计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2018.20170756
55(12):2794-2809,2018

基于联合树的隐私高维数据发布方法

张啸剑[1] 陈 莉[2] 金凯忠[1] 孟小峰[3]
[1](河南财经政法大学计算机与信息工程学院 郑州 450002)
[2](河南财经政法大学网络信息安全研究所 郑州 450046)
[3](中国人民大学信息学院 北京 100872)
(xjzhang82@ruc.edu.cn)

Private High-Dimensional Data Publication with Junction Tree

Zhang Xiaojian[1], Chen Li[2], Jin Kaizhong[1], and Meng Xiaofeng[3]
[1](*College of Computer & Information Engineering, He'nan University of Economics and Law, Zhengzhou* 450002)
[2](*Institute of Network Information Security, He'nan University of Economics and Law, Zhengzhou* 450046)
[3](*School of Information, Renmin University of China, Beijing* 100872)

Abstract The problem of differentially private data publishing has attracted considerable research attention in recent years. The current existing solutions, however, cannot effectively handle the release of high-dimensional data. That is because these methods suffer from curse of dimensionality and various domain sizes, which will lead to the lower utility of publication. To address the problems, this paper presents PrivHD (differentially private high dimensional data release) with junction tree, a differentially private method for publishing high-dimensional data. PrivHD firstly generates a Markov network with exponential mechanism, which employs the high-pass filter technique to reduce the candidate space in the sampling process. After that, based on the network, PrivHD obtains a complete cluster graph in terms of full triangulation and node elimination, and then relies on the cluster graph and maximum spanning tree method to construct a differentially private junction tree. Finally, PrivHD uses the post-processing technique to boost the noisy counts of marginal tables in each cluster in junction tree, and based on the boosted result, PrivHD produces the high-dimensional synthetic dataset. PrivHD is compared with the existing approaches such as PrivBayes, JTree on the different real datasets. The experimental results show that PrivHD is better than its competitors on k-way query and SVM classification.

Key words high-dimensional data; differential privacy; Markov network; junction tree; marginal distribution

摘 要 基于差分隐私的数据发布已得到研究者的广泛关注.然而,现有的发布方法却不能有效地处理高维数据,其原因在于维度灾难和值域多样会引入极大的噪音值,进而使得发布结果的可用性比较低.

收稿日期:2017-10-11;**修回日期**:2018-03-20
基金项目:国家自然科学基金项目(61502146,91646203,91746115,61772131);河南省自然科学基金项目(162300410006);河南省科技攻关项目(172102310713);河南省教育厅高等学校重点科研项目(16A520002);河南财经政法大学青年拔尖人才资助计划项目
This work was supported by the National Natural Science Foundation of China (61502146, 91646203, 91746115, 61772131), the Natural Science Foundation of He'nan Province of China (162300410006), the Key Technologies Research and Development Program of He'nan Province (172102310713), the Research Program of the Higher Education of He'nan Provincial Education Department (16A520002), and the Young Talents Fund of He'nan University of Economics and Law.

基于此,提出一种基于联合树的隐私高维数据发布方法 PrivHD(differentially private high dimensional data release),该方法通过指数机制构造 Markov 网,引入满足差分隐私的高通滤波技术缩减指数机制搜索空间.结合充分三角化操作和顶点消除操作对 Markov 网分割来获得完全团图,采用最大生成树方法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.基于真实的高维数据集比较 PrivHD 算法与 PrivBayes(private Bayesian network),JTree(junction tree)算法的精度,实验结果表明:PrivHD 算法的 k-way 查询和 SVM(support vector machine)分类精度优于同类算法.

关键词 高维数据;差分隐私;Markov 网;联合树;边缘分布

中图法分类号 TP392

近年来,基于差分隐私的数据发布已成为隐私保护领域重要研究点.目前大多数研究均聚焦于低维数据或者一维数据的发布.高维数据在现实生活中普遍存在,例如个人购物数据、移动轨迹数据等.通过对这些高维数据发布与分析可以产生更加有意义的模式.由于高维数据通常蕴含大量个人隐私信息(例如购物记录中的敏感商品),直接发布将泄露个人敏感信息.然而如何在高维数据中得到统计结果的同时,又要保护原始数据隐私是个非常具有挑战性问题.其主要原因是随着维度与维度值域的增加,所形成的发布空间呈指数增长.例如设某数据集拥有 1 000 万条记录,每条记录的维度(或者属性)为 10,每个维度有 20 个取值,对于全分布的域大小则有 $20^{10}\approx 10$TB 个单元.如果直接利用噪音机制对 10TB 的输出单元添加噪音,则所需隐私预算、存储空间以及计算代价非常高.此外,每个单元的信息量非常小,假设 1 000 万条记录的大小为 10 MB,则每个单元的真实信息量为 10 MB/10TB$=10^{-6}$.若设定隐私预算 $\varepsilon=0.01$,利用拉普拉斯机制[1]产生的噪音约等于 125($lap(1/0.01)\approx 125$).如果直接为每个单元添加 125 大小的噪音,则该噪音值极大地扭曲了每个单元中的真实值(即 $125+10^{-6}$),发布结果的可用性极差.

目前已有少数基于概率图模型[2-3]、阈值过滤技术[4]以及投影技术[5]的高维数据发布方法,利用维度转换达到降维效果以及减少噪音对发布结果的影响,这些方法存在 2 个问题:

1) 文献[2-3]中的基于概率图模型方法通常利用指数机制[6]与稀疏向量[7]技术挑选属性关系对.然而,文献[2]中基于指数机制的属性对挑选方法受到候选空间大小的影响.如果候选空间非常大,所挑选的属性对越来越不准确,进而导致基于所选属性对构建的概率图不准确;文献[3]中采用稀疏向量技术挑选属性对,然而该技术并不满足差分隐私,进而导致文献[3]中整个高维数据发布方法不满足差分隐私.

2) 文献[4-5]中的基于阈值过滤与投影技术没有顾及到属性之间的依赖关系,利用阈值对维度直接截断达到降维效果.然而高维数据的属性之间普遍存在依赖关系,文献[4-5]中方法的实际可用性比较低.

总而言之,目前还没有一个行之有效的高维数据发布方法同时克服上述 2 种问题带来的不足.为此,本文基于联合树提出一种高效的高维数据发布技术.本文主要贡献包括 3 个方面:

1) 为了挑选出彼此关联的候选属性对,提出一种基于高通滤波的阈值过滤技术来缩减候选空间;

2) 结合缩减后的属性对候选空间,利用指数机制提出一种属性依赖图生成方法.结合属性依赖图,提出一种基于最大生成树技术的联合树生成方法.利用联合树推理出整体高维数据的联合分布,进而得到满足差分隐私的高维数据发布结果.

3) 理论证明所提出的高维数据发布方法满足 ε-差分隐私.在 5 种真实数据集上进行了可用性评估实验.实验结果表明:PrivHD 算法均优于 PrivBayes 与 JTree 算法.

1 相关工作

目前,基于差分隐私的高维数据发布方法主要分为 2 类:数据独立发布方法与数据相关发布方法.PriView[8]是数据独立发布方法的典型代表.该方法均等地处理所有的属性对,通过选择 $k(1<k<d)$个属性对来构建相应的边缘分布,进而估计出高维数据的联合分布.然而,该方法假设所有属性是相互独立,该假设在实际的应用中并不完全成立.此外,该方法

缺乏比较正式的推理机制;PrivBayes[2]与JTree[3]方法是数据相关发布方法的典型代表.其中PrivBayes利用基于指数机制的贝叶斯网络来推理属性之间的关联性,进而得到近似的联合概率分布.然而,由于该方法采用指数机制选择属性对,很难系统性地探索所有的属性关联关系.当属性对繁多时,指数机制选择精度会越来越低.在PrivBayes基础,JTree方法利用Markov网络处理高维数据问题.该方法利用基于稀疏向量技术的抽样方法来探求所有的关联属性对,通过联合树构建相应的联合分布于边缘分布.然而JTree方法存在的问题包括:1)方法中稀疏向量技术不满足差分隐私要求[7];2)方法中的团组合方法可能是局部最优,进而使得总体噪音误差未必最小.H_b[9]方法采用层次树与直方图技术发布高维数据,然而数据维度大于3时该方法失效.DPSense[4]方法利用阈值技术选择部分属性来达到控制发布敏感性,进而减少最终的发布误差,然而该方法没有顾及到属性之间的关联性.PrivPfC[5]方法利用投影直方图与卡方关联测试实现高维数据发布,然而卡方关联测试的敏感性比较大,进而导致发布精度低.DPCopula[10]方法利用高维属性可以分解成高斯Copula函数与多个一维边缘分布,该方法利用高斯Copula函数达到降维效果,然而该方法却要求边缘分布是均匀分布的多元正态分布.DP-SUBN[11]方法与LoPub[12]方法分别基于PrivBayes方法解决了分布式环境于众包环境下的高维数据的发布问题,然而这2种方法却继承了PrivBayes方法的不足.总的来说,目前差分隐私下还没有一个行之有效的高维数据发布方法.基于上述分析,本文提出一种基于联合树的高维数据方法PrivHD,该方法利用概率图模型-联合树对高维数据进行降维,利用拉普拉斯机制与指数机制确保PrivHD满足差分隐私.

2 定义与问题

2.1 差分隐私

差分隐私保护模型[13-14]的精髓是确保在数据集中插入或者删除一条记录的操作不会影响任何计算的输出.相比于传统的隐私保护模型,差分隐私保护模型具有2个显著的特点:1)不依赖于攻击者的背景知识;2)具有严谨的统计学模型,能够提供可量化的隐私保证.

定义1. 设高维数据D与D'彼此相差1条记录,互为近邻关系.给定一个高维数据发布算法H,$Range(H)$为H的输出范围,若算法H在D和D'上任意输出结果$X\in Range(H)$满足:

$$Pr[H(D)=X]\leqslant \exp(\varepsilon)\times Pr[H(D')=X], \quad (1)$$

则H满足ε-差分隐私.式(1)中,ε表示隐私预算,其值越小则算法H的隐私保护程度越高;$Pr[H(D)=X]$表示算法H基于D输出X的概率.

从定义1可以看出,ε-差分隐私限制了任意一个记录对算法H输出结果的影响.实现差分隐私保护需要噪音机制的介入,拉普拉斯机制[1]是实现差分隐私的主要技术.而所需要的噪音大小与其响应查询函数f的全局敏感性密切相关.

定义2. 设f为某一个查询,且$f:D\rightarrow\mathbb{R}^d$,$f$的全局敏感性为

$$\Delta f=\max_{D,D'}\| f(D)-f(D')\|_1. \quad (2)$$

其中,D与D'互为近邻,$\mathbb{R}$为映射的实数空间,d为f的查询维度.

文献[6]提出的拉普拉斯机制可以取得差分隐私保护效果,该机制利用拉普拉斯分布产生噪音,进而使得发布方法满足ε-差分隐私,如定理1所示.

