doi:10.11959/j.issn.1000-436x.2015204

基于投影区域密度划分的 *k* 匿名算法

王超，杨静，张健沛，吕刚
(哈尔滨工程大学 计算机科学与技术学院，黑龙江 哈尔滨 150001)

摘 要：在数据发布的隐私保护中，现有的算法在划分临时匿名组时，没有考虑临时匿名组中相邻数据点的距离，在划分过程中极易产生许多不必要的信息损失，从而影响发布匿名数据集的可用性。针对以上问题，提出矩形投影区域，投影区域密度和划分表征系数等概念，旨在通过提高记录点的投影区域密度来合理地划分临时匿名组，使划分后的匿名组产生的信息损失尽量小；并提出基于投影区域密度划分的 *k* 匿名算法，通过优化取整划分函数和属性维选择策略，在保证匿名组数量不减少的同时，减少划分过程中不必要的信息损失，进一步提高发布数据集的可用性。通过理论分析和实验验证了算法的合理性和有效性。

关键词：隐私保护；临时匿名组；矩形投影区域；投影区域密度；划分表征系数
中图分类号：TP391.7 **文献标识码**：A

Algorithm for *k*-anonymity based on projection area density partition

WANG Chao, YANG Jing, ZHANG Jian-pei, LV Gang
(College of Computer Science and Technology, Harbin Engineering University, Harbin 150001, China)

Abstract: In data publishing privacy preserving, while classifying temporary anonymous groups, the existing algorithms didn't consider the distance between adjacent data points, and could easily produce a lot of unnecessary information loss, thus affecting the availability of released anonymous data sets. To solve the above problem, the concept of rectangular projection area, the projection area density and partition coefficient characterization were presented, aim to increase the recording points's projection area density to divide temporary anonymous group reasonably, and to make the information loss of divided anonymous groups as small as possible. And presents the algorithm for *k*-anonymity based on projection area density partition, by optimizing the rounded partition function and properties dimension selection strategy, to reduce unnecessary information loss and to further improve the availability of released data sets, without reducing the number of anonymous groups. The rationality and validity of the algorithm are verified by theoretical analysis and multiple experiments.

Key words: privacy preserving; temporary anonymous group; rectangular projection area; projection area density; partition coefficient characterization

1 引言

随着网络、数据存储技术的快速发展，数据库中存储的数据呈爆炸式增长，大量的个人数据能够使用数据挖掘的方法进行分析，虽然知识发现和数据挖掘等数据分析技术充分地挖掘了信息资源，但是也极有可能造成个人的隐私信息泄露。因此，如何在保护数据隐私的同时不影响数据的可用性已经成为信息安全与数据挖掘领域的一个重要研究方向[1]。

收稿日期：2014-03-20；**修回日期**：2015-01-07
基金项目：国家自然科学基金资助项目（61370083，61073043，61073041）；高等学校博士学科点专项科研基金资助项目（20112304110011, 20122304110012）；黑龙江省自然科学基金资助项目（F200901）；哈尔滨市科技创新人才研究专项（优秀学科带头人）基金资助项目（2011RFXXG015）
Foundation Items: The National Natural Science Foundation of China(61370083, 61073043, 61073041); The Research Fund for the Doctoral Program of Higher Education of China (20112304110011,20122304110012); The Natural Science Foundation of Heilongjiang Province (F200901); The Harbin Special Funds for Technological Innovation Research(2011RFXXG015)

2011 年，全球著名应用程序和数据安全解决方案厂商 Imperva 在公布的十大安全威胁趋势预测中，着重提到了社交网络中隐私和安全，以及数据安全和隐私法规的问题；2012 年 12 月底，十一届全国人大常委会第三十次会议审议通过了关于加强网络信息保护的决定草案，草案的核心内容为网络主体的行为约束，这标志着我国对网络行为的管理已经上升到了立法高度。我国首个个人信息保护国家标准——《信息安全技术公共及商用服务信息系统个人信息保护指南》已于 2013 年 2 月 1 日起实施。目前，许多国家都将信息安全的研究提升到了国家战略层次，隐私保护已成为不可或缺的重要研究方向。

隐私保护数据发布在保护数据隐私和维持数据可用性间寻求折衷。目前，学术界对隐私保护的技术做了比较深入的研究[2,3]，大致可以分为 3 类: 基于数据失真的技术[4,5]、基于数据加密的技术[6,7]和基于限制发布的技术[8~11]。基于限制发布的技术不仅能保证较高的隐私保护度，还能保证较低的数据依赖性、数据损失以及计算开销，因此得到了广泛的应用。

k 匿名模型是最早被提出的一种隐私保护机制[12]，满足 k 匿名安全要求的数据包含许多匿名组，且每个匿名组内部的记录（至少 k 条）是无法区分的。一般来说，最终发布的匿名数据集中包含的匿名组越多，该数据集包含的信息越丰富；且数据集的平均匿名组规模越小，信息损失越小，匿名化的数据越接近原来的真实数据，该数据集的可用性越高。

现有的许多基于限制发布的技术在实现 k 匿名算法时，采用基于分治策略的概化技术[13,14]，即使用二划分的方法，每次将一个大的临时匿名组划分为 2 个容量尽可能相等的较小匿名组。这种“局部贪心”的策略不具备“全局眼光”，按此策略进行划分，可减少潜在的匿名组数量；吴英杰等[15]提出了基于取整划分函数的 k 匿名算法，对原算法“局部贪心”的策略进行改进，提出基于取整划分函数的划分策略，避免了“可能减少潜在匿名组数量”这一情况的发生，增加了匿名数据集中匿名组的数量，提高了数据集的可用性。

然而，基于取整划分函数的划分策略，在划分临时匿名组时，没有考虑临时匿名组中相邻数据点的距离，在划分过程中极易产生许多不必要的信息损失，从而影响发布的匿名数据集的可用性。针对以上问题，本文提出基于投影区域密度划分的 k 匿名算法，实现对数据隐私保护和保持数据可用性的兼顾。

本文的主要贡献如下。

1) 提出矩形投影区域，投影区域密度和划分表征系数等概念，通过提高记录点的投影区域密度来合理地划分临时匿名组，使划分后的匿名组产生的信息损失尽量小，提高数据可用性。

2) 提出基于投影区域密度划分的 k 匿名算法，通过优化取整划分函数和属性维选择策略，在保证匿名数据集中匿名组数量不减少的同时，减少划分过程中不必要的信息损失，进一步提高发布数据集的可用性。

