基于朴素贝叶斯的差分隐私合成数据集发布算法

陈 旋 刘 健 冯新淇 赵雪美
(南京理工大学计算机科学与工程学院 南京 210094)

摘 要 差分隐私保护模型中,非交互式数据发布是一个研究的热点。提出了一个基于朴素贝叶斯的差分隐私合成数据集发布算法。该算法首先采用朴素贝叶斯的条件独立假设来计算原数据集的联合分布,然后采用指数机制生成发布的数据集。仿真实验表明,随着隐私预算的增加,使用合成数据集训练得到的分类器在测试数据集时分类正确率逐渐提高,并且趋于稳定。

关键词 差分隐私,朴素贝叶斯,数据发布,指数机制
中图法分类号 TP309 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.1.052

Differential Private Synthesis Dataset Releasing Algorithm Based on Navie Bayes

CHEN Xuan LIU Jian FENG Xin-qi ZHAO Xue-mei
(School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China)

Abstract Non-interactive data releasing has been a hotspot in differential privacy preservation model. A synthesis dataset releasing algorithm based on navie bayes was proposed. This algorithm computes the joint distribution of the original dataset based on the hypothesis of conditional independences in navie bayes firstly, then employs exponential mechanism to generate the synthesis dataset. The experiment results show that the accuracy of classifiers trained by the synthesis dataset improves and tends to be stable with privacy budget increasing.

Keywords Differential privacy, Navie bayes, Data release, Exponential mechanism

1 引言

随着信息化技术的普及,企事业单位、政府机构等积累了大量的数据。对这些数据进行挖掘可以得到大量的关于现实世界的知识。然而,这些数据中通常包含很多个人隐私信息,不能直接发布。因此,如何在保护个人隐私的前提下进行数据的发布和挖掘成为了一个研究的热点。

差分隐私模型[1]是 Dwork 在 2006 年提出的针对统计数据库的保护模型。差分隐私保护模型作为一个定义严格的、可证明的隐私保护模型,近年来受到越来越多的关注。在差分隐私的定义下,数据集的统计计算结果对单个数据的变化不敏感。

数据集发布是差分隐私研究中的难点。早期的差分隐私理论认为,在符合差分隐私定义的前提下,很难发布一个可响应任意查询的净化数据集,因为会引入过大的噪声,从而导致数据集不可用。因此,研究集中于对数据集统计特征的发布机制,如直方图发布[2]。由于这些机制仅能描述数据集的部分特征,在应用场景上存在很大的局限性。现实的需求促进了研究者对净化数据集发布的研究。例如,Blum[3]等在理论上证明了对于一个数据集 D,在不考虑计算复杂度的前提下,遍历整个数据域,可以找到满足差分隐私的数据集 D'。Mohammed 等人[7]提出了一种针对决策树分析的差分隐私数据发布算法 DiffGen。Zhang[8]等人提出了基于贝叶斯网络的差分隐私数据集合成算法 PrivBayes。

本文提出了一个基于朴素贝叶斯的差分隐私数据集发布算法。该算法首先采用朴素贝叶斯的条件独立假设来计算数据集的联合分布,然后采用指数机制,生成发布的数据集。

2 差分隐私保护理论基础

根据文献[1],ε-差分隐私定义如下。

定义 1(ε-差分隐私) 对于任意两个相差一个元组的数据集 D_1 和 D_2,如果随机算法 G 的任意输出 O 均满足:

$$P[G(D_1)=O] \leqslant e^{\varepsilon} \cdot P[G(D_2)=O] \tag{1}$$

则称算法 G 满足 ε-差分隐私。其中 $P[\cdot]$ 表示事件的概率。

此处,ε 通常取一个较小的数字,如 0.1,1.5 等。ε 取值越小,提供的隐私保护水平越高。

根据文献[6],全局敏感度、局部敏感度和指数机制定义如下。

定义 2(全局敏感度) 对于任意仅相差一个元组的数据集 D_1 和 D_2,当实数 $S(f)$ 满足

$$\| f(D_1)-f(D_2) \|_1 \leqslant S(f) \tag{2}$$

的最小数时,则称 $S(f)$ 是函数 $f: D^n \rightarrow R^d$ 的全局敏感度。

定义 3(局部敏感度) 对于给定的数据集 D_1 和任意与其相差一个元组的数据集 D_2,若有:

