智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

差分隐私的高维数据发布研究综述

张兴， 陈昊

引用本文:
张兴, 陈昊. 差分隐私的高维数据发布研究综述[J]. 智能系统学报, 2021, 16(6): 989–998.
ZHANG Xing, CHEN Hao. A research review of high-dimensional data publishing based on a differential privacy model[J]. *CAAI Transactions on Intelligent Systems*, 2021, 16(6): 989–998.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202104023

DOI: 10.11992/tis.202104023
网络出版地址: https://kns.cnki.net/kcms/detail/23.1538.tp.20210910.0938.002.html

差分隐私的高维数据发布研究综述

张兴, 陈昊
(辽宁工业大学 电子与信息工程学院, 辽宁 锦州 121001)

摘 要: 大数据时代的到来, 使得信息量暴增的同时, 数据维度也呈现几何式增长。在保护用户隐私的前提下, 如何充分挖掘高维数据的可用信息, 成为了大数据发布领域的研究热点和难点。差分隐私作为一种强大的隐私保护模型, 被越来越多地应用到高维数据发布中。本文归纳了差分隐私及其相关方法在高维数据发布的应用, 重点分析了差分隐私和特征降维、特征抽取、贝叶斯网络、树模型以及最新提出的粗糙集和随机投影等方法在高维数据发布中结合应用的优缺点, 梳理了各个方法在高维数据方面的应用和对比, 最后对未来差分隐私在高维数据发布中的应用方向进行了展望。

关键词: 大数据发布; 隐私保护; 数据挖掘; 高维数据; 特征降维; 贝叶斯网络; 粗糙集; 随机投影; 差分隐私
中图分类号: TP309.2 **文献标志码**: A **文章编号**: 1673−4785(2021)06−0989−10

中文引用格式: 张兴, 陈昊. 差分隐私的高维数据发布研究综述 [J]. 智能系统学报, 2021, 16(6): 989–998.
英文引用格式: ZHANG Xing, CHEN Hao. A research review of high-dimensional data publishing based on a differential privacy model[J]. CAAI transactions on intelligent systems, 2021, 16(6): 989–998.

A research review of high-dimensional data publishing based on a differential privacy model

ZHANG Xing, CHEN Hao
(School of Electronics & Information Engineering, Liaoning University of Technology, Jinzhou 121001, China)

Abstract: With the advent of the era of big data, the amount of digitally-generated information has increased dramatically, and the data dimension has also shown geometric growth. How to fully mine high-dimensional data while maintaining the user's privacy has become a focus and a difficult research topic in the field of big data publishing. As a powerful privacy protection model, differential privacy is increasingly in use in high-dimensional data publishing. This paper summarizes the application of differential privacy and its related methods in high-dimensional data publishing, focusing on an analysis of the advantages and disadvantages of differential privacy and feature dimension reduction, feature extraction, the Bayesian network, tree model, and the latest rough set and random projection methods in high-dimensional data publishing. Moreover, we survey the application and comparison of each method in high-dimensional data and finally discuss the future application of differential privacy in high-dimensional data publishing.

Keywords: big data publishing; privacy protection; data mining; feature dimension reduction; bayesian network; rough set; random projection; high dimensional data; differential privacy

在大数据时代的背景下, 数据呈现大量 (数据量大)、多样 (数据样式多)、高速 (数据产生迅速)、低价值密度 (数据有用性低) 等特点, 数据维度也出现井喷式剧增, 同时也带来了一些隐私安全问题。2020 年 4 月, Cyble 安全公司报道称有约 2.67 亿 Facebook 用户信息被黑客盗取, 并公开在暗网销售, 造成大量用的敏感信息泄露, 同年 6 月, 我国郑州某民办高校约 2 万名学生身份信息泄露, 多名学生及家长接到大量骚扰和威胁电话。2021 年 1 月 8 日, 某国外论坛公开贩卖国内某银行 1 679 万条数据, 并公开大量敏感数据样本。同年 3 月, 安全研究人员发现某印度政府网站上存在安全问题导致孟加拉邦百万人次核酸检测结果泄露。报告中含有姓名、年龄、居住地址、检测时间、婚姻状况等大量敏感信息。大量的例

收稿日期: 2021−04−12. **网络出版日期**: 2021−09−10.
基金项目: 国家自然科学基金项目 (61802161); 辽宁省教育厅科学研究经费项目 (JZL202015402, JZL202015404).
通信作者: 张兴. E-mail: 1123361380@qq.com.

子都说明大数据时代的背景下，数据安全问题变得尤为突出。

隐私数据泄露的同时也加剧了隐私保护的发展，相对于传统的数据匿名化保护方法 (如 k-匿名[1], l-多样性[2], t-紧密性[3], (α,k)-匿名[4]) 存在不能抵御全部背景知识攻击[5] 的缺点，差分隐私因具有严格的数学定义和逻辑证明更加受到人们的广泛关注。差分隐私在隐私保护数据发布 (privacy-preserving data publishing, PPDP) 应用广泛，可以对隐私进行量化分析，在降低数据泄露风险的同时能更好地保障数据的效用性。

大数据呈现数据量大、数据样式多、数据产生迅速、数据有用性低特点，数据规模越来越大，数据相比于以往更加复杂多变，高维数据存在“维度灾难”，导致传统的分析手段失效，对于高维数据的挖掘变得更加困难。为了解决上述问题，常利用数据降维的思想，使高维数据映射到低维空间，在低维空间进行数据分析。本文综述了常见的对于高维数据的研究和差分隐私的结合方法，保证了在挖掘高维数据的同时可以降低数据泄露的风险。最后对未来的研究方向进行展望。

1　特征降维和差分隐私

特征降维[6-7] 是处理高维数据最常用的手段，如图 1 所示，一般特征降维分为特征抽取与特征选择两方面。前者是将原始数据的特征重新组合，生成包含原始信息的更精炼更具有代表性的特征；后者则是对数据进行筛选，筛选出原始数据的部分特征，生成特征子集，特征子集反映出原始数据的绝大部分规律，子集的数据结构更加清晰，便于分析。