3) 通过理论分析和多个数据集上的实验验证，文中方法产生匿名组的规模在最坏的情况下小于 $2k$；在发布数据表足够大时，产生的匿名组的平均规模将足够趋近于 k，并且在数据质量没有降低的前提下，以较低的时间消耗为代价，减少了划分匿名组的信息损失，提高了数据可用性。

2 问题描述

本文用 $T(Q_1, Q_2,\cdots,Q_d, S_1, S_2,\cdots,S_m)$描述待发布的数据表，简称为 $T(d)$。其中，d 是准标识符的个数，m 是敏感属性的个数。k 匿名模型要求数据表中任意一条记录至少与其他 k–1 条记录在准标识符上完全相同。假设 $\Pi_{QI}(T)$为表 $T(d)$在属性集合 QI 上的投影，表 $T(d)$在属性集合 QI 下满足 k 匿名当且仅当 $\Pi_{QI}(T)$中的任意一条记录至少重复出现 k 次。在 Π 运算下，所有相同值的记录构成一个匿名组，如果该匿名组的元素个数不小于 k，则该匿名组是一个 k 匿名组，如果数据表中的所有匿名组都是 k 匿名组，那么该数据表满足 k 匿名模型。

如果能够为表 $T(d)$的每个属性域定义一个顺序，那么，$\Pi_{1\leqslant i\leqslant d}Q_i$ 可以映射到一个多维空间中，$T(d)$的每条记录都看作是多维空间的一个点。此时，构造原始数据表满足 k 匿名模型的数据表等价于寻找与其对应的多维空间中某个多维矩形区域的一个划分[15]。在二维情况下，这个矩形区域是一个平面矩形；在三维情况下，它是一个长方体；在多维情况下，该矩形区域在任意平面的投影都是一个平面矩形。不失一般性，该矩形区域可以取在该多维空间中能够覆盖所有记录点的最小矩形区域，数据

表的 k 匿名组等价于该矩形区域中的某个划分子区域所包含的记录点，每个匿名组的规模为其对应划分子区域中记录点的个数。

文献[15]提出了基于取整划分函数的 k 匿名算法，采用基于取整划分函数的划分策略，避免了常规二划分方法中“可能减少潜在匿名组数量”这一情况的发生，增加了匿名数据集中匿名组的数量，提高了发布数据集的可用性；同时从理论上证明了该算法在非平凡数据集中总能取得比 $2k-1$ 更低的上界；并且当数据集的大小超过 $2k^2$ 时，算法所产生的匿名化数据的匿名组规模必然不会超过 $k+1$。然而，基于取整划分函数的划分策略，在划分临时匿名组时，没有考虑临时匿名组中相邻数据点的距离，在划分过程中极易产生许多不必要的信息损失，从而影响发布的匿名数据集的可用性。如图 1 所示。

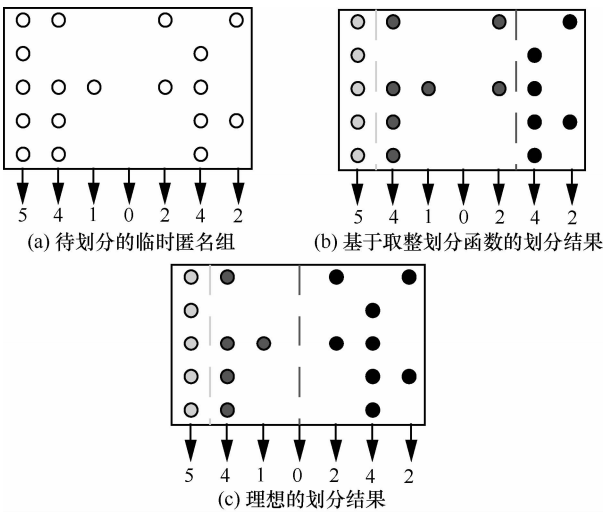


图 1　临时匿名组的划分情况

图 1 是某一临时匿名组在不同划分策略下的不同划分结果，该临时匿名组中共有 18 条记录，k 取值为 5。图 1(a)是待划分的临时匿名组，图 1(b)是基于取整划分函数的划分策略所产生的划分结果。该结果虽然满足 k 匿名模型，但形成的匿名组包含的记录在泛化时会产生较大的信息损失（主要是中间匿名组中记录间距离过大导致）。图 1(c)是一个比较理想的划分结果，该划分结果不仅满足 k 匿名模型，还会避免泛化过程中不必要的信息损失，提高匿名数据集的可用性。对比图 1(b)和图 1(c)的划分结果，可见图 1(b)的划分结果导致记录点的分布过于稀疏，图 1(c)的划分结果则使记录点的分布相对稠密，因此记录点分布的稀疏程度可以作为判断泛化产生的信息损失多少的标准：记录点分布的稀疏，则泛化产生的信息损失较大；反之，记录点分布的稠密，泛化产生的信息损失较小。

本文将记录点分布的稀疏程度进行量化，提出投影区域密度等概念，通过提高记录点的投影区域密度来合理地划分临时匿名组，使划分后的匿名组产生的信息损失尽量小。

3　相关概念

本文主要研究数值型属性的划分策略，因为任意 2 个相同的分类型[16]的属性值都是相同的，对分类型属性的划分不会产生额外的信息损失。

定义 1　临时匿名组：给定表 $T(d)$和 k，经过若干次划分之后，$T(d)$被分成 m 个部分，则划分后的任意一个部分都是一个临时匿名组。

为了清楚地阐述后文的基于投影区域密度划分的策略，本文定义了矩形投影区域、投影区域密度和划分表征系数的概念，具体表述如下。

定义 2　矩形投影区域：给定表 $T(d)$，$T(d)$的每条记录都可以看作是 d 维正交空间的一个点。P 表示所有记录点形成的集合，i、j 为 d 维正交空间的任意 2 个维度，则 d 维正交空间中的点集 P 在 i、j 维上的投影即形成矩形投影区域，用 $\Pi_{i,j}(P)$表示。

定义 3　投影区域密度：给定表 $T(d)$，i、j 为 d 维正交空间的任意 2 个维度，Q_i 和 Q_j 为对应的有序域，对于任意临时匿名组 X，$Q_i\max$、$Q_j\max$ 和 $Q_i\min$，$Q_j\min$ 分别为对应有序域在临时匿名组 X 中的最大值和最小值。则其在维度 i，j 上的投影区域密度 $\rho_{i,j}(X)$ 表示为

$$\rho_{i,j}(X)=\frac{|X|}{(Q_j\max-Q_j\min)(Q_i\max-Q_i\min)} \quad (1)$$

其中，$|X|$表示临时匿名组 X 中包含的记录条数。

定义 4　划分表征系数：给定表 $T(d)$，X 是任意临时匿名组，将其进行划分后得到 2 个临时匿名组 X_1 和 X_2，则对于此次划分，其划分表征系数 $Pcc_{i,j}(X)$表示为

$$Pcc_{i,j}(X)=\frac{\rho_{i,j}(X_1)+\rho_{i,j}(X_2)}{2\rho_{i,j}(X)} \quad (2)$$

