

基于动态分类树构造的集值型数据 差分隐私保护方法^{*}

郑剑, 黄奚芳[†], 刘聪

(江西理工大学信息工程学院, 江西赣州 341000)

摘要: 基于分类树划分的差分隐私方法能有效地对集值型数据的发布进行保护,但在构造分类树时该方法没有充分利用集值型数据集自身的特征。通过对添加噪声量的影响因素分析,提出了一种基于数据集特征的集值型数据发布方法,该方法首先对数据集进行分析,然后根据数据集中记录的种类数占总输出域的比例以及只出现一次的记录种类数占总输出域比例,动态构造分类树。实验结果表明:当数据集满足 $IOR \leq 40\%$ 且 $SIOR = (5\%, 20\%]$ 时,通过有效利用集值型数据集的特征,构造较优的分类树,可以添加少于 10% 的噪声。

关键词: 分类树; 差分隐私保护; 集值型数据; 数据集特征

中图分类号: TP391

文献标志码: A

文章编号: 1001-3695(2015)08-2420-05

doi:10.3969/j.issn.1001-3695.2015.08.042

Constructing taxonomy tree based dynamic method for differential privacy preserving set-valued data

Zheng Jian, Huang Xifang[†], Liu Cong

(School of Information Engineering, Jiangxi University of Science & Technology, Ganzhou Jiangxi 341000, China)

Abstract: Taxonomy tree partitioning based method for differential privacy could protect the effective releasing of set-valued data. However, taxonomy tree does not take the characteristics of set-valued datasets into consideration of tree construction. By analyzing the influence factors of added noise, this paper proposed a novel method that releases set-valued data based on the characteristics of datasets. This method firstly analyzed the datasets, and then dynamically formed taxonomy tree structure according to the types of records in the dataset and the proportion between the total output of a single record field and the total number of species appeared in proportional output fields. The experimental results show that the proposed method can effectively utilize the characteristics of set-valued datasets, when the datasets conditions satisfy $IOR \leq 40\%$ and $SIOR = (5\%, 20\%]$, constructing superior taxonomy tree and reducing noise to less than 10%.

Key words: taxonomy tree; differential privacy; set-valued data; datasets characteristics

0 引言

随着社会信息化和网络化的蓬勃发展,越来越多的数据信息在网络中被公开发布。预计到 2020 年,全球制造、复制出的数字信息量将达到 40 ZB。作为信息资料的一种,集值型数据(set-valued data)(如医院诊断记录、在线查询记录、证券交易数据等)的发布将有利于数据挖掘等研究,然而这些数据资料中包含个人敏感信息,在网上发布和共享这些信息资料会给个人隐私带来严重威胁。因此,数据发布中的隐私保护研究具有重要意义。

差分隐私^[1~5](differential privacy)定义了极为严格的攻击模型,对隐私泄露的风险给出了严谨、量化的表示和证明,并从根本上解决了传统方法在背景知识攻击和攻击模型方面的不足。差分隐私作为一种新的保护模型不关心攻击者拥有多少背景知识,通过向查询或者是分析结果中添加噪声以达到隐私保护的效果。其最大的优点是虽基于数据失真技术,但所加入

的噪声量与数据集的大小无关,对于大型的数据集,仅通过添加极少量的噪声就能达到高级别的隐私保护。这种保护模型大大降低隐私泄露的风险,同时极大地保证了数据的可用性。因此该方法在 2006 年一经提出就在国外掀起一股研究热潮^[6],但在国内还处于起步阶段。

传统隐私保护下的集值型数据通常是使用 k -anonymity 模型来保护数据的。Terrovitis 等人^[7]提出的 (k, m) -anonymity 隐私原则是通过泛化层次树对集值型元数据中的一些或是全部进行泛化处理来阻止隐私攻击。毛云青等人^[8]提出的 (k, l) -anonymity 隐私模型对集值型数据的匿名处理进行更严格的限制,从而保护了用户隐私。Chen 等人^[9,10]借鉴自顶向下的划分方式,提出了几种树型结构来支持集值型数据的发布和挖掘。在基于数据独立的树划分中, Cormode 等人^[11]提出的 Quad-Post 方法结合差分隐私并借鉴完全四分树的思想提出了数据独立的划分方法,采用自顶向下的均匀预算分配(uniform budget)和几何预算分配(geometric budget)策略来添加拉普拉

收稿日期: 2014-06-17; 修回日期: 2014-07-28 基金项目: 江西省教育厅科学技术研究项目(GJJ13415); 江西理工大学科研基金重点课题(NSFJ2014-K11)

作者简介: 郑剑(1977-),男,湖北武汉人,副教授,博士,主要研究方向为隐私保护、可信软件;黄奚芳(1990-),女(通信作者),上海人,硕士研究生,主要研究方向为集值型数据的差分隐私保护(huangxifang2008@126.com);刘聪(1990-),男,江西人,学士,主要研究方向为差分隐私。