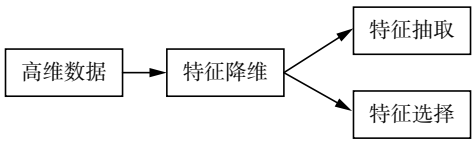

图 1　特征降维

Fig. 1　Feature dimension reduction

而差分隐私是通过对输出的结果添加适当的噪声 (多数为拉普拉斯机制、高斯机制和指数机制) 实现对原始数据集的扰动，从而实现数据的隐私保护。差分隐私严格的数学定义如定义 1。

定义 1　ε-差分隐私 (ε-differentail prvac)：对兄弟数据集 D_1 和 D_2(相差一条记录)，及其所有子集 S，满足随机算法 A：

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

则算法 A 满足 ε-差分隐私，其中 e 为自然对数，ε 为隐私预算 (衡量隐私保护程度)。一般情况下，实现差分隐私可以通过拉普拉斯机制 (数值型) 和指数机制 (离散型)。因为两种机制都依靠全局敏感度，故先定义全局敏感度。

定义 2　全局敏感度 (global sensitivity)：给定查询函数 $f:D\to R^d$，D 为输入数据集，R^d 为输出数据集。在任意一对 D_1 和 D_2 上，函数 f 的全局敏感度为

$$\Delta f=\max_{D_1,D_2}\|f(D_1)-f(D_2)\|_1$$

其中 $\|f(D_1)-f(D_2)\|_1$ 是 $f(D_1)$ 和 $f(D_2)$ 间的一阶距离。全局敏感度度量了当输入数据集变化时，对相应输出结果的影响。有了全局敏感度的概念后，给出 Laplace 机制和指数机制定义。

定义 3　拉普拉斯机制 (Laplace mechanism)：给定数据集 D 和隐私预算 ε，函数 f 的全局敏感度为 Δf，当 f 的输出满足：

$$A(D)=f(D)+\mathrm{Lap}(\Delta f/\varepsilon)$$

则称算法 A 满足 ε-差分隐私，其中 $\mathrm{Lap}(\Delta f/\varepsilon)$ 为满足 Laplace 分布的随机噪声，如图 2 所示。

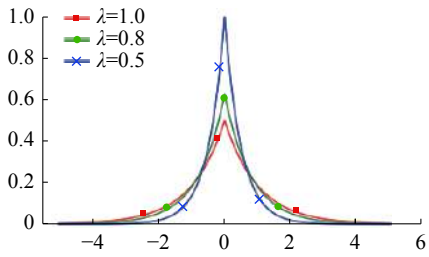

图 2　拉普拉斯分布

Fig. 2　Laplace distributions

定义 4　指数机制 (index mechanism)：给定数据集 D，输出为一个实体对象 $\gamma\in R$，$\mu(D,R)$ 为可用性函数，$\Delta\mu$ 为函数 $\mu(x,R)$ 的敏感度，若以正比于 $\exp\left(\dfrac{\varepsilon\cdot u(x,r)}{2\,\Delta\,u}\right)$ 的概率从输入中选择并输出 r，则算法 A 满足 ε-差分隐私。其中 $\Delta u=\max\limits_{r\in R,\ \|x-y\|_1\leqslant 1}|u(x,r)-u(y,r)|$。

经过特征降维后的数据，仍具有原始数据的绝大部分信息，再经过差分隐私处理后用于数据发布，在保证数据分析有效性的同时，可以大大提高数据的隐私保护程度。

1.1　基于线性特征抽取的差分隐私

特征抽取[8] 一般分为线性特征抽取和非线性特征抽取两类。由于线性特征抽取应用较多，这里则着重介绍。它是根据数据之间的线性关系进行数据的特征转换，产生的新特征复杂度更小，维数更低。

代表的方法是主成分分析方法[9]，由于差分

隐私直到 2006 年才由 Dwork 首次提出[10-12], 所以主成分分析差分隐私算法近几年才慢慢成为研究的热点, 它采用高维数据的主成分降维与加噪机制融合的方式, 有效对高维数据进行隐私保护的同时还可以保证数据的高可用性。2012 年, Chaudhuri 等[13] 首次提出可用于数据发布的隐私保护算法, 该算法采用指数机制来添加噪音实现了 PCA 满足差分隐私的要求, 相较于子线性查询 (SULQ) 框架, 采用此种方式可以得到数据的更多方差。但该算法缺乏收敛时间保证。针对这个问题, 2013 年, Kapralov 等[14] 在此基础上提出一种用于低秩近似矩阵的混合规划算法, 解决了因收敛时间保证从而造成隐私泄露的问题, 但该算法时间复杂度较高, 在用于高维数据发布时执行效率较低。

而 Jiang 等[15] 则是采用拉普拉斯机制来实现主成分分析差分隐私保护, 它将一部分隐私预算去构建噪声协方差矩阵, 剩余的隐私预算用于对投影矩阵的加噪。戚名钰等[16] 和徐亚红等[17] 分别在 2017 年和 2018 年对此方法进行改进, 两者同样采用拉普拉斯机制, 直接分解原始数据的协方差矩阵, 然后在投影矩阵上加噪音, 相比原来更为简便。但前者针对数据发布后面临的数据挖掘工作, 又提出了一种基于线性判别分析的差分隐私数据发布算法, 该算法是针对分类问题的优化, 提高了发布数据分类精度和高可用性。后者通过设计的 3 种不同的加噪方式, 得出了均分加噪效果更好的结论。

2020 年, 彭长根等[18] 根据传统的主成分分析差分隐私算法大多只能用于捕获线性关系, 提出最大信息系数 (MIC) 实现主成分分析的差分隐私算法 (MIC-PCA-DP), 该算法利用 MIC 构建主成分分析, 在捕获线性关系的同时也可以捕获非线性关系和多函数关系, 利用特征值占比加噪的方式对降维后的数据进行合理加噪, 优化了噪声的分配方式, 使算法执行效率进一步提升。2021 年, 顾贞等[19] 提出概率主成分分析的差分隐私数据发布方法, 即由概率主成分分析的生成模型生成数据集发布, 使得发布的数据集满足差分隐私。该方法在 SVM 分类准确率方面可以保持良好的数据效用。

