doi:10.11959/j.issn.1000-436x.2016100

基于聚类匿名化的差分隐私保护数据发布方法

刘晓迁，李千目
（南京理工大学计算机科学与工程学院，江苏 南京 210094）

摘 要：基于匿名化技术的理论基础，采用 DBSCAN 聚类算法对数据记录进行聚类，实现将个体记录匿名化隐藏于一组记录中。为提高隐私保护程度，对匿名化划分的数据添加拉普拉斯噪声，扰动个体数据真实值，以实现差分隐私保护模型的要求。通过聚类，分化查询函数敏感性，提高数据可用性。对算法隐私性进行证明，并实验说明发布数据的可用性。

关键词：差分隐私；隐私保护；聚类；数据发布；匿名化
中图分类号：TP392 **文献标识码：**A

Differentially private data release based on clustering anonymization

LIU Xiao-qian, LI Qian-mu
(School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China)

Abstract: Based on the theory of anonymization, the DBSCAN method was applied to divide all the data records into different groups to cover individuals. To provide priv enhancement, the Laplace noise was added to the anonymized partitioned data to perturb the real value of data record so that the requirements of differential privacy model were satisfied. With the clustering operation, the sensitivity of the query function has been partitioned to improve data utility. The proof of privacy has been given and experimental results have been provided to evaluate the utility of the released data.
Key words: differential privacy, privacy preservation, clustering, data release, anonymization

1 引言

互联网、传感技术和大数据的迅猛发展促使数据以指数增量急速暴涨，这些数据为政府部门和研究机构提供了重要的分析资源，促进了相关服务优化和产品升级。数据挖掘与分析技术在发现知识的同时，也带来了个体隐私泄露的问题，容易招致法律争端和道德争议。知识发现是数据挖掘和分析技术的首要任务，然而，在隐私保护问题日趋得到重视的情形下，如何保护数据隐私，构建隐私保护数据发布模型成为研究热点。隐私保护数据发布的任务是通过对数据记录进行扰动，确保隐私信息不被泄露，同时保证发布数据的可用性。换言之，隐私保护数据发布致力于数据发布给使用者和查询者之前的数据清洁工作，通过隐私保护手段对原始数据进行扰动，同时又注重加强数据查询和分析的准确性，力求构建算法以实现数据隐私性和可用性的平衡。

在数据分析领域，隐私保护技术大致可以分为数值扰动技术、查询限制技术、匿名分组技术以及数据分布技术等[1]。数值扰动技术通过添加随机噪声对数据原始值进行扰动，从而掩藏真实数值。查询限制技术从 2 个角度对数据查询者进行限制：1) 对查询数

收稿日期：2015-07-22；**修回日期：**2015-10-16
基金项目：中央高校基本科研业务专项基金资助项目（No.3091605104**）；国家自然科学基金资助项目（**No.61272419**）；江苏省未来网络前瞻性研究基金资助项目（**No.BY2013095-3-02**）；江苏省产学研前瞻性基金资助项目（**No.BY2014089, No.BY2013039, No.BY2013037**）；江苏省普通高校研究生创新计划基金资助项目（**No.KYLX15_0384**）**
Foundation Items: The Fundational Research Funds for the Central Universities (No.3091605104), The National Natural Science Foundation of China (No.61272419), The Future Network Prospective Study Project of Jiangsu Province (No. BY2013095-3-02), The Industry-University-Research Perspective Project of Jiangsu Province (No. BY2014089, No. BY2013039, No. BY2013037), Graduate Students Research Innovation Plan of Jiangsu Province (No.KYLX15_0384)

