基于分类树的动态集值型数据发布的隐私保护

石秀金 胡艳玲
(东华大学计算机科学与技术学院 上海 201620)

摘 要 基于分类树的差分隐私保护方法有效地对静态集值型数据进行了保护,但对于动态集值型数据却没有相应的保护方法,因此提出一种基于分类树的差分隐私保护下的动态集值型数据发布的算法。该算法首先根据数据集中项的全集构造关系矩阵,挑选关系最紧密的项集构造分类树;然后设定一个边界值来限制数据的增量更新,并将新增的记录添加到分类树的根节点中,按照初始分类树的分配法迭代分配每个记录;最后根据拉普拉斯机制向叶子节点中加入噪音,保证整个算法满足差分隐私的要求。相对已有算法,所提算法优化了分类树,使所发布数据建立的分类树模型有少量的叶子节点产生,减少了噪音的添加。实验用两组真实的数据集验证了所提算法的有效性和相对于其他算法的优越性。

关键词 隐私保护,分类树,动态集值型数据,增量更新
中图法分类号 TP309.7 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.05.022

Privacy Preserving Based on Taxonomy Tree for Dynamic Set-valued Data Publishing

SHI Xiu-jin HU Yan-ling
(School of Computer Science and Technology, Donghua University, Shanghai 201620, China)

Abstract Differential privacy protection method based on taxonomy tree is effective for static set-valued of data protection, but it doesn't have corresponding protection method for dynamic set-valued data. So, this paper presented a classification tree based analysis of differential protection of privacy under dynamic set-valued data released. Firstly, according to the complete set structure relation matrix of the data set, this algorithm chooses the most closely related item set, and constructs the taxonomy tree. And then a boundary value is set to limit the incremental data update, and the new record is added to the root node of the taxonomy tree, in accordance with the initial taxonomy tree distribution method iteratively assigns each record. Finally, according to the Laplace mechanism, noise is added into leaf node to ensure that the algorithm satisfies differential privacy requirements. Compared with the existing algorithms, this algorithm optimizes the taxonomy tree, so that the release of data taxonomy tree model is established with small leaf nodes, reducing the noise added. Experiment with two real datasets shows that the algorithm is effective and performs better than existing algorithms.

Keywords Privacy protection, Taxonomy tree, Dynamic set-valued data, Incremental update

1 引言

随着信息技术的飞速发展,各类数据相继发布,采集和分析变得方便快捷。作为信息资料的一种,集值型数据(如信用卡的交易记录、超市中用户的购物记录、医院电子病例记录)的发布和分享将有利于数据挖掘等研究。例如,分析用户的购物记录可以给超市提供有用的采购信息;采集病人的健康记录将有利于医院分析疾病信息和药品购买信息。然而,这些数据都包含个人敏感信息,直接发布会给个人隐私造成威胁。因此对集值型数据发布的隐私保护研究具有重要的意义。

目前,差分隐私[1-5]已经成为一种新的隐私保护模型,该模型不关心攻击者拥有多少背景知识,其通过向查询或分析结果中添加适当的噪音来达到隐私保护的效果。其最大的优点就是对于大型数据集,仅通过添加极少量的噪音就能达到高级别的隐私保护。这种保护模型降低了隐私泄露的风险,同时保证了数据的可用性。

集值型数据隐私保护通常使用 k-anonymity 模型进行数据保护。Terrovitis 等人[6]于 2008 年提出了 (k,m)-anonymity 隐私模型,通过泛化层次树对集值型元数据中的一些或全部进行泛化处理实现隐私保护。(k,m)-anonymity 隐私原则虽然保护了用户隐私,但很多时候很难确定攻击者拥有多少背景知识,导致 m 值很难确定。毛云青等人[7]于 2011 年提出了 (k,l)-anonymity 隐私模型,对集值型数据的匿名处理进行了严格的限制,实现用户隐私保护。(k,l)-anonymity 隐私原则提高了泛化程度,防止未泛化元素的披露,但是其为集值型数据强制引入了敏感属性,使适用范围受到了损失,对元素

到稿日期:2016-04-11 返修日期:2016-07-17
石秀金(1975—),男,博士,副教授,主要研究方向为个性化推荐、隐私保护、数据分析,E-mail:sxj@dhu.edu.cn;胡艳玲(1991—),女,硕士,主要研究方向为差分隐私保护。

的进一步泛化也造成了更高的信息损失。Chen[8] 等人于 2012 年通过自顶向下的划分方法，提出了分类树结构，向叶子节点中添加噪音来支持集值型数据的发布。这种方法在保护隐私的同时也提高了数据的可用性，但此方法局限于静态数据集，不支持增量更新。目前存在的解决方法是 Chan 和 Dwork 等人提出的数据流方法。由于数据流本身的特点，该方法不适合发布增量更新的集值型数据。