1.2　基于特征选择的差分隐私

特征选择[20] 是数据预处理的一种常用手段, 它能有效降低高维数据的特征维数, 过滤重复数据和噪声数据, 输出满足要求的特征子集。特征选择通常包含特征子集生成、特征子集评价、停止准则和验证过程 4 个过程, 如图 3 所示。

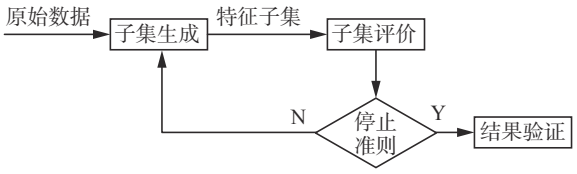

图 3　特征选择

Fig. 3　Feature selection

2013 年, 万文强[21] 首次将差分隐私机制融入到特征选择之中, 将差分隐私与基于统计理论 (基尼指数和误分类增益) 的特征选择相结合, 并采用分布式计算框架 (Map-Reduce) 实现分布式环境下的差分隐私特征选择方法。DPDFS 算法能在选取重要特征的同时, 有效保护用户数据隐私性不被泄露。

但上述算法并未考虑进行差分隐私的先后顺序对数据安全性的影响, 且对于隐私预算 ε 的选取比较随机, 效率较低。DPDFS 算法只适用于数据的预处理阶段, 对后续的数据挖掘 (分类和聚类等) 并不适用。下述方法是针对上述问题的解决方案。

2014 年, Yang 等[22], 首先验证了基于差分隐私的集成特征选择算法的性能, 得出在相同的环境下, 先加入差分隐私保护的效果优于先进行特征选择的效果, 并利用 Objective Perturbation 策略增加了特征选择算法的隐私保护强度。

2018 年, 高原秀男[23] 针对概率攻击问题, 提出来基于特征选择和离群点检测的差分隐私算法, 并针对医疗数据进行分析实验, 在满足安全性要求的前提下, 提高数据发布的可用性。DPFSOD 是一种聚类算法, 可以用于数据分析阶段, 且对于隐私预算 ε 的选择, 采用了数据集分簇后最小簇中元组数目相结合的方式, 得到隐私预算 ε 值的取值范围。2018 年, 刘中锋[24], 针对多值的特征选择, 提出了基于局部学习的带有输出干扰策略的差分隐私集成特征选择算法, 该算法采用了基于输出干扰的策略 (向输出结果添加噪声) 来保护隐私。并证明了局部差分隐私特征选择算法性能优于全局差分隐私特征选择算法。

以上提出的 3 种方法, 均从不同角度改进了 DPDFS 算法的不足, 在同一数据集上, 均能达到更优的隐私保护效果。

2　贝叶斯网络和差分隐私

2.1　贝叶斯网络模型

1988 年, 慕春棣等[25] 首次提出贝叶斯网络 (Bayesian network) 的概念, 其本质是一种概率图模型, 即通过先验知识, 建立随机变量间的依赖

关系，进而求取条件概率。贝叶斯网络是由结点(代表变量) 和连接结点的边 (代表依赖关系) 组成，本质为一个有向无环图 (directed acyclic graph, DAG)，可以有效处理变量间的依赖关系。图 4 是一个简单的贝叶斯网络模型，其中包含 6 个属性结点和 5 条有向边。

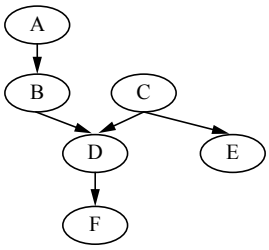

图 4　贝叶斯网络模型

Fig. 4　Bayesian network model

相比于特征降维方法，贝叶斯网络用图模型来表示数据间的依赖关系，更加直观且更容易被理解。高维数据因具有“维度灾难”和数据稀疏性等特点，使得数据的效用性很低，使得一般的隐私保护方法失去原有的保护效果。但贝叶斯网络能更容易确保数据属性间的一致性和完备性，因此在高维数据中有更广泛的应用空间，同时贝叶斯网络采用属性节点间的互信息大小来表示属性间的依赖程度，它将先验知识和样本知识结合，更加适用于稀疏性的高维数据。

2.2　基于贝叶斯网络的差分隐私

对高维数据的隐私保护方法研究中，很多没有考虑“维度灾难”问题直接加噪；也有一部分研究是对高维数据降维再对降维后的数据集注入噪声，但是这些方法仍然存在一些问题。为解决这些问题，2014 年，Zhang 等[26] 提出了一种基于贝叶斯网络的差分隐私保护发布方法 (PrivBayes)。该方法利用构建的贝叶斯网络实现对高维数据的降维，然后采用拉普拉斯机制对构建好的贝叶斯网络进行加噪，实现了差分隐私保护。但在构建贝叶斯网络时，首个属性字段节点选择方式随机，这种构造方式导致构造出来的贝叶斯网络具有不确定性，容易遭到攻击者模拟攻击，造成隐私的泄露。由于指数机制对属性对进行选择，随着属性对的增多，选择精度会明显下降。针对属性对选择随机的问题，2016 年王良等[27] 提出了一种基于加权贝叶斯网络的隐私数据发布方法 (加权 PrivBayes)，该方法通过属性字段结点在发布数据集中的重要程度，为每一个属性字段结点增加一个权重值。并将 K_2 评分函数和互信息相结合，所构建的贝叶斯网络模型的分类精度更高。进一步优化属性字段结点加入噪声的顺序，相比于 PrivBayes 模型在发布数据集精确性和算法效率性上有明显提升。

上述方法同样存在构建的贝叶斯网络不唯一，精度不够高的问题，如果对原始数据集多次构建，可能会泄露隐私。同时该方法的隐私预算需要手动分配，有可能导致噪声的加入不合理，对数据的效用性和隐私性有影响。下面的研究就上面出现的问题给出了较合理的解决方案。