划分表征系数可以用来衡量此次划分的合理程度：划分表征系数大于 1，表示划分之后新的临时匿名组中的记录点分布呈稠密状态，相应地，在泛化过程中信息损失减少；划分表征系数小于 1，

表示划分之后新的临时匿名组中的记录点分布呈稀疏状态，相应地，在泛化过程中信息损失增大。因此，在划分过程中，本文目标是划分表征系数能尽可能的大，以减少信息损失，提高数据可用性。

4 基于投影区域密度划分的 k 匿名算法

本文提出基于投影区域密度划分的 k 匿名算法(AA-PADP, algorithm for k-anonymity based on projection area density partition)，通过提高记录点的投影区域密度来合理地划分临时匿名组，并针对取整划分策略的不足，通过优化取整划分函数和属性维选择策略，在保证匿名数据集中匿名组数量不减少的同时，减少划分过程中不必要的信息损失，进一步提高发布数据集的可用性。

4.1 取整划分策略存在的问题

给定临时匿名组 X，$|X|=\alpha k+\beta$，其被划分成 2 个子匿名组，规模分别为 $\alpha_1 k+\beta_1$ 和 $\alpha_2 k+\beta_2$，显然，$\alpha_1+\alpha_2\leqslant\alpha$。在取整划分策略中，为了最大化可能产生的匿名组数量，提出如下的划分函数，其划分后 2 个匿名组的规模分别为

$$\begin{cases} X_1:|X_1|=\left\lfloor\dfrac{\alpha}{2}\right\rfloor k+\left\lfloor\dfrac{\beta}{2}\right\rfloor \\ X_2:|X_2|=\left\lceil\dfrac{\alpha}{2}\right\rceil k+\left\lceil\dfrac{\beta}{2}\right\rceil \end{cases} \tag{3}$$

其中，$\lfloor\ \rfloor$ 是向下取整，$\lceil\ \rceil$ 是向上取整，$\alpha_1=\left\lfloor\dfrac{\alpha}{2}\right\rfloor$，$\alpha_2=\left\lceil\dfrac{\alpha}{2}\right\rceil$。

然而，该取整划分策略属于硬划分，由式(3)可知，给定表 $T(d)$和 k，在一次划分后，$\beta_1=\left\lfloor\dfrac{\beta}{2}\right\rfloor$，$\beta_2=\left\lceil\dfrac{\beta}{2}\right\rceil$。即划分后匿名组 X_1 和 X_2 的规模是固定的，则划分的结果也是固定的，因此导致了图 1(b)划分结果中记录点的分布呈稀疏状态。

4.2 基于投影区域密度划分的策略

针对取整划分策略导致划分结果相对固定，易产生较大的信息损失的问题，本文对取整划分策略进行改进，提出基于投影区域密度划分的策略，放松了对式(3)中导致硬划分 β 值的限制，在保证匿名组数量不减少的前提下，降低泛化过程中的信息损失。

给定临时匿名组 X，$|X|=\alpha k+\beta$，其被划分成 2 个子匿名组，规模分别为 $\alpha_1 k+\beta_1$ 和 $\alpha_2 k+\beta_2$，显然，$\alpha_1+\alpha_2\leqslant\alpha$。在基于投影区域密度划分策略中，本文提出如下的划分函数，其划分后 2 个匿名组的规模分别为

$$\begin{cases} X_1:|X_1|=\left\lfloor\dfrac{\alpha}{2}\right\rfloor k+\beta_1 \\ X_2:|X_2|=\left\lceil\dfrac{\alpha}{2}\right\rceil k+\beta_2 \end{cases} \tag{4}$$

其中，$\lfloor\ \rfloor$ 是向下取整，$\lceil\ \rceil$ 是向上取整，$\alpha_1=\left\lfloor\dfrac{\alpha}{2}\right\rfloor$，$\alpha_2=\left\lceil\dfrac{\alpha}{2}\right\rceil$，$\beta_1\geqslant 0$，$\beta_2\geqslant 0$，$\beta_1+\beta_2=\beta$。

基于投影区域密度划分属于软划分，它可以根据不同的 β_1、β_2 值来调节划分结果，并且调节的效果随着 k 值的增加而增加。因此，在划分过程中，可以通过改变 β_1、β_2 值来调整划分后的临时匿名组，使其内部的记录点呈现更加稠密的状态。

下面给出算法的详细描述。

算法 1 AA-PADP

输入：原始数据表 $T(d)$, 参数 k, P,Ω //P 表示所有记录点的集合，Ω 表示该 d 维空间中能覆盖 P 的最小多维矩形区域

输出：匿名数据表 $T(d)^*$

BEGIN

step1 令 $Q=\Phi$, $S=\Omega$, $|P|=\alpha k+\beta$, 其中，β 是比 k 小的非负整数。 //对于给定的表 $T(d)$和 k、α 和 β 的取值是一定的。

step2 遍历Ω的所有维i, 并找到合适的正整数w，满足：$\left|\bigcup_{p\in P\wedge\prod p\geqslant q(i,w)}\right|\geqslant\left\lfloor\dfrac{\alpha}{2}\right\rfloor k+\beta_1$,$\left|\bigcup_{p\in P\wedge\prod p\leqslant q(i,w)}\right|\geqslant\left\lceil\dfrac{\alpha}{2}\right\rceil k+\beta_2$, $\beta_1\geqslant 0$, $\beta_2\geqslant 0$, $\beta_1+\beta_2=\beta$，并计算该可能的划分表征系数。对于每个维度 i，求其最大的划分表征系数及其对应的划分点 w //$\prod_i(p)$ 表示点 p 在 d 维空间中第 i 维的投影，$q(i, w)$表示第 i 维有序域中的元素值。

step3 在 step2 获得的划分表征系数中，选择划分表征系数最大值对应的维度 i'，及其对应的合适的正整数 w'。

step4 在第 i'维 w'处，将 S 划分成为 2 个多维矩形区域 S_1 和 S_2。

step5 将 P 划分成 2 个点集 P_1 和 P_2，满足：$|P_1|=\left\lfloor\dfrac{\alpha}{2}\right\rfloor k+\beta_1$，$|P_2|=\left\lfloor\dfrac{\alpha}{2}\right\rfloor k+\beta_2$，$\beta_1\geqslant 0$，$\beta_2\geqslant 0$，