量的严格限制；2) 对可能实现组合推断的连续性查询进行控制。匿名分组技术以 *k*-anonymity 机制[2]为代表，*k*-anonymity 机制实际上是一种分组策略，将准标识符相同的数据相组合以实现总体数据记录分组，每组中至少有 *k* 条数据，从而将一条数据记录隐藏在 *k* 条数据当中。数据分布技术是指通过对数据进行垂直或水平划分，从物理存储角度对数据进行分布化，从而达到数据隐藏的目的。以上的这些隐私保护技术，往往不对攻击者所能获得的数据背景知识做定义，因此在处理复杂多变的攻击模型中，随着攻击者掌握背景知识的增加，往往会生成很多攻击变体，如联合性攻击、一致性攻击等。该类隐私保护算法在通用性上的限制，使数据管理者不得不针对个性化的攻击模式设计出新的隐私保护算法。如 *k*-anonymity 机制及其扩展模型往往难以应对组合式攻击、一致性攻击等模型。攻击者通过将用户的出生日期、性别、邮编等准标识符数据进行组合，常常能推断并锁定特定个体，进而获取该个体其他重要的隐私信息。Dwork[3]在 2011 年提出了差分隐私保护模型，该模型提供了顽健性的隐私证明，该模型不对攻击者的背景知识做限定，假设攻击者拥有全部的背景知识，因此克服了背景知识不断扩大引起的隐私保护模型不再适用的缺点。然而，差分隐私模型往往提供较差的数据可用性。

针对以上问题，本文提出了一种适用于数值属性数据的数据发布方法，该方法在满足差分隐私保护模型要求的同时，提升了数据可用性。

2 相关工作及概念定义

2.1 相关工作

近年来，在数据隐私保护领域取得了很多研究成果。Vassilios 等[4]从数据分布、数据修改、挖掘算法改造、数据隐藏等角度对数据隐私保护研究成果进行了综述，Dwork 等[5]研究了多属性数据库和垂直划分数据库的隐私保护问题。Friedman 等[6]提出了一种兼顾算法精确性和隐私性的差分隐私保护决策树分类算法，Kamalika 等[7]通过经验风险最小化实现了差分隐私保护的逻辑斯蒂回归分类方法和支持向量机分类方法。

作为数据挖掘领域的重要方法和实现数据分组的重要手段，隐私保护聚类算法研究引起了巨大的研究热情。为了避免数据扰动技术过多影响聚类算法中的距离指标，造成聚类结果失真严重的问题，文献[8]提出了一种几何数据转换方法，该方法提供了隐私保护的聚类分析方法，但是并没有进行十分严格的数学证明。文献[9]对 *k* 邻域替换的方法实现数据隐藏，但只对数据聚类可用性进行了分析，没有给出隐私性证明。文献[10]给出了一种差分隐私保护的 *k*-means 聚类方法，但是该方法评价的是加噪扰动后数据聚类的准确性，旨在对数据是否正确划分到某个聚类进行评估，并没有从数据发布的角度对总体信息损失做出量化表示。

2.2 聚类在隐私保护中的应用

从挖掘方法角度分析，聚类算法依据数据记录的簇内相似性和簇间相异性对数据集进行聚类划分，力求使簇内数据相似性更大，簇间数据相异性更大。隐私保护聚类算法的动机在于保护个体敏感信息的同时，不丧失聚类的准确性。如药品制造公司希望通过对用户的购买行为数据进行聚类分析，其动机是获得准确的聚类划分结果以辅助产品定位或服务优化，同时又要保障客户的个人隐私信息，即不能泄露某个特定客户曾买过治疗艾滋病的药物等。文献[11~13]对划分共享计算场景的聚类任务设计了隐私保护的 DBSCAN 聚类算法，该类算法通过加密协议构造处理垂直和水平划分形式下数据集的隐私保护问题。然而，以上算法仅对特定场景的隐私保护问题加以解决，难以抵御变体模型的攻击。

在数据发布领域，聚类算法可以作为一种分组算法，依据相似度和相异性指标对数据进行分组预处理，从而实现单条数据到组数据的匿名化隐藏。依据文献[3]提出的查询函数敏感性加噪方法，经过分组操作后，函数查询敏感性会由单一个体分化到一组个体，从而实现扰动噪声量到组数据的分化。从组中单一个体数据的角度考虑，添加在其上的噪声量会大大减小，进而信息损失量减少，数据可用性得到提升。