斯噪声。

本文主要针对分类树和自顶向下^[12] (top-down partitioning) 分割方式相结合的集值型数据差分隐私保护方法进行研究,该类方法的关键是如何构造分类树,如何根据分类树对发布树进行分割以及如何进行隐私预算分配。该类方法最典型的代表是Chen等人^[13]所提出的DiffPart算法。该方法首先根据扇出值(fan-out) F 来构造一棵分类树,然后根据分类树泛化数据集,再从层次分割的根节点开始迭代生成不同的子分割和叶子分割,最后发布叶子分割的统计结果。为增加空节点的生成,该方法中将只含有1条记录的非叶子节点也定义为空节点(本文称为伪空节点),并且不再继续进行分割,这样对该类记录也就不再添加噪声。

基于这一重要信息,本文研究在数据集确定时,如何结合数据集的特征尽可能构造出更多的伪空节点,从而减少噪声的添加。通过对发布树的分割过程的分析,发现数据集中记录的种类数占总输出域的比率(简记为IOR)以及单个记录出现的种类数占总输出域比率(简记为SIOR)不同,对伪空节点的生成有较大影响。为此,设计了一种基于数据集特征的分类树动态构造方法,该方法首先分别计算IOR和SIOR,当 $\text{IOR} \leq 40\%$ 且 $5\% < \text{SIOR} \leq 20\%$ 时就根据设计的算法动态构造分类树,否则就随机构造分类树。实验表明,动态构造的分类树比随意构造分类树的平均相对误差平均值降低了10%。

1 基本问题描述

定义1 集值型数据。定义 $I = \{I_1, I_2, \dots, I_{|I|}\}$ 为项(item)的全集, $|I|$ 表示全集的大小, $I_i \in I$ 表示全集中的一个项。多重集 $D = \{D_1, D_2, \dots, D_{|D|}\}$ 表示一个集值型的数据集,每条记录 D_i 是由 I 的非空子集构成,其中 $D_i \in D$ 。

在一条记录中任取两项组合在一起,称为二项集,即为二项集。例如,对于给定的项的全集 $I = \{I_1, I_2, I_3, I_4\}$,其数据集如表1所示,记录号为 t_7 的记录 D 中 $\{I_1, I_2\}, \{I_2, I_4\}, \{I_1, I_4\}$ 就称为记录 $\{I_1, I_2, I_4\}$ 中的二项集。记录的种类数是表示数据集中所有的记录种类。如表1中 t_1 的记录表示一种记录, t_3 和 t_8 表示另一种记录,所以表1的数据集的记录种类数为6。

表1 简单的数据集

记录号	记录	记录号	记录
t_1	$\{I_1, I_4\}$	t_5	$\{I_2, I_4\}$
t_2	$\{I_1, I_2, I_3, I_4\}$	t_6	$\{I_4\}$
t_3	$\{I_1\}$	t_7	$\{I_1, I_2, I_4\}$
t_4	$\{I_4\}$	t_8	$\{I_1\}$

给定项的全集 I ,集值型数据集的输出域 o 是全集中所有项集 I 的组合,其输出的长度为 $|o| = \sum_{k=1}^{|I|} \binom{|I|}{k} = 2^{|I|} - 1$ 。例如,给定项集全集 $I = \{I_1, I_2\}$,其输出域 $o = \{\{I_1\}, \{I_2\}, \{I_1, I_2\}\}$,输出域的长度为 $2^2 - 1 = 3$ 。

定义2 差分隐私^[4]。在非交互式系统下,差分隐私的定义如下所示:

对于给定的两个数据集 D_1 和 D_2 ,它们之间至多相差一条记录。给定的隐私机制 A , $\text{Rang}(A)$ 表示 A 的取值范围,对于任意处理过的数据集 $D \in \text{Rang}(A)$,则隐私机制 A 满足 ϵ -差分隐私。

$$P_r[A(D_1) = \tilde{D}] \leq e^\epsilon \times P_r[A(D_2) = \tilde{D}] \quad (1)$$

其中,算法的概率由 A 的随机性控制。实现差分隐私有两种噪声机制:拉普拉斯机制和指数机制,本文主要采用的是拉普拉斯噪声机制。

定义3 计数查询。计数查询^[14]对于集值型数据的数据挖掘,如挖掘频繁模式和关联规则,是至关重要的。因此,本文研究解决在非交互式环境中,面对计数查询时,发布的集值型数据的可用性的问题。计数查询的定义描述如下:

对于给定的项集 $U \in o$,计数查询 Q 对数据集 D 上的查询 $Q(D) = |\{D \in D; U \subseteq D\}|$ 。

定义4 分类树^[9]。对于给定的数据集,把数据集中的项作为分类树的叶子节点,泛化叶子节点成为分类树的节点,分类树的根节点是所有叶子节点的集合。如图1所示表1数据集的一个分类树, $I_{\{1,2,3,4\}}$ 根节点, I_1 和 I_2 是数据集中的项,可以泛化成 $I_{\{1,2\}}$ 作为分类树的节点。

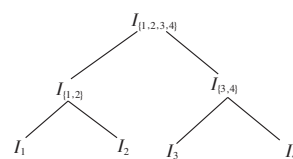


图1 简单数据集的分类树

定义5 平均相对误差。对于计数查询 Q ,本文用平均相对误差^[15,16]来衡量相对于原始数据集 D 来说,处理过的数据集 $\tilde{D}$ 中数据的可利用性。其平均相对误差的计算公式如式(2)所示。

$$\text{error} = \frac{|Q(\tilde{D}) - Q(D)|}{\max_s |Q(D), s|} \quad (2)$$

其中: s 为理智约束,为防止在极小计数查询的情况下,其分母为零。

定义6 伪空节点。在划分过程中,子分割中记录数为1的不可再分的非叶子分割节点称为伪空节点(pseudo blank nodes)。例如,在图2为表1的一种划分过程。在图1中, $I_{\{1,2,3,4\}}$ 是根节点,它的子分割是 $I_{\{1,2\}}, I_{\{3,4\}}$ 和 $\{I_{\{1,2\}}, I_{\{3,4\}}\}$, $\{I_{\{1,2\}}, I_3\}$ 为 $\{I_{\{1,2\}}, I_{\{3,4\}}\}$ 的子分割,由于 $\{I_{\{1,2\}}, I_3\}$ 中只有一条记录,那么 $\{I_{\{1,2\}}, I_3\}$ 就是一个伪空节点。

2 数据发布精度影响因素分析

在DiffPart算法^[12]中,分类树是实现隐私保护的前提。其算法实现是通过分类树将数据集的所有记录泛化到层次分割的根节点,然后再根据分类树,从层次分割的根节点开始迭代生成不同的子分割和叶子分割,最后对分割结束后的叶子节点添加拉普拉斯噪声。图2就是根据图1的分类树对表1数据集的划分过程。对于同一数据集来说,改变分类树,那么迭代生成的子分割和叶子分割也会改变,伪空节点的数目也会随之改变。因此,所构造的分类树的优劣很大程度上会影响到伪空节点的数目。

从平均相对误差的定义可以看出,影响平均相对误差的因素有两个方面:一方面是分子上的 $|Q(\tilde{D}) - Q(D)|$;另一方面是分母上的 $\max_s |Q(D), s|$ 。对于已知的数据集, $Q(D)$ 和 s 都是相对固定的值,其对平均相对误差的影响可以说是非常小的。因此,影响平均相对误差的因素主要是 $|Q(\tilde{D}) - Q(D)|$ 。差分隐私保护模型是一种基于数据失真的保护模式,经过处理后的数据集,也可以认为是添加了噪声的数据集。那么

$$|Q(\tilde{D}) - Q(D)| = |\sum_{i=1}^m (Q(I_i) + N_i) - \sum_{i=1}^m Q(I_i)| = |\sum_{i=1}^m N_i| \quad (3)$$

由式(3)可知,影响平均相对误差的因素主要是 $|\sum_{i=1}^m N_i|$,即在原始数据集中所添加的拉普拉斯噪声量。在 DiffPart 算法中,仅对所有叶子分割添加噪声,那么影响噪声量大小的因素是分割结束后所产生的叶子分割的数目。叶子分割的数目越多,在原始数据集中添加的拉普拉斯噪声也就越多。因此,要减少噪声量就要减少到达叶子分割的数目。在 DiffPart 算法中,当分割层次中存在伪空节点,这些分割层次将不会划分至叶子分割,那么到达叶子分割的数目就会减少。由此可以推断,在 DiffPart 算法中,伪空节点数越多,到达叶子分割的数量就越少。由于 DiffPart 算法中伪空节点和不存在的项集是不添加噪音的,所以随着伪空节点数量的增大,添加的噪声量就越小,平均相对误差就越小。因此,除了在叶子分割添加拉普拉斯噪声对平均相对误差有较大影响外,伪空节点也是影响平均相对误差的一个主要因素。由伪空节点的定义可知,伪空节点是由只出现一次的记录构成,同时每个分割节点由多个叶子节点构成。因此,分类树的划分、记录的种类数和单个出现记录的种类数都会影响伪空节点的产生,进而影响平均相对误差和数据的发布精度。