定理1[6]. 设f为某个查询函数,且$f:D\rightarrow\mathbb{R}^d$,若方法$H$符合:

$$H(D)=f(D)+[lap_1(\Delta f/\varepsilon),lap_2(\Delta f/\varepsilon),\cdots,lap_d(\Delta f/\varepsilon)], \quad (3)$$

则H满足ε-差分隐私.式(2)中,$lap_i(\Delta f/\varepsilon)(1\leqslant i\leqslant d)$为相互独立的拉普拉斯噪音变量,噪音量大小与$\Delta f$成正比、与$\varepsilon$成反比.因此,查询$f$的全局敏感性越大,所需的噪音越多.

文献[6]提出的指数机制主要处理抽样算法的输出为非数值型的结果.该机制的关键技术是如何设计打分函数$u(e_i)$.设H为指数机制下的某个隐私方法,则H在打分函数作用下的输出结果为

$$H(e_i)=\left\{e_i:|Pr[e_i\in X]\propto\exp\left(\frac{\varepsilon u(e_i)}{2\Delta u}\right)\right\}. \quad (4)$$

其中,Δu为打分函数$u(e_i)$的全局敏感性,X为采用算法的输出域.由式(4)可知,e_i的打分函数越高,被选择输出的概率越大.

2.2 联合树

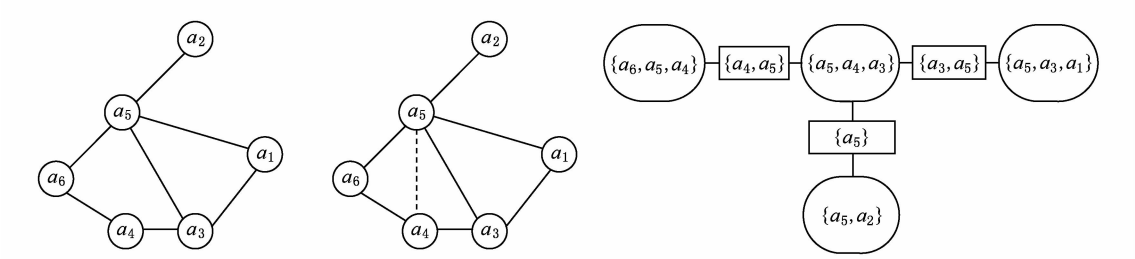
给定高维数据集D,且D具有d个属性,即$A=\{a_1,a_2,\cdots,a_d\}$,每个属性的值域为$\Omega_i(1\leqslant i\leqslant d)$.因此,整个值域的输出域为$\Omega_1\times\Omega_2\times\cdots\times\Omega_d$.在实际应用中,条件独立性在高维属性之间普遍存在,概率图模型是探索这种关系的主要技术,因此,本文采用联合树发掘低维属性构成的团和分割顶点的边缘分布,

进而推理出来整体属性的联合分布.高维属性 A 的联合分布:

$$Pr(A)=\frac{\prod_{C_i\in T}Pr(C_i)}{\prod_{S_{ij}=C_i\cap C_j,\ S_{ij}\in T}Pr(S_{ij})},\tag{5}$$

其中,$Pr(A)$表示属性集合 A 的联合分布概率;T 表示联合树,C_i 表示 T 中第 i 个团,$Pr(C_i)$表示团 C_i 的边缘分布概率;S_{ij}表示团 C_i 与团 C_j 之间的分割顶点,$Pr(S_{ij})$表示 S_{ij}的边缘分布概率.

采用 Markov 网表示属性对之间的关联关系,对 Markov 网进行三角化操作后获得团图.基于团图进行顶点消除来构建联合树.例如设 $A=\{a_1,a_2,a_3,a_4,a_5,a_6\}$,结合 A 所对应的 Markov 网 G 如图 1(a)所示,三角化操作结果如图 1(b)所示.依据属性下标顺序进行顶点消除,所得到的联合树如图 1(c)所示.由上述例子可以看出,结合联合树获得团和分割顶点的边缘分布之后,可以利用式(5)推理出整个高维属性的联合分布.



(a) Markov network G (b) Triangulation of G (c) Junction tree on G

Fig. 1 Construction of junction tree

图 1 联合树构建

2.3 问题描述

给定具有高维属性 A 的数据集 D,结合属性集合 A 构建 Markov 网 G.基于 G 构建联合树 T,结合 T 生成团和分割顶点的边缘分布.基于式(5)生成属性 A 的联合分布 $Pr(A)$.为了使高维数据发布满足差分隐私,联合树的构建过程需要满足差分隐私.即:Markov 网 G 的构建、团和分割顶点的边缘分布生成过程均需要满足差分隐私.因此,本文的问题是满足差分隐私的情况下如何发布比较精确的高维数据.

3 基于联合树的精确发布方法

本节主要介绍 PrivHD 算法的概述以及该算法的具体实现细节,其中包括满足差分隐私的 Markov 网构建方法、Markov 网的三角化方法、联合树的生成方法、边缘分布表的后置处理方法以及判断 PrivHD 算法是否满足 ε-差分隐私.

3.1 PrivHD 算法概述

PrivHD 算法利用 Markov 网对高维数据进行降维,利用满足差分隐私的联合树生成团和分割顶点的噪音边缘分布,最后生成合成高维数据集进行发布,该算法的描述如算法 1 所示:

算法 1. PrivHD 算法.

输入:高维数据集 D、属性 $A=\{a_1,a_2,\cdots,a_d\}$、隐私预算 ε;

输出:合成的高维数据集 $\widetilde{D}$.

① $\varepsilon=\varepsilon_1+\varepsilon_2$;

② $G\leftarrow Build\text{-}Noisy\text{-}MN(D,A,\varepsilon_1)$; /* 生成噪音 Markov 网 */

③ $T\leftarrow Build\text{-}Noisy\text{-}JunctionTree(G,\varepsilon_2)$; /* 生成联合树和团的边缘分布 */

④ $Pr(\cdot)\leftarrow Build\text{-}Noisy\text{-}Marginals(T)$; /* 生成分割顶点的边缘分布 */

⑤ $\widetilde{D}\leftarrow Clique\text{-}Join(Pr(\cdot))$; /* 生成合成高维数据集 */

⑥ return $\widetilde{D}$.

给定一个具有高维属性 A 的数据集 D 和相应的参数,PrivHD 算法首先利用 *Build-Noisy-MN* 方法生成满足差分隐私的噪音 Markov 网(行②).然后,通过对噪音 Markov 网三角化操作、完全团图构建操作,以及利用最大生成树方法生成联合树和团的边缘分布(行③).结合联合树所产生的团和分割顶点,生成每个团和分割顶点的边缘概率分布(行④).最后,通过对每个团的噪音边缘分布进行连接,生成最终的合成高维数据集进行发布.

3.2 Build-Noisy-MN 算法

结合属性集合构建 Markov 网 $G=(V,E)$,顶点集合 V 代表属性,边集合 E 代表 2 个属性之间的联系.本文利用互信息度量 2 个属性之间的关联性.给定属性 a_k,a_l,两者之间的互信息为 $I(a_k,a_l)$:

$$I(a_k,a_l)=\sum_{i\in dom(a_k)}\sum_{j\in dom(a_l)}Pr(a_k=i,a_l=j)\times \text{lb}\frac{Pr(a_k=i,a_l=j)}{Pr(a_k=i)Pr(a_l=j)},\tag{6}$$

其中,$i\in dom(a_k)$,$j\in dom(a_l)$,$dom(a_k)$,$dom(a_l)$分别表示属性a_k和a_l的取值范围;$Pr(a_k=i,a_l=j)$表示属性的联合分布,$Pr(a_k=i)$,$Pr(a_l=j)$分别表示a_k与a_l的边缘分布.

由上述分析可知,G中2个属性的互信息决定着两者之间是否存在一条边.然而,计算2个属性之间的互信息是数据相关的,因此,需在满足差分隐私的情况下获得该信息.由于噪音机制的介入,首先需要计算出互信息的敏感度ΔI:

$$\Delta I=\begin{cases}\dfrac{1}{n}\text{lb}\,n+\dfrac{n-1}{n}\text{lb}\dfrac{n}{n-1}, & \text{属性为二进制},\\ \dfrac{2}{n}\text{lb}\dfrac{n+1}{2}+\dfrac{n-1}{n}\text{lb}\dfrac{n+1}{n-1}, & \text{其他},\end{cases}\tag{7}$$

其中,$n=|D|$.

为了使所构建的Markov网G满足差分隐私,本文利用指数机制在候选空间$O\left(|O|=\binom{d}{2}\right)$中挑选边构建$G$.由式(4)可知,利用指数机制的挑选精度直接受到候选空间O与ΔI的制约.例如,$d=64$,则$O=2016$,在2016个属性对中挑选边可能造成最终生成的G不准确.为了克服大的候选空间带来的问题,本文采用高通滤波压缩O.首先利用拉普拉斯机制对所有属性对的互信息$I(a_k,a_l)$添加拉普拉斯噪音,然后利用式(8)进行过滤,进而得到压缩后的候选空间.

$$\tilde{I}(a_k,a_l)=\begin{cases}\tilde{I}(a_k,a_l), & \tilde{I}(a_k,a_l)\geqslant\theta,\\ 0, & \text{其他},\end{cases}\tag{8}$$

其中,$\theta=\eta\,\text{lb}\binom{d}{2}\Big/\varepsilon_{11}$,$\eta$为调节参数.

结合上述分析,Build-Noisy-MN的实现细节如算法2所示:

算法2. Build-Noisy-MN算法.

输入:高维数据集D、属性$A=\{a_1,a_2,\cdots,a_d\}$、隐私预算ε_1、过滤阈值θ;

输出:生成满足差分隐私的Markov网G.

① $\varepsilon_1=\varepsilon_{11}+\varepsilon_{12}$;

② $Eset\leftarrow\varnothing$;

③ 初始化$G=(V,E)$,$V=\{a_1,a_2,\cdots,a_d\}$,$E\leftarrow\varnothing$;

④ for $i=1$ to $|\tilde{O}|$ do

⑤ for each (a_k,a_l) do

⑥ $\tilde{I}(a_k,a_l)\leftarrow I(a_k,a_l)+lap(2|\tilde{O}|\Delta I/\varepsilon_{11})$;

⑦ if $\tilde{I}(a_k,a_l)\geqslant\theta$ then

⑧ $Eset\leftarrow Eset\cup(a_k,a_l)$;

⑨ endif

⑩ endfor

⑪ endfor

⑫ 以概率$Pr[(a_k,a_j)\in Eset]\propto\exp\left(\dfrac{\varepsilon_{12}I(a_k,a_j)}{2|\tilde{O}|\Delta I}\right)$随机选择一条边$(a_k,a_j)$;

⑬ 把边(a_k,a_j)添加到G中;

⑭ return G.