本文主要将分类树与自顶向下的分割方式相结合来实现动态集值型数据隐私保护方法的研究。静态集值型数据隐私保护中最典型的代表是 Chen 等人提出的 DiffPart[9] 算法。DiffPart 算法的思想是将所有数据泛化成分类树的根节点，再从根节点开始迭代生成不同的子分割，最后在叶子节点中添加拉普拉斯噪音。本文将根据 DiffPart 算法来实现动态集值型数据的隐私保护，首先处理增量更新的数据使其满足差分隐私，该方法是将更新的数据作为一张新的表然后使用 DiffPart 算法进行处理，其缺点是随着时间的推移，会有更多的噪音产生，最终导致数据的利用率降低；然后整合新增的数据与初始数据，并使用现有的方法发布新的版本，其缺点是数据更新得越多，每次数据发布添加的噪音也越多。此外，如果更新的数量没有限制，隐私预算 ε 会被耗尽，导致隐私机制不满足 ε-差分隐私。

基于这一问题，随着更新数量的增加，误差也会提高，其原因是每次更新时噪音增加，这就意味着数据集不能无限更新。因此，本文的主要工作是，根据集值型数据本身的特性找出两组频繁项集，将其作为分类树的两个分支来构造一棵较优的分类树，对于新增的数据首先给定一个边界值来限制新增数量，将新增的记录添加到分类树的根节点中，按照初始分类树的分配方法把新增的记录迭代分配到每个子分割中，最后在叶子节点中添加拉普拉斯噪音。

2　问题描述

集值型数据就是一条记录的标识符和一个元素集合相关联的数据，其形式为 $\{personid,\{item_1,item_2,\cdots,item_n\}\}$。一个典型的数据实例就是超市中的用户购物记录，其中 $\{item_1,item_2,\cdots,item_n\}$ 被用来表示编号为 personid 的用户所购买的商品集合。例如，用户 A 在超市里买了香蕉、牛奶、可乐，用户 B 买了苹果、可乐，用户 C 买了香蕉、苹果、牛奶。表 1 列出了 3 个用户在超市购物数据的示例。

表 1　超市购物数据

用户	记录号	记录
用户 A	t_1	香蕉、牛奶、可乐
用户 B	t_2	苹果、可乐
用户 C	t_3	香蕉、苹果、牛奶

动态集值型数据是指将已有的数据作为初始数据集，将新增的数据集追加到初始数据集中。同样以用户在超市的购物记录为应用背景，表 1 中超市购物数据为初始数据集，将用户 D、用户 E 等用户的购物记录追加到初始数据集中，从而构成动态集值型数据，其中超市中物品的种类数保持不变。

定义 1（动态集值型数据）　定义 $I=\{I_1,I_2,\cdots,I_{|I|}\}$ 为项的全集，$|I|$ 表示全集的大小。$T=\{t_1,t_2,\cdots,t_{|T|}\}$ 表示初始化的集值型数据表，每条记录 t_i 由 I 的非空子集构成，其中 $t_i\in T$。$\Delta T_1,\Delta T_2,\cdots$ 表示新增数据的表，假设表 $T,\Delta T_1,\Delta T_2,\cdots$ 的项是固定大小的。

设项的全集 $I=\{I_1,I_2,I_3,I_4\}$（在超市购物记录中香蕉、苹果、牛奶、可乐为项的全集），表 2 是简单的初始化集值型数据集 T，表 3 和表 4 是新增的数据集 $\Delta T_1,\Delta T_2$。

表 2　初始数据集

记录号	记录
t_1	$\{I_1,I_2\}$
t_2	$\{I_2\}$
t_3	$\{I_1\}$

表 3　新增的数据集 ΔT_1

记录号	记录
t_4	$\{I_1,I_2,I_3,I_4\}$
t_5	$\{I_2,I_4\}$
t_6	$\{I_2\}$

表 4　新增的数据集 ΔT_2

记录号	记录
t_7	$\{I_1,I_2,I_3,I_4\}$
t_8	$\{I_2,I_3,I_4\}$
t_9	$\{I_1,I_2\}$

定义 2（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。图 1 示出了表 1—表 3 数据集的一个分类树，$I_{\{1,2,3,4\}}$ 是根节点，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。

