·数据挖掘·

采用熵的多维*K*-匿名划分方法

晏 华，刘贵松

(电子科技大学计算智能实验室 成都 610054)

【摘要】*K*-匿名是数据发布应用场景下重要的隐私保护模型。近年来数据集*K*-匿名化的算法得到广泛的研究，Median Mondrian算法是目前唯一的多维*K*-匿名划分方法。文中研究了Median Mondrian算法，指出其不能有效地平衡数据划分精度与数据隐私安全性之间的矛盾，由此提出基于熵测度机制的多维*K*-匿名划分方法以及评估*K*-匿名化结果安全性的测量标准。实验表明该算法是可行的，能有效地提高数据安全性。

关 键 词 熵； *K*-匿名； 多维划分； 准标识符
中图分类号 TP309.2 文献标识码 A

Multidimensional *K*-anonymity Partition Method Using Entropy

YAN Hua，LIU Gui-song

(Computational Intelligence Laboratory, University of Electronic Science and Technology of China Chengdu 610054)

Abstract *K*-anonymity is an important privacy preserving model in the data publishing scenario. The algorithms on dataset *K*-anonymization are researched extensively in recent years, Median Mondrian algorithm is the only multidimensional *K*-anonymity partition method. However, our research shows that Median Mondrian algorithm is not well-balanced on dealing with the contradiction between data partition precision and data privacy preserving. In this paper, we propose an entropy-based multidimensional *K*-anonymity partition method and a new evaluation measure on *K*-anonymization results. The experimental results show that our new method is feasible and preserves the privacy much more efficiently than Median Mondrian algorithm.

Key words entropy; *K*-anonymity; multidimensional partition; quasi-identifier

近年来，数据的安全与隐私问题已经成为数据挖掘领域研究的重要课题之一。由于个人数据极容易被商业用途的应用收集与分析，所以越来越多的数据拥有者不愿意提供个人信息，除非个人信息中的敏感信息的隐私能得到保障。保护个人隐私信息最直接的方法是将能唯一标识一个人的属性信息(Identifier)隐藏，如姓名和身份证号码。但这种方法无法解决另一种隐私威胁问题，即链接攻击[1]。

链接攻击是指用户通过对发布的数据和其他渠道获得的数据进行链接处理，推演出隐私数据，从而造成隐私泄露。文献[1]提出的*K*-匿名数据模型正是为了解决链接攻击问题。为了实现数据集的*K*-匿名化，最有代表性的一类算法[2-4]是通过用户定义的概念层次结构实现*K*-匿名化，都属于单维的划分方法，其质量取决于使用的概念层次结构。文献[5-6]提出了目前唯一的一个多维*K*-匿名划分方法，即Median Mondrian算法。实验结果表明该近似的贪心算法能有效地实现数据集的*K*-匿名化。

K-匿名化后的数据以数值范围替代原始数据的精确值，并且至少有*K*个数据具有相同的表示，即数据的*K*-匿名化以损失数据的精确度实现数据的隐私保护。数据的精确度和数据的隐私安全性是相互矛盾的，现有的算法很难在上述两个指标上获得平衡。理想的数据*K*-匿名化结果应该是在尽可能地减少数据信息损失的情况下，实现数据的隐私安全的最大化。如果两组数据分布具有相同数值范围但数据分布不同，那么数据分布离散程度高的数据安全性高于数据分布相对集中的数据。Median Mondrian算法的设计是在简单满足*K*-匿名模型要求的前提下，追求数据划分精度的最大化，而在数据隐私安全性方

收稿时间：2007– 08– 17
基金项目：国家自然科学基金(60471055)
作者简介：晏 华(1970 –)，女，在职博士生，讲师，主要从事数据挖掘及其应用方面的研究.

面考虑不够。

熵是最能反映数据点多样性和不确定性的度量机制，因此，基于熵的概念，本文提出一种多维*K*-匿名划分方法，并针对数据隐私安全性提出新的*K*-匿名划分结果评价标准。

1　*K*-匿名模型中的基本概念

K-匿名模型相关概念的定义如下：已知数据集 $T=\{t_1,t_2,\cdots,t_n\}$ 属于一个更大的数据分布 Ω，拥有属性集 $A_1, A_2,\cdots, A_m$, A_i 为数据记录的第i个属性，$t[A_i]$ 为数据记录t属性 A_i 的值。