在Build-Noisy-MN算法中,行⑤~⑩部分实现对原始候选空间的压缩.获得候选边集合$Eset$后,利用指数机制挑选合适的边添加到G中(行⑫~⑬),进而形成满足差分隐私的Markov网G.

定理2. Build-Noisy-MN算法满足ε_1-差分隐私.

证明. 给定2个近邻D与D',令lap为噪音变量.给定一个属性对(a_k,a_l),在D与D'上所对应的噪音互信息分别是:

$$\tilde{I}_D(a_k,a_l)=I_D(a_k,a_l)+lap,$$

$$\tilde{I}_{D'}(a_k,a_l)=I_{D'}(a_k,a_l)+lap.$$

$$Pr(I_D(a_k,a_l)+lap\geqslant\theta)=Pr(lap\geqslant\theta-I_D(a_k,a_l))=\int_{\theta-I_D(a_k,a_l)}^{\infty}Pr(lap=x)\,\text{d}x\leqslant\int_{\theta-I_{D'}(a_k,a_l)-\Delta I}^{\infty}Pr(lap=x)\,\text{d}x=\int_{\theta-I_{D'}(a_k,a_l)}^{\infty}Pr(lap=x-\Delta I)\,\text{d}x\leqslant\int_{\theta-I_{D'}(a_k,a_l)}^{\infty}\exp\left(\frac{\varepsilon_{11}}{2|\tilde{O}|}\right)\times Pr(lap=x)\,\text{d}x=\exp\left(\frac{\varepsilon_{11}}{2|\tilde{O}|}\right)\times Pr(lap\geqslant\theta-I_{D'}(a_k,a_l))=\exp\left(\frac{\varepsilon_{11}}{2|\tilde{O}|}\right)\times Pr(I_{D'}(a_k,a_l)+lap\geqslant\theta).$$

根据上述推理过程可知不等式(9)与不等式(10)成立:

$$Pr(\tilde{I}_D(a_k,a_l)\geqslant\theta)\leqslant\exp\left(\frac{\varepsilon_{11}}{2|\tilde{O}|}\right)\times Pr(\tilde{I}_{D'}(a_k,a_l)\geqslant\theta),\tag{9}$$

$$Pr(\tilde{I}_D(a_k,a_l)\geqslant\theta)\geqslant\exp\left(-\frac{\varepsilon_{11}}{2|\tilde{O}|}\right)\times Pr(\tilde{I}_{D'}(a_k,a_l)\geqslant\theta).\tag{10}$$

令$M(Eset|D)$表示在边集合$Eset$中选择边操作.根据指数机制以及不等式(9)和不等式(10)可证明不等式成立:

$$Pr(M(Eset \mid D)=(a_k,a_j))=\frac{\exp\left(\frac{\varepsilon_{12} I_D(a_k,a_j)}{2 \mid \widetilde{O} \mid \Delta I}\right) Pr(\tilde{I}_D(a_k,a_l)\geqslant\theta)}{\sum\limits_{(a_m,a_n)\in Eset}\exp\left(\frac{\varepsilon_{12} I_D(a_m,a_n)}{2 \mid \widetilde{O} \mid \Delta I}\right) Pr(\tilde{I}_D(a_m,a_n)\geqslant\theta)}\leqslant$$

$$\frac{\exp\left(\frac{\varepsilon_{12} I_D(a_k,a_j)}{2 \mid \widetilde{O} \mid \Delta I}\right)\times\exp\left(\frac{\varepsilon_{11}}{2 \mid \widetilde{O} \mid}\right)\times Pr(\tilde{I}_{D'}(a_k,a_l)\geqslant\theta)}{\sum\limits_{(a_m,a_n)\in Eset}\exp\left(\frac{\varepsilon_{12} I_D(a_m,a_n)}{2 \mid \widetilde{O} \mid \Delta I}\right)\times\exp\left(-\frac{\varepsilon_{11}}{2 \mid \widetilde{O} \mid}\right)\times Pr(\tilde{I}_{D'}(a_m,a_n)\geqslant\theta)}. \tag{11}$$

根据不等式

$$\left|\exp\left(\frac{\varepsilon_{12} I_D(a_k,a_j)}{2|\widetilde{O}|\Delta I}\right)\Big/\exp\left(\frac{\varepsilon_{12} I_{D'}(a_k,a_j)}{2|\widetilde{O}|\Delta I}\right)\right|\leqslant\exp\left(\frac{\varepsilon_{12}}{2|\widetilde{O}|\Delta I}\right)$$

可知,不等式(11)满足条件:

$$Pr(A(Eset|D)=(a_k,a_j))\leqslant\frac{\exp\left(\frac{\varepsilon_{12}}{2 \mid \widetilde{O} \mid}\right)\times\exp\left(\frac{\varepsilon_{12} I_{D'}(a_k,a_j)}{2 \mid \widetilde{O} \mid \Delta I}\right)\times}{\sum\limits_{(a_m,a_n)\in Eset}\exp\left(-\frac{\varepsilon_{12}}{2 \mid \widetilde{O} \mid}\right)\times\exp\left(\frac{\varepsilon_{12} I_{D'}(a_m,a_n)}{2 \mid \widetilde{O} \mid \Delta I}\right)\times}\rightarrow$$

$$\leftarrow\frac{\exp\left(\frac{\varepsilon_{11}}{2|\widetilde{O}|}\right)\times Pr(\tilde{I}_{D'}(a_k,a_l)\geqslant\theta)}{\exp\left(-\frac{\varepsilon_{11}}{2|\widetilde{O}|}\right)\times Pr(\tilde{I}_{D'}(a_m,a_n)\geqslant\theta)}\leqslant\exp\left(\frac{\varepsilon_1}{|\widetilde{O}|}\right)\times Pr(A(Eset|D')=(a_k,a_j)). \tag{12}$$

证毕.

根据不等式(12)可知,从 $Eset$ 集合中每选出一条边 (a_k,a_j) 满足 $\exp\left(\frac{\varepsilon_1}{|\widetilde{O}|}\right)$-差分隐私,$Eset$ 中共 $|\widetilde{O}|$ 条边. 根据差分隐私的序列组合性质可知,Build-Noisy-MN 算法满足 ε_1-差分隐私.

3.3 Build-Noisy-JunctionTree 算法

基于 Markov 网 G 的联合树构建通常包括 G 的三角化(如图 1(b)所示)、生成完全团图、生成联合树 3 个过程. 结合 Build-Noisy-MN 算法产生的 Markov 网 G,三角化操作过程如算法 3 所示:

算法 3. *Triangulation-Partition* 算法.

输入:Markov 网 $G=(V,E)$、顶点消除顺序 φ;

输出:包含所有团的集合 CL、三角化后 Markov 网.

① $G'\leftarrow G, CL\leftarrow\varnothing$;
② for each node $u\in\varphi$ do
③ if 不存在边$(v_1,v_2)\in neibors(u)$ then
/ * 三角化过程 * /
④ $(v_1,v_2).mark=$true;
⑤ $G'.E\leftarrow G'.E\cup(v_1,v_2)$;
⑥ endif
⑦ $C\leftarrow neibors(u)\cup u$; / * 根据顶点消除顺序找团 * /
⑧ $CL\leftarrow CL\cup C$;
⑨ endfor
⑩ return G', CL.

在 Triangulation-Partition 算法中,行②~⑨是根据顶点消除顺序寻找所有团的过程. Markov 网 $G=(V,E)$ 中的团是一个顶点子集,其中每一对顶点之间均有一条边相连,即一个团是 G 的一个完全子图. 例如给定 $\varphi=\{a_1,a_2,a_3,a_4,a_5,a_6\}$,按照所设定的顶点消除顺序,所形成的如图 1(b)所示. 所形成的团分别是 $C_1=\{a_1,a_3,a_5\}$,$C_2=\{a_2,a_5\}$,$C_3=\{a_3,a_4,a_5\}$,$C_4=\{a_4,a_5,a_6\}$. 行③~⑥是对 G 三角化的过程. 所谓三角化操作是对所有长度大于 3 的环引入弦的过程. 例如,图 1(b)中虚线即为三角化后的弦.

根据所获得的团构建完全团图 $CG=(CL,E)$,其中 CL 表示所有团的集合,E 表示团之间存在的边. 每一条边 $(C_i,C_j)\in E$,都有一个权重 $w(C_i,C_j)$,并且 $w(C_i,C_j)=size(C_i\cap C_j)$. 即 2 个团的交集大小为边的权重. 例如根据 Triangulation-Partition 算法获得 $CL=\{C_1,C_2,C_3,C_4\}$,相应的权重分别为 $w(C_1,C_2)=1, w(C_1,C_3)=2, w(C_1,C_4)=1, w(C_2,C_3)=1, w(C_2,C_4)=1, w(C_3,C_4)=2$,所产生的完全团图如图 2(a)所示:

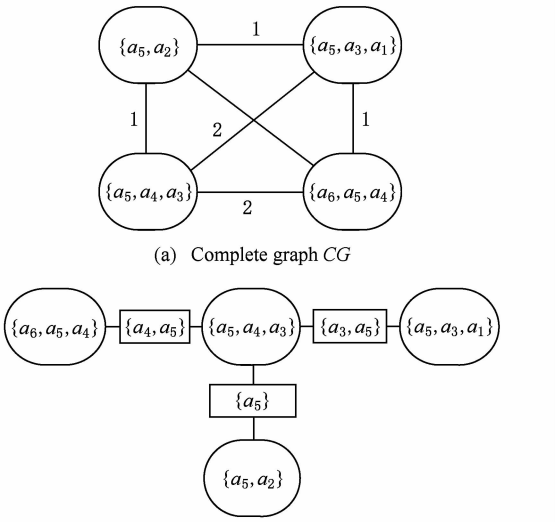


Fig. 2 Construction of junction tree on complete clique graph

图 2 基于完全团图的联合树构建

结合上述生成的完全团图 CG 构建联合树. 构建联合树的性能瓶颈在于团的大小,而基于最小数量的团构建联合树是个 NP 难问题. 因此,本文基于 CG 设计一种类似于最大生成树的贪心算法,具体细节如算法 4 所示:

算法 4. Build-Noisy-JunctionTree 算法.

输入:完全团图 $CG=(CL,E)$、隐私预算 ε_2;

输出:联合树 T.

