隐私保护k-匿名算法研究

王平水[1],马钦娟[2]
WANG Pingshui[1],MA Qinjuan[2]
1.安徽财经大学 管理科学与工程学院,安徽 蚌埠 233030
2.安徽财经大学 商学院,安徽 蚌埠 233041
1.College of Management Science and Engineering,Anhui University of Finance & Economics,Bengbu,Anhui 233030,China
2.Business Institute,Anhui University of Finance & Economics,Bengbu,Anhui 233041,China

WANG Pingshui,MA Qinjuan.Research on *k*-anonymity algorithm for privacy preservation.Computer Engineering and Applications,2011,47(28):117-119.

Abstract: Privacy preservation has been an essential issue for individuals or organizations.*k*-anonymity is one of the primary techniques realizing privacy protection in data dissemination environment.Current *k*-anonymity solutions based on generalization and suppression techniques suffer from high information loss and low usability mainly due to reliance on pre-defined generalization hierarchies or order imposed on each attribute domain.It develops a new *k*-anonymity algorithm based on clustering technology.Experimental results show that the method can improve the usability of the released data while preserving privacy.
Key words: data dissemination;privacy-preserving;anonymization;*k*-anonymity;clustering

摘 要: 隐私保护已成为个人或组织机构关心的基本问题,*k*-匿名是目前数据发布环境下实现隐私保护的主要技术之一。鉴于多数*k*-匿名方法采用泛化和隐匿技术,严重依赖于预先定义的泛化层或属性域上的全序关系,产生很高的信息损失,降低了数据的可用性,提出了一种基于聚类技术的*k*-匿名算法。实验结果表明,该算法在保护隐私的同时,提高了发布数据的可用性。
关键词: 数据发布;隐私保护;匿名化;*k*-匿名;聚类
DOI:10.3778/j.issn.1002-8331.2011.28.032 **文章编号:** 1002-8331(2011)28-0117-03 **文献标识码:** A **中图分类号:** TP311

1 引言

随着Internet技术、大容量存储技术和数据处理技术的迅猛发展以及数据共享范围的逐步扩大,数据的自动收集和发布越来越方便。然而,在数据发布过程中隐私泄露问题也日益突出,因此实施隐私保护就显得尤为重要。数据发布中隐私保护对象主要是用户敏感数据与个体身份之间的对应关系。通常使用删除标识符的方式发布数据是无法真正阻止隐私泄露的,攻击者可以通过链接攻击获取个体的隐私数据。匿名化是解决链接攻击所带来的隐私泄露问题的主要技术之一。自从P.Samarati和L.Sweeney首次提出k-匿名模型以来[1],国内外研究人员对匿名化技术开展了广泛而又深入的研究工作以寻求保护隐私的有效方法,取得了一系列相关研究成果[2-9]。然而,目前多数k-匿名方法是基于泛化和隐匿技术,由于其严重依赖于预先定义的泛化层或属性域上的序关系,使得匿名结果产生很高的信息损失,从而降低了数据的可用性[10]。为此,本文提出了基于聚类的匿名化技术,以期在保护隐私的同时进一步提高数据的可用性。

2 相关概念

2.1 准标识符

准标识符(Quasi-Identifiers,QI)是指与其他外部数据表进行链接以标识个体身份的属性或属性组合,如性别、出生日期、邮政编码等。准标识符的选择取决于进行链接的外部数据表。

2.2 链接攻击

链接攻击是从发布的数据表中获取隐私数据的常见方法。其基本思想为:攻击者通过对发布的数据和其他渠道获取的外部数据进行链接操作,以推理出隐私数据,从而造成隐私泄露。例如文献[2],通过将医疗信息表与选民登记表进行链接,几乎可以唯一确定就诊病人的医疗诊断结果,然而,病人的医疗诊断结果正是需要保护的隐私数据。