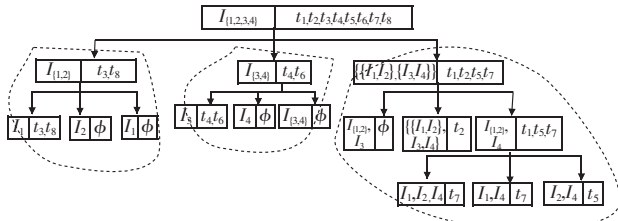


图 2 划分过程

3 分类树动态构造算法的基本思想

通过上述分析可以看出在数据集确定时,划分过程中产生的伪空节点越多,平均相对误差就越小。从对 DiffPart 算法分析可以看出,算法根据 F 把项集的全集 $|I|$ 随机分成两组的形式,两组之间没有相交的项集。在划分过程中,根据分类树可划分成三种子分割,分配记录时把两组间相交的数据分配到第三种子分割中。如图 1 中画虚线部分为三种子分割, $I_{1,2}$ 和 $I_{3,4}$ 和 $I_{1,2,3,4}$ 为子分割。一般来说,第三种子分割的子分割更多,可划分层次也就更深,所产生的叶子分割也更多。因此,可以认为只要控制第三种子分割的可划分的子分割的数目,就会使得伪空节点越多,那么只要两组分割中组内数据关系尽可能紧密,两组之间的数据关系尽量少,那么分类树也就更优。

3.1 算法的基本思想

基于上述思想,为了有效地增加伪空节点的出现,减少叶子分割的产生,减少拉普拉斯噪声的添加,从而起到降低平均相对误差的作用。本文提出一种基于数据集特征的分类树动态构造算法 CDTT(construct dynamic taxonomy tree)来构造较优的分类树,其基本思想是:首先选出数据集中只出现一次的记录,通过二项集出现的次数,首先挑选出一个数据,其所对应的项集作为一个中心点,然后挑出的项集的两行中挑出最小数的项集,在这个项集所在的行中选出最大的数,作为第二个中心。然后迭代地挑选其他项集与这两个中心点组合,直到所有的项

集完全被挑出,该方法结束。这两个分组就是分类树,分类树就构造好了。该算法的思想是使得组内项集高聚合,组间项集低耦合。

3.2 算法的描述

算法 CDTT 如算法 1 和 2 所示。

算法 1 关系矩阵构造算法

输入:原始集值型数据集 D 。

输出:矩阵 C 。

- (1) 计算只出现一次记录 R 的个数
- (2) 获取数据集 $|I|$ 的长度为 m
- (3) 构造 $m \times m$ 的矩阵 $C[m, m]$
- (4) $C[m, m] = 0$;
- (5) for each $\{I_p, I_q\} \in R_i$ do 其中 $1 \leq p, q \leq m$;
- (6) for $i = 1$ to n do
- (7) $C[p, q] = C[p, q] + 1, C[q, p] = C[q, p] + 1$;
- (8) end for
- (9) end for
- (10) return C

关系矩阵构造算法的主要目的是生成一个二项集的关系矩阵。其处理过程是对原始集值型数据集进行处理,首先遍历原始数据集选出只出现一次的记录 R 的个数 n ,并根据项集全 $|I|$ 的长度 m 来构造矩阵;然后查找每条记录中存在的二项集,对矩阵中数据不断增加;最后,直到记录中的二项集关系全部添加到矩阵中,生成矩阵。

算法 2 分类树构造算法

输入:矩阵。

输出:分类树。

- (1) $\text{temp}[m, m] = C$
- (2) $\text{first}[1:m] = 0, \text{second}[1:m] = 0$
- (3) for $Z = 1$ to do
- (4) $K = \text{mod}(Z, 2)$;
- (5) if $K = 1$ then
- (6) if $Z = 1$ then
- (7) 从矩阵 C 中得到最大元素 $\max\{C[i, j]\}$;
- (8) $\text{first} = \text{first} \cup \{i, j\}$;
- (9) else
- (10) $\text{temp} = C$
- (11) $\text{temp}[p, m] = 0, \text{temp}[m, p] = 0$; 其中 $p \in \text{second}$
- (12) 从 $\text{temp}[q, m]$ 中取最大元素 $\max\{\text{temp}[i, j]\}$, 其中 $q \in \text{first}$
- (13) $\text{first} = \text{first} \cup \{j\}$
- (14) end if
- (15) $C[i, j] = 0, C[j, i] = 0$
- (16) else
- (17) $\text{temp} = C$;
- (18) $\text{temp}[p, m] = 0, \text{temp}[m, p] = 0$; 其中 $p \in \text{first}$
- (19) if $Z = 2$ then
- (20) 从 C 中取出最小元素 $\min\{C[r, j]\}$, 其中 $r \in \text{first}, j \notin \text{first}$
- (21) $\text{second} = \text{second} \cup \{j\}$;
- (22) 从 temp 中取到最大元素 $\max\{\text{temp}[i, j]\}$;
- (23) $\text{second} = \text{second} \cup \{j\}$;
- (24) else
- (25) 从 $\text{temp}[q, m]$ 中取得最大值 $\max\{\text{temp}[i, j]\}$, 其中 $q \in \text{second}$
- (26) $\text{second} = \text{second} \cup \{j\}$;
- (27) end if
- (28) $C[i, j] = 0, C[j, i] = 0$
- (29) end if
- (30) end for
- (31) return $T(\text{first}, \text{second})$

