties of random data perturbation techniques[C]//Proceedings of the 3rd IEEE International Conference on Data Mining. Washington D. C., USA: IEEE Press, 2003: 99.

[3] CHEN K, LIU L. Privacy preserving data classification with rotation perturbation[C]//Proceedings of the 5th IEEE International Conference on Data Mining. Washington D. C., USA: IEEE Computer Society, 2005: 589-592.

[4] XIAO X, TAO Y. Personalized privacy preservation[C]//Proceedings of 2006 ACM SIGMOD International Conference on Management of Data. New York, USA: ACM Press, 2006: 229-240.

[5] SWEENEY L. *k*-anonymity: a model for protecting privacy[J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10(5): 557-570.

[6] MACHANAVAJJHALA A, GEHRKE J, KIFER D. ℓ-diversity: privacy beyond *k*-anonymity[C]//Proceedings of the 22nd International Conference on Data Engineering. Washington D. C., USA: IEEE Press,

2006:24.
[7] DWORK C. Differential privacy[C]//Proceedings of the 33rd International Conference on Automata, Languages and Programming. Berlin, Germany: Springer, 2006: 1-12.
[8] DWORK C. Differential privacy: a survey of results[C]//Proceedings of the 5th International Conference on Theory and Applications of Models of Computation. Berlin, Germany: Springer, 2008: 1-19.
[9] DWORK C. A firm foundation for private data analysis[M]. New York, USA: ACM Press, 2011.
[10] 张啸剑,孟小峰. 面向数据发布和分析的差分隐私保护[J]. 计算机学报,2014,37(4):927-949.
[11] YAROSLAVTSEV G, PROCOPIUC C M, CORMODE G, et al. Accurate and efficient private release of datacubes and contingency tables[C]//Proceedings of 2013 IEEE International Conference on Data Engineering. Washington D. C., USA: IEEE Computer Society, 2013: 745-756.
[12] DWORK C, MCSHERRY F, NISSIM K. Calibrating noise to sensitivity in private data analysis[M]. Berlin, Germany: Springer, 2006: 637-648.
[13] ZHANG J, CORMODE G, PROCOPIUC C M, et al. PrivBayes: private data release via bayesian networks[M]. New York, USA: ACM Press, 2014.
[14] XU J, ZHANG Z, XIAO X, et al. Differentially private histogram publication[J]. VLDB Journal, 2013, 22(6): 797-822.
[15] LI N, LI T, VENKATASUBRAMANIAN S. t-closeness: privacy beyond k-anonymity and l-diversity [C]// Proceedings of 2007 IEEE International Conference on Data Engineering. Washington D. C., USA: IEEE Press, 2007: 106-115.
[16] WONG R C, LI J, FU A W, et al. (α, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing[C]//Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM Press, 2006: 754-759.
[17] FRIEDMAN A, SCHUSTER A. Data mining with differential privacy[C]//Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM Press, 2010: 493-502.
[18] 熊平,朱天清,王晓峰. 差分隐私保护及其应用[J]. 计算机学报,2014,37(1):101-122.
[19] MCSHERRY F, TALWAR K. Mechanism design via differential privacy [C]//Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science. Washington D. C., USA: IEEE Press, 2007: 94-103.
[20] SPIRTES P, GLYMOUR C, SCHEINES R. Causality from probability [EB/OL]. [2017-12-30]. https://philpapers.org/rec/SPICFP-2.
[21] COOPER G F, HERSKOVITS E. A Bayesian method for the induction of probabilistic networks from data [J]. Machine Learning, 1992, 9(4): 309-347.
[22] 曾千千,曾安,潘丹,等. 基于最大信息系数的贝叶斯网络结构学习算法[J]. 计算机工程,2017,43(8):225-230.
[23] 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[J]. 计算机研究与发展,2016,53(10):2343-2353.

编辑　司森森

(上接第 123 页)

[3] 蔡铭,谢晓玲,王雪畅,等. 智能电网脆弱性分析及对策研究[J]. 信息工程大学学报,2013,14(3):376-379.
[4] LIANG G, WELLER S R, ZHAO J, et al. The 2015 Ukraine blackout: implications for false data injection attacks [J]. IEEE Transactions on Power Systems, 2017, 32(4): 3317-3318.
[5] Communication networks and systems for power utility automation: IEC/TR 61850 [EB/OL]. [2017-11-20]. https://webstore.iec.ch/publication/6028.
[6] LIN H, SLAGELL A, SAUER P W, et al. Semantic security analysis of SCADA networks to detect malicious control commands in power grids [C]//Proceedings of the 1st ACM Workshop on Smart Energy Grid Security. New York, USA: ACM Press, 2013: 29-34.
[7] LIN H, SLAGELL A, KALBARCZYK Z, et al. Runtime semantic security analysis to detect and mitigate control-related attacks in power grids [J]. IEEE Transactions on Smart Grid, 2018, 9(1): 163-178.
[8] PERDISCI R, ARIU D, FOGLA P, et al. McPAD: a multiple classifier system for accurate payload-based anomaly detection [J]. Computerand Telecommunications Networking, 2009, 53(6): 864-881.
[9] 费金龙,王禹,王天鹏,等. 基于云模型的网络异常流量检测[J]. 计算机工程,2017,43(1):178-182.
[10] 凌骏,尹博学,李晟,等. 基于监控数据的 MySQL 异常检测算法[J]. 计算机工程,2015,41(11):41-46.
[11] MANNING C D, SCHÜTZE H. Foundations of statistical natural language processing [M]. Cambridge, USA: MIT Press, 1999.
[12] PACCANARO A, HINTON G E. Learning distributed representations of concepts using linear relational embedding [J]. IEEE Transactions on Knowledge and Data Engineering, 2001, 13(2): 232-244.
[13] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality [C]//Proceedings of the 26th International Conference on Neural Information Processing Systems. [S. l.]: Curran Associates Inc., 2013: 3111-3119.
[14] LIU F T, KAI M T, ZHOU Z H. Isolation Forest [C]// Proceedings of the 8th IEEE International Conference on Data Mining. Washington D. C., USA: IEEE Press, 2008: 413-422.
[15] Supervisory control and data acquisition: IEC 60870 [EB/OL]. [2017-11-20]. https://webstore.iec.ch/publication/3755.

编辑　赵　辉