2.3 k-匿名

为解决链接攻击所导致的隐私泄露问题,L.Sweeney等提出k-匿名($k-$anonymity)方法[1-3]。k-匿名通过概括和隐匿技术,发布精度较低的数据,使得每条记录至少与数据表中其他$k-1$条记录具有完全相同的准标识符属性值,从而减少链接攻击所导致的隐私泄露。如表2是表1的一个2-匿名化表。

基金项目: 国家自然科学基金(the National Natural Science Foundation of China under Grant No.71071001);安徽省自然科学基金项目(No.11040606M140)。
作者简介: 王平水(1972—),男,博士研究生,副教授,研究方向为数据挖掘与隐私保护;马钦娟(1975—),女,讲师。E-mail:pshwang@163.com
收稿日期: 2010-07-23;**修回日期:** 2010-10-11

表1 医疗信息表

Name	Race	Birth	Sex	Zip	Disease
Alice	Blank	1965-3-18	M	02141	Flu
Bob	Blank	1965-5-1	M	02142	Cancer
David	Blank	1966-6-10	M	02135	Obesity
Helen	Blank	1966-7-15	M	02137	Gastritis
Jane	White	1968-3-20	F	02139	HIV
Paul	White	1968-4-1	F	02138	Cancer

表2 医疗信息表的一个2-匿名化表

Race	Birth	Sex	Zip	Disease
Blank	1965	M	0214*	Flu
Blank	1965	M	0214*	Cancer
Blank	1966	M	0213*	Obesity
Blank	1966	M	0213*	Gastritis
White	1968	F	0213*	HIV
White	1968	F	0213*	Cancer

3 聚类 k-匿名算法

算法的基本思想是将 k-匿名问题视为聚类问题，将数据对象分成若干类或簇，使同一簇中的对象之间关于已定义的相似性标准具有很高的相似度，而不同簇中的对象之间高度相异。

3.1 k 成员聚类问题

传统的聚类过程要求指定具体的簇数目，然而，k-匿名问题并不限制簇的数目，而是要求每个簇至少包含 k 条记录。因此，可以将 k-匿名问题视为聚类问题，通常称为 k 成员聚类问题[8]。

定义1（k 成员聚类问题） k 成员聚类问题是将包含 n 条记录的集合划分成一系列簇，使得每个簇至少包含 k 条记录，并且要求簇内间距总和最小。形式地，令 S 为包含 n 条记录的集合，k 为具体的匿名化参数，则 k 成员聚类问题的最优解是产生满足以下条件的簇的集合 $E=\{e_1, e_2, \cdots, e_m\}$：

(1) $\forall i\neq j\in\{1,2,\cdots,m\}$，$e_i\cap e_j=\varnothing$；

(2) $\bigcup_{i=1,2,\cdots,m} e_i=S$；

(3) $\forall e_i\in E, |e_i|\geq k$；

(4) $\sum_{l=1,2,\cdots,m}|e_l|\cdot \underset{i,j=1,2,\cdots,|e_l|}{MAX}\Delta(p(l,i),p(l,j))$ 是最小的。

其中，$|e|$ 表示簇 e 的大小，$p(l,i)$ 表示簇 e_l 中的第 i 个数据点（将记录视为数据点），$\Delta(x,y)$ 表示数据点 x 和 y 之间的距离。

3.2 距离和代价度量

聚类问题的核心是定义距离函数用以度量数据点间的相似性，定义代价函数以使聚类问题代价最小化。距离函数通常由数据点的数据类型（如数值型或分类型）决定，而代价函数则由聚类问题的具体目标来定义。由于 k-匿名问题所涉及的数据中可能既包含数值型属性，又包含有分类型属性，因此，需要定义能够处理不同类型数据的距离函数。以下描述适用于 k-匿名问题的距离和代价函数[8]。