$\beta_1+\beta_2=\beta$，并且对于 P_1 中任意元素 p，都有 $\prod_{i'}(p)\leqslant q(i',w')$；对于 P_2 中任意元素 p，都有 $\prod_{i'}(p)\geqslant q(i',w')$。$P_1$ 的点属于 S_1、P_2 的点属于 S_2。

step6　如果$|P_1|\geqslant 2k$, 则利用参数 k, P_1 和 S_1 继续递归执行。

step7　如果$|P_1|<2k$, 则 $Q=Q\cup P_1$。

step8　如果$|P_2|\geqslant 2k$, 则利用参数 k、P_2 和 S_2 继续递归执行。

step9　如果$|P_2|<2k$, 则 $Q=Q\cup P_2$。

step10　依次处理 Q 中的每个集合，对其进行泛化操作，得到匿名数据表 $T(d)^*$。

END

算法 AA-PADP 首先遍历 Ω 的所有维，在每次遍历时，找到一个合适的正整数 w 作为划分点，并计算此次可能的划分表征系数。在遍历结束后，选择最大划分表征系数对应的维度 i'，及其对应的正整数 w'；然后根据维度 i'和正整数 w'将 S 划分成为 2 个多维矩形区域 S_1 和 S_2。并根据 S_1 和 S_2 所含记录点的个数决定继续递归执行，还是将其添加至集合 Q 中；递归结束后，对 Q 中的每个集合进行泛化操作，得到匿名数据表 $T(d)^*$。

4.3　算法的合理性和复杂性分析

4.3.1　算法的合理性分析

定理 1　给定表 $T(d)$和 k，若$|T(d)|=\alpha k+\beta$，则基于投影区域密度划分的 k 匿名算法产生的匿名化数据恰好包含 α 个匿名组，并且满足 k-匿名模型。

证明　首先，当 $\alpha=1$ 时，结论显然成立。当 $\alpha\geqslant 2$ 时，根据算法思想，$T(d)$会被分成 2 个部分，每个部分的大小分别是 $\left\lfloor\frac{\alpha}{2}\right\rfloor k+\beta_1$ 和 $\left\lceil\frac{\alpha}{2}\right\rceil k+\beta_2$，并且 $\beta_1\geqslant 0$，$\beta_2\geqslant 0$，$\beta=\beta_1+\beta_2$，每个部分的大小均大于等于 k。这 2 个部分的 k 系数之和为 α。同样，对于任意临时匿名组 X，当它被划分成 2 个部分之后，其 2 个子匿名组的 k 系数之和仍等于 X 的 k 系数。并且，对于任意一个临时匿名组 X，必然可以递归地划分 k 系数为 1 的子临时匿名组的并。因此，定理 1 得证。

由定理 1 可知，如果表 $T(d)$中记录的个数不小于 k，那么经过 AA-PADP 算法处理后的匿名数据表 $T(d)^*$一定满足 k-匿名模型，并且恰好包含 α 个匿名组。

定义 5　匿名组层次[15]:给定表 $T(d)$，若称 $T(d)$是第 0 层匿名组，则称第 $i-1(i>0)$层的临时匿名组划分后形成的子临时匿名组为第 i 层匿名组。

定理 2　给定表 $T(d)$和 k，若$|T(d)|=\alpha k+\beta$, $[\mathrm{lb}\alpha]=x$，则任意第 i 层匿名组 X 的 k 系数 α_i 必然满足：$2^{x-i}\leqslant\alpha_i<2^{x-i+1}$。

证明　本文采用与文献[15]相似的数学归纳法证明。当 $i=0$ 时，显然有 $2^{x-i}=2^x=2^{[\mathrm{lb}\alpha]}\leqslant 2^{\mathrm{lb}\alpha}=\alpha=\alpha_i<2^{[\mathrm{lb}\alpha_i]+1}<2^{x-i+1}$。

不妨设对于第 i 层匿名组 X 的 k 系数 α_i 必然满足：$2^{x-i}\leqslant\alpha_i<2^{x-i+1}$。那么对于第 $i+1$ 层的临时匿名组 X, 其 k 系数为 α_{i+1}, 根据式(4)有 $\left\lfloor\frac{\alpha_i}{2}\right\rfloor\leqslant\alpha_{i+1}\leqslant\left\lceil\frac{\alpha_i}{2}\right\rceil$。因为 $2^{x-i}\leqslant\alpha_i<2^{x-i+1}$，有 $\left\lfloor\frac{\alpha_i}{2}\right\rfloor\geqslant\left\lfloor\frac{2^{x-i}}{2}\right\rfloor=2^{x-(i+1)}$，$\left\lceil\frac{\alpha_i}{2}\right\rceil<\left\lceil\frac{2^{x-i+1}}{2}\right\rceil=2^{x-(i+1)+1}$，所以 $2^{x-(i+1)}\leqslant\alpha_{i+1}<2^{x-(i+1)+1}$。

综上，定理 2 得证。

定理 3　给定表 $T(d)$和 k，$|T(d)|=\alpha k+\beta$，当$|T(d)|\geqslant 2k$，且 $k>3$ 时，基于投影区域密度划分的 k 匿名算法产生的匿名组的规模在最坏情况下小于 $2k$；在$|T(d)|$足够大时，产生匿名组的平均规模将足够趋近于 k。

证明　由定理 1 可知，基于投影区域密度划分的 k 匿名算法产生的匿名化数据恰好包含 α 个匿名组，且满足 k-匿名模型。因此，根据式(4)，在最坏情况下，包含最多纪录的匿名组的规模为 $\alpha k+\beta-(\alpha-1)k=k+\beta<2k$。

匿名组的平均规模为$|T(d)|/\alpha=k+\beta/\alpha$。在$|T(d)|$足够大时，即$|T(d)|\to\infty$ 时，$\alpha k+\beta\to\infty$。在 k 取值一定，$\beta<k$ 的情况下，$\alpha\to\infty$。因此 $k+\beta/\alpha\to k$，定理得证。

以上定理说明，在非平凡条件下，基于投影区域密度划分的 k 匿名算法能产生一个较优的匿名组规模上界，并且在海量数据条件下，产生的匿名组的平均规模将足够趋近于 k，保证了较高的数据质量。