定义 1　标识符(Identifier)。数据集 T 的标识符 I是指能唯一标识数据记录的属性。

如美国人口普查信息中的社会保险号码。通常标识符信息在数据发布时，为了保护隐私信息被直接隐去。

定义 2　准标识符(Quasi-identifier)。准标识符 $Q_I=\{Q_1,Q_2,\cdots,Q_d\}$，满足 $Q_I\subseteq(A_1, A_2,\cdots, A_m)$，且 Q_I 与外部数据连接至少可重新识别 Ω 中的一个数据记录t。

定义 3　敏感属性(Sensitive attributes)。属性 A_i 如果是不允许竞争对手通过I关联的属性，则称为敏感属性。

如医疗数据中病人记录的HIV状态值。

定义 4　*K*-匿名(*K*-Anonymity)已知数据集T及其准标识符 Q_I，如果每一个数据记录 $t\in T$ 都存在 $K-1$ 个其他数据记录 $t_{i_1},t_{i_2},\cdots,t_{i_{k-1}}\in T$，满足 $t[Q_I]=t_{i_1}[Q_I]=t_{i_2}[Q_I]\cdots=t_{i_{k-1}}[Q_I]$，则数据集$T$是*K*-匿名的。

定义 5　等价类(Equivalence Class)。数据集T中，在属性集 $A_1, A_2,\cdots, A_d$ 上的一个等价类是指等价类中的所有数据记录在属性集 $A_1, A_2,\cdots, A_d$ 上有相同的值。

信息理论中出现的Shannon熵、Kolmogorov熵及拓扑熵等都是测量随机变量不确定性的一种数学度量[7]。本文的数据对象是离散型的类别数据，因此计算离散型随机变量熵的定义如下：

定义 6　熵(Entropy)。如果离散型的随机变量 X 可能出现的值的集合为 $S(X)$，并且 $p(x)$ 是 X 的概率函数，则 X 的熵为：

$$E(X)=-\sum_{x\in S(X)}p(x)\log_2(p(x)) \tag{1}$$

如当 $S(X)$ 为 $\{1,2\}$，且 $\{1,2\}$ 的概率分别为50%，则 $E(X)=2\times(-0.5\times\log_2 0.5)=1$。如果 X 仅有可能出现一种取值，即取值的概率为100%，则 $E(X)=-(1\times\log_2 1)=0$。由此可见，随机变量 X 的熵越大，表明 X 取值的离散程度越大；反之离散程度越小。

在本文所述的*K*-匿名划分方法中，数据集在每一维上的熵作为划分过程中，维选择的标准如下：

(1) 如果总是选择数据离散程度最高的维对数据进行划分，则数据集的可分性强于其他方法，特别是对畸形的数据分布；

(2) 采用熵作为划分原则，每组划分结果中的数据点分布相对离散，则减小了竞争对手正确猜测数据点实际值的概率，从而进一步提高划分结果的安全性。

2　基于熵的多维*K*-匿名划分算法

基于熵的多维*K*-匿名划分算法采用自顶向下的贪心方法对准标识符空间不断地划分，直至所有的子标识符空间不可再分。标识符空间不可分定义为：当子空间中的数据记录数小于2*K*时，该子空间是不可再分的。上述的贪心划分过程借用*K*-d tree[8]的数据结构以及构建过程实现。事实上，Median Mondrian 算法完全采用了*K*-d tree实现多维*K*-匿名。*K*-d tree中的*K*与*K*-匿名模型中的*K*的定义是不同的，*K*-d tree中的*K*是数据空间的维数，而*K*-匿名模型中的*K*是数据集 T 在准标识符空间 Q_I 上的等价类所包含的最小数据记录数。

2.1　*K*-d tree

K-d tree称为*K*-维二叉搜索树，包括根结点、内部结点和外部结点。根结点代表初始的整个空间；每个内部结点包含两个后续的结点，后续的结点指向划分后的两个子空间；外部结点是不可再分的子空间，所有外部结点的集合构成整个*K*-维空间。

K-d tree构建过程是一个递归的空间划分过程：

(1) 从整个*K*-维空间开始；

(2) 选择划分维；

(3) 中点划分；

(4) 对得到的子空间重复步骤2～4，直至所有的子空间不可再分。

K-d tree是一种非常有效的多维划分方法，其构建过程的时间复杂度为 $O(KN\log_2 N)$，其中，K 为数据的维数；N 为数据点数。对于划分维的选择，*K*-d tree的选择具有最大取值范围的维。在确定划分维以后，确定划分值如下：对子空间内的数据点以划分维的值排序，以中点的值作为划分值。如果数据分布是均匀的，则选取具有最大取值范围的维作

