计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2017,26(8):206−211 [doi: 10.15888/j.cnki.csa.005910]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 KD 树的信息发布隐私保护①

林国滨[1], 姚志强[1,2], 熊金波[1,2], 林铭炜[1,2]

[1](福建师范大学 软件学院, 福州 350108)
[2](福建省公共服务大数据挖掘与应用工程技术研究中心, 福州 350108)
通讯作者: 姚志强, E-mail: yzq@fjnu.edu.cn

摘　要: 随着医疗信息共享服务的发展, 越来越多的患者病历信息被发布出来, 敌手通过患者属性推断患者的隐私信息, 从而造成患者隐私泄露. 基于上述需求, 提出基于 KD 树的隐私保护数据发布算法. 利用 KD 树的性质, 对每一维所在属性的泛化值进行分解, 直到所有属性的泛化值不能分解, 以确保每个叶子节点的所有属性的泛化值的区域达到最小, 以减少信息损失. 在对等价元组属性分解期间, 对每个节点敏感属性值个数做 *l* 多样性约束, 以降低隐私泄漏风险. 实验结果表明, 方案可以减少隐私泄露风险和信息损失.

关键词: 隐私保护; 数据发布; *k*-匿名; *l*-多样性; 信息损失

引用格式: 林国滨,姚志强,熊金波,林铭炜.基于 KD 树的信息发布隐私保护.计算机系统应用,2017,26(8):206–211. http://www.c-s-a.org.cn/1003-3254/5910.html

KD Tree-Based Privacy Protection of Data Publishing

LIN Guo-Bin[1], YAO Zhi-Qiang[1,2], XIONG Jin-Bo[1,2], LIN Ming-Wei[1,2]

[1](Faculty of Software, Fujian Normal University, Fuzhou 350108, China)
[2](Fujian Engineering Research Center of Public Service Big Data Mining and Application, Fuzhou 350108, China)

Abstract: With the development of regional health information sharing services, an increasing number of patient records are released. However, the adversary can infer the patient's privacy information through the patient's attributes, thereby causing the patient's privacy leakage. Based on the above requirements, a privacy protection data publishing algorithm based on KD tree is proposed. By the properties of KD-tree, the generalized value of each attribute is decomposed until the generalized value of all attributes cannot be decomposed to ensure that the generalized value of all attributes of each leaf node is minimized to reduce the information loss. During the decomposition of equivalent tuple attributes, the number of sensitive attribute values for each node is made to be a diversity constraint to reduce the risk of privacy leakage. The experimental results show that this scheme can reduce the risk of leakage of privacy, and information loss.

Key words: privacy protection; data releasing; *k*-anonymous; *l*-diversity; information loss

引言

为了分析和挖掘医疗信息, 数据接收者通过向病历拥有者请求病历, 以分析病历信息. 而数据拥有者发布医疗病历信息, 即便对病历部分信息处理, 敌手可以通过获取外部信息, 与发布的病历信息关联, 从而获取到患者的隐私信息. 文献[1]表明 87% 的美国公民可以通过邮政编码、性别、生日日期来唯一确定, 而通过收集选民信息表, 和医疗信息表, 之后通过邮政编码、性别、生日日期等属性链接, 可以获取美国公民信息, 造成隐私泄露.

敌手获取外部信息, 与发布的医疗信息表的进行关联, 而获取患者隐私信息. 所以, 文献[1,2]提出了 *k*-

① 基金项目: 国家自然科学基金(61370078, 61402109, 61502102)
收稿时间: 2016-12-11; 采用时间: 2017-01-04