① $T=(V,E)$;

② $T.V\leftarrow\varnothing, T.E\leftarrow\varnothing$;

③ $visited(CG.E)=$false; /* CG 的边均未被访问过 */

④ $T.V\leftarrow C, C\in CG.CL$; /* 选择最大的团 */

⑤ $m\leftarrow size(CG.CL)$; /* m 为 CG 中团的个数 */

⑥ $noisy\text{-}count(Tab(C)\leftarrow count(Tab(C))+lap(m/\varepsilon_2)$ /* $Tab(C)$表示团 C 的边缘分布表 */

⑦ while $T.V\neq CG.CL$ do /* 构建联合树 */

⑧ for each $(C_i,C_j)\in CG.E$ and $visited(C_i,C_j)=$false do

⑨ $(C_k,C_l)\leftarrow\max(w(C_i,C_j))$; /* 选择权重最大的边 */

⑩ if $C_k\in T.V$ and $C_l\notin T.V$ then

⑪ $T.V\leftarrow T.V\cup C_l$;

⑫ $T.E\leftarrow T.E\cup(C_k,C_l)$;

⑬ $visited((C_k,C_l))=$true;

⑭ $noisy\text{-}count(Tab(C_l))\leftarrow count(Tab(C_l))+lap(m/\varepsilon_2)$;

⑮ 生成新的分割结点 $S_{kl}=C_k\cap C_l$;

⑯ endif

⑰ endfor

⑱ endwhile

⑲ return T.

在 Build-Noisy-JunctionTree 算法中,首先选择 CG 中规模最大的团开始联合树的构建(行④). 针对所挑选的团,构建该团的边缘分布表,利用拉普拉斯机制对该表添加噪音(行⑤~⑥). 在行⑧~⑰的 for 循环中,先选择权重最大的边,再判断团 C_k 与 C_l 是否分别满足条件. 如果是,把团 C_l 添加联合树的顶点集合中,把边 (C_k,C_l) 添加联合树的边集合中. 对于贪心搜索到的团 C_l,利用拉普拉斯机制对其边缘分布表添加噪音(行⑭). 例如 $CL=\{C_1,C_2,C_3,C_4\}$. 选择 $C_3=\{a_3,a_4,a_5\}$ 作为开始顶点,以最大生成树方法贪心地生成联合树,如图 2(b)所示,使得最终的权重之和最大.

定理 3. Build-Noisy-JunctionTree 算法满足 ε_2-差分隐私.

证明. 由 Build-Noisy-JunctionTree 算法可知,产生 m 个团之后,分别对这些团添加独立的拉普拉斯噪音. 在原始数据 D 中去掉或者添加 1 条记录,最多影响 m 个团的计数,因此,为每个团添加噪音的大小为 $lap(m/\varepsilon_2)$. 根据定理 1 和差分隐私序列组合性质可知,Build-Noisy-JunctionTree 算法满足 ε_2-差分隐私. 证毕.

由 Build-Noisy-JunctionTree 算法中的行⑥与行⑭可知,最终形成的联合分布精度如何,直接受到联合树 T 中团个数 m 的约束. 如果采用噪音方差 $(2(\Delta f/\varepsilon_2)^2)$ 度量每个团产生的误差,则 m 个团产生的误差总和:

$$Lerror(CL)=\sum_{i=1}^{m}\frac{2m^2}{\varepsilon_2^2}\prod_{a_j\in C_i}|\Omega_j|. \tag{13}$$

其中,Ω_j 表示属性 a_j 的值域,$|\Omega_j|$ 表示该值域大小.

例如由图 2(b)可知 $CL=\{C_1,C_2,C_3,C_4\}$,即 $m=4$. 设置 A 中 6 个属性的值域大小分别为{2,2,3,3,2,2}. 根据式(13)可知 $Lerror(CL)=2304/\varepsilon_2^2$.

由式(13)可知,每个团的边缘分布表噪音误差的高低直接与团的个数 m 相关,m 越少最终的联合分布越精确. Triangulation-Partition 算法产生的团通常只包含 3 个属性顶点,而仔细观察可以发现,该算法的三角化操作并不充分,而三角化操作要求所

有长度大于 3 的环均要引入弦.为此,本文设计一种 Bigger-Clique 算法来构建更大的团图,具体细节如算法 5 所示:

算法 5. Bigger-Clique 算法.

输入:由 Triangulation-Partition 算法产生的 $G'=(V,E)$;

输出:包含所有团的集合 CL.

① $VS\leftarrow\varnothing, UV\leftarrow G'.V$; /* VS 表示访问过的顶点集合,UV 表示未被访问过的顶点集合 */

② $max\_label=1$;

③ 随机选择一个顶点 $u\in G'.V$, $labeled(u)=max\_label$;

④ $VS\leftarrow VS\cup u$;

⑤ $UV\leftarrow UV-u$;

⑥ while $UV\neq\varnothing$ do

⑦ for each $v\in G'.V$ and $v\in neibors(VS)$ do

⑧ $n\_v\leftarrow max\_labeled\_neibors(v)$;

⑨ $labeled(n\_v)\leftarrow max\_label+1$;

⑩ endfor

⑪ if 不存在(v_1,v_2),且$(v_1,v_2)\in neibors(n_v)$ then /* 充分三角化过程 */

⑫ $(v_1,v_2).mark=$true;

⑬ $VS\leftarrow VS\cup n\_v, UV\leftarrow UV\text{-}n\_v$;

⑭ $max\_label\leftarrow max\_label+1$;

⑮ endif

⑯ for each $node(i), i\in[1,max\_label]$ do

⑰ $C\leftarrow maximal\_subgraph(node(i))$; /* 包含 $node(i)$的极大完全子图 */

⑱ $CL\leftarrow CL\cup C$;

⑲ $G'.V\leftarrow G'.V\text{-}node(i)$;

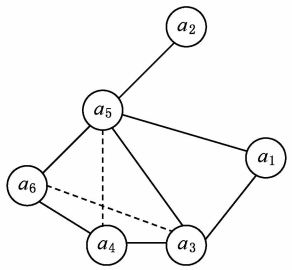
⑳ $G'.E\leftarrow G'.E\text{-}E\{node(i), neibors(node(i))\}$; /* 删除所有与 $node(i)$相连的边 */

㉑ endfor

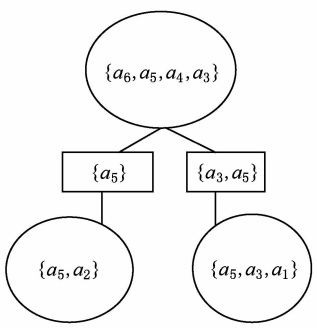
㉒ return CL.

在 Bigger-Clique 算法中,行⑪~⑮是充分三角化操作的过程.行⑯~㉑是发现极大团的过程,不同于 Triangulation-Partition 算法,该过程能够发现多个更大的团(顶点个数大于 3).由 Bigger-Clique 算法生产 CL 之后,直接送入 Build-Noisy-JunctionTree 算法生产联合树.为了说明 Bigger-Clique 算法的效果,给出如下例子.

例如在图 3(a)所示的 Markov 网中,为了充分三角化操作,顶点 a_6 与顶点 a_3 之间也引入了弦.依据顶点消除顺序 ,获得更大的团$\{a_3,a_4,a_5,a_6\}$.最终形成的团 $CL=\{C_1=\{a_1,a_3,a_5\},C_2=\{a_2,a_5\},C_3=\{a_3,a_4,a_5,a_6\}\}$,此时 $m=3$.根据式(13)可知 $Lerror(CL)=1296/\varepsilon_2^2$.相比 $2304/\varepsilon_2^2$,充分三角化操作能够比较明显地降低噪音误差.然后利用 Build-Noisy-JunctionTree 算法获得的联合树如图 3(b)所示.



(a) Markov network after triangulation



(b) Junction tree on maximum spanning tree

Fig. 3 Construction of junction tree on bigger complete graph

图 3 基于更大完全团图的联合树构建

3.4 Build-Noisy-Marginals 算法

由 Build-Noisy-JunctionTree 算法行⑥与行⑭可知,对每一个团的边缘分布表均添加 $lap(m/\varepsilon_2)$,进而保证每个边缘分布表满足差分隐私.根据联合树的执行相交性质(如性质 1)可知,对于任意分割顶点 S_{ij},则满足 $S_{ij}=C_i\cap C_j$.因此,对于联合树中的任意分割顶点,通过团边缘分布表投影的方法可以获得它们的噪音边缘分布表.当获得所有团和所有分割顶点的边缘分布之后,利用式(5)可以计算出所有属性的联合分布 $Pr(A)$.

性质 1. 执行相交性质.给定 1 棵联合树 T 以及 T 中任意顶点对 C_i 与 C_j,如果有变量 X 满足 $X\in C_i\cap C_j$,则 X 也在 C_i 与 C_j 之间唯一路径的每个顶点中.那么 T 具有执行相交性质.

例如在图 2(b)中获得 $Tab(C_2=\{a_5,a_2\})$和 $Tab(C_3=\{a_5,a_4,a_3\})$之后,通过投影可以获得分割

顶点$\{a_5\}$的边缘分布.假设a_2,a_3,a_4,a_5均为二进制属性,相应的噪音边缘分布如图4中(a)部分与图4中(c)部分所示.由于分割顶点$\{a_5\}=C_2\bigcap C_3$,可以利用团C_2或者团C_3来获得$\{a_5\}$的边缘分布.然而,通过图4中(b)部分与图4中(d)部分的噪音值*ncnt*可以发现,分割顶点$\{a_5\}$在团C_2与团C_3中获得的边缘分布互相不一致.而互不一致性直接导致最终所有属性的联合分布不准确.为了处理这种不一致性,本文基于文献[3,8,15]中的互一致性方法来后置处理团和分割顶点上的边缘分布不一致性.

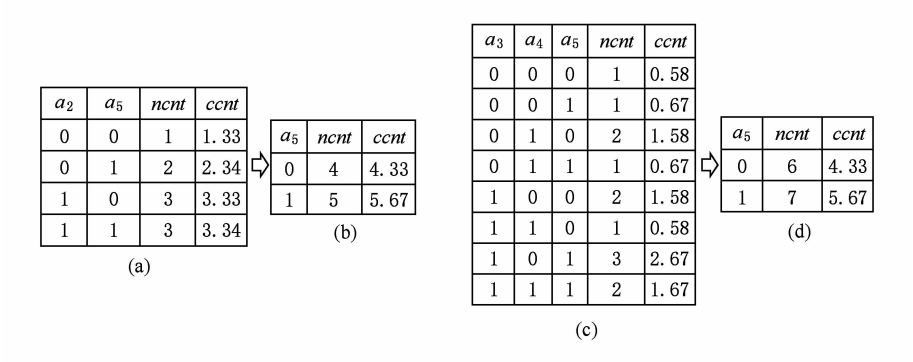


Fig. 4 Mutual consistency processing on clique and separator

图4 团与分割顶点的互一致性处理

设$Tab(C_1),Tab(C_2),\cdots,Tab(C_k)$为团$C_1,C_2\cdots,C_k$所对应的噪音边缘分布表.令$I=C_1\bigcap C_2\bigcap\cdots\bigcap C_k$.后置处理的目标是对于任意2个团$C_x,C_y\in\{C_1,C_2,\cdots,C_k\}$,使得$Tab_{C_x}(i)=Tab_{C_y}(i)$成立,其中$i$是$I$值域中一个可能的取值.令$ccnt_I(i)$表示$I$经过处理后的噪音计数(例如图4(b)中*ccnt*的值).$ccnt_I(i)$表示形式为

$$ccnt_I(i)=\frac{\sum_{l=1}^{k}Tab_{C_l}(i)\Big/\prod_{I_j\in(C_l-I)}|\Omega_j|}{\sum_l\left(1\Big/\prod_{I_j\in(C_l-I)}|\Omega_j|\right)},\quad(14)$$

其中,I_j表示C_l中除I之外的属性,$|\Omega_j|$表示I_j的值域大小.