到稿日期:2014-03-05 返修日期:2014-06-10

陈 旋(1990-),男,硕士生,主要研究方向为隐私保护;刘 健(1990-),男,硕士生,主要研究方向为信息安全;冯新淇(1991-),女,硕士生,主要研究方向为信息安全;赵雪美(1988-),女,硕士生,主要研究方向为推荐系统。

$$LS_f(D_1)=\max_{D_2}\| f(D_1)-f(D_2)\|_1 \quad (3)$$

则称 $LS_f(D_1)$ 为函数 $f:D^n\to R^d$ 的局部敏感度。

在差分隐私中使用局部敏感度代替全局敏感度通常会泄露数据集的某些特征。然而，当泄露的信息是模型中假设攻击者已知的信息时，则可以进行替换。

定义 4(指数机制) 设算法 G 输入为数据集 D，输出为一实体对象 $r\in Range$，$q(D,r)$ 为打分函数，Δq 为函数 $q(D,r)$ 的敏感度。若算法 G 以正比于 $\exp(\frac{\varepsilon q(D,r)}{2\Delta q})$ 的概率从 $Range$ 中选择并输出 r，那么算法 G 提供 ε 差分隐私保护。

表 1 指数机制应用举例

候选人	得票数	概率		
		$\varepsilon=0$	$\varepsilon=0$	$\varepsilon=0$
A	50	0.25	0.928	0.999
B	20	0.25	0.046	9.36E-14
C	10	0.25	0.017	4.25E-18
D	4	0.25	0.009	1.05E-20

指数机制应用举例。假如进行一次选举，候选人有 A、B、C、D，需根据票数进行决策，并且要求决策过程满足 ε-差分隐私保护要求。以所得票数作为打分函数，则 $\Delta q=1$。根据指数机制，在给定的隐私保护预算 ε 情况下，可以计算出 4 人胜出的概率。

3 朴素贝叶斯法

3.1 基本原理

朴素贝叶斯法[5]是基于贝叶斯定理和特征条件独立假设的分类方法。对于给定的训练数据集，首先基于特征条件独立假设学习输入和输出的联合概率分布；然后基于此模型，对于给定的输入 x，利用贝叶斯定理求出后验概率最大的输出 y。

设输入空间 $\chi\subseteq R^n$ 为 n 维向量集合，输出空间为类标记集合 $\gamma=\{c_1,c_2,\cdots,c_k\}$。输入为特征向量 $x\in\chi$，输出为类标记 $y\in\gamma$。X 是定义在输入空间 χ 上的随机向量，Y 是定义在输出空间 γ 上的随机变量。$P(X,Y)$ 是 X 和 Y 的联合分布。训练数据集：

$$T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$$

由 $P(X,Y)$ 独立同分布产生。

朴素贝叶斯法通过训练数据集学习联合概率分布 $P(X,Y)$。具体地，分别学习先验概率分布及条件概率分布：

$P(Y=c_k),k=1,2,\cdots,K$

条件概率分布：

$$P(X=x|Y=c_k)=P(X^{(1)}=x^{(1)},\cdots,X^{(n)}=x^{(n)}|Y=c_k),\quad k=1,2,\cdots K \quad (4)$$

于是学习到了联合概率分布 $P(X,Y)$。

条件概率分布 $P(X=x|Y=c_k)$ 有指数级数量的参数，其估计实际上是不可行的。假设 $x^{(j)}$ 有 S_j 个取值，$j=1,2,\cdots,n$，Y 有 K 个取值，那么参数个数为 $K\prod_{j=1}^{n}S_j$。