2017 年汤诗一[28] 提出利用小波变换的方式实现差分隐私下的数据发布，同时用改良的 F 函数 (捕获更多的互信息量) 替换互信息，得到更准确的依赖关系，从而使构建的贝叶斯网络的精度更高。算法的核心思想是使用低维边缘分布去刻画贝叶斯网络，随后对低维边缘分布中的每一项均添加相同的拉普拉斯噪声，抽样形成带噪声的合成数据集 D^*。虽然小波变换实现了范围查询精度的提升，但该算法只能在单机环境下运行，且未将敏感字段作为第一考虑对象，导致算法普适性不足。

针对 PriveBayes 算法中加入噪声不合理，影响数据效用性的问题，2019 年 Li 等[29] 提出 SS-PrivBayes(smooth sensitivity privacy Bayes) 算法，该算法引入了平滑敏感度的概念，即在分析实际的数据集的同时考虑局部数据集变化，从而得到不同属性敏感度，有利于在进行差分隐私时减少噪声的摄入，来提高数据发布的效用性。并针对 PriveBayes 算法在构建贝叶斯网络搜索空间效率低的问题，提出了一种高效的贝叶斯网络搜索空间算法 PBCPC(privacy Bayesian candidate parents and children)，该算法通过启发式方法得到目标变量的候选父子集，在属性数量过多时，算法运行效率明显高于 PriveBayes。2018 年，Wei 等[30] 针对高维数据差分隐私发布中存在的数据稀疏性问题，提出基于马尔可夫网的高维数据差分隐私发布的方法 (PrivMN)，即采用马尔可夫模型度量属性间的依赖性，然后结合图形近似推理算法计算高维数据差分隐私的分布，该算法能有效降低噪声摄入量，同时还解决了在精确推理中存在的计算复杂性高的问题。

针对现有的在基于构建贝叶斯网络差分隐私模型中，存在的贝叶斯网络不唯一和隐私预算分配不合理的情况，2019 年唐雨薇[31] 提出了一种优化的贝叶斯差分隐私 (OBDP) 模型。该模型通过互信息构建网络、评分函数和组合优化算法得到候选稀疏节点集，采用爬山算法计算两个节点的净增益，确保构建的网络是唯一的。同时划分敏

感属性和非敏感属性及不同的属性簇, 对不同的属性簇添加不同程度的噪声, 从而确保隐私运算分配更合理。

上述方法从不同角度优化了贝叶斯网络, 但都不适用于群智感知系统的高维数据发布, 因此, 任雪斌等[32] 针对群智感知系统采用本地差分隐私保护方法, 即用感知服务器接收用户的隐私信息, 然后计算其联合概率分布和互信息值从而构建出贝叶斯网络模型, 最后按不同属性信息降维, 合成新的数据集, 经过大量仿真实验对比, 验证了该合成数据的有效性。2020 年, 肖彪等[33] 利用属性段首选机制并利用聚类的方法取代等宽法对数据进行离散化处理, 该算法避免了对首个属性段属性的随机化选择, 提高了数据的可用性。

3　树模型和差分隐私

树模型[34], 即通过构建特征子树实现高维数据的降维。最早, 在 PrivBayes 的基础上 Chen 等[35] 在 2015 年提出一种基于抽样推理的差分隐私发布方法 (JTree 算法), 该方法采用基于抽样的框架和阈值机制相融合的方式构建属性依赖图, 然后利用联合树最小化误差, 在依赖图中识别出一组边缘表逼近联合分布, 最终实现高维数据发布。

但上述方法中构建的团图容易出现局部最优。针对这些问题, 2018 年张啸剑等[36] 提出了一种基于联合树的差分隐私高维数据发布算法 (PrivHD 算法), 该算法利用马尔可夫网络对高维数据进行降维并结合高通滤波的方式对生成的属性对进行过滤, 最后采用联合树生成完全团图, 该过程符合差分隐私要求, 在数据查询和分类上均高于 JTree 算法。同年, 苏炜航等[37] 采用隐树模型对高维数据的关联属性对进行过滤, 该算法 (LTMDP 算法) 在生成隐变量过程中通过互信息值提前分组, 在分组过程中引入指数机制使该过程满足差分隐私, 在隐树结构学习中引入拉普拉斯机制并采用自顶向下的采样方式生成高维数据集, 该算法采用隐树模型在执行效率和分析精度上高于 PrivBayes 算法。2019 年, 郝志峰等[38] 引入格雷码和语义树对 PrivBayes 算法改进, 提出了基于语义树和贝叶斯网络的发布算法 (PrivBayes-Hierarchical 算法), 该算法用格雷码使插入的噪声鲁棒性更好, 提高数据的高可用性, 同时采用语义层次树减少数据误差, 提高分类精度。

上述 3 种方法均采用不同的树模型, 对从不同角度对 PrivBayes 算法进行改进, 因此在执行效率上均优于传统的 PrivBayes 算法。

针对在高维数据中搜索日志泄露的问题, 陆叶等[39] 利用前缀树的思想, 提出一种满足差分隐私的前缀树方法, 该算法通过前缀树把项集中小于 k 的剪枝, 然后关联频繁项, 对用户的 ID 进行差分隐私保护, 保证了用户的敏感信息不被泄露, 但该算法执行效率较低。同样采用前缀树的实例, 2017 年王晓男[40] 针对多维顺序数据, 采用改良的前缀树对数据集采用分层加噪, 减少了隐私预算的消耗, 在降低相对误差的同时使模型的执行效率和安全性得到提升。2020 年, 邓蔚等[41] 采用指数机制和 Laplace 机制分别处理连续型特征和离散型特征并准确找到最佳分裂特征和分裂点,采用等差预算分配加噪策略。该算法充分利用了隐私预算, 提高了分类准确度。

4　粗糙集及随机投影的差分隐私

4.1　基于粗糙集的差分隐私

随着数据维度的增加, 数据之间关联程度也会提高, 而差分隐私对于关联数据的保护水平往往很低, 原因是维数的增加导致噪声的增多, 如果采用传统的加噪方式, 可能会破坏属性间的关联性, 从而降低数据的效用性。虽然很多方法针对这个问题, 采用互信息去计算属性间的相关性, 然后对不同数据采用分别加噪的方式去降低噪声的添加量, 但对敏感数据和非敏感数据的区分往往存在较大误差。