分类树构造算法是把分类树分成两个分支。从矩阵 C 中分别选出最大值和次大值对应的项,并把项存放到 first 数组和 second 数组中,把已经选出的数值置为零。对 $K = \text{mod}(Z, 2)$ 进行奇偶数判定,当 Z 为奇数时,把 second 数组中的项所在的

行和列都置为零,再从 first 数组中项所在的行中选出最大值的项,选出的项存放在 first 数组中;当 Z 为偶数时,把 first 数组中的项所在的行和列都置为零,再从 second 数组中项所在的行中选出最大值的项,选出的项存放在 second 数组中,直到所有的项都被选出后,算法结束。算法 2 的分类树是二分支结构,若要改变成多分支形式,只需改变 $K = \text{mod}(Z, i)$ 就可以了,所以这个算法具有一定的可扩展性。

3.3 算法的复杂性分析

算法 1 和 2 的运行时间最大复杂性为 $O(n \cdot |D|)$ 。本文中最主要的计算代价是算法 1 的关系矩阵算法,它需要遍历整个数据集,在其中选出只出现一次的记录,组合成二项集,生成矩阵。在数据集中选出只出现一次的记录的时间复杂性是 $O(n \cdot |D|)$,这里的 $|D|$ 是集值型数据集的长度, n 为给定的项集全集 $|I|$ 的输出域 o , n 的取值不超过输出域,因为它需要查询整个数据集,通过对比每条数据才能选出。

4 实验设计与结果分析

为了明确影响数据发布精度的因素,本文实现了基于分类树的 DiffPart 算法。通过对数据集特征以及平均相对误差定义的分析,从分类树、记录的种类数以及只出现一次的记录的种类数三个方面分别设计实验方案,并对结果进行了分析。本实验使用 C# 语言实现,编程环境是 Microsoft Visual Studio 2010,实验环境是: Intel® Core™ i5 CPU 2.67 GHz; 4 GB 内存; Win 7 操作系统。

4.1 实验设计

为了验证分类树、记录种类数占输出域的比率和只出现一次的记录种类数占输出域的比率和平均相对误差的关系,本实验分别使用两个数据集:手动构造的数据集和 MSNBC 数据集。两个数据集的特征如表 2 所示。手动构造的数据集中,项的全集 $I = \{I_1, I_2, I_3, I_4, I_5, I_6, I_7, I_8\}$, 含有 1 000 条记录,其中 100 条记录是只出现一次的记录。MSNBC^[16] 数据集是一个真实的数据集,这个数据集是一个记录在一定时间段内,用户访问 URL 网站类别数。本实验忽略其数据集中的序列性,把它转变成可用的集值型数据,其每条记录中包含的是一组一名用户访问 URL 网站类别。

表 2 实验数据集统计表

数据集	$ D $	$ I $	$\max D $
MSNBC	989818	17	17
手动构造的数据集	1000	8	8

本文通过以下三个方面来验证:

(a) 不同的分类树对平均相对误差的影响

本次实验使用手动构造的数据集,分别取隐私预算 $\epsilon = \{0.5, 0.75\}$, 构造不同的分类树。实验取 $F = 4$ 对其平均相对误差的影响。本文对不同长度的项集随机产生 5 000 次计数查询。为便于和 DiffPart 算法进行比较,采用文献[12]中查询长度,其查询的记录长度为 $[1, (i \cdot \max |D|/5)]$, 其中 $i = \{1, 2, 3, 4, 5\}$ 。

(b) IOR 对伪空节点的影响

本实验从 MSNBC 的数据集中取其前 8 项 $I = \{I_1, I_2, I_3, I_4, I_5, I_6, I_7, I_8\}$ 的所有记录、前 12 项 $I = \{I_1, I_2, I_3, I_4, I_5, I_6, I_7, I_8, I_9, I_{10}, I_{11}, I_{12}\}$ 的所有记录。分别从 8 项集和 12 项集取出 $\text{IOR} = \{20\%, 40\%, 50\%\}$ 的记录,对其所有分类树产生的所有

伪空节点数相加,取其平均值。

(c) SIOR 对平均相对误差的影响

本实验从 MSNBC 这个集值型数据集中取其前 8 项的所有记录前 12 项的所有记录。取 $F = 4$ 和 $F = 6$, 选择 IOR 对伪空节点有较大影响时取 SIOR 不同值时,伪空节点数对其平均相对误差的影响。