朴素贝叶斯法对条件概率分布作了条件独立性的假设。这是一个较强的假设，朴素贝叶斯法也因此得名。具体地，

$$P(X=x|Y=c_k)=\prod_{j=1}^{n}P(X^{(j)}=x^{(j)}|Y=c_k) \quad (5)$$

朴素贝叶斯法分类时，对给定的输入 x，通过学习到的模型计算后验概率分布 $P(Y=c_k|X=x)$，将后验概率最大的类作为 x 的类输出。根据贝叶斯定理计算后验概率：

$$P(Y=c_k|X=x)=\frac{P(X=x|Y=c_k)P(Y=c_k)}{\sum_k P(X=x|Y=c_k)P(Y=c_k)} \quad (6)$$

3.2 拉普拉斯平滑

在朴素贝叶斯法中，参数的估计通常使用极大似然估计，即使用样本频度估计先验概率和条件概率。然而，这样通常会出现条件概率的情况，从而影响后验概率的估计。解决的方法是采用贝叶斯估计。

$$P_\lambda(X^{(j)}=a_{jl}|Y=c_k)=\frac{\sum_{i=1}^{N}I(x_i^{(j)}=a_{jl},y_i=c_k)+\lambda}{\sum_{i=1}^{N}I(y_i=c_k)+S_j\lambda} \quad (7)$$

N 为样本个数，$I(x)$ 为指示函数。贝叶斯估计等价于在随机变量各个取值的频数上赋予一个正数 λ。当 $\lambda=1$ 时，称为拉普拉斯平滑。下文中的条件概率均是进行拉普拉斯平滑后的条件概率。

4 差分隐私数据集合成算法

4.1 算法思想

监督学习中，通常假定输入和输出的随机变量 X 和 Y 遵循联合概率分布 $P(X,Y)$。生成学习即可看作对联合概率 $P(X,Y)$ 的学习过程。朴素贝叶斯是一种常见的生成式模型。若一个朴素贝叶斯分类器对数据的分类效果良好，则可以认为它对联合概率密度 $P(X,Y)$ 有较好的估计。因此，对于一随机生成的特征向量 x，若分类器将其分到某个类别 c_j，则可以认为样本 (x,c_j) 是 $P(X,Y)$ 中随机抽取的样本。

4.2 数据集合成算法

算法 1

输入：原始数据集 D，合成数据集大小 S

输出：合成数据集 D′。

1. 令 $D'=\{\}$
2. 由原始数据集学习到先验概率 $p(c_k)$ 和条件概率 $p(x^{(j)}|c_k)$，$k=1,2,\cdots,K$，$j=1,2,\cdots,n$。
3. 随机生成 $x\in S$，S 是输入空间。
4. 计算 x 属于各类别的后验概率 $p(c_i|x)$，选择后验概率最大的类别 c_m。
5. 将 (x,c_m) 添加到 D′中
6. 重复步骤 4、5，直到 $|D'|=S$。

算法 1 首先计算各类别的后验概率，然后选择后验概率最大的类别 c_m，并且认为 (x,c_m) 是 $P(X,Y)$ 中随机抽取的样本。

4.3 差分隐私决策算法

若将算法 1 中的步骤 4、5 选择类别 c_m 的过程视为根据条件概率和先验概率进行投票的过程，则可根据指数机制设计一个符合差分隐私保护的决策算法。

算法的输出是一类别变量，此处采用指数机制。效用函数的设计是指数机制的核心问题。这里首先给出效用函数的定义：

$$q(D,x,c_r)=\log(p(c_r))+\sum_{i=0}^{n}\log(p(x^{(i)}|c_r)) \quad (8)$$

定理 1 对于任意两个只相差一个属性的随机样本 x 和 x'，$q(D,x,C_r)$ 的局部敏感度为 $\log(2N)$，N 为数据集 D 中元组的个数。

证明：首先，在给定数据集中，进行拉普拉斯平滑后的条件概率 $p(x_i^{(j)}|c_k)$ 是有界的。

由于 $p(x_i^{(j)}|c_k)$ 表示一个概率，因此其上界为 1。式(7)