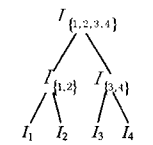


图 1　简单数据集的分类树

定义 3（差分隐私[10]）　对于给定的两个数据集 T_1 和 T_2，它们之间至多相差一条记录。给定一个隐私算法 A，$Rang(A)$ 表示 A 的取值范围，若算法 A 在数据集 T_1 和 T_2 上的任意输出结果 $O(O\in Rang(A))$ 满足下列不等式，则 A 满足 ε-差分隐私。

$$\Pr[A(T_1)=O]\leqslant\exp(\varepsilon)\cdot\Pr[A(T_2)=O] \tag{1}$$

其中，ε 表示隐私预算，$\Pr[A(T_i)=O]$ 表示算法的概率，其由 A 的随机性控制。

实现差分隐私保护常用的噪音机制有两种：拉普拉斯机制和指数机制。本文主要采用的是拉普拉斯噪音机制。

定义 4（全局敏感性[11]）　对于任意一个函数 $f:T\to R^d$，函数 f 的全局敏感性为：

$$\Delta f=\max_{T_1,T_2}\|f(T_1)-f(T_2)\|_p \tag{2}$$

其中，T_1 和 T_2 至多相差一条记录，R 表示映射的实数空间，d 表示函数的查询维度，p 表示度量 Δf 使用的 L_p 距离，通常使用 L_1 度量。

定理 1（拉普拉斯机制[11]）　对于任意一个函数 $f:T\to R^d$，若算法 A 的输出结果满足下列不等式，则 A 满足 ε-差分隐私。

$$A(T)=f(T)+\langle Lap_1(\Delta f/\varepsilon),Lap_2(\Delta f/\varepsilon),\cdots,Lap_i(\Delta f/\varepsilon)\rangle \tag{3}$$

其中，$Lap_i(\Delta f/\varepsilon)(1\leqslant i\leqslant d)$是相互独立的拉普拉斯变量，噪音量大小与 Δf 成正比，与 ε 成反比。算法 A 的全局敏感性越大，所需噪音越大。

定义 5(计数查询[12]) 计数查询对于集值型数据的挖掘是至关重要的。因此本文主要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问题。

对于给定的数据集 T_i 和项集 $I'\in I$，计数查询 Q 对数据集 T_i 上的查询 $Q(T_i)=|\{t\in T_i, I'\in t\}|$。

定义 6(平均相对误差[13-14]) 对于计数查询 Q，本文采用平均相对误差来衡量处理过的数据集 $\tilde{T}_i$ 中数据的利用率，如式(4)所示：

$$RE=\frac{|Q(\tilde{T}_i)-Q(T_i)|}{\max\{Q(T_i),b\}} \tag{4}$$

其中，$Q(\tilde{T}_i)$表示处理过的数据集 $\tilde{T}_i$ 的查询计数，$Q(T_i)$表示原始数据集 T_i 的查询计数，b 为理智约束，防止在极小计数查询的情况下其分母为零。

3 算法及性能分析

本文主要研究解决动态集值型数据发布的隐私保护问题，由于集值型数据是增量更新的，当数据更新时，隐私保护机制也必须更新。因此，传统的差分隐私保护机制不能应用到动态集值型数据集上，会导致隐私预算耗尽等问题，最终降低了数据的利用率。

DiffPart 算法通过对分类树分割结束后的叶子节点添加拉普拉斯噪音来实现隐私保护。分类树中会产生两种节点，即半结构化的空节点(子分割记录数为 1 的不可再分的非叶子分割节点)和叶子节点，这两种节点是影响噪音量的关键因素。在 DiffPart 算法中，仅对叶子节点添加噪音，叶子节点的数目越多，在原始数据集中添加的拉普拉斯噪音也越多，所以要减少噪音量，就必须减少叶子节点的数目。在 DiffPart 算法中，当分类树的分割层次中存在半结构化的空节点，这些分割层次将不会划分至叶子分割，那么叶子节点的数目就会减少。由此可知，随着半结构化的空节点数量的增多，叶子节点就会减少，添加的噪音量就越少。

DiffPart 算法[9]是根据扇出值将项的全集随机分成没有相交项的两组，在分类树划分过程中，将两组间相交的记录划分成第三种子分割。第三种子分割的划分层次越深，产生的叶子节点就越多。只要两组分割中各组组内间数据关系尽可能紧密，两组间的数据关系尽量少就可以减少第三种子分割的划分层次，减少叶子节点数量，降低噪音量。