匿名保护模型, 通过泛化用户的准标识符使其达到 k 组元组(k≥2), 来达到隐藏用户的隐私数据. 文献[3]提出 *LKC-privacy* 隐私保护模型, 其可以应用于大规模的匿名数据集. 文献[4]针对中国电子病历文档, 提出 *PHI* 机制, 以提取文档信息内容, 应用自底向上匿名化方法, 对文档匿名化操作, 以保护患者隐私. 而在做隐私保护的时候, 由于泛化操作, 引起信息损失. 为了解决这个问题, 文献[5]提出了 *KMCSSA* 算法, 采用 *k-member* 聚类算法, 保证每一组集合聚类含有 k 条记录, 以便对患者记录做匿名化处理达到减少信息损失的目的. 文献[6]采用 *K-means* 和 *C-means* 对医疗病历进行聚类, 利用 *LKC-privacy* 隐私保护模型, 对其做隐私保护, 以提高数据的通用性, 降低信息损失.

在满足 k 匿名时, 有可能因一组等价类的敏感属性值出现相同, 而引发信息泄露. 为了解决问题, 文献[7]提出 l-多样性技术, 这 k 组元组上的敏感属性出现不同的个数至少满足 l(l≥2)个, 以此降低泄露用户隐私的可能性. 文献[8]提出 *t-Closeness* 技术, 其要求在每一个等价类敏感属性的分布接近的在整体表中的属性分布, 以此来降低隐私泄露的概率. 文献[9]在对电子病历文档, 提出了增强型的电子病历隐私保护模型, 即在每一组等价类, 敏感属性值的个数不少于 l 个, 以此降低患者隐私泄露风险. 文献[10]在对患者病历发布时, 提出了针对敏感属性聚类的方法, 以减少隐私泄露的风险. 然而患者的隐私泄露风险降低了, 却提高了数据信息损失, 降低了数据的通用性.

综上所述, 为了降低电子病历中存在的患者隐私信息泄露的风险, 并且降低匿名化过程中所造成信息损失, 降低数据的通用性, 本文提出了基于 KD 树的隐私算法, 采用 KD 树, 将患者属性作为一个节点, 并将所有患者记录泛化到最大, 利用 KD 树的性质, 对节点的每一维属性的泛化值进行分解, 直到其所有属性的泛化值不能分解, 以确保等价类的信息损失达到最低. 期间, 对每个节点, 也就是等价元组的敏感属性值做 l 多样性约束, 以降低隐私泄露风险. 由此形成一颗 KD 树, 最后发布树的叶子节点(等价类).

1 相关概念

定义 1. 准标识符, 是指一张表内属性能够与外部数据的属性相连接, 而获取用户的隐私信息, 我们称为准标识符. 如表 1 所示, Sex、Age、Job 属于准标识符

表 1 原始数据

ID	Sex	Age	Job	Surgery
1	M	34	Janitor	Transgender
2	M	58	Doctor	Plastic
3	M	34	Mover	Transgender
4	M	24	Lawyer	Vascular
5	M	58	Mover	Urology
6	M	44	Janitor	Plastic
7	M	24	Doctor	Urology
8	M	58	Lawyer	Plastic
9	M	24	Doctor	Vascular
10	F	63	Carpenter	Vascular
11	F	63	Technician	Plastic

定义 2. 敏感属性, 是指用户不希望公布的信息, 如表 1 所示, Surgery 等, 我们称为敏感属性.

定义 3. 泛化, 是指将属性值扩大形成一个区间, 为泛化操作. 如 Age 25, 泛化成[21, 30]. 泛化树[11], 是指一个属性的属性域 D(D 为有限集), 一个树的节点集合 S={R, $M1$, …, Mn, $L1$, …, Ln}(R 为根节点, $M1$, …, Mn 为根节点到叶子节点之间的中间节点, 其中不包括根节点和叶子节点, $L1$, …, Ln 为叶子节点), 函数 f 为 D 到 S 的映射, S 中存在父子关系的节点 a 和 b, 其中 $f(b)f(a)$. 对根节点和叶子节点有: $f(L1)\cup f(L2)\cup\ldots\cup f(Ln)=f(R)$, 而且 $f(R)S$.