中，$\sum_{i=1}^{N} I(x_i^{(j)}=a_{jl}, y_i=c_k)\geqslant 0$，$\sum_{i=1}^{N} I(y_i=c_k)\leqslant N$，$S_j\leqslant N$，取 $\lambda=1$，当不等式取等号时，条件概率取到最小值 $p(x_i^{(j)}\mid c_k)=\frac{1}{2N}$，综上，

$$p(x_i^{(j)}\mid c_k)\in[\frac{1}{2N},1) \tag{9}$$

给定数据集 D，对于只有第 j 维不同的 n 维特征向量 x 和 x'，不相等的项只有第 j 维的条件概率。即有 $q(D,x,c_k)-q(D,x',c_k)=\log(p(x^{(j)}\mid c_k))-\log(p(x'^{(j)}\mid c_k))$。根据式(9)，有：

$$|q(D,x,c_k)-q(D,x',c_k)|<\log(2N) \tag{10}$$

由定理1可知，效用函数的敏感度 Δq 仅与数据集的大小 N 有关，采用局部敏感度代替全局敏感度会在一定程度上泄露与数据集大小相关的信息。然而，这里可取 $\Delta q=\lceil\log(2N)\rceil$，由于对数函数是一个增长缓慢的函数，因此向上取整操作可以很大程度地减小数据集大小信息泄露的可能。

算法 2

输入：由数据集 D 学习到的先验概率 $p(c_k)$ 和条件概率 $p(x^{(j)}\mid c_k)$，隐私预算 ε，随机样本 x

输出：类别标签 c_m

1. 对于每一个类别 c_r，计算 $q(D,x,c_r)$

2. 以 $p_{c_r}\propto\exp(\frac{\varepsilon q(D,x,c_r)}{2\Delta q})$ 的概率选取输出标签 c_m

5 实验过程及分析

本文的数据集合成算法采用 python 2.7 实现。分类器训练和测试部分使用开源机器学习工具 Weka 3.7.11。实验环境为 Windows 7 双核 2.40GHz，内存 4G。实验数据集采用 UCI adult[9]。UCI adult 包含 data 集和 test 集。

将定义 1 中的 ε 取值由 0 逐渐调到 1.5。对于 ε 的每一个取值，重复如下操作 5 次：首先使用 data 集生成合成数据集，然后用合成数据集训练贝叶斯网络分类器和逻辑斯特分类器，记录分类的正确率。