3.1 算法的基本思想

基于以上问题，为了解决集值型数据的不断更新，减少隐私预算的消耗和拉普拉斯噪音的添加等问题，达到降低平均相对误差的效果，本文选出两组频繁项集构造一棵较优的分类树。其思想是，根据每条记录中任意两项出现的次数，首先选出次数最多的其所对应的项集作为第一组，然后在挑出的项集所在的两行挑出次数最小的项集，再在这个项集所在的行挑出最大的数作为第二组，迭代地挑选其他项集放入这两组中，直到所有的项集被选出，这两个分组就是分类树。引用一种更新边界机制，限制数据集的更新数量，并向构造好的分类树中添加新增的记录，首先设定一个更新数量 U，初始化数据集 T 和隐私预算 ε，为 T 构造分类树 $TBP\_Tree_{(0)}$，然后当新增的数据集 ΔT_i 到达时，先将 ΔT_i 中所有记录添加到 $TBP\_Tree_{(i-1)}$ 的根节点并递归到不相交的子集中，如果某些记录被添加到 $TBP\_Tree_{(i-1)}$ 的非叶子节点中，就根据 $TBP\_Tree_{(i-1)}$ 的分类方法处理此记录；如果某些记录被添加到 $TBP\_Tree_{(i-1)}$ 的叶子节点中且该叶子节点只有一条记录，就重新分配该节点的隐私预算，继续分割该节点，该方法结束后，生成新的分类树 $TBP\_Tree_{(i)}$。最后向 $TBP\_Tree_{(i)}$ 的叶子节点添加拉普拉斯噪音。

3.2 算法的描述

动态数据集的隐私保护算法首先构造一棵较优的分类树，然后根据初始数据集中记录的分配方法将新增数据集中的记录添加到分类树中。分类树构造算法和增量更新算法如算法 1 和算法 2 所示。

算法 1 分类树构造算法

输入：初始集值型数据集 T

输出：分类树

```
1. n=数据集全集|I|
2. 构造 n*n 的矩阵 M
3. m[n,n]=任意两项出现的次数
4. first[],second[];
5. for Z=1 to do
6. k=mod(Z,2)
7.   if k=1then
8.     从矩阵中取出最大的元素 max{m[i,j]}
9.     first=first∪{j};m[i,j]=0;m[j,i]=0
10.  else
11.    从 i,j 所在行取最小元素 min{m[p,q]}
12.    从 p,q 所在行取最大元素 max{m[x,y]}
13.    second=second∪{y}; m[x,y]=0;m[y,x]=0
14. return T(first,second)
```

算法 2 增量更新算法(IncrePart)

输入：初始化数据集 T 和一系列增加的数据集 $\Delta T_1,\Delta T_2,\cdots,\Delta T_U$，数据集更新的数量 U；隐私预算 ε

输出：加噪音的数据集 $\tilde{T},\tilde{T}_i,\cdots$

```
1. 构造一个扇出为 f 的分类树 H
2. ε'=ε/(U+1)
3. 构造初始化分类树 TBP_Tree(0)
4. for each ΔTi(1≤i≤U)do
5.   V=T,ΔT1,ΔT2,…,ΔT(i-1)中所有的节点记录;
6.   P=ΔTi 中所有记录;
7.   p.cut=H 的根节点
8.   p.ε'_V=ε'/2;p.α=p.ε'_V/|InternalNodes(p.cut)|
9.   将 P 添加到 TBP_Tree(i-1)的根节点中;
10.  分配 ΔTi 中所有的记录
11.  for each pi∈P do
12.    if pi∈V then
13.      if pi∈NoEmSet and pi 不是叶子节点 then
14.        根据 TBP_Tree(i-1)分类方法分配此节点
15.      if pi∈SemiEmSet then
16.        pi·ε'_V=pi·ε'-pi·α
```

17. $p_i \cdot \alpha = \frac{p_i \cdot \varepsilon'^{V_i}}{|\text{InternalNodes}(p.\text{cut})|}$
18. $V = p_i \cup V$
19. else
20. if $p_i \in$ NoEmSet and p_i 不是叶子节点 then
21. 重复第 16－第 18 行
22. if p_i 是叶子节点 then
23. $V = p_i \cup V$
24. return $\text{TBP\_Tree}_{(i)}$
25. 发布 $\text{TBP\_Tree}_{(i)}$ 中叶子节点的信息
26. Return $\hat{T}, \hat{T}_i, \cdots$

分类树构造算法是将分类树分成两个分支，首先根据任意两项生成关系矩阵，当 Z 为奇数时从矩阵中挑出最大值放入第一个数组，并把选出来的数置为 0；当 Z 为偶数时，在选出数据所在的行中取最小值，再从这两行中选取最大值放入第二个数组，直到所有的项被选出。

增量更新算法(IncrePart)是将更新的数据集中的记录分配到构造好的分类树中。当第 i 个更新的数据 ΔT_i 到达时，首先将 ΔT_i 中的记录插入到 $\text{TBP\_Tree}_{(i-1)}$ 的根节点中，迭代分配到不同的子分割中。如果 ΔT_i 中的记录被添加至非空的非叶子节点中，则根据 $\text{TBP\_Tree}_{(i-1)}$ 分类方法分割此记录；如果 ΔT_i 中的记录被添加到 $\text{TBP\_Tree}_{(i-1)}$ 中只存在一条记录的非叶子节点中，则重新分配隐私预算，继续分割此节点直到生成叶子节点；如果 ΔT_i 中的记录被添加到 $\text{TBP\_Tree}_{(i-1)}$ 不存在的节点中，判断是否继续分割此节点；如果 ΔT_i 中的记录是叶子节点则直接发布此节点。直到所有的记录被分配完，算法结束。图 2 示出了向初始分类树中添加 ΔT_1，ΔT_2 中记录的划分过程。ΔT_2 中的数据如表 4 所列，隐私预算为 ε'。