2.3 差分隐私

差分隐私保护[3]是一种加密机制驱动的、具有很强顽健性的隐私保护模型。该模型假定攻击者掌握了任何关于数据的背景知识，并且对隐私保护的程度提供严格的数学证明。

对于能够满足 e -差分隐私保护的随机算法而言，其输出随机变量的概率分布显然是要以数据集的先验特征为前提条件的。从随机变量的概率分布角度进行分析，如果一个随机算法满足 e -差分隐私保护模型的要求，那么其输出结果的概率分布不会因为在数据集中添加或减少一条数据记录而产生很大变化，即某个个体存在与否仅仅会对数据总体的概率分布产生很小的影响，这种影响的程度通过

2016100-2

隐私保护预算[14]进行估计。在差分隐私保护模型中，数据分析者可以获得数据总体的统计属性或模型特征，但对任意特定数据记录的信息则无法获取。因此，这种机制能够对数据集中特定个体的敏感信息进行保护，又不至于引起数据分布的巨大变化。

定义 1　差分隐私模型定义了概率分布层面的隐私量化。假设随机算法 M 满足 e -差分隐私模型的要求，那么当其满足式(1)的概率约束时，随机算法 M 提供 e -差分隐私。

$$\Pr[M(D)=S] \leq \mathrm{e}^{e} \Pr[M(D')=S] \tag{1}$$

其中，e 表示隐私保护预算，D 表示原始数据集，D' 为 D 的邻近数据集，指在 D 中添加或删除任意一条数据记录，S 表示任意输出结果集合。差分隐私保护机制要求将随机算法 M 作用在邻近数据集上后，所得到输出结果相同的概率比值上界为 e^{e}。通过限制 e 的大小，使随机算法作用于邻近数据集上输出相同结果的概率尽量接近。e 越小，隐私性越强，引入的噪声也越大。当 e 为 0 时，输出结果相同的概率也相同。文献[15]提出了针对数值数据进行差分隐私保护的拉普拉斯机制，即通过对数据属性值添加拉普拉斯噪声的形式，实现 e -差分隐私保护。将查询函数表示为 f，为原始数据添加均值为 0，尺度参数为 $\frac{\Delta f}{e}$ 的拉普拉斯噪声以实现数值型数据的 e -差分隐私保护，加噪操作的形式化表示为

$$M(D)=f(D)+\mathrm{Lap}\left(\frac{\Delta f}{e}\right) \tag{2}$$

在上述表示中，Δf 表示查询函数的敏感性，指的是查询函数 f 作用于邻近数据集时产生的最大距离差。

3　差分隐私保护的聚类匿名化数据发布方法

本文提出的聚类匿名化数据发布方法，首先依据准标识符对数据集进行聚类划分，使数据满足 k-anonymity 模型的要求。在匿名化的过程中，所有数据属性都被当作准标识符。然后对匿名分组后的数据添加拉普拉斯噪声，扰动数据记录真实值，从而实现差分隐私保护模型的隐私性要求。相对于常规拉普拉斯机制而言，聚类操作主要用于减小查询函数的敏感性，敏感性减小促使添加噪声量的减小，从而大大增强数据可用性。

3.1　基于密度的聚类匿名化方法

DBSCAN（density-based spatial clustering of applications with noise）算法是一种基于密度的聚类算法，与划分和层次聚类方法不同，它将簇定义为密度相连的点的最大集合，能够把具有足够高密度的区域划分为簇。相对于其他聚类算法而言，DBSCAN 的适用性更强，能够发现任意形状的聚类，因此在数据分组通用性上，具有很大优势。

本文提出了一种密度聚类机制（DCM, density-based clustering mechanism），利用 DBSCAN 算法对数据集进行聚类操作，按照密度分布将数据划分到不同的等价组，每组数据记录的数量至少为 k。在实现数据匿名化的同时，将查询函数的敏感性分化到每组数据的 k 条记录上，以降低查询函数的敏感性。在聚类划分中，假定所有的属性都是准标识符，那么通过聚类划分得到的数据集就满足 k-anonymity 机制的要求。