定义 4. k-匿名约束, 是指在每个元祖的准标识符内出现重复次数 k(k≥2), 则称为满足 k 匿名约束. 如表 2 所示.

定义 5. l-多样性, 是指在每一元组内, 敏感属性值出现不同次数 l(l≥2), 根据用户需求, 指定 l, 则其满足 l-多样性. 如表 2 所示, 其满足 2-l-多样性.

定义 6. 等价元组, 指由多个相似数据, 构成一组数据. 如表 2, {“M”, “[30, 60]”, “Non-Technical”}, 一组等价元组性质(k, l)-*Diversity* 等价元组内的元组个数在[k, 2k-1]为最优.

表 2 2-匿名表

Sex	Age	Job	Surgery
M	[30, 60]	Non-Technical	Transgender
M	[30, 60]	Professional	Plastic
M	[30, 60]	Non-Technical	Transgender
M	[1, 30]	Professional	Vascular
M	[30, 60]	Non-Technical	Urology
M	[30, 60]	Non-Technical	Plastic
M	[1, 30]	Professional	Urology
M	[30, 60]	Professional	Plastic
M	[30, 60]	Professional	Vascular
F	[60, 90]	Technical	Vascular
F	[60, 90]	Technical	Plastic

证明: 文献[12]提出了满足 *k*-匿名的每个等价元组的元组个数不超过 2*k*-1 为最优, 而(*k*, *l*)-*Diversity* 满足 *k*-匿名对等价元组个数的约束, 其等价元组的元组个数不超过 2*k*-1 也是最优的.

2 方案构造

当大量医疗数据要发布的时候, 对数据做隐私保护是至关重要. 而当数据发布时, 敌手可以通过属性链接, 得到患者敏感信息; 当等价组内的敏感属性相同时, 敌手直接可以推断出患者的隐私信息. 所以为了抵抗链接攻击和同质攻击, 本文提出了基于 KD 树的隐私保护算法. 算法利用 KD 树的性质, 将患者的属性当作一个节点, 不断对其进行分解, 直到不满足匿名要求, 并且控制每个叶子节点的元组个数在[*k*, 2*k*-1]之类, 以确保信息损失得到降低的目的.

如对表 1 数据进行基于 KD 树的隐私保护算法处理, 将{“age”, “sex”, “job”, “k”, “L”}构成一个节点, 根节点为最大等价元组, 包含所有数据集. 从第一维开始, 对根节点进行分解, 并判断两个节点是否满足 *k*, *l* 约束, 即 *k* 至少是 2, *l* 至少是 2. 通过判断可知, 两个节点满足要求. 接着第二维, 对两个节点进行分解, 并判断分解到的节点是否满足 *k*, *l* 约束, 由此得到第三层的两个节点. 以此类推, 通过不断的对每一维的属性, 进行分解, 直到不满足 *k*, *l* 约束, 由此得到如图 1 所示的 KD 树, 其叶子节点满足 *k*, *l* 约束, 并且每个节点的元组个数在 2*k*-1 之内, 达到降低信息损失目的. 而通过对 *l* 进行约束, 以此达到提高隐私保护的目的.