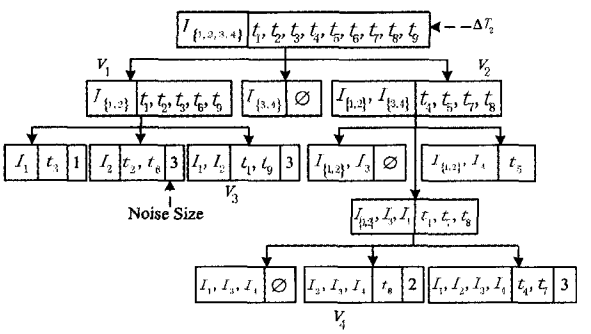


图 2　新增记录的划分过程

如图 2 所示，首先将记录 t_9 添加到节点 V_1 中，记录 t_7，t_8 添加到节点 V_2 中。由于节点 V_1 在 $\text{TBP\_Tree}_{(2)}$ 中的划分过程与在 $\text{TBP\_Tree}_{(1)}$ 中的划分过程一样，因此直接将记录 t_9 添加到叶子节点 V_3 中，噪音数就等于第一次添加的噪音加上此次添加的噪音。节点 V_2 在 $\text{TBP\_Tree}_{(1)}$ 中是一个半结构化的空节点，添加记录 t_7，t_8 后变为非空节点，所以需深层次地划分 V_2 节点，直到生成叶子节点(如节点 V_4)。

3.3　算法性能分析

3.3.1　隐私性分析

增量更新算法根据数据集更新量 U 将全部隐私预算 ε 分为 $\varepsilon'=\varepsilon/(U+1)$ 份，再将 ε' 分为两部分 $B_1=\varepsilon'/2$ 和 $B_2=\varepsilon'/2$。其中，ε' 是初始数据集合每个新增数据集的隐私预算，B_1 用于分类树迭代分割过程中以拉普拉斯机制选择细分方案，B_2 用于每个叶子分割的计数添加 Laplace 噪音。按照差分隐私的序列组合性，多次迭代消耗的隐私预算为各次迭代的累加，当 B_1 耗尽后($B_1<\varepsilon'/2$)迭代过程结束；按照差分隐私的并组合性，叶子分割的总体隐私预算仍为 B_2。所以每次新增数据集消耗的隐私预算不大于 $B_1+B_2=\varepsilon'$，即增量更新算法的全部隐私预算不大于 ε，它具有 ε-差分隐私。

3.3.2　复杂度分析

分类树构造算法的运行时间的最大复杂度为 $O(|T|\cdot|I|)$，增量更新算法的运行时间的最大复杂度都为 $O(|D|\cdot|I|)$，$|T|$是初始集值型数据集的长度，$|D|$是新增集值型数据和初始数据集的总长度，$|I|$是给定项的全集。分类树构造算法中最主要的计算代价是需要遍历整个初始数据集，得到任意两项出现的次数构成$|I|*|I|$的矩阵，因此时间复杂度是 $O(|T|\cdot|I|)$。增量更新算法中最主要的计算代价是根据分类树将所有的记录划分成单个子分割，其时间复杂度是 $O(|D|)$；分类树的扇出 $f\geqslant 2$，子分割数量是$(|I|-1)/(f-1)$，所以总的时间复杂度是 $O(|D|\cdot|I|)$。

4　实验分析

为了验证算法对计数查询的有效性，在相同的测试条件下将本文算法与 Stra-Solu1 算法[15]和 Stra-Solu2 算法[16]进行比较。实验分别使用两组数据集，即 MSNBC[17]和 Kosarak[18]，这两组数据集中的记录都是在一定时间段内用户访问 URL 网站的类别数，实验中忽略数据集中的序列性，将其转变成可用的集值型数据，其每一条记录中包含一名用户访问 URL 网站类别。实验环境为：Inter(R) Core(TM) i5-2450M CPU @ 2.50GHz，8GB 内存，Win7 操作系统，编程环境是 MyEclipse。