算法 1　$DCM(D, r, k)$

输入： D 为原始数据集，r 为非敏感聚类邻域半径，k 为最小聚类尺寸

输出： $\bar{D}$ 为经聚类算法划分之后的数据集

1) 使用 $DBSCAN(D, r, k)$ 对数据集 D 进行聚类，得到聚类结果数据集 D_c；

2) 将 D_c 中无法划分到聚类中的噪音异常点使用总体均值进行替换；

3) 对于划分出的每个簇，使用簇质心替换簇内各条数据；

4) 返回经替换之后的数据集 $\bar{D}$。

定理 1　定义原始数据集为 D，定义查询函数 f_i，返回数据集中的第 i 条记录。那么对于聚类机制与查询函数而言，$\Delta(f_i \cdot DCM) \leq \frac{\Delta(f_i)}{k}$ 成立。

证明　将聚类机制 DCM 作用于数值型数据集得到 $\bar{D}$，最小簇尺寸为 k。显然，当查询函数 f_i 作用于数据集时，由于临近数据集之间的差异被分化到 k 条数据记录上。那么，当 $f_i \cdot DCM$ 操作作用于数据集时，将返回第 i 条数据记录所在簇的质心（对于数值型数据而言，质心用均值来表示），其敏感性至多为 $\frac{\Delta(f_i)}{k}$。对于数据集中未被划分到某个簇的噪音异常点来说，通过整体均值进行替换就意味着平均分化的范围为整个数据集，因此查询函数敏感性会远小于已经划分到聚类中的点。综上，聚类函数作用后，在可用性上实现了 $\Delta(f_i \cdot DCM) \leq \frac{\Delta(f_i)}{k}$。

因此，通过聚类操作对数据集进行分组，实现

了将单条记录匿名隐藏于一组数据当中的目的。同时通过均值替换，实现了查询函数的敏感性分化，从而减小了查询函数的敏感性。

3.2 差分隐私保护数据发布方法

为增强数据隐私性，需要对数据原始信息添加噪声扰动，以达到差分隐私保护模型的要求。本文提出了一种噪声扰动的隐私保护数据发布机制（DCMDP, density-based clustering mechanism with differential privacy），对经密度聚类划分得到的数据添加拉普拉斯噪声，使其满足差分隐私模型的需求。

算法 2 $DCMDP(\bar{D}, e)$

输入： $\bar{D}$ 为聚类划分之后的数据集，e 为隐私保护预算

输出： D_e 为添加噪声后满足 e -差分隐私模型的数据集

1) 查询函数 f_i 作用于数据集 $\bar{D}$，返回 $\bar{D}$ 中第 i 条数据记录 $f_i(\bar{D})$，其中，$i=1,2,\mathrm{L},|D|$；

2) 对查询到的每一条数据记录添加拉普拉斯噪声 $N_e(f_i(\bar{D}))$，并将加噪后的数据记录加入 D_e；

3) 返回满足 e -差分隐私的数据集 D_e。

定理 2 经噪声扰动后，数据集满足 e -差分隐私。

证明 加噪操作的原理是对查询出的每一个数据记录添加拉普拉斯噪声。依据隐私保护模型的并行性规则[16]，对于不相交数据记录施以隐私保护预算为 e 的随机算法，那么整个数据集满足 e -差分隐私。在本文的算法中，针对的是互不相交的簇，因此满足差分隐私保护的并行性规则。

综上所述，本文提出的基于聚类匿名化的差分隐私保护数据发布方法满足 e -差分隐私模型的要求。

4 实验评估

4.1 实验环境及数据集描述

本算法使用基于 Python 语言的 scikit-learn 框架进行实现，实验环境为 Windows 8.1，内存 4 GB。实验中的数据来源于 UCI Knowledge Discovery Archive Database（http://archive.ics.uci.edu/ml/）。