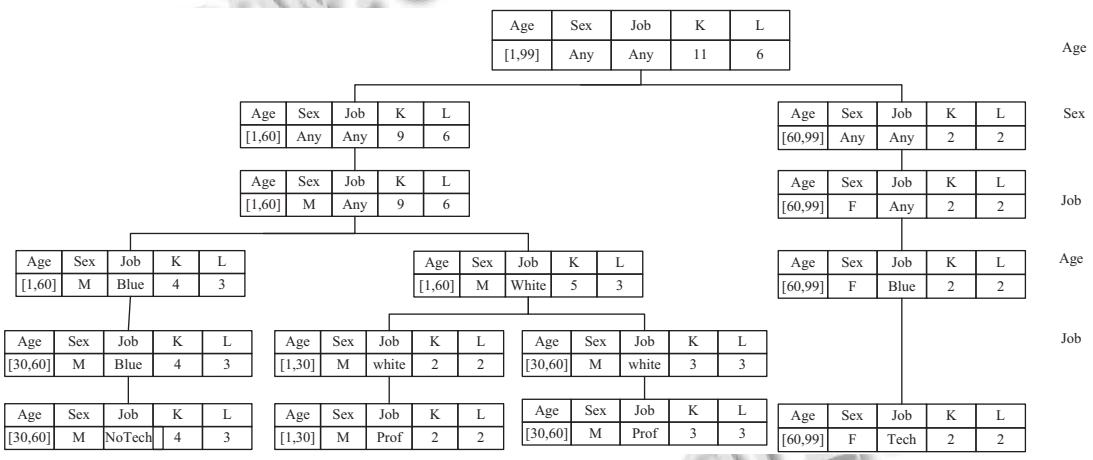

图 1 元组分解

2.1 算法实现

构造一个 KD 树的节点, 用于存储患者属性信息、*K*、*L*, 再对其最大匿名化, 形成最大的等价元组. 根据 KD 树的性质, 对每个节点的属性的泛化值进行分解, 直到所有属性的泛化值所分解的等价元组不满足匿名要求, 则树的叶子节点, 即为满足匿名化要求的等价元组.

输入: 病历集和根节点

输出: 叶子节点

createTree(TreeNode node, *k*, *l*, |S|)

{//node 节点, *k* 匿名, *l* 多样性, |S|记录集

while 当 node 节点有子节点时 do

createNode=获取 node 的子节点;

if 判断 createNode 内所有的维内属性是否都执行 then do

return;

else

attributeValue=获取 createNode 维上属性值

list=用 attributeValue 获取该属性的泛化树所对应节点的子节点;

newNodeList=list 与 createNode 合并成新的节点, 设置父节点 createNode;

while newNodeList 不为空 do

newNode=从 newNodeList 获取一个节点;

if 判断 newNode>*K*, *L* then do

createNode.split++;

```
        createNode 添加节点 newNode;
      end if;
      end;
      添加子节点 newNodeList;
      createTree(createNode, , K, L, |S|);
    end if;
    end;
  }
```

2.2 算法分析

设数据集|s|中的记录数目=n, 患者的属性个数=d, 树的节点总数=m.

算法迭代的次数与节点总数 m 相关, 而在对每个节点分解时, 要判断节点是否满足匿名要求, 满足, 则分解, 迭代, 不满足, 则判断下一个属性. 所以判断节点是否满足匿名要求时, 消耗的时间是 O(n), 算法总消耗时间是 O(dmn), 因为 d 远小于 n, 所以算法消耗时间是 O(mn).

3 实验结果

3.1 实验环境

本文实验的验证采用 UCI 机器学习库中的 adult 数据集, bank 数据集, student 数据集[13], 以及数据堂中的 credit 数据集[14]. 这几个数据集是目前数据发布实验使用的事实标准. 对这几个数据集做无效值处理, adult 数据集剩下 45222 个记录, bank 数据集剩下 4301 个记录, student 数据集剩下 649 个记录, credit 数据集剩下 131068 个记录. 在这个方案中, 选择 Adult 数据集中的{age, work class, education-num, marital-status, race, sex, native-country}作准标识符, 其中 age, education-num 是数值型属性; 选择 bank 数据集中的{age, balance, job, education, marital}作准标识符, 其中 age, balance 是数值型属性; 选择 credit 数据集中{sheetAmount, frequency, sex, marital-status, education, job}作准标识符, 其中 sheetAmount 是数值型属性. 同时为这三个数据集增加一列{disease}作敏感属性, 该列属性值{HIV, Cancer, Bronchitis, Pneumonia, Gastric-Ulcer, Gastritis, Indigestion, Flu, Heart-Disease, None}中的任意一个. 而选择 student 数据集中的{ age, sex, mJob, fJob, reason, traveltime, studytime}作准标识符, 其中 age, traveltime, studytime 是数值型属性, 将{health}作为 student 数据集的敏感属性, 该列属性值{好, 良好, 一般, 差, 很差}中的任意一个.