粗糙集理论是由 Pawlak 教授[42] 提出的, 它是用于处理不确定关系的一种数学工具。

定义 5　决策系统 (decision system): 在粗糙集理论中, 知识的表达是用决策系统表示的, 一个决策系统由一个四元组表示:

$$S=(U,A,V,f)$$

式中: U 为非空有限集, 称为论域; A 为非空有限属性集, 由条件属性集 C 和决策属性集 D 组成; $V=U_{\alpha\in A}V_{\alpha}$, V_{α} 是属性 α 的值域; $f: U\times A\to V$ 是信息函数 (样本空间到属性空间) 数据属性间的关联性, 可以用粗糙集中属性依赖度 k 进行衡量 (k 为 1 表示 D 完全依赖 C), 然后再针对不同的属性添加合理的噪声。属性集 D 在 C 上的依赖度公式为

$$k=\frac{\overset{\circ}{a}_{X\in U/D}\left|\underline{C}(X)\right|}{|U|}$$

Pawlak 属性约简算法[43] 可以有效把高维数据中不相关的冗余信息剔除, 得到分类规则。该算法的主要思想是求出核属性并启发式扩展, 最后得到一个约简。但该方法计算核属性过程较繁琐, 因此王一斌[44] 利用差别矩阵来构成差分函

数，通过析取和合取运算去简化区分函数，从而可以快速得到所有的约简。2017 年孙志鹏[45]针对高维数据的属性约简利用粗糙集理论去改进 K 均值算法，提出自适应的粗糙集 K 均值算法，该算法有效提高了聚类的精度。2019 年 Li 等[46]针对医疗数据，首次将差分隐私和粗糙集提取规则相结合，提出了一种挖掘医疗数据中隐藏模式、保障患者隐私的新方法 (DPRers)。该算法利用拉普拉斯机制在数据挖掘过程中增加噪声，同时使数据的效用最大化。2019 年 Luo 等[47]利用粗糙集对高维数据的降维的同时，引入信息熵衡量数据的敏感度达到更合理的加噪方式，提高了数据的效用性。

粗糙集具有的优势和高维数据的特点十分契合，比普通的降维方法有着天然的优势，因此两者的结合往往能够起到很好的隐私保护效果。

4.2　基于随机投影的差分隐私

随机投影[48]能有效解决高维数据“维度灾难”问题。它的理论依据是 Johnson-Lindenstrauss 引理。

定义 6　Johnson-Lindenstrauss 引理：任意给定一个整数 $d>0$，存在 $\varepsilon\in(0,1)$ 和 $\delta<1/2$，$k=O\left(\frac{1}{\varepsilon^2}\log\left(\frac{1}{d}\right)\right)$，存在一个映射 $f:\mathbf{R}^d\to\mathbf{R}^k$，$x\in\mathbf{R}^d$，

$$P_r\left[\left|\|f(x)\|_2^2-\|x\|_2^2\right|\geqslant\varepsilon\|x\|_2^2\right]\leqslant\delta$$

Johnson-Lindenstrauss 引理[49]表明将 d 维空间的 n 个点映射到 k 维 ($k<d$)，可以保证原空间中任意两点间的距离是近似不变的。因此通过随机投影技术能够实现高维数据的有效降维，同时经过数据扰动后也能有效保留数据的可用性。

当投影矩阵构建完成后，投影矩阵泄露可能也会导致数据的泄露。针对这个问题，2013 年杨静等[50]利用哈希函数产生随机投影矩阵并引入安全子空间去保障构建投影矩阵的安全性。同年，针对高维数据的稀疏性问题，Hardt 等[51]利用随机投影方法使输入矩阵满足差分隐私的低秩近似逼近，同时给出误差边界，并针对误差边界为空的情况，通过幂迭代法替换原来生成的特征向量，增强扰动可靠性的同时，该方法还能减少由于添加噪声所引起的误差边界，消除了数据矩阵秩具有的依赖关系。2014 年赵家石[52]就高维稀疏数据易遭到攻击者通过重建原始数据，导致隐私泄露的问题，采用噪声投影扰动和差分隐私结合，即先进行投影转换，在此基础上添加噪声 (相互独立的随机数)，可有效确保扰动数据的可用性。

两者都是针对高维数据的稀疏性问题，但从不同角度给出了不同的处理方案，共同点都是采用投影转换的思考方式。

近年，在保护数据据隐私的前提下去发布高维数据集引起了数据库界越来越多的关注。2017 年 Xu 等[53]针对现有的差分隐私无法解决由于干扰误差和计算复杂度增加对高维数据发布失效的问题，提出了一种基于随机投影的数发布算法 (DPPro)，主要思想采用随机投影将数据集从高维空间投影到随机选择的低维空间中，以保持欧式距离不变，从而使用户根据距离进行分割。然后对每个矢量添加高斯噪声，这样，可以在最小化添加噪声量的同时最大化提高效用性，确保发布数据的隐私性。针对高维数值型数据 (多个数值型属性)，在处理较大属性个数时误差较大的问题，2020 年孙慧中等[54]满足本地差分隐私的高维数值型数据收集算法 (Multi-RPHM),与 DPPro 算法不同之处的，用户自己通过随机投影对原始高维数据降维，然后发送给数据收集者，数据收集者再进行维度还原，生成高维扰动数据集用于数据发布。

5　方法对比

本节将以上提到的技术及算法进行汇总比较，列举出了各种方法代表算法的优缺点，如表 1 所示，基于特征抽取、贝叶斯网络、树模型、粗糙集、投影和差分隐私结合的方法比较。

为了更直观比较各个技术的内在联系和主要区别，给出各个算法的针对不同场景的对比表格，包括 PPCA 和 DPPro、PrivBayes 和 DPRers、PrivMN 和 PrivHD 在不同场景下的优缺点和数据可用性分析，如表 2 所示。