例如给定团$C_2=\{a_2,a_5\}$和团$C_3=\{a_3,a_4,a_5\}$,以及二者的交集$I=\{a_5\}$,假设a_2,a_3,a_4,a_5均为二进制属性.$Tab_{C_2}(i=0)=4$,$Tab_{C_2}(i=1)=5$(如图4中(b)部分中的*ncnt*所示),$Tab_{C_3}(i=0)=6$,$Tab_{C_3}(i=1)=7$(如图4中(d)部分中的*ncnt*所示).利用式(14)可以计算出$ccnt_I(i=0)=4.33$,$ccnt_I(i=1)=5.67$(如图4中(b)部分和图4中(d)部分所示).

利用式(14)可以获得I在$C_1,C_2\cdots,C_k$中的一致性值.接下来是根据$ccnt_I(i)$来调整团$C_1,C_2\cdots,C_k$中的*ncnt*值进行调整.令$ccnt_{C_l}(c|i)$表示I取值为i时团C_l的后置处理值,具体表示形式为

$$ccnt_{C_l}(c|i)=ncnt(c|i)+\frac{ccnt_I(i)-Tab_{C_l}(i)}{\prod_{I_j\in(C_l-I)}|\Omega_j|}.\quad(15)$$

例如给定团C_2,C_3,其边缘分布表如图4中(a)部分与图4中(c)部分所示.由式(15)可知$ccnt_{C_2}(C_1|i=0)=1.33$,$ccnt_{C_2}(C_2|i=0)=3.33$,$ccnt_{C_2}(C_1|i=1)=2.34$,$ccnt_{C_2}(C_2|i=1)=3.34$,具体结果如图4中(a)部分所示.同理可以对图4中(c)部分中的*ncnt*值进行调整.

因此,根据式(14)和式(15)可以对团和分割顶点的边缘分布表进行一系列后置处理.然而在调整过程中,后置处理后的值有可能发生冲突.例如,$C_1\bigcap C_3=\{a_3,a_5\}$,$C_3\bigcap C_4=\{a_4,a_5\}$.如果第1步先对团$C_1$和$C_3$进行一致性处理,则$Tab(C_1)$和$Tab(C_3)$在属性$\{a_4,a_5\}$达成一致性.第2步对团$C_3$和$C_4$一致性处理,则$Tab(C_3)$和$Tab(C_4)$在属性$\{a_4,a_5\}$达成一致性.然而,第2步的一致性处理可能导致第1步中计数的不一致性,其原因是C_3和C_4一致性处理改变了$\{a_5\}$的分布.如果首先$C_1\bigcap C_3\bigcap C_4$在$\{a_5\}$上进行了互一致性处理,则第2步不会改变第1步中的一致性结果[2].

根据上述的例子可知,互一致性处理的先后顺序至关重要.因此,本文结合分割顶点上存在的偏斜关系给出一个全序的拓扑序列,进而获得互一致性处理的先后关系.令$S=\{s_1,s_2,\cdots,s_k\}$为部分分割顶点的集合,根据S上的偏序关系$\langle S,\preccurlyeq\rangle$形成哈斯图,然而结合哈斯图进行拓扑排序.本文利用分割顶点之间的覆盖关系表示$\preccurlyeq$.对于任意2个分割顶点s_i与s_j,若s_j覆盖s_i,则$s_i\prec s_j\wedge\neg\exists s_l(s_l\in S\wedge s_i\prec s_l\prec s_j)$成立.$s_j$覆盖$s_i$使得$s_i$与$s_j$之间存在一条无向边,且$s_i$在$s_j$的下方.

例如根据图 2(b)所示,$S=\{\{a_5\},\{a_3,a_5\},\{a_4,a_5\}\}$,根据$\langle S,\preccurlyeq\rangle$所形成的哈斯图如图 5 所示:

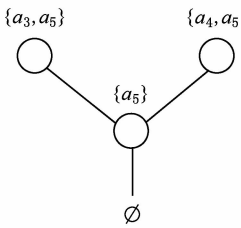


Fig. 5 Construction of Hasse diagram

图 5 哈斯图构建

从哈斯图的空顶点开始遍历,根据偏序$\langle S,\preccurlyeq\rangle$中存在的覆盖关系,删除空顶点以及所有边.再选择入度为 0 的顶点,删除该顶点以及与其相连接的边.直到不存在入度为 0 的顶点为止,输出的顶点序列就是一种拓扑序列.例如图 5 中偏序关系$\langle S,\preccurlyeq\rangle$上的拓扑序列为$\{\{a_5\},\{a_3,a_5\},\{a_4,a_5\}\}$.根据所找到的拓扑序列,利用式(14)与式(15)即可对所有的团和分割顶点执行互一致性后置处理.

3.5 Clique-Join 算法

当获得团与分割顶点的噪音边缘分布概率$Pr(\cdot)$之后,即可计算出$Pr(A)$.为了生成合成高维数据集 ,可以直接基于$Pr(A)$进行抽样来获取$\tilde{D}$中的每条记录.然而,该方法不但耗时,而且由于直接抽样会造成比较大的误差.为此,本文结合联合树的执行相交性与所生成的团,利用 Merge-Join 方法对所有的团进行迭代 Merge-Join 操作.例如图 4 中(a)部分与图 4 中(c)部分中,团C_2和团C_3的边缘分布表分别基于属性a_5进行排序,然后基于排序结果进行 *Merge-Join*,进而能够得到包含$\{a_2,a_3,a_4,a_5\}$的数据表T_{new},再T_{new}由与所剩余团的边缘分布表进行 *Merge-Join* 操作.最后可以获得高维数据集$\tilde{D}$.

定理 4. PrivHD 算法满足ε-差分隐私.

证明. 在 PrivHD 算法中,行②利用指数机制挑选属性对来形成 Markov 网,定理 3 已证明该步骤满足ε_1-差分隐私.行③利用拉普拉斯机制为每个团的边缘分布表添加噪音,根据定理 1 与定理 4 可知,行③满足满足ε_2-差分隐私.根据定理 6 差分隐私的顺序组合性质可知,PrivHD 满足$(\varepsilon_1+\varepsilon_2)$-差分隐私.由于$\varepsilon=\varepsilon_1+\varepsilon_2$,则 PrivHD 算法满足$\varepsilon$-差分隐私. 证毕.

定理 5[16]. 设D为隐含个人隐私的数据集,$H_1,H_2,\cdots,H_n$为n个随机算法,且H_i满足ε_i-差分隐私.$\{H_1,H_2,\cdots,H_n\}$在D上操作的顺序组合满足ε-差分隐私,且$\varepsilon=\sum_{i=1}^{n}\varepsilon_i$.

4 实验结果与分析

为了对 PrivHD 算法的可用性进行分析,本节将通过具体的实验进行验证与说明.实验平台是 4 核 Intel i7-4790 CPU(4 GHz),8 GB 内存,Windows 7 操作系统,部分实验在 Intel E5-2600 CPU,32 GB 内存,Linux 平台上运行.所涉及代码用 R 语言及 Matlab 实现.实验所采用的 5 个数据集 BR2000,Adult,NLTCS,TPC-E,TIC 均被广泛使用于高维数据发布.其中 BR2000 来自于 2000 年巴西全国人口普查数据,该数据包含 38 000 条记录;Adult 源自于美国人口普查中心,包含 45 222 条个人信息;NLTCS 源自于美国护理调查中心,记录了 21 574 名残疾人不同时间段的日常活动;TPC-E 来自于 TPC 开发的在线事务处理程序,记录了包含交易、交易类型、证券、证券状态等信息的 40 000 条数据;TIC 为 Coil 网站数据分析竞赛所用数据集,记录了某保险公司包括客户购买产品信息以及客户社交信息在内的 98 220 条信息.数据集的具体细节如表 1 所示:

Table 1 Description of Five Datasets

表 1 5 种数据集信息描述

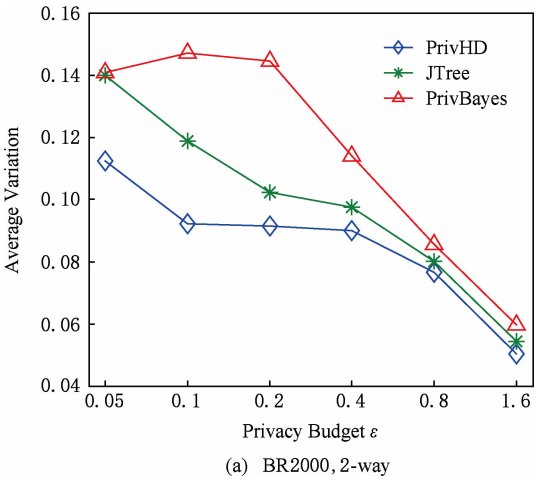
Datasets	Data Type	Dataset Size	Dimension	Domain Size
BR2000	Non-binary	38 000	14	$\approx 2^{32}$
Adult	Non-binary	45 222	15	$\approx 2^{52}$
NLTCS	Binary	21 574	16	2^{16}
TPC-E	Non-binary	40 000	24	$\approx 2^{77}$
TIC	Non-binary	98 220	86	$\approx 2^{86}$

结合上述 5 种数据集,分别用 k-way 查询与 SVM 分类度量 PrivHD,PrivBayes,JTree 算法发布高维数据的准确性与可用性.k-way 查询 k 的取值为 2,3,5,6,并使用平均方差(average variation)度量查询结果的准确性;在合成数据集上构建 SVM 分类模型并做出预测,使用误分类率(misclassification rate)度量分类结果的准确性.隐私预算参数 ε 的取值为 0.05,0.1,0.2,0.4,0.8,1.6.分配策略为 $\varepsilon_1=0.1$,$\varepsilon_2=\varepsilon-0.1$,即隐私预算 $\varepsilon_1=0.1$ 用于构建联合树,剩余的隐私预算用于产生噪音边缘分布,当 ε 取

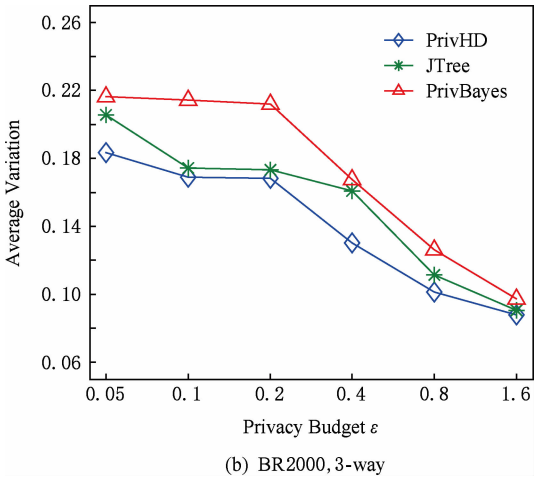
值为0.05,0.1时,$\varepsilon_1=1/2\varepsilon,\varepsilon_2=1/2\varepsilon$.每个算法重复实验50次,取度量指标的平均值作为实验结果.