本文实验硬件环境 Intel(R) Core(TM) i5-4590 CPU @ 3.30 GHz, 8.00 GB RAM, 操作系统是 Window 7, 开发环境是 MyEclipse 2014, 数据库环境是 MySql 5.1.8. 本文的实验步骤如下:

(1) 对两个数据集的每一个记录的“disease”属性进行随机赋值.

(2) 验证算法的安全性能. 为了验证算法的安全性能, 分别采用 k-匿名模型[1], l-多样性模型[2]和本文算法对两个数据集进行隐私保护, 对比分析三种模型的隐私泄露风险的差异.

(3) 验证模型的实际性能. 为了验证模型的实际性能, 分别采用 k-匿名模型[1], l-多样性模型[2]和本文算法对两个数据集进行隐私保护, 对比分析三种模型在信息损失上的差异.

3.2 隐私泄露风险分析

目前, 针对匿名化隐私风险泄漏风险评估的具体模型, 而对于匿名化隐私保护模型的重要动机, 是为了抵抗敏感信息泄漏的风险, 以达到保护患者隐私的目的. 通过分析比较每个等价元组的元组个数和不同敏感属性值的个数, 可以发现等价元组之间的元组个数, 不同敏感属性值的个数之间存在的差异. 一般说等价元组内记录个数越多, 敏感属性值越多, 则隐私泄露风险越低. 利用每个敏感属性值的个数在等价元组内所占的比例来表示, 单个敏感属性值在等价元组内泄露的概率; 计算所有等价元组内单个敏感数值泄露概率之和, 之后求平均数, 来表示单个敏感属性隐私泄露的风险.

图 2 表示, 三种模型对三个不同敏感值的隐私保护程度. 其中横坐标表示三个不同的敏感属性值, 纵坐标表示三个敏感属性值的隐私泄漏风险概率.

由图 2 可知, 本文算法隐私泄露的风险比较小.

本文对等价组的元组个数和敏感属性的多样性分布做了约束, 以减少隐私泄露风险. 而 K-Anonymous 没有对敏感属性的多样性约束, 导致等价元组的敏感属性出现唯一, 提高隐私泄露风险. L-Diversity 约束了敏感属性的分布, 但对等价元组的个数不作处理, 使得元组个数越少, 隐私泄露风险越高.

3.3 信息损失

信息损失, 指对等价元组的准标识符属性进行泛化, 造成的元组信息损失. 单个属性信息损失, 指在属性泛化树中, 属性值到其泛化节点的高(h)与属性值到

根节点的高(H)之比$\frac{h}{H}$, 即. 元组信息损失, 是所有准标识符内属性的信息损失之和$\sum_n^i \frac{h_i}{H_i}$, 其中 n 是准标识符属性个数. 本文计算所有等价元组的信息损失之和, 之后计算元组信息损失的平均数.