定义2（数值型数据间的距离） 令 D 为有限数值域，任意数值 $v_i, v_j\in D$ 间的标准距离定义为：

$$\delta_N(v_i,v_j)=\frac{|v_i-v_j|}{|D|}$$

其中 $|D|$ 表示域 D 的最大值与最小值之间的差值。

对于分类型属性，由于大多数分类域不具有完整的序关系，因此，上述定义并不适用于分类型数据。一种简单直观的解决办法是，假定域中各值互不相同，若两个属性值相同，则距离为0，否则距离为1。然而，有些域中值之间可能存在某种语义关系，在这些域中，希望基于这种语义关系定义距离函数。分类树通常可以反映出这种语义关系，假定一个域上的分类树是一棵平衡树，其叶节点代表域中所有不同的分类值。例如，图1示意了“Country”属性的分类树。对于有些属性，如Occupation，其分类值之间可能不存在任何可用于分类的语义关系，对于这种域，将构造一棵扁平的平凡分类树，如图2所示。于是，定义分类型数据间的距离函数。

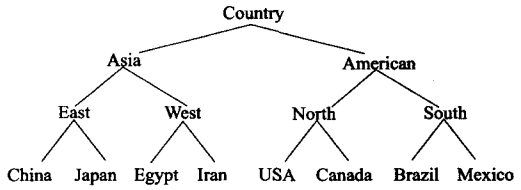


图1 “Country”属性分类树

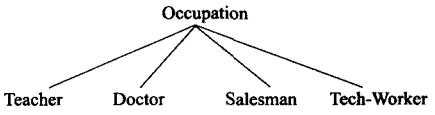


图2 “Occupation”属性分类树

定义3（分类型数据间的距离） 令 D 为分类域，T_D 为 D 上的分类树，任意分类值 $v_i, v_j\in D$ 间的标准距离定义为：

$$\delta_C(v_i,v_j)=\frac{H(\Lambda(v_i,v_j))}{H(T_D)}$$

其中，$\Lambda(x,y)$ 代表分类树中以 x 和 y 的最小公共祖先为根的子树，$H(T)$ 表示分类树 T 的高度。

结合数值域和分类域上的距离函数，来定义两记录间的距离如下。

定义4（记录间的距离） 令 $Q_T=\{N_1,N_2,\cdots,N_m,C_1,C_2,\cdots,C_n\}$ 为数据表 T 的准标识符，其中 $N_i(i=1,2,\cdots,m)$ 为数值型属性，$C_j(j=1,2,\cdots,n)$ 为分类型属性，则任意记录 $r_1,r_2\in T$ 间的距离定义为：

$$\Delta(r_1,r_2)=\sum_{i=1,2,\cdots,m}\delta_N(r_1[N_i],r_2[N_i])+\sum_{j=1,2,\cdots,n}\delta_C(r_1[C_j],r_2[C_j])$$

其中 $r_i[A]$ 表示记录 r_i 的属性 A 的值。

既然 k 成员聚类问题的最终目标是实现发布数据的 k-匿名，构造代价函数来表示泛化处理所产生的数据扭曲程度。由于每个簇中的记录被泛化成相同的准标识符值，假定数值型数据泛化成区间[最小值，最大值]，分类型数据泛化成不同属性值的集合。

定义5（信息损失） 令 $e=\{r_1,r_2,\cdots,r_k\}$ 是一个簇（等价类），其准标识符包含数值型属性 $N_1,N_2,\cdots,N_m$ 和分类型属性 $C_1,C_2,\cdots,C_n$，T_{C_i} 为分类型属性 C_i 域的分类树，MIN_{N_i} 和 MAX_{N_i} 分别为簇 e 中数值型属性 N_i 的最小值和最大值，$\cup_{C_i}$ 表示簇 e 中分类型属性 C_i 不同属性值的集合，对簇 e 进行泛化处理所产生的信息损失 $IL(e)$ 定义为：

$$IL(e)=|e|\cdot\left(\sum_{i=1,2,\cdots,m}\frac{(MAX_{N_i}-MIN_{N_i})}{|N_i|}+\sum_{j=1,2,\cdots,n}\frac{H(\Lambda(\cup_{C_j}))}{H(T_{C_j})}\right)$$