表 1　降维方法比较

Table 1　Comparison of dimensionality reduction methods

方法	代表算法	优点	缺点
线性特征抽取	PPCA MIC-PCA-DP	实现方便简单，有多重变体和扩展方法	特征值分解存在局限性，在非高斯分布得到主元可能不是最优解
特征选择	DPDFS DPFSOD	能有效排除非关键特征噪声	需手动设置调整相关阈值，算法时间复杂度高

续表 1

方法	代表算法	优点	缺点
贝叶斯网络	PrivBayes PrivMN	对稀疏性问题有较好的处理效果, 对缺损数据不太敏感, 隐私保护性强	有时构建网络不唯一, 依赖先验概率
树模型	JTree PrivHD	可解释性强、直观, 能有效处理不相关特征	容易过拟合, 团图有时会出现局部最优
粗糙集	DPRers	不需附加先验知识, 能有效区分不同属性重要度, 更合理添加噪声	不能用于连续数据, 要先离散化
投影技术	DPPro Multi-RPHM	短时高效把高维数据映射成低维	有时变换矩阵的生成未考虑数据的固有结构, 导致误差较大

表 2　降维算法比较

Table 2　Comparison of dimensionality reduction algorithms

算法	针对场景	数据可用性
PPCA	高维数据线性结构较明显, 没有多函数性结构	通过主成分分析实现降维, 能保留绝大部分信息数据可用性较强
DPPro	线性结构不明显, 存在多函数性结构, 需要更快速降维	通过随机投影变化实现降维, 投影矩阵的选择至关重要, 误差有时较大
PrivBayes	依赖互信息选择属性重要度, 要求属性可以区分	构建贝叶斯网络, 实现高维数据降维, 能较好地保留原始数据的信息, 可用性较好
DPRers	不依赖先验条件, 通过粗糙集区分属性重要度	对离散型数据可用性强, 对连续型数据可用性较差
PrivMN	高维数据的稀疏性问题, 采用马尔可夫网实现降维	通过构建马尔可夫网度量属性的依赖关系, 减少噪声的输入, 数据可用性强
PrivHD	针对贝叶斯网络构建团图过程中容易出现局部最优的情况	通过马尔可夫网降维的同时, 应用高通滤波和联合树优化生成的团图, 数据可用性较强

6　结束语

大数据时代下, 对高维数据的挖掘比以往任何时候都更加迫切, 采用非常规手段往往会导致用户隐私的泄露。差分隐私是一种被严格证明的隐私保护模型, 可有效抵制各种安全攻击。因此把差分隐私应用到高维数据上可以有效实现数据的隐私安全保障。本文就高维数据的差分隐私模型作出详细归纳总结, 包括将特征降维、贝叶斯网络、树模型等方法与其结合应用, 分析了不同方法的优缺点和应用场景。

未来, 数据将会更加复杂多变, 因此必定会产生更多问题。

1) 属性关联性强。数据属性之间往往存在关联性, 随着数据维度的增加, 属性间的关联性更加复杂。近期提出一些研究方法, 如通过粗糙集的属性依赖度去度量属性的关系程度, 然后针对不同属性添加不同的噪声。但相关研究的文章仍然较少, 未来有较大的空间亟待研究。

2) 算法复杂度高。目前, 差分隐私在高维数据的应用大多停留在实验阶段, 往往伴随着较高的时间复杂度, 导致算法的效率较低, 无法高效地应用到生产实践中。因此, 亟需对算法进行优化, 以满足实际应用的需要。

参考文献:

[1] SWEENEY L. 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]//Proceedings of the 22nd International Conference on Data Engineering. Atlanta, USA, 2006: 24.

[3] LI Ninghui, LI Tiancheng, VENKATASUBRAMANIAN S. *t*-closeness: Privacy beyond *k*-anonymity and *l*-diversity[C]//2007 IEEE 23rd International Conference on Data Engineering. Istanbul, Turkey: IEEE, 2007: 106−115.

[4] WONG R C W, LI Jiuyong, FU A W C, 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, 2006: 754−759.

[5] 杨高明, 杨静, 张健沛. 聚类的 (α,k)-匿名数据发布 [J]. 电子学报, 2011, 39(8): 1941–1946.
YANG Gaoming, YANG Jing, ZHANG Jianpei. Achieving (α,k)-anonymity via clustering in data publishing[J]. Acta electronica sinica, 2011, 39(8): 1941–1946.

[6] 胡洁. 高维数据特征降维研究综述 [J]. 计算机应用研究, 2008, 25(9): 2601–2606.
HU Jie. Survey on feature dimension reduction for high-dimensional data[J]. Application research of computers, 2008, 25(9): 2601–2606.

[7] 史庆伟, 从世源, 唐晓亮. LSI__LDA: 一种混合特征降维方法 [J]. 计算机应用研究, 2017, 34(8): 2269–2273.
SHI Qingwei, CONG Shiyuan, TANG Xiaoliang. LSI_LDA: mixture method for feature dimensionality reduction[J]. Application research of computers, 2017, 34(8): 2269–2273.

[8] 吴晓婷, 闫德勤. 数据降维方法分析与研究 [J]. 计算机应用研究, 2009, 26(8): 2832–2835.
WU Xiaoting, YAN Deyue. Analysis and research on method of data dimensionality reduction[J]. Application research of computers, 2009, 26(8): 2832–2835.

[9] 杜子芳. 多元统计分析 [M]. 北京: 清华大学出版社, 2016: 240−241.

[10] DWORK C. Differential privacy[C]//Proceedings of the 33rd International Colloquium on Automata, Languages, and Programming. Venice, Italy: Springer, 2006: 1−12.

[11] DWORK C, MCSHERRY F, NISSIM K, et al. Calibrating noise to sensitivity in private data analysis [C]//Proceedings of Theory of Cryptography Conference. New York, USA: Springer, 2006: 265−284.

[12] DWORK C. Differential privacy: A survey of results[C]// Proceedings of the 5th International Conference on Theory and Applications of Models of Computation. Berlin, Heidelberg: Springer, 2008: 1−19.

[13] CHAUDHURI K, SARWATE A D, SINHA K. A near-optimal algorithm for differentially-private principal components[J]. Journal of machine learning research, 2013, 14(1): 2905–2943.