表 5　实验数据集

数据集	N	$\|I\|$	$Avg\|t\|$
MSNBC	989818	17	1.72
Kosarak	990002	41270	8.1

其中，N 是数据集中的记录条数，$|I|$是数据集中项的数量，$Avg|t|$是平均每条记录的长度。

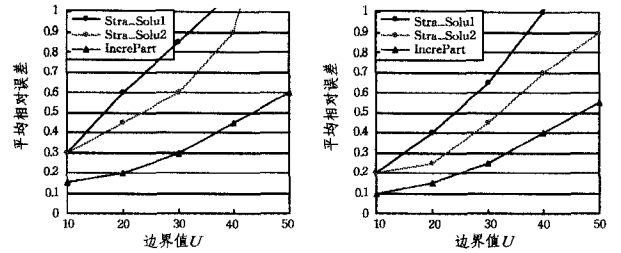
基于这两个数据集，从中随机选择 400000 条记录作为初始数据集 T，另外选择 10000 条记录作为增量更新数据集 ΔT_i。

在 $\varepsilon=0.5, 0.75, 1.0, 1.25$ 以及不同的边界值 $U=10, 20, 30, 40, 50$ 设置下进行了多组实验。每组实验在给定的扇出值 $f=2$ 进行泛化并构造分类树，计数查询为从 I 中随机选一条记录的数量。每组实验进行 10 次，以 10 次结果的平均值作为该组实验的最终结果。

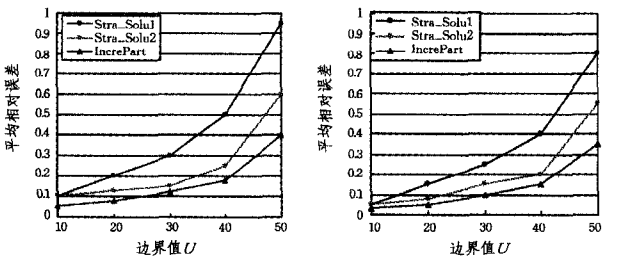
图 3 和图 4 分别示出了数据集为 MSNBC 和 Kosarak 时的平均相对误差。

从图 3 和图 4 可以看出，在不同数据集中，当边界值增加时，查询记录数量的平均相对误差也随之增加；当 ε 越来越大时，同一算法中平均相对误差会减少。相比另外两种算法，本文算法的误差相对较低，因为 IncrePart 算法构造了较优的分类树，将总隐私预算平均分配到新增数据集中，防止了隐私预

算过早耗尽。

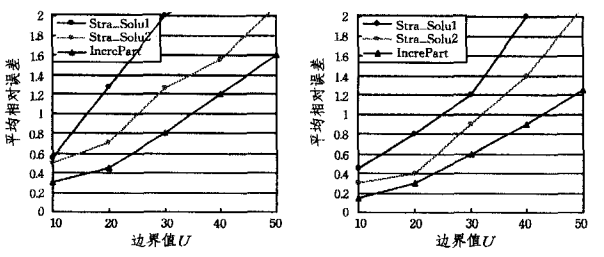


(a)ε=0.5 (b)ε=0.75

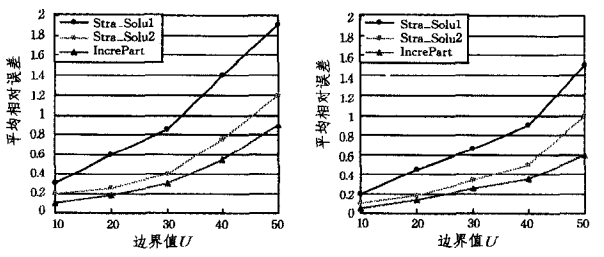


(c)ε=1.0 (d)ε=1.25

图 3 数据集为 MSNBC 时的平均相对误差



(a)ε=0.5 (b)ε=0.75



(c)ε=1.0 (d)ε=1.25

图 4 数据集为 Kosarak 时的平均相对误差

为了与 Stra_Solu1 算法、Stra_Solu2 算法的性能进行比较，取不同数据集记录数，比较各算法的运行时间。实验中边界值 $U=50$，$f=2$，$\varepsilon=1.0$，数据集记录数为 500k～900k，如图 5 和图 6 所示。从图 5 和图 6 可以看出，在不同的数据集中，当数据集记录数增大时，运行时间也随之增加；在相同记录数的情况下，本文算法的运行时间相对较少。