定理 4　给定表 $T(d)$和 k，若$|T(d)|=\alpha k+\beta$，则基于投影区域密度划分的 k 匿名算法产生的匿名化数据，其投影区域密度增大，信息损失减小。

证明　假设：对于任意临时匿名组 X，覆盖它的最小 d 维矩形区域是 S，经第 i 维 w 处的一次划分产生 2 个子临时匿名组 X_1 和 X_2，其对应的 d 维矩形区域分别为 S_1 和 S_2。由基于投影区域密度划分的策略可知，在第 i 维 w 处的划分表征系数最大，即划分产生的 2 个矩形区域 S_1 和 S_2 中的投影区域

密度最大。增大投影区域密度，会增加单位区域中记录点的数量，在记录条数不变的情况下意味着每个记录点占据较小的矩形区域。因此，在泛化过程中，信息损失减小，定理得证。

4.3.2 算法的复杂性分析

定理 5 基于投影区域密度划分的 k 匿名算法的时间复杂度是 $O(dn\text{lb}\frac{n}{k})$。

证明 给定表 $T(d)$和 k，$|T(d)|=n=\alpha k+\beta$，对于第 j 层的任意临时匿名组 X，根据划分策略，在选择合适的维度 i 时，需要遍历表 T 的全部 d 个维度，时间复杂度为 $O(d)$;在选择合适的 w 分割线时，最多需要遍历临时匿名组 X 的全部记录，时间复杂度为 $O(|X|)$，由于第 j 层的所有临时匿名组规模之后必然不大于 n，因此整个第 j 层的临时匿名组划分的时间复杂度不大于 $O(dn)$。由定理 2 分析可知，表 T 最多可以划分到 $x=[\text{lb}\alpha]=[\text{lb}(n/k)]$层，因此整个算法的时间复杂度为 $O(dn\text{lb}\frac{n}{k})$。

5 实验及结果分析

5.1 实验数据及参数

本节对 AA-PADP 算法的性能进行实验分析，实验数据来源于 UCI knowledge discovery archive database(http://archive.ics.uci.edu/ml/datasets.html)，具体如表 1 所示。

选择 Blood 数据集中的 3 个属性，删除重复的记录，最终选择了 499 个记录作为实验数据集。选择 Image 数据集中恒为正的 6 个属性，将其中小数部分的记录做整数化处理，并删除重复的记录，得到 2 100 个记录作为实验数据集；选择 Recognition 数据集中的前 10 个属性，删除其中的重复记录作为实验数据集，其中 Recognition_2 000 数据集中包含 2 171 个记录，Recognition_20 000 数据集中包含 18 431 个记录，用来测试算法在较大规模数据集下的执行情况。

硬件环境为:Intel(R) Core 2 Quad CPU Q8 400 @2.66 GHz 2.67 GHz, 2.00 GB 内存，操作系统为 Microsoft Windows 7，算法均在 Matlab R2012a 下实现。

为了全面地衡量发布匿名数据的可用性，从匿名数据的质量以及信息损失两方面来进行分析。

5.1.1 匿名数据质量的度量

对于满足 k 匿名模型的匿名数据而言，可以通过最大匿名组的规模、匿名组的平均规模以及包含匿名组的总量来度量匿名数据的质量。一般来说，匿名数据包含的匿名组越多，平均匿名组规模越小，匿名化的数据越接近原来的真实数据，可用性也越高。显然，原始数据表包含最多的匿名组，匿名组规模最小，其可用性也最高。

根据文献[14,15,17]，最常见的质量度量函数是可辨别度量(DM, discernibility metric)，当表中没有元组被删除时，DM 可以定义为

$$DM(T^*)=\sum_{\forall E\in T^*}|E|^2 \tag{5}$$

其中，T^*表示 $T(d)$经匿名处理的发布数据表，E 是任意的匿名组。显然，DM 的函数值越小，匿名数据的可用性越大。

5.1.2 匿名数据信息损失的度量

为了度量匿名数据的信息损失，使用匿名组所占多维空间之和与整个多维空间的比值作为信息损失的度量，定义如下

$$IL=\frac{\sum_{i=1}^{n}S_i}{G} \tag{6}$$

其中，G 表示覆盖整个 d 维空间的最小多维矩形区域，S_i 表示划分结束后的覆盖第 i 个匿名组的最小多维矩形区域，n 表示匿名组的数量。信息损失越小，匿名数据的可用性越高。

5.2 实验分析

本节选择与 AA-PADP 算法相关的几个著名算法（flexible partition[15]、mondrian (relaxed)[13]、mondrian (strict)[13]）进行实验对比，关于数据质量、

表 1 实验数据信息

数据集名称	别名	属性数目	记录数目	数据类型
输血服务中心数据集	blood	3	499	实数
图像数据集	image	6	2 100	实数
字符识别数据集	recognition_2 000	10	2 171	实数
	recognition_20 000	10	18 431	实数

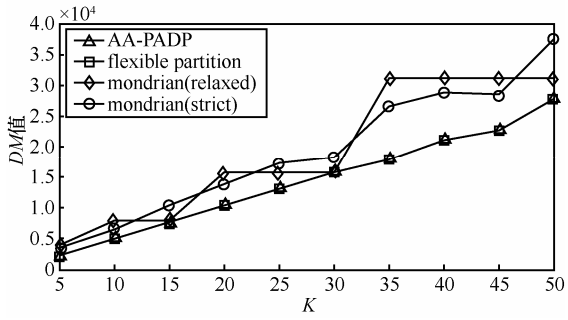
信息损失、匿名组数量以及算法执行时间进行分析。

图 2 显示了 4 个算法在 Blood 数据集上的执行结果。图 2(a)是关于 *DM* 的数据质量比较，其中，AA-PADP 和 flexible partition 的 *DM* 值较低，且几乎完全重合；mondrian (relaxed)和 mondrian (strict)的 *DM* 值相对较高。这是因为本文的 AA-PADP 算法和 flexible partition 算法，采取了较相似的划分策略，在划分的过程中，产生的匿名组数量最大，每个匿名组的大小在[k, 2k−1]之间，因此 *DM* 值最小，数据质量最高；mondrian (relaxed)和 mondrian (strict)算法采用二划分的策略，减少了潜在的匿名组数量，因此 *DM* 值较大，数据质量低。分析图 2(a)还发现，随着 K 值的增加，4 种算法的 *DM* 值都在增加，这是因为增大 K 值，虽然降低了匿名组的数量，却增加了每个匿名组的规模，造成了 *DM* 值的增加。