4.2 结果与分析

实验分别从不同分类树、不同 IOR、不同 SIOR 与伪空节点数、平均相对误差间的关系对实验结果进行了分析。

(1) 不同分类树对平均相对误差的影响

实验分别构造了如表 3 所示的四种不同分类树,它们产生的伪空节点个数也不尽相同。图 3 分别对应了隐私预算取不同值时不同查询长度对应的平均相对误差数,从图 3 中可以看出当取值越大时,四种不同分类树构造方法间对应的平均相对误差的差值越小,但四种方法中都是方法 1 所对应的平均相对误差较小。因此不难看出,分类树不同所产生的平均相对误差不同,分类树不同所产生的伪空节点数也不同,而且伪空节点数越多,所产生的平均相对误差也越小。

通过对平均相对误差定义的分析不难看出,因为在伪空节点上是不添加任何噪声,所以随着伪空节点数的不断增加,平均相对误差就会减小。

表 3 分类树表

分法序号	分类树(两个分支)	伪空节点数目/个
分法 1	$\{I_1, I_3, I_4, I_5\}, \{I_2, I_6, I_7, I_8\}$	3
分法 2	$\{I_1, I_2, I_3, I_5\}, \{I_4, I_6, I_7, I_8\}$	1
分法 3	$\{I_1, I_2, I_4, I_6\}, \{I_3, I_5, I_7, I_8\}$	0
分法 4	$\{I_1, I_2, I_3, I_6\}, \{I_4, I_5, I_7, I_8\}$	0

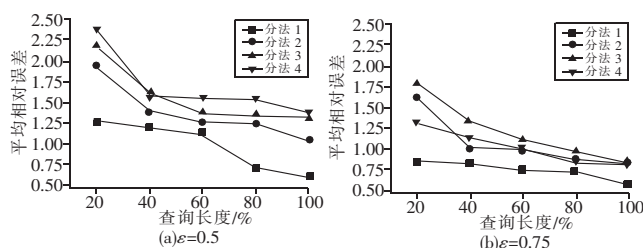


图 3 平均相对误差与隐私预算的关系

(2) IOR 对伪空节点的影响

图 4 表示 $\text{IOR} = \{20\%, 40\%, 50\%\}$ 时, 8 项集的 35 种分类树和 12 项集的所有分类树的伪空节点数之和的平均值。从图 3 中可以看出, 当 $\text{IOR} = 40\%$ 时, 12 项集产生的伪空节点数最多为 60 个; 当 $\text{IOR} = 20\%$ 时, 12 项集的伪空节点数稍有减少为 50 个; 但当 $\text{IOR} = 50\%$ 时, 12 项的伪空节点数急剧下降为 5 个, 说明 IOR 不同取值对伪空节点有较大影响。

因为记录的种类数较多时, 无论分类树如何划分, 最终几乎所有的子分割都能到达叶子分割, 那么出现伪空节点的情况就会很少, 此时分类树并无优劣之分, 可以随机选择一种分类树。当数据集中的记录种类数占输出域不大于 40% 时, 由于很多子分割没有分配到记录, 出现伪空节点的更多, 此时就需要选择一种较优的分类树。

(3) SIOR 对平均相对误差的影响

从实验(2)看出, 当 $\text{IOR} = 40\%$ 时, 对伪空节点影响比较大。图 5 和 6 分别表示 $\text{IOR} = 40\%$ 时, $\text{SIOR} = \{5\%, 10\%, 15\%, 20\%, 30\%\}$, 取 8 项集和 12 项集的平均相对误差的负对数值(不同长度的查询平均相对误差很大), 负对数值越大, 平

均相对误差就越小。图 7 表示当 SIOR 不同时,伪空节点数的变化。

结合图 5~7 可以看出,当数据集中的 IOR 不大于 40% 时,当 $SIOR = (5\%, 20\%]$ 时,数据的敏感度很高,不同的数据集会出现不同的伪空节点,数目各不同。这是因为只出现一次的记录的种类数控制着伪空节点的出现情况,若只出现一次的记录种类数特别少时,伪空节点几乎不会出现。因此, $SIOR = (5\%, 20\%]$ 时选择更优的分类树有利于增加伪空节点的出现,同时减小平均相对误差。