其中$|e|$表示簇e中的记录数，$|N_i|$表示数值域N_i的最大值与最小值之差，$A(\cup_{C_j})$表示分类树中以$\cup_{C_j}$内所有值的最小公共祖先为根的子树，$H(T)$表示分类树T的高度。

基于上述定义，匿名化数据表的总信息损失定义如下。

定义6(总计信息损失) 令E为匿名表AT的等价类的集合，则AT的总信息损失定义为：

$$Total-IL(AT)=\sum_{e\in E} IL(e)$$

由于k成员聚类问题的代价函数是所有簇内距离总和，其中簇内距离定义为簇内最远数据点间的距离，于是，对簇内记录进行泛化处理时，最小化信息损失就等同于k成员聚类问题中最小化代价函数，因此，聚类处理时需最小化的代价函数即为$Total-IL$。

3.3 k成员聚类算法

在大多数聚类问题中，对k成员聚类问题最优解的穷尽搜索法具有指数级的复杂度，针对问题的困难性，提出一个简单高效的贪婪式算法，其基本思想为：对于给定的n条记录，首先从中随机地选取一条记录r_i并将其单独作为一个簇e_1，然后选取记录r_j使得$IL(e_1\cup r_j)$最小，重复上述操作直到$|e_1|=k$。当簇e_1大小达到k时，再从剩余记录中随机选取一条记录重复上述聚类过程，直到剩余记录数小于k为止。将剩余记录依次插入到已有簇中，使得增加的信息损失最小。可以证明贪婪式k成员聚类算法生成的簇的大小至少为k，最多为$2k-1$，其时间复杂度为$O(n^2)$。算法过程如下：

算法 k-成员聚类算法

输入 数据集S和匿名参数k

输出 簇的集合$result$，其中每个簇至少包含k条记录

1.如果数据集中记录个数小于k，则返回；

2.令簇集$result=\phi$；

3.从数据集S中随机选取一条记录r；

4.当数据集S的记录个数不小于k时，循环执行：

(1)从数据集S中随机选取一条记录r；

(2)$S=S-\{r\}$；

(3)$c=\{r\}$；

(4)当簇c中记录个数小于k时，循环执行：

a. 从数据集S中选取记录r，使其加入簇c产生的信息损失最小；

b. $S=S-\{r\}$；

c. $c=c\cup\{r\}$；

(5)$result=result\cup\{c\}$；

5.当数据集S中有剩余记录时，循环执行：

(1)从数据集S中随机选取一条记录r；

(2)$S=S-\{r\}$；

(3)从簇集$result$中选取簇c，使r加入其中后产生的信息损失最小；

(4)$c=c\cup\{r\}$；

6.返回簇集$result$。

4 实验结果

实验的目标是考查本算法的性能，如数据质量、执行效率等。为客观地评估本文的k-成员聚类算法，与基于中心点的划分算法做比较。

4.1 实验环境

实验采用UCI机器学习数据库中的Adult数据集验证本算法的性能，该数据集包含部分美国人口普查数据，在数据匿名隐私保护研究中被广泛使用，已成为该领域事实上的标准测试数据集，采用文献[4]中的数据预处理方法，删除含有缺失值的数据记录，得到的实验数据集包含45 222条数据记录。将属性age，work class，education，martial status，race，gender，native country作为准标识符，其中age和education为数值型属性，其余5个属性为分类型属性，将occupation作为敏感属性。

实验运行环境：Intel® Core™ 2 Duo CPU T5450 1.67 GHz，2.0 GB RAM，Windows XP，MATLAB7.0，Visual C++ 6.0。

4.2 数据质量

图3对两种算法在k值递增情况下的信息损失代价做了比较。从图中可看出，贪婪式k成员算法对于所有k值具有较小的代价。本文算法优于中心点划分算法原因在于中心点划分算法仅考虑数据点在单一维上相似性。