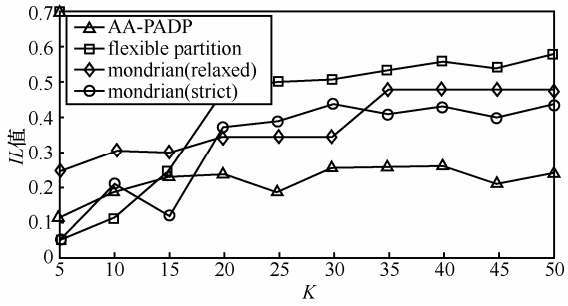
图 2(b)是关于 *IL* 的信息损失的比较，其中，AA-PADP 的信息损失相对较低，flexible partition 的信息损失相对较高，并且随着 K 值的增加，这种趋势越发得明显。这是因为 Flexible Partition 算法在划分的过程中仅仅要求划分的匿名组数量最大；而 AA-PADP 算法提出投影区域密度等概念，通过提高记录点的投影区域密度来合理地划分临时匿名组，在保证匿名数据集中匿名组数量不减少的同时，减少划分过程中不必要的信息损失；并且在优化的划分函数中，随着 K 值的增加，可调节的 β 值增加，有更大机会避免不必要的信息损失，因此随着 K 值的增加，AA-PADP 算法在 IL 方面的优势越发明显。分析图 2(b)还发现，随着 K 值的增加，4 种算法的 *IL* 值都在增加，这是因为增大 K 值，增加了每个匿名组的规模，使得匿名组在概化时产生了更大的信息损失。

图 2(c)是关于匿名组数量的比较，其中，AA-PADP 和 flexible partition 的匿名组数量最大，且几乎完全重合；mondrian (relaxed)和 mondrian (strict)的匿名组数量相对较低。这是因为本文的 AA-PADP 算法和 flexible partition 算法，采取了较相似的划分策略，在划分的过程中，保证匿名组数量最大；mondrian (relaxed)和 mondrian (strict)算法采用二划分的策略，减少了潜在的匿名组数量。分析图 2(c)还发现，随着 K 值的增加，4 种算法的产生的匿名组数量都在减少，这是因为增大 K 值，增加了每个匿名组的规模，造成了匿名组数量减少。

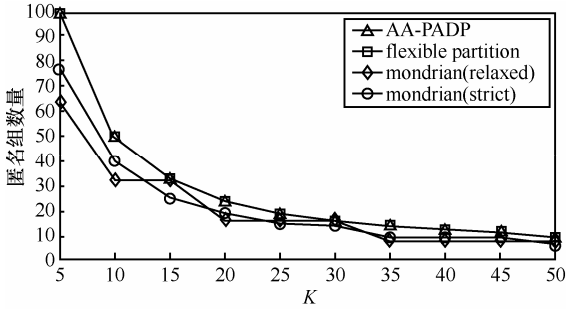
图 2(d)是关于算法执行时间的比较，其中，AA-PADP 算法的执行时间最大，但也在可接受的范围之内，flexible partition 的执行时间相对较低。这是因为 AA-PADP 算法在划分过程中，不仅要保证匿名组数量不减少，还要保证较小的信息损失，因此增加了算法的复杂度，消耗了更多的执行时间。分析图 2(d)还发现，随着 K 值的增加，4 种算法的执行时间都呈减少的趋势。这是因为增大 K 值，减少了匿名组的数量，减少了划分次数，因此算法的执行时间减少。



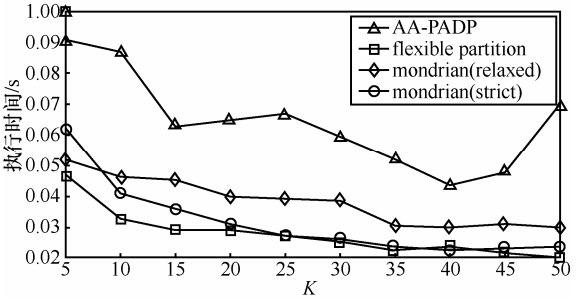
(a) 关于 *DM* 的数据质量比较



(b) 关于 *IL* 的信息损失比较



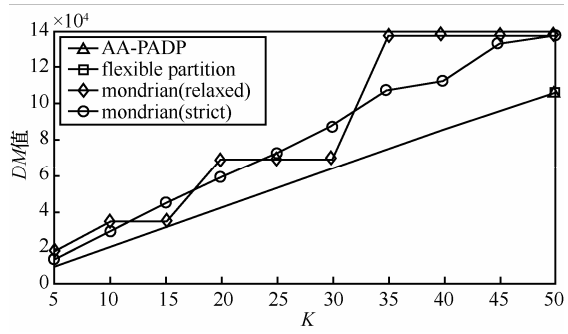
(c) 匿名组数量比较



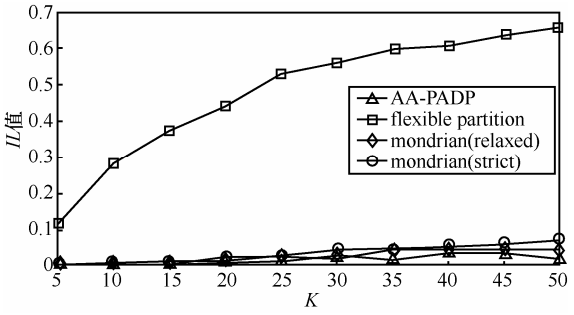
(d) 执行时间比较

图 2　在 Blood 数据集上的实验结果

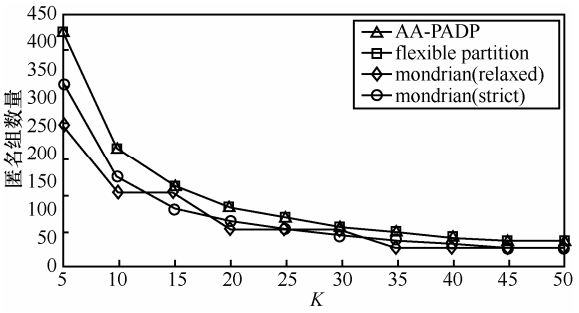
图 3~图 5 分别给出了 4 个算法在 image、recognition_2 000 和 recognition_20 000 数据集上的执行结果，结果呈现了与 blood 数据集上相似的实验效果，鉴于以上数据集代表了不同种类的数据，说明本文的算法具有一定的普适性；在拥有较大数据量的 recognition_20 000 数据集上的执行结果表明本文的算法的执行时间虽然有所增加（时间消耗在可接受范围之内），但是在保证数据质量的前提下，进一步减少了信息损失，提高了数据的可用性。