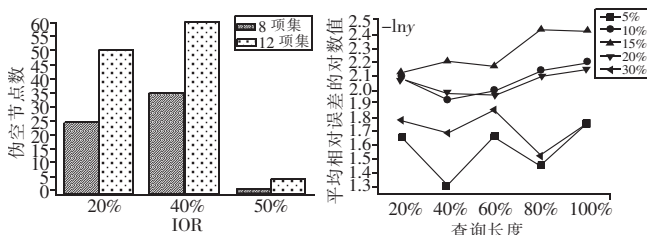


图 4 记录种类数/输出域与伪空节点关系

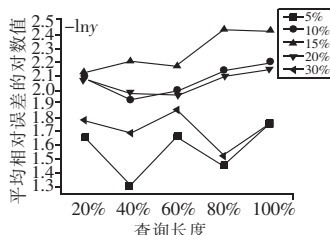


图 5 8 项集平均相对误差曲线

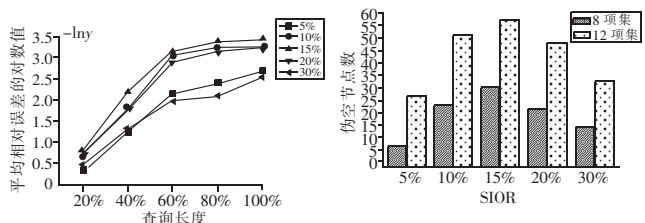


图 6 12 项集的平均相对误差曲线

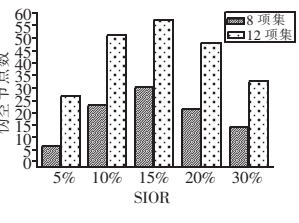


图 7 只出现一次记录的种类数/输出域与伪空节点关系

上述实验从分类树、IOR 和 SIOR 三个方面对数据集进行实验分析。从实验结果得出,对于一个已知的集值型数据集,在 $IOR = 40\%$ 并且 $SIOR = (5\%, 20\%]$ 时,构造一个较优的分类树,其平均相对误差会有明显减小,从而数据的发布精度得到了提高。

通过采用 MSNBC 数据集中 8 项集的记录对实验结果进行验证。取 $IOR = 40\%$ 且 $SIOR = [5\%, 30\%]$ 时,对其平均相对误差改进率进行分析。平均相对误差改进率 = (优化值 - 平均值) / 平均值,其中优化值是优化的分类树取平均相对误差值,平均值是表示通过对 35 种分类树实验后的平均相对误差取其平均值。图 8 所示为 8 项集的平均相对误差改进率。从图 8 中可知,在 $IOR = (5\%, 20\%]$ 时,平均相对误差改进率更好。

为进一步验证 IOR 和 SIOR 对 MSNBC 数据集的影响,通过采用 $F=9$ 对整个的 MSNBC 数据集对实验结果进行验证。分法 1 为构造较优的分类树,分法 2 是表示取 20 种不同的分类树所得的平均值。从实验图 9 中可知,分法 1 和 2 几乎没有差别。由于 $IOR = 8\%$ 且 $SIOR = 4\%$,所以,分类树的优劣不会影响该数据集。因此,针对于此数据集来说,随机选择一种分类树即可。

通过上述实验结果分析,对于一个已知的数据集来说,只有存在只出现一次的记录时,在划分时才有可能会产生伪空节点。由此可以推断,只出现一次的记录存在与否,对伪空节点起到了重要作用。但是对于同一个子分割所划分的叶子分割,即使每个叶子分割仅包含一条记录,此子分割仍能到达叶子分割。当记录的种类数过多时,不管其中包含多少种只出现一次的记录,几乎所有的子分割都能到达叶子分割。因此,记录的

种类数和只出现一次的记录的种类数都会影响伪空节点的产生。所以,在 IOR 和 SIOR 满足条件时,选择一种更优的分类树更有利于数据的利用率。

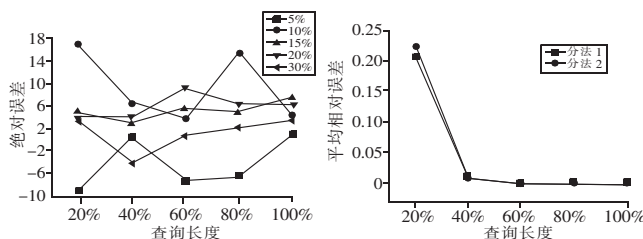


图 8 8 项集平均相对误差曲线

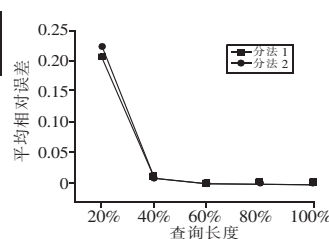


图 9 MSNBC 数据集的平均相对误差和隐私预算关系

5 基于数据集特征的集值型数据发布算法设计

从理论角度推断,对于一个已知的集值型数据集,当记录的种类数和单个出现记录的种类数占输出域一定比率时,构造较优的分类树从一定程度上可以增加伪空节点的出现,从而大大减小平均相对误差,很大程度上可以平衡数据利用率和隐私保护之间的关系,即在保护个人隐私安全性的同时,增加数据的可利用性。从实验的角度看,对于同一数据集,当 $IOR \leq 40\%$ 且 $SIOR = (5\%, 20\%]$ 时,随着分类树的不断优化,伪空节点数也随之增大,而平均相对误差则会减小。因此,在数据集满足一定条件下,构造一种较优的分类树更有利于 DiffPart 算法的实现,更能够在保护个人隐私的同时,增大数据的利用率。当然,这对于树型结构的隐私保护非常重要。因此,根据理论依据和实验结果,设计了如图 10 所示的基于数据集特征的分类树的动态构造算法。对于给定的数据集 D ,首先判断数据集是否满足 $IOR \leq 40\% \&\& SIOR = (5\%, 20\%]$ 的条件,若满足条件,则使用优化分类树的算法来动态构造分类树;若不满足条件,则随机选择一种分类树。