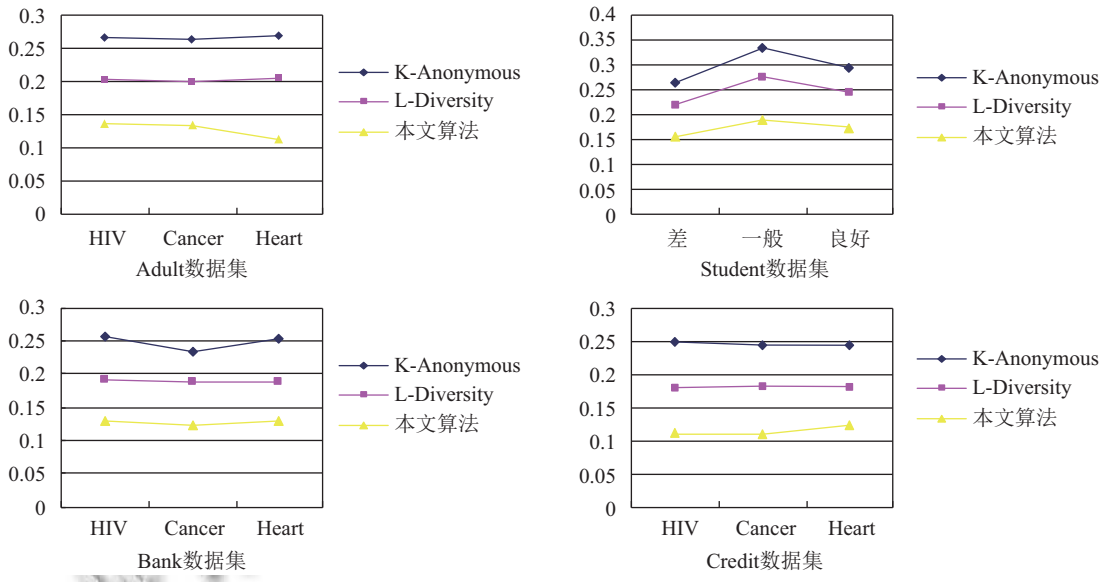

图 2 隐私风险

图 3 表示, 三种模型在准标识符数量的变化下, 造成等价元组的信息损失的比率. 其中横坐标表示准标识符数量, 纵坐标表示信息损失的比率.

由图 3 可知, 当准标识符数量较小时, 三种模型的信息损失相近, 而当准标识符数量较大时, 本文算法信息损失明显小于 *K*-Anonymous 模型和 *L*-Diversity 模型. 这是因为当准标识符数量较小时, 三种模型等价元组之间的准标识符的值差异性小, 使得三种模型的信息损失差异小; 当准标识符数量较大时, 泛化层次复杂迅速增大, 而且由于本文算法对 k, l 的约束较大, 并且对最大等价元组不断分解, 并保证分解后等价元组中记录 高度相似, 从而达到减少信息损失的目的.