[14] KAPRALOV M, TALWAR K. On differentially private low rank approximation[C]//Proceedings of the twenty-fourth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA 2013). New York, USA: Society for Industrial and Applied Mathematics, 2013: 1395−1414.

[15] JIANG Xiaoqian, JI Zhanglong, WANG Shuang, et al. Differential-private data publishing through component analysis[J]. Transactions on data privacy, 2013, 6(1): 19–34.

[16] 戚名钰, 黄刘生, 陆潇榕, 等. 采用成分分析的差分隐私数据发布算法 [J]. 小型微型计算机系统, 2017, 38(3): 437–443.
QI Mingyu, HUANG Liusheng, LU Xiaorong, et al. Differential privacy data publish algorithm with compont analysis[J]. Journal of Chinese computer systems, 2017, 38(3): 437–443.

[17] 徐亚红, 杨庚, 白云璐, 等. 面向主成分分析的差分隐私数据发布算法 [J]. 网络空间安全, 2018, 9(10): 74–82.
XU Yahong, YANG Geng, BAI Yunlu, et al. A differential privacy data publishing algorithm for principal component analysis[J]. Cyberspace security, 2018, 9(10): 74–82.

[18] 彭长根, 赵园园, 樊玫玫. 基于最大信息系数的主成分分析差分隐私数据发布算法 [J]. 信息网络安全, 2020, 20(2): 37–48.
PENG Changgen, ZHAO Yuanyuan, FAN Meimei. A differential private data publishing algorithm via principal component analysis based on maximum information coefficient[J]. Netinfo security, 2020, 20(2): 37–48.

[19] 顾贞, 张国印, 马春光, 等. 基于概率主成分分析的差分隐私数据发布方法 [J]. 哈尔滨工程大学学报, 2021, 42(8): 1217−1223.
GU Zhen, ZHANG Guoyin, MA Chunguang, et al. Differential privacy data publishing method based on the probabilistic principal component analysis[J]. Journal of Harbin Engineering University, 2021, 42(8): 1217−1223

[20] 姚旭, 王晓丹, 张玉玺, 等. 特征选择方法综述 [J]. 控制与决策, 2012, 27(2): 161–166, 192.
YAO Xu, WANG Xiaodan, ZHANG Yuxi, et al. Summary of feature selection algorithms[J]. Control and decision, 2012, 27(2): 161–166, 192.

[21] 万文强. 分布式的隐私保护特征选择研究 [D]. 南京: 南京邮电大学, 2013.
WAN Wenqiang. Privacy preserving feature selection in distributed environment[D]. Nanjing: Nanjing University of Posts and Telecommunications, 2013.

[22] YANG Jun, LI Yun. Differentially private feature selection[C]//2014 International Joint Conference on Neural Networks (IJCNN). Beijing: IEEE, 2014: 4182−4189.

[23] 高原秀男. 数据发布中的隐私保护关键技术研究 [D]. 北京: 北京邮电大学, 2018.
GAO Yuanxiunan. Research on the key technologies of privacy preserving data publishing[D]. Beijing: Beijing University of Posts and Telecommunications, 2018.

[24] 刘中锋. 基于局部学习的差分隐私集成特征选择算法 [J]. 计算机技术与发展, 2018, 28(10): 79–82.
LIU Zhongfeng. An ensemble feature selection algorithm with differential privacy based on local learn-

ing[J]. Computer technology and development, 2018, 28(10): 79–82.

[25] 慕春棣, 戴剑彬, 叶俊. 用于数据挖掘的贝叶斯网络 [J]. 软件学报, 2000, 11(5): 660–666.
MU Chundi, DAI Jianbin, YE Jun. Bayesian network for data mining[J]. Journal of software, 2000, 11(5): 660–666.

[26] ZHANG Jun, CORMODE G, PROCOPIUC C M, et al. PrivBayes: private data release via Bayesian networks [C]//Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data. New York, USA, 2014: 1423−1434.

[27] 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.
WANG Liang, WANG Weiping, MENG Dan. Privacy preserving data publishing via weighted Bayesian networks[J]. Journal of computer research and development, 2016, 53(10): 2343–2353.

[28] 汤诗一. 基于贝叶斯网络差分隐私发布算法的研究 [D]. 大连: 大连海事大学, 2017.
TANG Shiyi. The research on data publication algorithms satisfy in differential privacy[D]. Dalian: Dalian Maritime University, 2017.

[29] LI Mingzhu, MA Xuebin. Bayesian networks-based data publishing method using smooth sensitivity[C]//2018 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Ubiquitous Computing & Communications, Big Data & Cloud Computing, Social Computing & Networking, Sustainable Computing & Communications (ISPA/IUCC/BDCloud/SocialCom/SustainCom). Melbourne, Australia: IEEE, 2018: 795−800.

[30] WEI Fengqiong, ZHANG Wei, CHEN Yunfang, et al. Differentially private high-dimensional data publication via Markov network[C]//International Conference on Security and Privacy in Communication Systems. Singapore, Singapore, 2018: 133−148.

[31] 唐雨薇. 高维数据的优化贝叶斯差分隐私方法研究 [D]. 桂林: 广西师范大学, 2019.
TANG Yuwei. Research on the optimization of Bayesian differential privacy method for high-dimensional data[D]. Guilin: Guangxi Normal University, 2019.

[32] 任雪斌, 徐静怡, 杨新宇, 等. 基于 Bayes 网络的高维感知数据本地隐私保护发布 [J]. 中国科学:信息科学, 2019, 49(12): 1586–1605.
REN Xuebin, XU Jingyi, YANG Xinyu, et al. Bayesian network-based high-dimensional crowdsourced data publication with local differential privacy[J]. Science China:Information Science, 2019, 49(12): 1586–1605.

[33] 肖彪, 闫宏强, 罗海宁, 等. 基于差分隐私的贝叶斯网络隐私保护算法的改进研究 [J]. 信息网络安全, 2020, 20(11): 75–86.
XIAO Biao, YAN Hongqiang, LUO Haining, et al. Research on improvement of Bayesizan network privacy protection algorithm based on differential privacy[J]. Information network security, 2020, 20(11): 75–86.