1)基于参数ε和k-way查询范围的变化,对比PrivHD,PrivBayes,JTree算法的k-way查询结果.

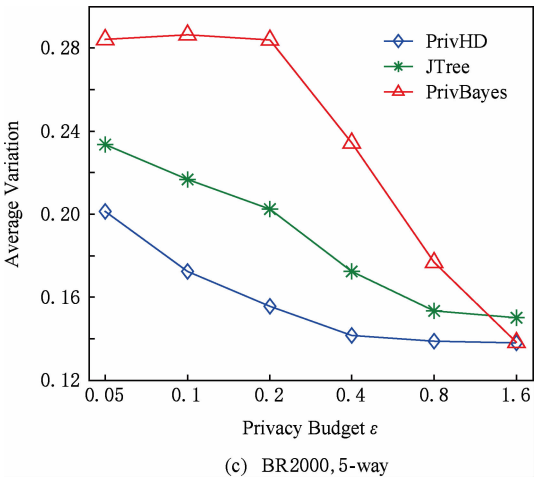
由图6可发现,在大多数情况下,PrivHD优于JTree和PrivBayes,在$\varepsilon<0.2$的情况下,PrivBayes的平均方差是PrivHD的2倍多.尤其是$k=5$或6的情况下,PrivHD明显优于JTree和PrivBayes,在维度比较大的TPC-E数据集上,JTree的平均方差是PrivHD的1倍多.在维度很大的TIC数据集上,由于PrivHD采用基于更大团的联合树构建方法的缘故,其平均方差同样小于JTree和PrivBayes,尤其是在$k=5$或$k=6$的情况下,PrivBayes的平均方差最坏情况下大概是PrivHD的1倍多.