为划分维。以中点的值作为划分值也是合理的，但如果数据是畸形分布的，则最大取值范围不能代表数据的离散程度。以中点值做划分值有时是不可分的，因此本文提出用熵作为维选择和划分值确定新的测度。

2.2 基于熵的划分算法描述

基于熵的多维划分算法(Entropy-based Multidimensional Partition，EMP)的主函数描述如下：

EMP(Q_I _space , K)

If ($|Q_I$ _space $|< 2*K$)

return Q_I _space ；

else

sortedVal ← sort(all_dim);

best_dim ← choose_max_entropy(sortedVal);

splitVal ← max_div_entropies(sortedVal[best_dim]);

lqs ← $\{t \in T : t.\text{best}\_A \le \text{splitVal}\}$;

rqs ← $\{t \in T : t.\text{best}\_A > \text{splitVal}\}$;

return EMP(lqs, k) and EMP(rqs, k)

主函数的输入为准标识符空间Q_I_space和K-匿名数k，递归划分过程为：

(1) 判断子空间是否可再分；如果不可再分，返回子空间。

(2) 对所有维上的数据点值进行排序。

(3) 选择具有最大熵值的维作为划分维。

(4) 在划分维上选择一个划分值，使待分的子区间的熵之和最大。

(5) 根据划分值求得左子空间和右子空间。

(6) 递归调用主函数，对左、右子空间进行K-匿名划分。由此可见，基于熵的划分方法的过程与K-d tree的构建过程基本类似，不同之处在于维选择子函数choose_max_entropy( )以及划分值确定子函数max_div_entropies( )。

维选择函数choose_max_entropy( )的操作过程如下：

(1) 计算每一维的熵值；

(2) 选择具有最大熵值的维作为划分维。

划分值确定子函数max_div_entropies( )的主要功能是在划分维上确定一个划分值，使划分后的两个子区间的熵值和最大。最直接的方法是尝试将每一个划分维上的值作为划分值，并计算以此作为划分值得到的子区间的熵值和；最后选择使子区间熵值和最大的值作为划分值。显然，这种查找方法能找到最优划分值，但效率不高。本文采用近似逼近最优值的方法选择划分值，过程如下：

(1) 假定在划分维上有m个独立的数值，从$\lfloor m/2 \rfloor$起始，计算待分子区间的熵值和；

(2) 左右移动划分值，如果使待分子区间的熵值和减少，停止移动；否则继续移动，直到找到最大待分子区间的熵值和对应的划分值。实验发现最优值通常在$\lfloor m/2 \rfloor$附近，如何从理论上证明上述逼近过程的正确性，将是下一步的工作。

3 实 验

本文通过UCI机器学习数据集中的真实数据集Adults[9]来验证基于熵的EMP方法的有效性与安全性。Adults是K-匿名化测试的基准测试数据集，采用与文献[2-6]中一样的数据预处理方法，得到实验数据集是30 162条人口普查的数据记录，且每条记录有八项常规属性，如年龄、性别、国籍等。

文献[10]提出了通用的K-匿名化质量测量标准—— 可辨别性测量标准(Discernability Metric，DM)。DM是每个等价类的数据点数平方之和，即$C_{\text{DM}} = \sum_{\text{EquivClasses}E} |E|^2$。文献[5]采用平均等价类大小(Average Equivalence class Size，AES)作为度量标准。AES是DM的一种替换标准，直接采用数据集的数据记录数除以K-匿名化产生的等价类的个数。AES或DM越小，表明等价类中的数据点数越接近K，划分精度越高。然而划分精度与隐私安全是相互矛盾的，划分的精度越高，数据的隐私安全性越低。

本文提出以等价类中数据点的熵密度来衡量K-匿名结果的数据安全性。一个等价类中数据点的熵值越大，即数据离散程度越高，则能破解K-匿名隐私安全的可能性越小。根据多维划分产生等价类的过程可知，一个等价类是用准标识符空间的一个子空间E^s标识的，即$E^s=\{[\min A_1, \max A_1], [\min A_2, \max A_2], \cdots, [\min A_d, \max A_d]\}$。子空间的超体积(Hyper-Volume)可以用子空间每一维的取值范围相乘计算，即$\text{HV}(E) = \prod_{i=1}^{d} (\max A_i - \min A_i)$。

对于具有相同子空间大小的两个等价类，熵值越大的等价类的安全性越高。一个等价类的熵密度可由等价类的熵值除以等价类子空间的超体积计算。因此，对于一个K-匿名划分结果的安全性，可用所有等价类的熵密度之和(SED)来度量，即：

$$\text{SED} = \sum_{\text{EquivClasses}E} \frac{\text{Entropy}(E)}{\text{HV}(E)} \qquad (2)$$