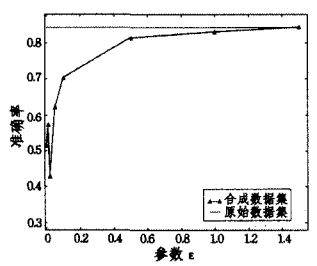

图 1 贝叶斯网络正确率均值

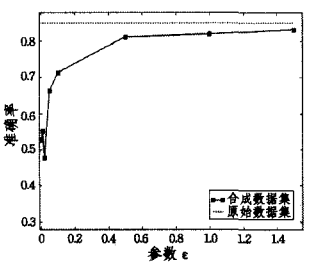

图 2 逻辑斯特正确率均值

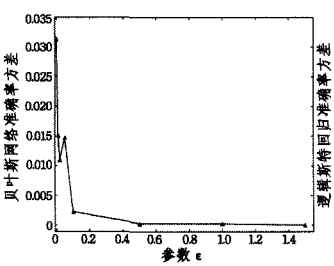

图 3 贝叶斯网络正确率方差

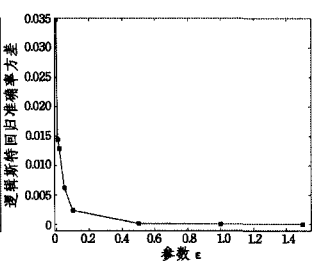

图 4 逻辑斯特正确率方差

图 1 和图 2 是 ε 取值不同的情况下，5 次分类的正确率均值。图中水平线表示基准正确率，即由 data 集训练的分类器的正确率。从中可以看出，当 ε 取值较小时，由合成数据集训练出的分类器在 test 集上的分类正确率较低；随着 ε 取值增大，正确率虽有波动，但总体趋势是逐渐提高的；分配的隐私预算足够大时，分类正确率接近基准正确率。

对于图 1 和图 2，当 ε 取值较小时，正确率出现波动的情况可结合图 3、图 4 给出如下解释：由于 ε 取值较小时，给定一个随机生成的特征向量 x，决策算法近似于随机决定其类别标签 c_j。当大量样本的标签恰好偏离其后验概率最大类别时，由生成数据集训练出的分类器在 test 集上的分类正确率较低；当大量样本的标签恰好等于其后验概率最大类别时，由生成数据集训练出的分类器在 test 集上的分类正确率较高；故当 ε 取值较小时，其正确率方差较高，正确率均值波动较大。随着 ε 取值逐渐增大，给定随机样本 x，决策算法的随机性减小，合成数据集中越来越多的样本接近假设的联合分布 $P(X,Y)$，因此，训练出的分类器正确率逐渐提高，并且趋于稳定，方差逐渐减小。例如，当 $\varepsilon=0$ 时，使用逻辑斯特分类器 5 次实验得到的正确率最高为 73.8%，最低为 23.3%；而当 $\varepsilon=1.5$ 时，最高为 83.6%，最低为 82.2%。

结束语 本文提出了一个基于朴素贝叶斯的差分隐私合成数据发布算法，给出了效用函数的定义，对其局部敏感度进行了证明。最后利用合成数据集训练贝叶斯网络和逻辑斯特分类器，通过比较其在测试数据集上分类的平均正确率和方差来评估算法的可行性。

在将来的工作中，我们将研究新的生成和决策算法，保留属性间的关联关系，同时添加后置约束，保留原数据集的其他统计信息。

参 考 文 献

[1] Dwork C. Differential privacy[C]// Proceedings of the 33rd international conference on Automata, Languages and Programming-Volume Part II. Springer-Verlag, 2006: 1-12

[2] Xu J, Zhang Z, Xiao X, et al. Differentially private histogram publication[J]. The VLDB Journal—The International Journal on Very Large Data Bases, 2013, 22(6): 797-822

[3] Blum A, Ligett K, Roth A. A learning theory approach to non-interactive database privacy[C]// STOC'08. 2008: 609-618

[4] McSherry F, Talwar K. Mechanism design via differential privacy[C]// 48th Annual IEEE Symposium on Foundations of Computer Science, 2007. FOCS'07. IEEE, 2007: 94-103

[5] Li Hang. Statistics learning method[M]. Beijing: Tsinghua university press, 2012: 47-52

[6] Xiong Ping, Zhu Tian-qing, Wang Xiao-feng. A survey on differential privacy and application[J]. Chinese Journal of Computers, 2014, 37(1): 101-122

[7] Mohammed N, Chen R, Fung B, et al. Differentially private data release for data mining[C]// Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2011: 493-501

[8] Zhang J, Cormode G, Procopiuc C M, et al. PrivBayes: Private Data Release via Bayesian Networks[C/OL]. [2014-7-8]. http://dimacs.rutgers.edu/~graham/pubs/papers/PrivBayes.pdf

[9] Bache K, Lichman M. UCI Machine Learning Repository[DB/OL]. [2014-7-8]. http://archive.ics.uci.edu/ml