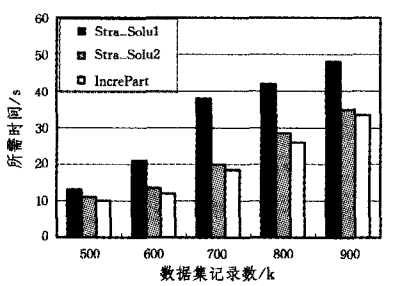


图 5 数据集为 MSNBC 时所需的时间

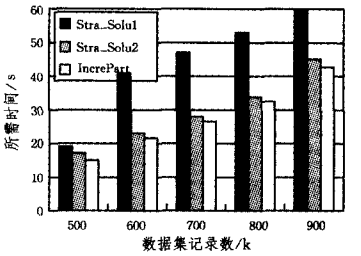


图 6 数据集为 Kosarak 时所需的时间

结束语 在差分隐私保护下，解决了基于分类树的动态集值型数据的发布问题。首先根据初始数据集中任意两项出现的次数生成关系矩阵，从中得到关系最紧密的两组项集，以此来构造较优的分类树；然后设定一个边界值来限制新增数据集的数量，将新增数据集中的每条记录添加到分类树的根节点中，动态分配新增的记录至每个子分割。此方法可以有效地保护个人隐私信息，提高数据的利用率。下一步将继续研究集值型数据增量更新的差分隐私保护。

参 考 文 献

[1] XIONG P, ZHU T Q, WANG X F. A Survey on Differential Privacy and Application [J]. Journal of Computers, 2014, 37(1): 101-122. (in Chinese)
熊平，朱天清，王晓峰. 差分隐私保护及其应用[J]. 计算机学报，2014，37(1)：101-122.

[2] ZHANG X J, MENG X F. Differential Privacy in Data Publication and Analysis [J]. Journal of Computers, 2014, 37(4): 101-122. (in Chinese)
张啸剑，孟晓峰. 面向数据发布和分析的差分隐私保护[J]. 计算机学报，2014，37(4)：101-122.

[3] LI Y, HAO Z F, WEN W, et al. Research on Differential Privacy Preserving K-means Clustering[J]. Journal of Computers, 2013, 40(3): 287-290. (in Chinese)
李杨，郝志峰，温雯，等. 差分隐私保护 k-means 聚类方法研[J]. 计算机科学，2013，40(3)：287-290.

[4] ZHOU S G, LI F, TAO Y F, et al. Privacy Preservation in Database Application: A Survey[J]. Journal of Computers, 2009, 32(5): 847-861. (in Chinese)
周水庚，李丰，陶宇飞，等. 面向数据库应用的隐私保护研究综述[J]. 计算机学报，2009，32(5)：847-861.

[5] TERROVITIS M, MAMOULIS N, KALNIS P. Privacy-preserving anonymization of set-valued data[J]. Proceedings of the Vldb Endowment, 2008, 1(1): 115-125.

[6] MAO Y Q. A Study of Efficiently Privacy Preserving Data Publishing of Set-valued Data[D]. Hanzhou: Zhejiang University, 2011. (in Chinese)
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.

[7] CHEN R, ACS G, CASTELLUCCIA C. Differentially Private Sequential Data Publication via Variable-Length N-Grams[C]// ACM Conference on Computer and Communications Security (CCS). 2012: 638-649.

[8] CHEN R, MOHAMMED N, FUNG B C M, et al. Publishing Set-Valued Data via Differential Privacy[J]. Proceedings VLDB Endowment, 2011, 4(11): 1087-1098.

(下转第 165 页)

[3] OSTROVSKY R, SAHAI A, WATERS B. Attribute-based encryption with non-monotonic access structures [C] // Proceedings of ACM Conference on Computer and Communication Security-CCS 2007. ACM Press, 2007: 195-203.

[4] LEWKO A, OKAMOTO T, SAHAI A, et al. Fully secure functional encryption: attribute-based encryption and (hierarchical) inner product encryption [C] // Advances in Cryptology-EUROCRYPT 2010, LNCS 6110. Springer-Verlag, 2010: 62-91.

[5] OKAMOTO T, TAKASHIMA K. Fully secure functional encryption with general relations from the decisional linear assumption [C] // Advances in Cryptology-CRYPTO 2010. Springer-Verlag, 2010: 191-208.

[6] HERRANZ J, LAGUILLAUMIE F, RAFOLS C. Constant-size ciphertext in threshold attribute-based encryption [C] // Proceedings of Public Key Cryptology-PKC 2010, LNCS 6056. Springer-Verlag, 2010: 19-34.

[7] WATERS B. Ciphertext-policy attribute-based encryption: An expressive, efficient, and provably secure realization [C] // Public Key Cryptography- PKC 2011. Springer Berlin Heidelberg, 2011: 53-70.

[8] YAMADA S, ATTRAPADUNG N, HANAOKA G, et al. Generic constructions for chosen ciphertext secure attribute based encryption [C] // Proceedings of Public Key Cryptology- PKC 2011, LNCS 6571. Springer-Verlag, 2011: 71-89.

[9] LEWKO A, WATERS B. New proof methods for attribute-based encryption: achieving full security through selective techniques [C] // Advances in Cryptology-CRYPTO 2012, LNCS 7417. Springer-Verlag, 2012: 180-198.