[34] 裘国永, 张娇. 基于二分 K-均值的 SVM 决策树自适应分类方法 [J]. 计算机应用研究, 2012, 29(10): 3685–3687, 3709.
QIU Guoyong, ZHANG Jiao. Adaptive SVM decision tree classification algorithm based on bisecting K-means [J]. Application research of computers, 2012, 29(10): 3685–3687, 3709.

[35] CHEN Rui, XIAO Qian, ZHANG Yu, et al. Differentially private high-dimensional data publication via sampling-based inference[C]//Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM, 2015: 129−138.

[36] 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 [J]. 计算机研究与发展, 2018, 55(12): 2794–2809.
ZHANG Xiaojian, CHEN Li, JIN Kaizhong, et al. Private high-dimensional data publication with junction tree[J]. Journal of computer research and development, 2018, 55(12): 2794–2809.

[37] 苏炜航, 程祥. 一种基于隐树模型的满足差分隐私的高维数据发布算法 [J]. 小型微型计算机系统, 2018, 39(4): 681–685.
SU Weihang, CHENG Xiang. Latent tree model based differentially private high-dimension data publishing algorithm[J]. Journal of Chinese computer systems, 2018, 39(4): 681–685.

[38] 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.
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.

[39] 陆叶, 卢菁. 基于差分隐私与前缀树的搜索日志隐私保护研究 [J]. 小型微型计算机系统, 2016, 37(3): 540–544.
LU Ye, LU Jing. Differential privacy and prefix tree based research for search log privacy protection[J]. Small and micro computer systems, 2016, 37(3): 540–544.

[40] 王晓男. 多维数据发布的差分隐私保护系统的研究与实现 [D]. 北京: 北京邮电大学, 2017.
WANG Xiaonan. Research and implementation of dif-

ferential privacy protection system for Multidimensional data publishing[D]. Beijing: Beijing University of Posts and Telecommunications, 2017.

[41] 邓蔚, 陈秀婷, 张清华, 等. 基于树模型的差分隐私保护算法 [J]. 重庆邮电大学学报(自然科学版), 2020, 32(5): 848–856.
DENG Wei, CHEN Xiuting, ZHANG Qinghua, et al. Differential privacy protection algorithms based on tree model[J]. Journal of Chongqing University of Posts and Telecommunications (natural science edition), 2020, 32(5): 848–856.

[42] PAWLAK Z. AI and intelligent industrial applications: the rough set perspective[J]. Cybernetics and systems, 2000, 31(3): 227–252.

[43] 王国胤, 姚一豫, 于洪. 粗糙集理论与应用研究综述 [J]. 计算机学报, 2009, 32(7): 1229–1246.
WANG Guoyin, YAO Yiyu, YU Hong. A survey on rough set theory and applications[J]. Chinese journal of computers, 2009, 32(7): 1229–1246.

[44] 王一斌. 基于属性重要度算法改进及应用 [D]. 西安: 西安科技大学, 2015.
WANG Yibin. Algorithm improvement and application based on attribute signifficance[D]. Xi'an: Xi'an University of Science and Technology, 2015.

[45] 孙志鹏. 高维数据聚类算法的研究及应用 [D]. 无锡: 江南大学, 2017.
SUN Zhipeng. Research and application of clustering algorithm on the high dimensional datasets[D]. Wuxi: Jiangnan University, 2017.

[46] LI Xianxian, LUO Chunfeng, LIU Peng, et al. Injecting differential privacy in rules extraction of rough set[C]// The International Conference on Healthcare Science and Engineering. Singapore: Springer, 2019, DOI: 10.1007/978-981-13-6837-0_13.

[47] LI Xianxian, LUO Chunfeng, LIU Peng, et al. Information entropy differential privacy: a differential privacy protection data method based on rough set theory[C]// 2019 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech). Fukuoka, Japan, 2019: 918–923.

[48] 张嵩, 景华炯. 基于 Log-Gabor 特征的非局部均值去噪算法及其加速方案研究 [J]. 模式识别与人工智能, 2015, 28(3): 266–274.
ZHANG Song, JING Huajiong. Log-gabor feature-based nonlocal means denoising algorithm and its acceleration scheme[J]. Pattern recognition and artificial intelligence, 2015, 28(3): 266–274.

[49] DASGUPTA S, GUPTA A. An elementary proof of the Johnson-Lindenstrauss Lemma[J]. Random structures and algorithms, 1999, 22(1).

[50] 杨静, 赵家石, 张健沛. 一种面向高维数据挖掘的隐私保护方法 [J]. 电子学报, 2013, 41(11): 2187–2192.
YANG Jing, ZHAO Jiashi, ZHANG Jianpei. A privacy preservation method for high dimensional data mining [J]. Acta electronica sinica, 2013, 41(11): 2187–2192.

[51] HARDT M, ROTH A. Beyond worst-case analysis in private singular vector computation[C]//Proceedings of the Forty-fifth Annual ACM Symposium on Theory of Computing. New York, USA: ACM, 2013: 331−340.

[52] 赵家石. 基于随机投影数据扰动的隐私保护技术研究 [D]. 哈尔滨: 哈尔滨工程大学, 2014.
ZHAO Jiashi. Research on privacy-preservation technique based on random projection data perturbation[D]. Harbin: Harbin Engineering University, 2014.

[53] XU Chugui, REN Ju, ZHANG Yaoxue, et al. DPPro: differentially private high-dimensional data release via random projection[J]. IEEE transactions on information forensics and security, 2017, 12(12): 3081–3093.

[54] 孙慧中, 杨健宇, 程祥, 等. 一种基于随机投影的本地差分隐私高维数值型数据收集算法 [J]. 大数据, 2020, 6(01): 3–11.
SUN Huizhong, YANG Jianyu, CHENG Xiang, et al. A high-dimensional numeric data collection algorithm for local difference privacy based on random projection[J]. Big data research, 2020, 6(01): 3–11.

作者简介：

张兴，教授，主要研究方向为大数据安全与隐私保护。获授权或公开发明专利 10 项，发表学术论文 60 余篇。

陈昊，硕士研究生，主要研究方向为大数据安全与隐私保护。