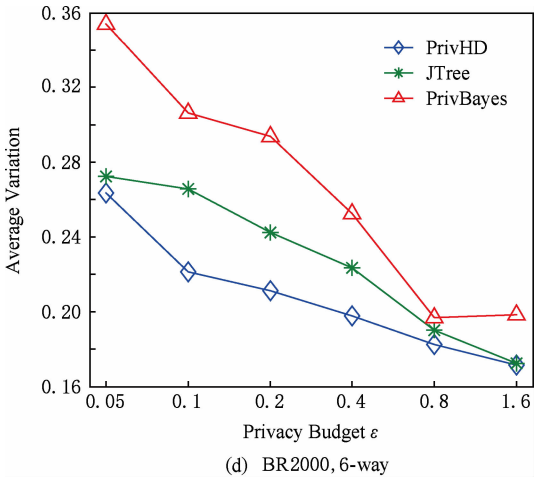
(a) BR2000,2-way



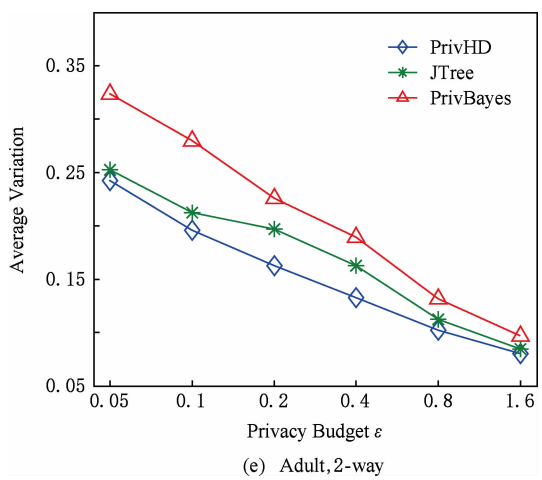
(b) BR2000,3-way



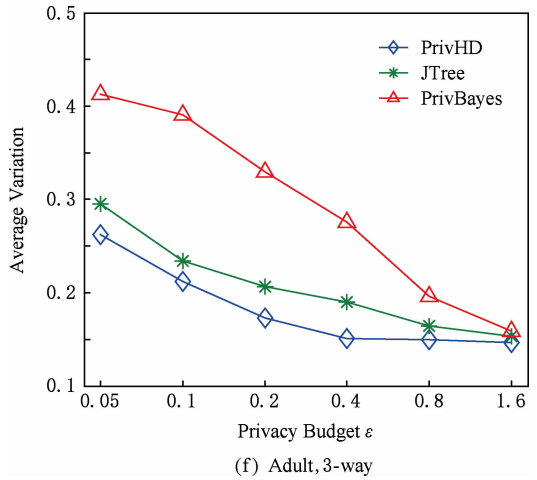
(c) BR2000,5-way



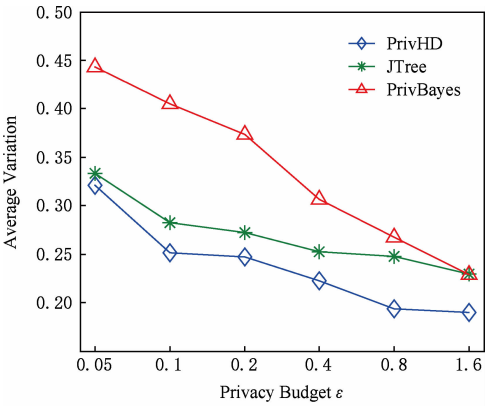
(d) BR2000,6-way



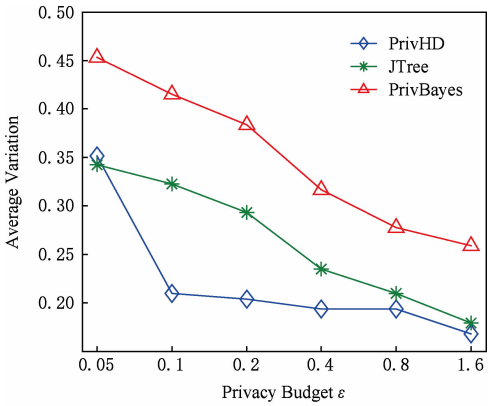
(e) Adult,2-way



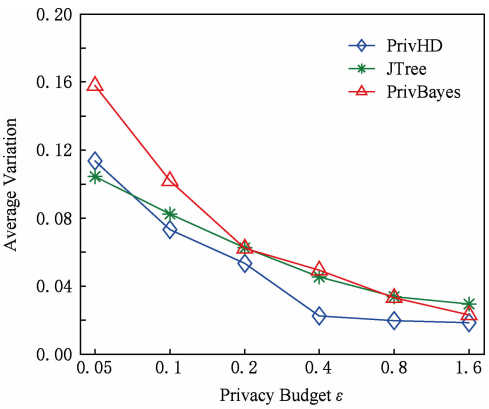
(f) Adult,3-way



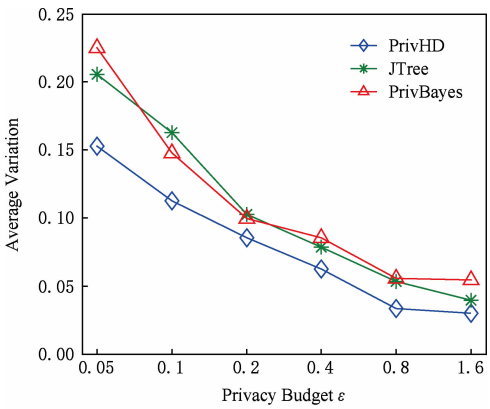
(g) Adult,5-way



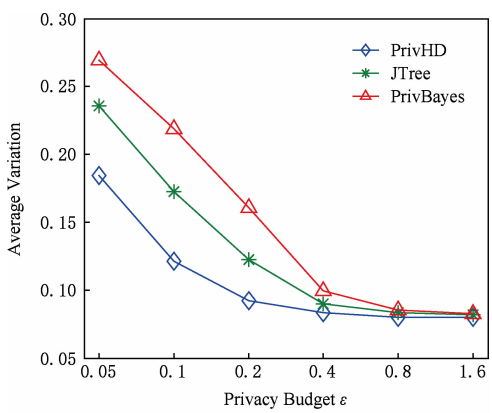
(h) Adult,6-way



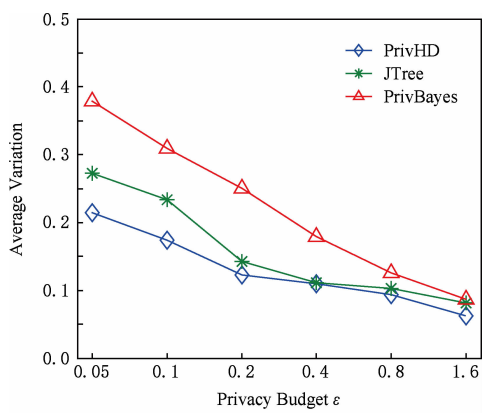
(i) NLTCS,2-way



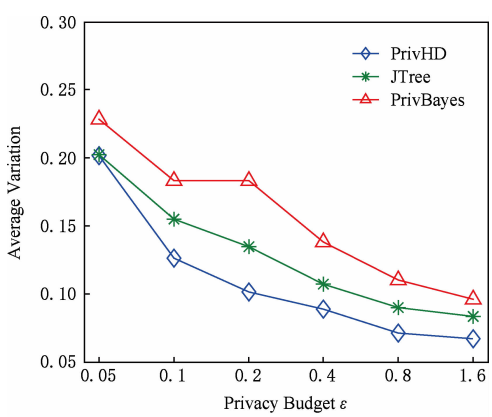
(j) NLTCS,3-way



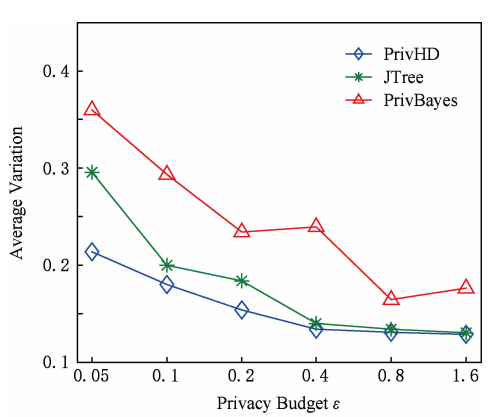
(k) NLTCS,5-way



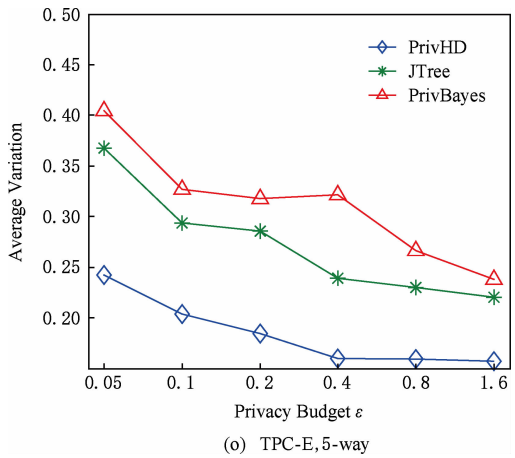
(l) NLTCS,6-way



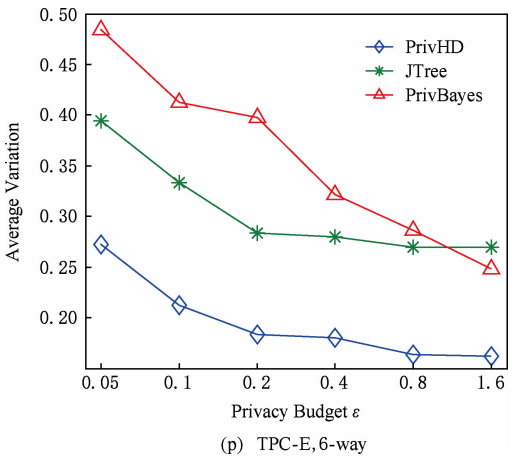
(m) TPC-E,2-way



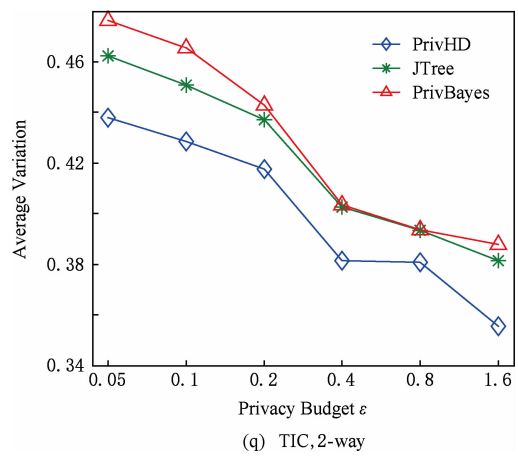
(n) TPC-E,3-way



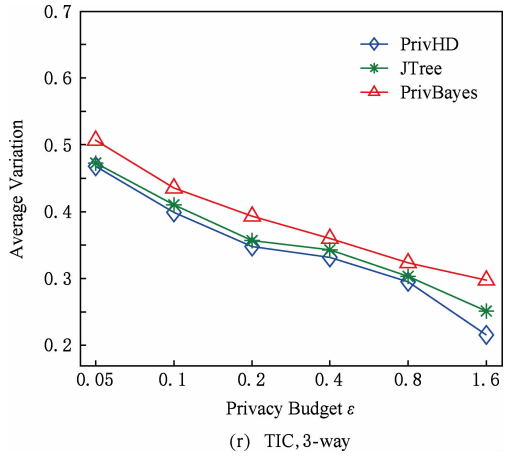
(o) TPC-E,5-way



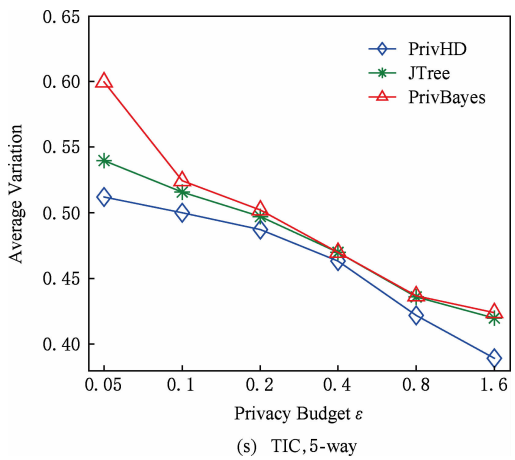
(p) TPC-E,6-way



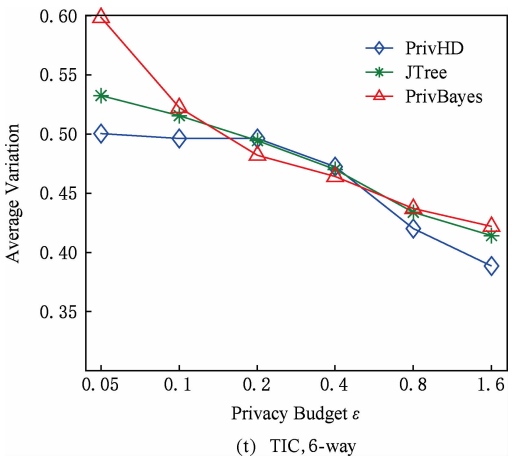
(q) TIC,2-way



(r) TIC,3-way



(s) TIC,5-way



(t) TIC,6-way

Fig. 6 Result of k-way marginals on five datasets

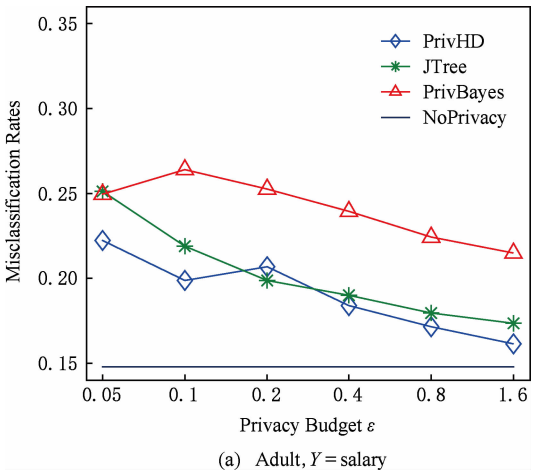
图 6 5 种数据集下 k-way 查询结果

k-way 查询的平均方差值越小表示发布高维数据的可用性越高,随着 k 值的增大,3 种算法的平均方差也越来越大,其原因是较大的 k-way 查询会造成拉普拉斯噪音的累积,进而造成可用性降低.但是在较大的 k-way 查询下,PrivHD 的表现仍然好于 JTree 和 PrivBayes,其原因是 PrivHD 采用了更优的联合树构建方法以及后置处理技术,进一步说明了 PrivHD 的高可用性.在数据集维度较低的 NLTCS 数据集上,PrivHD 表现仍好于 JTree,PrivBayes,而 3 种算法差别不大的原因是 NLTCS 数据集维度比较小,且为二进制数据.

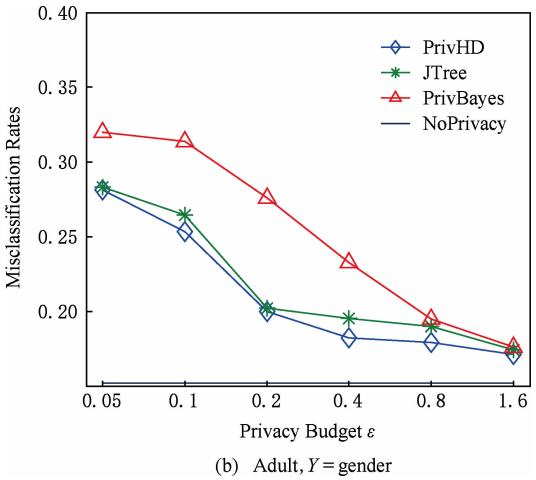
2) 基于参数 ε 的变化,对比 PrivHD,PrivBayes,JTree,NoPrivacy 算法的 SVM 分类结果.

本组实验在 Adult,NLTCS 和 TIC 数据集上进

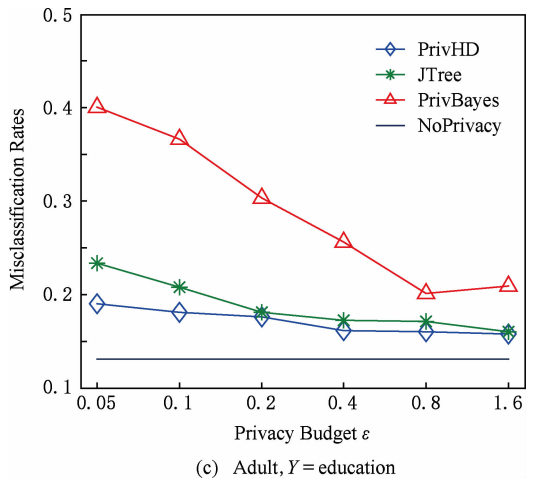
行分析.首先分别根据 PrivHD,PrivBayes,JTree 算法产生合成数据集,然后在合成数据集上构建 SVM 分类模型.对于 Adult 数据集,依据某个人:1)是否是男性(如图 7(b)中 Y=gender);2)是否有大专学位;3)年薪是否大于 5 万;4)是否已婚做出预测.对于 NLTCS 数据集,依据某个人:1)是否能够外出;2)是否能够管理资金;3)是否能够游泳;4)是否能够旅行做出预测.对于 TIC 数据集,依据某个人:1)是否拥有轿车;2)是否拥有房子;3)家庭平均年收入是否大于 3 万;4)是否已婚做出预测.其中 NoPrivacy 算法是在原始数据集上构建 SVM 分类模型并做出预测.本文将数据的 20%作为测试集,将 80%作为训练集.