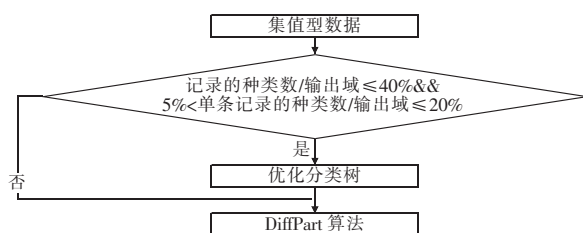


图 10 集值型数据发布算法思想流程图

6 结束语

在 ϵ -差分隐私的保护下,基于 DiffPart 算法,对数据集特征进行研究与分析,在记录的种类数和只出现一次的记录的种类数满足条件后,使用分类树动态构造方法添加的噪声量更少。与 DiffPart 算法相比较,其方法可以有效地保护个人敏感信息的同时提高集值型数据的利用率。下一步将继续深入研究在同一级别隐私保护下,根据集值型数据集本身的特征来提高数据的发布精度,并且将继续研究集值型数据的差分隐私保护方法。

参考文献:

- [1] 张啸剑,王森,孟小峰. 差分隐私保护下一种精确挖掘 top-k 频繁模式方法[J]. 计算机研究与发展, 2014, 51(1): 104-114.
- [2] 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): 103-120.

(下转第 页)

(上接第 页)

- [3] 康海燕. 面向大数据的个性化检索中用户匿名化方法[J]. 西安电子科技大学学报, 2014, 41(5): 170-176.
- [4] 张啸剑, 孟晓峰. 面向数据发布和分析的差分隐私保护研究[J]. 计算机学报, 2014, 37(4): 927-949.
- [5] 熊平, 朱天清, 金大卫. 一种面向决策树构建的差分隐私保护算法[J]. 计算机应用研究, 2014, 31(9): 3108-3112.
- [6] Dwork C. Differential privacy[M]//Automata, Languages and Programming. Berlin Heidelberg: Springer, 2006: 1-12.
- [7] Terrovitis M, Mamoulis N, Kalnis P. Privacy-preserving anonymization of set-valued data[J]. *Proceedings of the VLDB Endowment*, 2008, 1(1): 115-125.
- [8] 毛云青. 高效的集值属性数据隐私保护发布技术研究[D]. 杭州: 浙江大学, 2011.
- [9] Chen R, Acs G, Castelluccia C. Differentially private sequential data publication via variable-length n-grams[C]//Proc of ACM Conference on Computer and Communications Security. [S. l.]: ACM Press, 2012: 638-649.
- [10] Chen Rui, Fung B C M, Desai B C, et al. Differentially private transit data publication: a case study on the montreal transportation system[C]//Proc of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. [S. l.]: ACM Press, 2012.
- [11] Cormode G, PROCOPIUC M, Shen Entong, et al. Differentially private spatial decompositions[C]//Proc of the 28th IEEE International Conference on Data Engineering. [S. l.]: IEEE, 2012.
- [12] He Yeye, Naughton J F. Anonymization of set-valued data via top-down, local generalization[J]. *Processing VLDB Endowment*, 2009, 2(1): 934-945.
- [13] Chen Rui, Mohammed N, Fung B C M, et al. Publishing set-valued data via differential privacy[J]. *Proceedings of the VLDB Endowment*, 2011, 4(11): 1087-1098.
- [14] Blum A, Ligett K, Roth A. A learning theory approach to noninteractive database privacy[J]. *Journal of the ACM*, 2013, 60(2): 12.
- [15] Xiao Xiaokui, Bender G, Hay M, et al. iReduct: differential privacy with reduced relative errors[C]//Proc of the ACM SIGMOD International Conference on Management of Data. [S. l.]: ACM Press, 2011: 229-240.
- [16] Xiao Xiaokui, Tao Yufei. Personalized privacy preservation[C]//Proc of the ACM SIGMOD International Conference on Management of Data. [S. l.]: ACM Press, 2006.
- [17] Heckerman D. MSNBC[EB/OL]. (1999-09-28). [http://archive.uci.edu/ml/datasets/MSNBC.com + Anonymous + Web + Data](http://archive.uci.edu/ml/datasets/MSNBC.com+Anonymous+Web+Data).