[10] HOHENBERGER S, WATERS B. Attribute based encryption: with fast decryption [C] // Proceedings of Public Key Cryptology-PKC 2013. Springer-Verlag, 2013: 162-179.

[11] ROUSELAKIS Y, WATERS B. Practical constructions and new proof methods for large universe attribute-based encryption [C] // Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security. ACM, 2013: 463-474.

[12] NARUSE T, MOHRI M, SHIRAISHI Y. Attribute-based encryption with attribute revocation and grant function using proxy re-encryption and attribute key for updating [M] // Future Information Technology. 2014: 119-125.

[13] QIAN H, LI J, ZHANG Y, et al. Privacy Preserving Personal Health Record Using Multi-Authority Attribute-Based Encryption with Revocation [J]. International Journal of Information Security, 2015, 14(6): 487-497.

[14] ZHANG K, GONG J, TANG S, et al. Practical and Efficient Attribute-Based Encryption with Constant-Size Ciphertexts in Outsourced Verifiable Computation [C] // Proceedings of the 11th ACM on Asia Conference on Computer and Communications Security. ACM, 2016: 269-279.

[15] SHI Y, ZHENG Q, LIU J, et al. Directly revocable key-policy attribute-based encryption with verifiable ciphertext delegation [J]. Information Sciences, 2015, 295: 221-231.

[16] HINEK M J, JIANG S, SAFAVI-NAINI R, et al. Attribute-based encryption with key cloning protection [J]. Bulltein of the Korean Mathematical Society, 2008, 2008(4): 803-819.

[17] YU S, REN K, LOU W, et al. Defending against key abuse attacks in KP-ABE enabled broadcast systems [C] // Security and Privacy in Communication Networks. Springer Berlin Heidelberg, 2009: 311-329.

[18] LI J, REN K, KIM K. A2BE: Accountable attribute-based encryption for abuse free access control [EB/OL]. [2009-03-11]. http://eprint.iacr.org/2009/118.

[19] KATZ J, SCHRODER D. Tracing insider attacks in the context of predicate encryption schemes [EB/ OL]. https:// www.usukita.org/node/1779.

[20] LIU Z, CAO Z, WONG D S. White-box traceable ciphertext-policy attribute-based encryption supporting any monotone access structures [J]. IEEE Transactions on Information Forensics and Security, 2013, 8(1): 76-88.

[21] NING J, DONG X, CAO Z, et al. White-Box Traceable Ciphertext-Policy Attribute-Based Encryption Supporting Flexible Attributes [J]. IEEE Transactions on Information Forensics and Security, 2015, 10(6): 1274-1288.

[22] ZHANG X, JIN C, WEN Z, et al. Attribute-Based Encryption without Key Escrow [C] // Cloud Computing and Security 2015, LNCS 9483. Springer-Verlag, 2015: 74-87.

[23] BONEH D, LYNN B, SHACHAM H. Short signatures from the Weil pairing [C] // Advances in Cryptology—ASIACRYPT 2001. Springer Berlin Heidelberg, 2001: 514-532.

[24] POINTCHEVAL D, STERN J. Security arguments for digital signatures and blind signature [J]. Journal of Cryptology, 2000, 13(3): 361-396.

（上接第 124 页）

[9] SWEENEY L. Achieving k-anonymity privacy protection using generalization and suppression [J]. International Journal on Uncertaint, Fuzziness and Knowledge-based Systems, 2002, 10(5): 571-588.

[10] LU W T, MIKLAU G, GUPTA V. Generating private Synthetic Databases for Untrusted System Evaluation [C] // Proceedings of the 30th International Conference on Data Engineerin (ICDE). Washington, USA, 2014: 654-663.

[11] DWORK C. Differential privacy in new settings [C] // Proc. Symposium on Discrete Algorithms (SODA), Society for Industrial and Applied Mathematics. 2010: 174-183.

[12] DWORK C. A firm foundation for private data analysis [J]. Communications of the ACM, 2011, 54(1): 86-95.

[13] XIAO X K, WANG G Z, GCHRKE J. Differential privacy via wavelet transforms [J]. IEEE Transactions on Knowledge and Data Engincering, 2011, 23(8): 1200-1214.

[14] NERGIZ M E, CLIFTON C, NERGIZ A E. Multirelational k-anonymity [J]. IEEE Trans on Knowledge and Data Engineering, 2009, 21(8): 1104-1117.

[15] Releasing Differentially Private DataCubes for Health Information [C] // Proceedings of the 28th International Conference on Data Engineerin (ICDE). Washington, USA, 2012: 1305-1308.

[16] HECKERMAN D. MSNBC [EB/OL]. (1999-09-28). http://archive.ics.uci.edu/ml/datasets.

[17] Frequent itemset mining dataset repository [OL]. http://fimi.ua.ac.be/data.