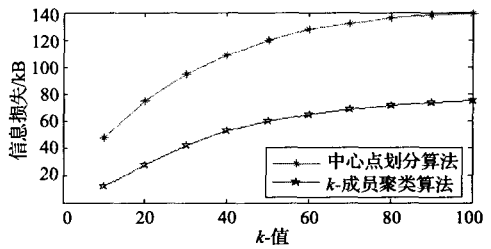


图3　k-值与信息损失度量

4.3 执行效率

图4对两种算法在k值递增情况下的执行时间做了比较。从图中可看出，贪婪式k成员算法对于所有k值均高于中心点划分算法的执行时间。然而，k-成员聚类通常在离线状态下执行，在多数情况下为了提高信息损失方面的性能，该算法的时间开销是可以接受的。

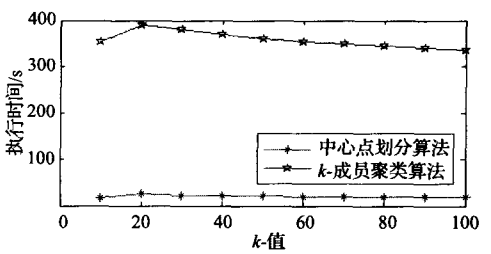


图4　k-值与执行时间

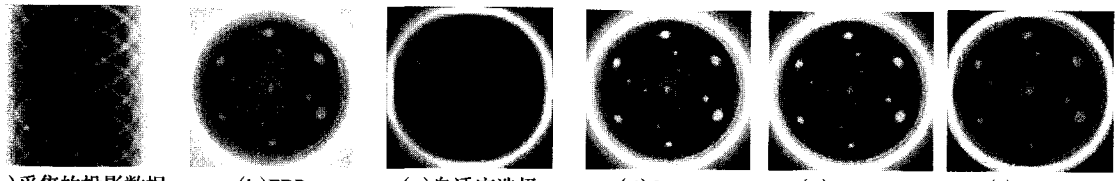
5 结论

通过将k-匿名问题转化为k成员聚类问题提出一个有效的k-匿名算法，同时还提出两个有关聚类的重要度量准则，即距离和代价度量。该度量准则能够很好地刻划泛化处理所导致的数据扭曲，也可用于k-匿名数据集的质量度量。所提算法虽然时间开销较大，但信息损失较小。今后的工作准备提高匿名算法的执行效率，并针对簇中敏感属性值的多样化展开深入研究，以进一步提高敏感数据的隐私保护强度。

参考文献：

[1] Samarati P，Sweeney L.Generalizing data to provide anonymity when disclosing information(abstract)[C]//Proc of the 17th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems.New York：ACM Press，1998.

(下转200页)



(a)采集的投影数据 (b)FBP (c)自适应选择正则项参数 (d)$\beta=0.01$ (e)ML-EM (f)$\beta=0.3$

图2 实际重建结果

表1 仿真结果的误差分析

重建方法	信噪比	归一化平均绝对距离
自适应正则	7.245 183	0.307 595
正则化参数为0.01	6.852 043	0.319 526
正则化参数为5	7.831 843	0.335 568
ML-EM	6.848 085	0.319 574
FBP	6.838 149	0.575 846

从仿真结果上看,当正则化参数太大时,虽然能很好地抑制噪声,图像的信噪比较高,但是不能保持图像的边缘,重建出来的图像质量较差。当正则化参数太小时,重建出来的图像有较大的噪声,信噪比较低。但当自适应选择正则化参数时,可以在抑制噪声和保持图像边缘上达到一个平衡,并与ML-EM算法、FBP算法相比较,其信噪比较高,归一化平均绝对距离较小,提高了图像的质量,减少了噪声,也保持了边缘。