本文采用Median Mondrain算法和EMP算法对Adults数据集进行K-匿名划分，K分别设置为2,3,4,5,

6,7,8,9,10。*K*-匿名的划分结果分别用AES和SED衡量，其结果分别如图2和图3所示。

图2反映了EMP算法在Adults上的*K*-匿名划分精度略低于Median Mondrain算法。

图3反映了EMP算法得到的*K*-匿名划分的熵密度和远大于Median Mondrain算法得到的划分结果。

对于Adults数据集而言，选择具有最大取值范围的维划分可获得最高的划分精度，而采用EMP算法能在略微降低划分精度的情形下，极大地提高数据隐私保护的安全性。

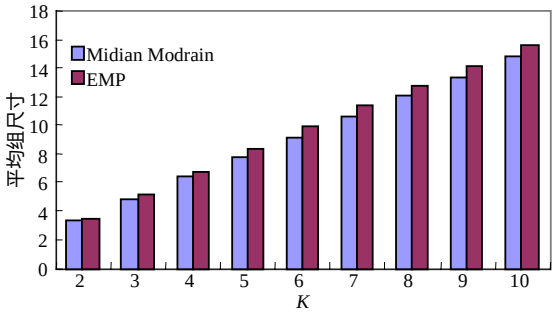

图2　采用AES的实验对比结果

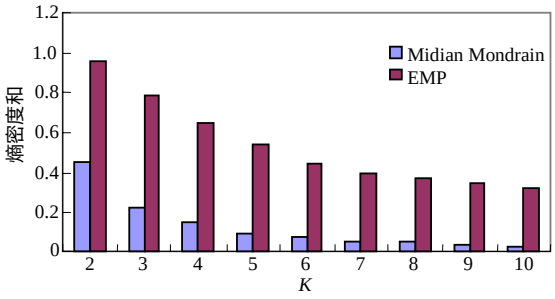

图3　采用SED的实验对比结果

4　结　论

针对Median Mondrain算法不能很好地平衡数据划分精度与数据隐私安全保护之间的矛盾，本文提出一种基于熵的多维*K*-匿名划分方法EMP，并在此基础上提出新的衡量数据*K*-匿名化结果隐私保护安全性的标准SED，即通过计算数据集熵值来控制多维划分过程，并用划分结果的熵值密度之和评估划分的隐私保护质量。在基准测试数据集Adults上的实验结果表明，与Median Mondrain算法相比，在略微降低数据划分精度的情形下，EMP算法极大地提高数据隐私保护的质量。

参　考　文　献

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

[2] SWEENEY L. Achieving *K*-anonymity privacy protection using generalization and suppression[J]. International Journal of Uncertainty, Fuzziness and Knowledge-based Systems, 2002, 10(5): 571-588.

[3] FUNG B, WANG K, YU P. Top-down specialization for information and privacy preservation[C]//Proceedings of the 21st ICDE. Los Alamitos, USA: IEEE Computer Society Press, 2005: 205-216.

[4] LEFEVRE K, DEWITT D, RAMAKRISHNAN R. Incognito: efficient full-domain *K*-anonymity[C]//Proceedings of the ACM SIGMOD. New York, USA: ACM Press, 2005: 49-60.

[5] LEFEVRE K, DEWITT D, RAMAKRISHNAN R. Mondrian multidimensional *K*-anonymity[C]//Proc of 22nd ICDE. Los Alamitos, USA: IEEE Computer Society Press, 2006: 25-34.

[6] LEFEVRE K, DEWITT D, RAMAKRISHNAN R. Workload-aware anonymization[C]//Proceedings of the ACM KDD'06. New York, USA: ACM Press, 2006: 277-286.

[7] BARBARA D, LI Y, COUTO J. Coolcat: an entropy-based algorithm for categorical clustering[C]//Proceedings Press of ACM CIKM. New York, USA: ACM Press, 2002: 582-589.

[8] FRIEDMAN J, BENTLEY J, FINKEL R. An algorithm for finding best matches in logarithmic time[J]. ACM Trans on Mathematical Software, 1977, 3(3): 209-226.

[9] BLAKE C, MERZ C. UCI repository of machine learning databases[DB/OL]. http://www.ics.uci.edu, 1998-11-07.

[10] BAYARDO R, AGRAWAL R. Data privacy through optimal *K*-anonymization[C]//ProceedingsPress of the 21st ICDE. Los Alamitos, USA: IEEE Computer Society Press, 2005: 217-228.

编　辑　黄　莘