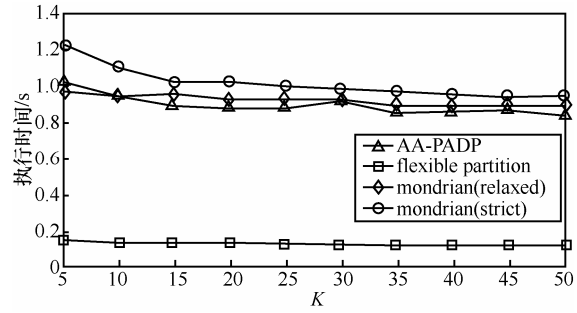
(a) 关于 *DM* 的数据质量比较



(b) 关于 *IL* 的信息损失比较

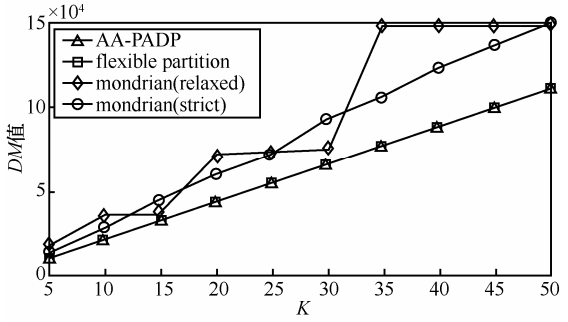


(c) 匿名组数量比较

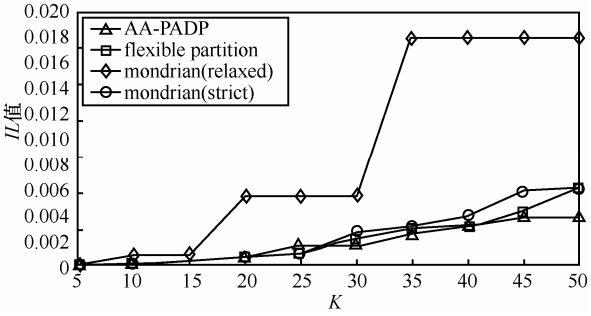


(d) 执行时间比较

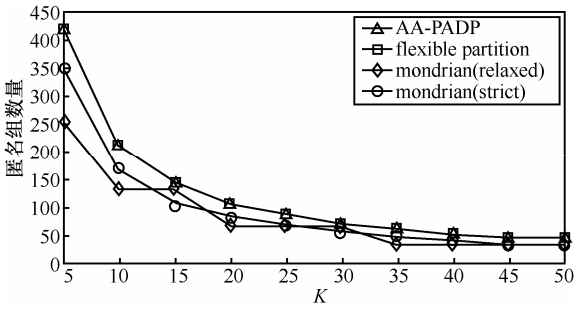
图 3　在 image 数据集上的实验结果



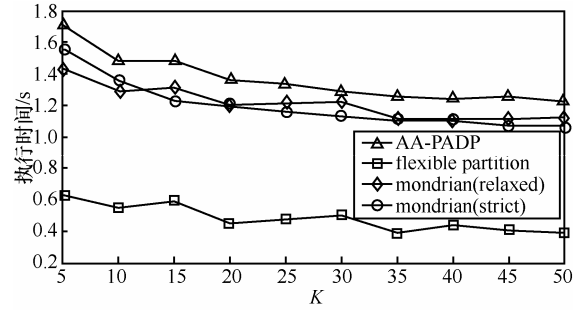
(a) 关于 *DM* 的数据质量比较



(b) 关于 *IL* 的信息损失比较

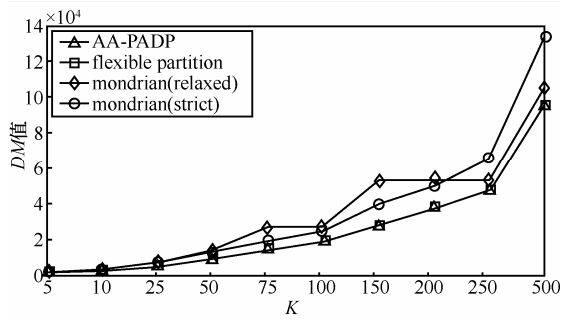


(c) 匿名组数量比较

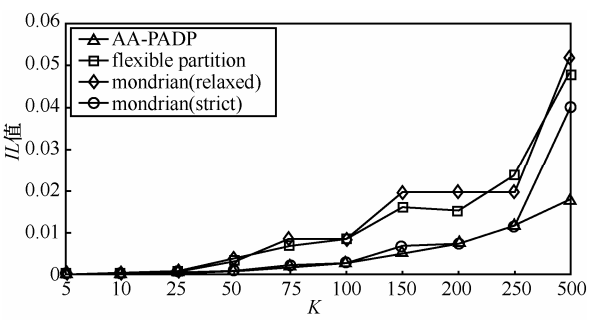


(d) 执行时间比较

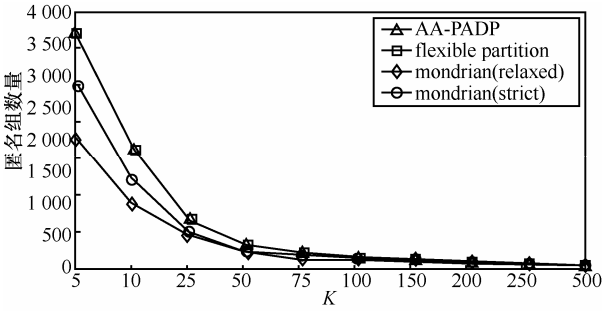
图 4　在 recognition_2 000 数据集上的实验结果



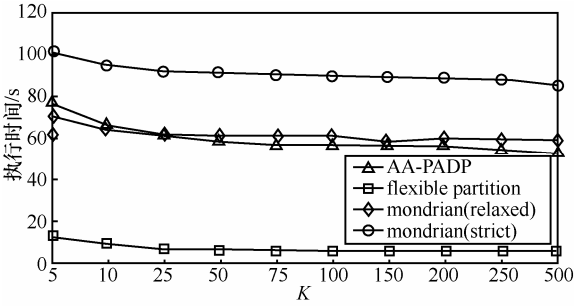
(a) 关于 *DM* 的数据质量比较



(b) 关于 *IL* 的信息损失比较



(c) 匿名组数量比较



(d) 执行时间比较

图 5　在 recognition_20 000 数据集上的实验结果

6　结束语

在数据发布的隐私保护中，现有的算法在划分临时匿名组时，没有考虑临时匿名组中相邻数据点的距离，在划分过程中极易产生不必要的信息损失，从而影响发布匿名数据集的可用性。针对以上问题，提出矩形投影区域、投影区域密度和划分表征系数等概念，旨在通过提高记录点的投影区域密度来合理地划分临时匿名组，使划分后的匿名组产生的信息损失尽量小；并提出基于投影区域密度划分的 *k* 匿名算法，通过优化取整划分函数和属性维选择策略，在保证匿名组数量不减少的情况下，减少划分过程中不必要的信息损失，进一步提高发布数据集的可用性。