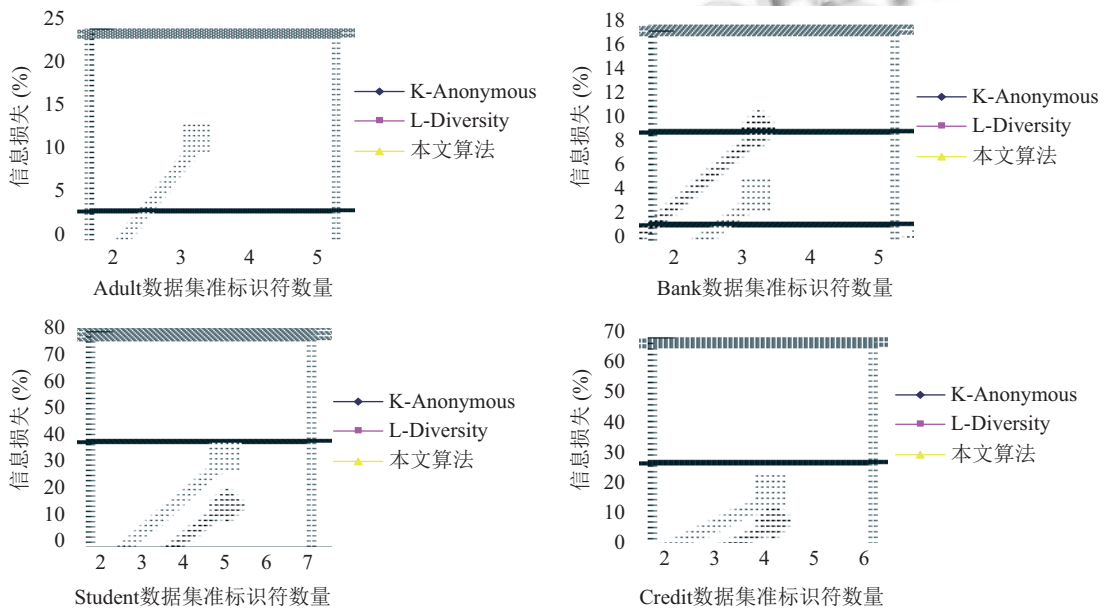

图 3 信息损失

4 总结

本文提出来基于 KD 树的隐私保护算法, 先将患者属性作为一个节点, 并对将属性值泛化到最大, 以包含数据集的所有病历信息, 根据 KD 树的性质, 不断对所有属性值分解, 直至所有属性值不能分解为止, 以确保信息损失得到降低. 在对等价元组分解时, 对敏感属性值做多样性约束, 来降低隐私泄漏风险. 实验表明, 本文算法在对 k, l 做约束下, 隐私泄露风险较低, 并且信息损失较低, 解决了病历信息发布时可能造成隐私泄露的风险, 提高了病历发布时的数据通用性. 在下一步工作, 研究如何构造一棵动态的泛化树, 以便减少信息的损失.

参考文献

1 Sweeney L. k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10(5): 557–570. [doi: 10.1142/S0218488502001648]

2 Sweeney L. Achieving k-anonymity privacy protection using generalization and suppression. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10(5): 571–588. [doi: 10.1142/S021848850200165X]

3 Mohammed N, Fung BCM, Hung PCK, *et al*. Anonymizing healthcare data: A case study on the blood transfusion service. Proc. of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA. 2009. 1285–1294.

4 Chen L, Yang JJ, Wang Q. Privacy-preserving data publishing for free text Chinese electronic medical records. Proc. of 2012 IEEE 36th Annual Computer Software and Applications Conference. Washington, DC, USA. 2012. 567–572.

5 Shin M, Yoo S, Lee KH, *et al*. Electronic medical records privacy preservation through k-anonymity clustering method. Proc. of 2012 Joint 6th International Conference on Soft Computing and Intelligent Systems (SCIS) and 13th International Symposium on Advanced Intelligent Systems (ISIS). Kobe, Japan. 2012. 1119–1124.

6 Hmood A, Fung BCM, Iqbal F. Privacy-preserving medical reports publishing for cluster analysis. Proc. of 2014 6th International Conference on New Technologies, Mobility and Security (NTMS). Dubai, Emirate. 2014. 1–8.

7 Machanavajjhala A, Kifer D, Gehrke J, *et al*. l-diversity: Privacy beyond k-anonymity. ACM Trans. Knowledge Discovery from Data (TKDD), 2007, 1(1): 3. [doi: 10.1145/1217299]

8 Li NH, Li TC, Venkatasubramanian S. t-closeness: Privacy beyond k-anonymity and l-diversity. Proc. of 2007 IEEE 23rd International Conference on Data Engineering. Istanbul, Turkey. 2007. 106–115.

9 王伟. 电子病历发布中的匿名化隐私保护方法研究[硕士学位论文]. 长沙: 中南大学, 2013.

10 陈炜, 陈志刚, 邓小鸿, 等. 抵抗背景知识攻击的电子病历隐私保护新算法. 计算机工程, 2012, 38(11): 251–253. [doi: 10.3969/j.issn.1000-3428.2012.11.076]

11 刘艺龙. 基于泛化树的 k-匿名数据集的挖掘算法研究[硕士学位论文]. 上海: 东华大学, 2013.

12 Xu J, Wang W, Pei J, *et al*. Utility-based anonymization using local recoding. Proc. of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Philadelphia, Pa, USA. 2006. 785–790.

13 UCI machine learning repository. http://archive.ics.uci.edu/ml/datasets.

14 数据堂. 保险客户理赔资料. http://more.datatang.com/data/45904.[2014-02-19].