下面将该方法用于重建X射线CT采集的投影数据,实验采用220 kV,10 mA的X射线源。探元大小为0.127 mm。采用的探测PAXSCAN 2520。工作模式:数字视频。数据类型:unsigned short。A/D:12 bit。射线源—检测工件-探测器间距:850 mm~200 mm,探测器大小为960×768。旋转一周采样间隔为1度。选取的投影数据的大小是360×256,得到的投影数据如图2(a),$h=0.002$,$\delta=0.001$,$\lambda=0.002\,5$。迭代次数为96次重建图像大小为256×256,如图2所示。

从实际重建结果上看,与其他方法相比,选用自适应选择正则项参数的方法重建出来的实际切片图像清晰,噪声小,边缘效果好。

5 结论

在CT图像重建中,正则化参数的选择直接影响着MAP类重建算法的重建图像质量,本文提出一种自适应动态确定正则化参数的方法,无需像传统方法那样通过选取几个候选值进行处理、比较,再选择一个最优结果,因此适用性更强。它的突出特点就是正则化参数不是固定不变的,充分利用每一次迭代的重建结果的信息进行动态调整,有效保存图像的高频部分,能较好地保持图像的边缘信息,提高图像信噪比。实验充分证明了这种自适应正则MAP的CT图像重建方法的正确性和有效性。

参考文献:

[1] Geman S,Geman D.Stochastic relaxation,Gibbs distribution,and the Bayesian restoration of imge[J].IEEE Transactions on Pattern Analysis and Machine Intelligence,1984,6(6):721-741.

[2] Li S Z.Markov random field modeling in image analysis[M].Berlin:Springer-Verlag,2001.

[3] 庄天戈.CT原理与算法[M].上海:上海交通大学出版社,1992.

[4] Elbakri I A,Fessler J A.Statistical image reconstruction for polyenergetic X-ray computed tomography[J].IEEE Transactions on Medical Imaging,2002,21(2):89-99.

[5] Chen Yang,Ma Jian-hua,Feng Qian-jin,et al.Nonlacal prior Bayesian tomographic reconstruction[J].Journal of Mathematical Imaging and Vision,2008,30(2):133-146.

[6] 沈焕峰,李平湘,张良培.一种基于正则化技术的超分辨影像重建方法[J].中国图象图形学报,2005,10(4):436-440.

[7] Green P J.Bayesian reconstructions from emission tomography data using a modified EM algorithm[J].IEEE Transactions on Medical Imaging,1990,9(1):84-93.

(上接119页)

[2] Sweeney L.K-anonymity:a model for protecting privacy[J].International Journal of Uncertainty,Fuzziness and Knowledge-based Systems,2002,10(5):557-570.

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

[4] Lefevre K,Dewittd J,Ramakrishnan R.Incognito:efficient full-domain *k*-anonymity[C]//Proc of the 2005 ACM SIGMOD International Conference on Management of Data.New York:ACM Press,2005:49-60.

[5] Fung B,Wang Ke,Yu P.Top-down specialization for information and privacy preservation[C]//Proc of the 21st IEEE International Conference on Data Engineering.Washington DC:IEEE Computer Society,2005:205-216.

[6] Wang Ke,Yu P,Chakraborty S.Bottom-up generalization:a data mining solution to privacy protection[C]//Proc of the 4th IEEE International Conference on Data Mining.Washington DC:IEEE Computer Society,2004:249-256.

[7] Machanavajjhala A,Gehrke J,Kifer D.l-diversity:privacy beyond *k*-anonymity[J].ACM Transactions on Knowledge Discovery from Data.New York:ACM Press,2007,1(1):24-35.

[8] Byun J W,Kamra A,Bertino E,et al.Efficient *k*-anonymization using clustering techniques[C]//LNCS 4443:Proceedings of DASFAA 2007.Berlin Heidelberg:Springer-Verlag,2007:188-200.

[9] 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.

[10] 王智慧,许俭,汪卫,等.一种基于聚类的数据匿名方法[J].软件学报,2010,21(4):680-693.