理论分析和实验证明，基于投影区域密度划分的 *k* 匿名算法产生的匿名组的规模在最坏的情况下小于 $2k$；在发布数据表足够大时，产生的匿名组平均规模将足够趋近于 k，并且在数据质量没有降低的前提下，以较低的时间消耗为代价，减少了划分匿名组的信息损失，进一步提高了发布数据的可用性。

参考文献：

[1] 韩建民, 岑婷婷, 虞慧群. 数据表 *k*-匿名化的微聚集算法研究[J]. 电子学报, 2008, 36(11): 2021-2029.
HAN J M, CEN T T, YU H Q. Research in microaggregation algorithms for *k*-anonymization[J]. Acta Electronica Sinica, 2008, 36(11): 2021-2029.

[2] 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.
ZHOU S G, LI F, TAO Y F, *et al*. Privacy preservation in database applications: a survey[J]. Chinese Journal of Computers, 2009, 32(5): 847-861.

[3] 朱青, 赵桐, 王珊. 面向查询服务的数据隐私保护算法[J]. 计算机学报, 2010, 33(8): 1315-1323.
ZHU Q, ZHAO T, WANG S. Privacy preservation algorithm for service-oriented information[J]. Chinese Journal of Computers, 2010, 33(8): 1315-1323.

[4] SAYGIN Y, VERYKIOS V S, ELMAGARMID A K. Privacy preserving association rule mining[A]. Proceedings of the 12th International Workshop on Research Issues in Data Engineering (RIDE)[C]. San Jose, USA, 2002. 151-158.

[5] AGGARWAL C C, YU P S. A condensation approach to privacy preserving data mining[A]. Proceedings of the 9th International Conference on Extending Database Technology (EDBT)[C]. Heraklion, Greece, 2004. 183-199.

[6] YAO A C. How to generate and exchange secrets[A]. Proceedings of the 27th IEEE Symposium on Foundations of Computer Science (FOCS)[C]. Toronto, Canada, 1986. 162-167.

[7] CLIFTON C, KANTARCIOGLOU M, LIN X, *et al*. Tools for privacy preserving distributed data mining[J]. ACM SIGKDD Explorations, 2002, 4(2): 28-34.

[8] 韩建民, 于娟, 虞慧群. 面向敏感值的个性化隐私保护[J]. 电子学报, 2010, 38(7): 1723-1728.

HAN J M, YU J, YU H Q. Individuation privacy preservation oriented to sensitive values[J]. Acta Electronica Sinica, 2010, 38(7): 1723-1728.

[9] 杨静, 王波. 一种基于最小选择度优先的多敏感属性个性化 l-多样性算法[J]. 计算机研究与发展, 2012, 49(9): 2603-2610.
YANG J, WANG B. Personalized l-diversity algorithm for multiple sensitive attributes based on minimum selected degree first[J]. Journal of Computer Research and Development, 2012, 49(9): 2603-2610.

[10] 韩建民, 于娟, 虞慧群. 面向数值型敏感属性的分级 l-多样性模型[J]. 计算机研究与发展, 2011, 48(1): 147-158.
HAN J M, YU J, YU H Q. A multi-level l-diversity model for numerical sensitive attributes[J]. Journal of Computer Research and Development, 2011, 48(1): 147-158.

[11] 杨静, 王超, 张键沛. 基于敏感属性熵的微聚集算法[J]. 电子学报, 2014, 42 (7): 1327-1337.
YANG J, WANG C, ZHANG J P. Micro-aggregation algorithm based on sensitive attribute entropy[J]. Acta Electronica Sinica, 2014, 42(7): 1327-1337.

[12] SWEENEY L. *k*-anonymity: a model for protecting privacy[J]. International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10(5): 557-570.

[13] LEFEVRE K, DEWITT D J, RAMAKRISHNAN R. Mondrian multidimensional *k*-anonymity[A]. Proceedings of the 22nd International Conference on Data Engineering[C]. Atlanta, Georgia, USA, 2006. 25-34.

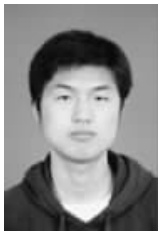
[14] HORE B, JAMMALAMADAKA R C, MEHROTRA S. Flexible anonymization for privacy preserving data publishing: a systematic search based approach[A]. Proceedings of the 7th SIAM International Conference on Data Mining[C]. Philadelphia, USA: Society for Industrial and Applied Mathematics, 2007. 497-502.

[15] 吴英杰, 唐庆明, 倪巍伟等. 基于取整划分函数的 *k* 匿名算法[J]. 软件学报, 2012, 23(8): 2138-2148.
WU Y J, TANG Q M, NI W W, *et al*. Algorithm for *k*-anonymity based on rounded partition function[J]. Journal of Software, 2012, 23(8): 2138-2148.

[16] 杨高明, 杨静, 张健沛. 半监督聚类的匿名数据发布[J]. 电子学报,2011, 32(11): 1489-1494.
YANG G M, YANG J, ZHANG J P. Semi-supervised clustering-based anonymous data publishing[J]. Acta Electronica Sinica, 2011, 32(11): 1489-1494.

[17] KIFER D. Attacks on privacy and deFinetti's theorem[A]. Proceedings of the 2009 ACM SIGMOD International Conference on Management of data[C]. New York, USA: Association for Computing Machinery, 2009. 127-138.

作者简介：



王超（1988-），男，河北沧州人，哈尔滨工程大学博士生，主要研究方向为数据库与知识工程、数据挖掘、隐私保护。

杨静（1962-），女，黑龙江哈尔滨人，哈尔滨工程大学教授、博士生导师，主要研究方向为数据库与知识工程、数据挖掘、隐私保护、软件理论等。

张健沛（1956-），男，黑龙江哈尔滨人，哈尔滨工程大学教授、博士生导师，主要研究方向为企业智能计算、数据库与知识工程、数据挖掘、社会网络、软件理论等。

吕刚（1988-），男，吉林白城人，哈尔滨工程大学硕士生，主要研究方向为数据库与知识工程、数据挖掘、隐私保护。