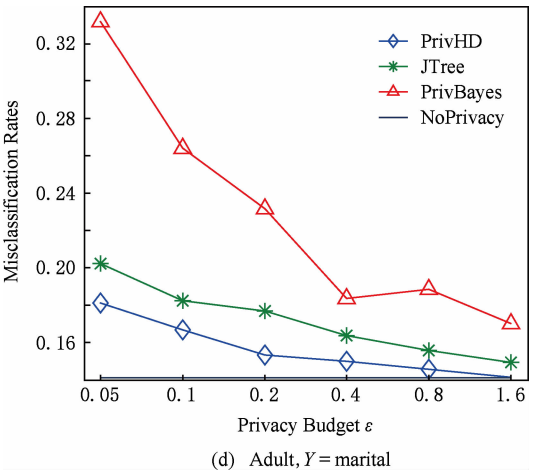
(a) Adult, Y = salary



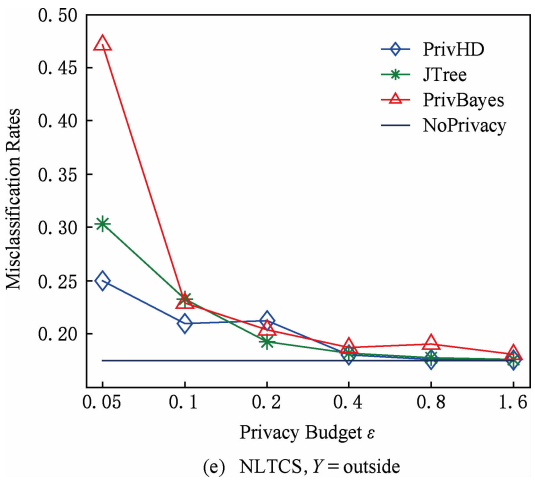
(b) Adult, Y = gender



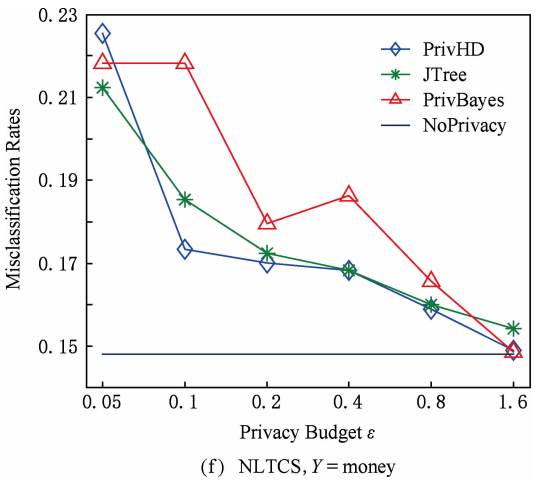
(c) Adult, Y = education



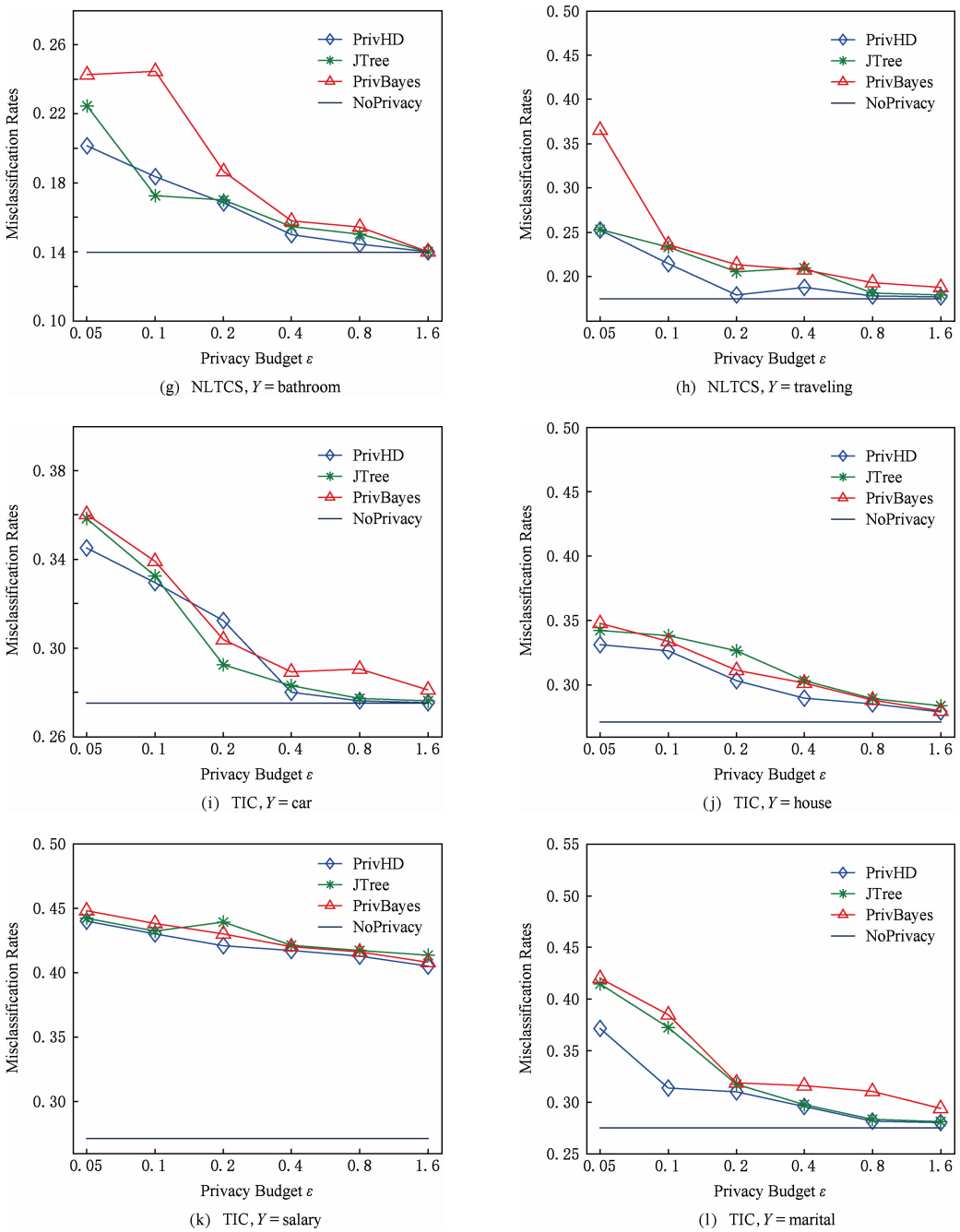
(d) Adult, Y = marital



(e) NLTCS, Y = outside



(f) NLTCS, Y = money



(g) NLTCS, Y = bathroom (h) NLTCS, Y = traveling

(i) TIC, Y = car (j) TIC, Y = house

(k) TIC, Y = salary (l) TIC, Y = marital

Fig. 7 Result of SVM classifiers on three datasets

图 7 3 种数据集下 SVM 分类结果

由图 7 可以发现,PrivHD 算法的 SVM 分类结果优于 JTree 和 PrivBayes 算法,在 Adult 数据集上,最坏情况下,PrivBayes 的误分类率是 PrivHD 的将近 3 倍,JTree 的误分类率也明显高于 PrivHD. 在数据维度很高的 TIC 数据集上,PrivBayes 和 JTree 的误分类率同样高于 PrivHD. 随着参数 ε 从 0.05 变化到 1.6,PrivHD,JTree,PrivBayes 算法的误分类率在相应降低,其原因是小的 ε 值会引入更多的拉普拉斯噪音,进而增加拉普拉斯误差.

5 结束语

针对差分隐私行下高维数据发布问题,本文首先对现有高维数据发布方法进行分析,并在此基础上提出基于联合树的发布算法 PrivHD,引入满足差分隐私的高通滤波与后置机制. 在过滤的基础上提出了基于指数机制的 Markov 网构建方法以及基于最大生成树的联合树构建方法. 从理论角度进行

对比分析的结果显示,PrivHD 优于同类算法.最后,通过真实数据集的实验结果表明 PrivHD 能够在满足差分隐私的同时输出比较精确的 k-way 查询与 SVM 分类结果.今后的工作需要考虑在数据流环境中发布高维数据.

参 考 文 献

[1] Dwork C, McSherry F, Nissim K, et al. Calibrating noise to sensitivity in private data analysis [C] //Proc of the 3rd Theory of Cryptography Conf (TCC 2006). Berlin: Springer, 2006: 363-385

[2] Zhang Jun, Cormode G, Procopiuc C M, et al. PrivBayes: Private data release via bayesian networks [C] //Proc of the 2014 ACM SIGMOD Int Conf on Management of Data. New York: ACM, 2014: 1423-1434

[3] Chen Rui, Xiao Qian, Zhang Yu, et al. Differentially private high-dimensional data publication via sampling-based inference [C] //Proc of the 21st ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2015: 129-138

[4] Day W Y, Li Ninghui. Differentially private publishing of high-dimensional data using sensitivity control [C] //Proc of the 10th ACM Symp on Information, Computer and Communications Security (ASIACCS 2015). New York: ACM, 2015: 451-462

[5] Su Dong, Cao Jianneng, Li Ninghui. Differentially private projected histograms of multi-attribute data for classification [OL]. [2015-06-01]. https://arxiv.org/pdf/1504.05997

[6] McSherry F, Talwar K. Mechanism design via differential privacy [C] //Proc of the 48th Annual IEEE Symp on Foundations of Computer Science. Piscataway, NJ: IEEE, 2007: 94-103

[7] Lü Ming, Su Dong, Li Ninghui. Understanding the sparse vector technique for differential privacy [J]. Proceedings of the Very Large Database Endowent, 2017, 10(6): 637-648

[8] Qardaji W H, Yang Weining, Li Ninghui. PriView: Practical differentially private release of marginal contingency tables [C] //Proc of the 2014 ACM SIGMOD Int Conf on Management of Data. New York: ACM, 2014: 1435-1446

[9] Qardaji W H, Yang Weining, Li Ninghui. Understanding hierarchical methods for differentially private histograms [J]. Proceedings of the Very Large Database Endowent, 2013, 6 (14): 1954-1965

[10] Li Haoran, Xiong Li, Jiang Xiaoqiang. Differentially private synthesization of multi-dimensional data using copula functions [C] //Proc of the 17th Int Conf on Extending Database Technology. New York: ACM, 2014: 475-486

[11] Su Sen, Tang Peng, Cheng Xiang, et al. Differentially private multi-party high-dimensional data publishing [C] //Proc of the 32nd IEEE Int Conf on Data Engineering. Piscataway, NJ: IEEE, 2016: 205-216

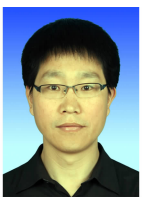
[12] Ren Xuebin, Yu C M, Yu Weiren, et al. LoPub: High-dimensional crowdsourced data publication with local differential privacy [OL]. [2017-02-01]. https://arxiv.org/pdf/1612.04350

[13] Dwork C. Differential privacy [C] //Proc of the 33rd Int Colloquium on Automata, Languages and Programming (ICALP 2006). Berlin: Springer, 2006: 1-12

[14] Dwork C, Jing Lei. Differential privacy and robust statistics [C] //Proc of the 41st Annual ACM Symp on Theory of Computing. New York: ACM, 2009: 371-380

[15] Hay M, Rastogi V, Miklau G, et al. Boosting the accuracy of differentially private histograms through consistency [J]. Proceedings of the Very Large Database Endowent, 2010, 3 (1): 1021-1032

[16] McSherry F. Privacy integrated queries: An extensible platform for privacy-preserving data qnalysis [C] //Proc of the 2009 ACM SIGMOD Int Conf on Management of Data. New York: ACM, 2009: 19-30



Zhang Xiaojian, born in 1980. PhD, associate professor in the College of Computer and Information Engineering, He'nan University of Economics and Law. His main research interests include differential privacy, data mining, and graph data management.

Chen Li, born in 1968. Professor in the Modern Educational Technology Center, He'nan University of Economics and Law. Her main research interests include information security, data mining.

Jin Kaizhong, born in 1991. Master candidate in the College of Computer and Information Engineering, He'nan University of Economics and Law. His main research interests include differential privacy, data mining.

Meng Xiaofeng, born in 1964. Professor and PhD supervisor at Renmin University of China. His main research interests include Cloud data management, Web data management, native XML databases, and flash-based databases, privacy-preserving, and etc.