在本文的研究中，对数值型数据进行差分隐私保护的数据发布，实验数据集为 MAGIC Gamma Telescope Data Set，数据集描述如表 1 所示。

表 1 数据集属性信息描述

数据集名称	数据集标识符	属性个数	数据记录数目	数据类型
MAGIC Gamma Telescope Data Set	*DS*	10	1 012	real

4.2 数据可用性度量

4.2.1 k 值选择

聚类函数作用后，在可用性上实现了 $\Delta(f_i \cdot DCM) \le \frac{\Delta(f_i)}{k}$。在实验过程中，从整体数据可用性角度衡量，聚类最小尺寸 k 的选择，需要满足 $\left(\frac{|D|}{k}\right)\left(\frac{\Delta f_i}{k}\right) \le \Delta f_i$，因此，理论上，当 k 值的选择在大于等于 $\sqrt{|D|}$ 时，数据的可用性会高于单独应用拉普拉斯机制。

4.2.2 实验结果度量

对于可用性的度量，本文通过计算信息损失来实现，即计算聚类匿名化及加噪操作后的扰动数据与原始数据之间的平方误差和。这种度量的公式为

$$SSE = \sum_{d_j \in D} \sum_{a_j^i \in d_j} (dist(a_j^i, (a_j^i)'))^2 \tag{3}$$

本文讨论数值型数据的隐私保护问题，因此使用欧几里得距离作为距离的衡量尺度。SSE 衡量的是信息损失量，因此 SSE 越大，扰动操作后的数据可用性越差。设置差分隐私保护预算为最常用的数值，即 e =0.01、0.1、1 和 10，加噪操作后通过式（4）对信息损失量进行相对评分。

$$SCORE = \frac{SSE_{\mathrm{DCMDP}}}{SSE_{\mathrm{DP}}} \tag{4}$$

其中，SSE_{DCMDP} 表示本文算法下的信息损失量，SSE_{DP} 表示直接使用拉普拉斯机制下的信息损失量。评分的过程是对信息损失量数据进行归一化的过程，SSE_{DCMDP} 相对于 SSE_{DP} 越小，表示发布数据可用性的提升越可观。

4.2.3 实验结果讨论

以 e =0.01 时直接添加拉普拉斯噪声引发的信息损失量为基准，不断调整聚类尺寸 k 的大小，以进行数据可用性对比实验，实验中可用性相对评分结果如图 1 所示。

在图 1 中，水平线表示不断变化隐私保护预算时，直接添加拉普拉斯噪声时的数据可用性相对评分，e =0.01 时直接添加拉普拉斯噪声生成基准可用性相对评分为 1。因为其评分并不会随 k 值变化，因此表现为水平。应用本文提出的 DCMDP 机制作用数据集，随着 k 值增大，可用性在逐渐的提高。总体而言，在 $k=\sqrt{1\,012}\approx 32$ 附

近，使用 DCMDP 机制发布的数据可用性超过原始拉普拉斯机制。

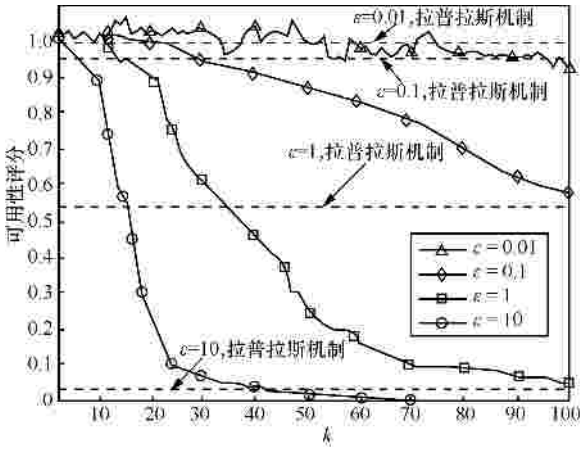

图 1　不同隐私保护预算下变换 k 值时拉普拉斯机制与 DCMDP 方法可用性评分对比

5　结束语

本文提出了一种基于密度聚类的数据匿名化机制，以提高数据可用性。同时，结合拉普拉斯机制对隐私保护程度进行提升，使该方法满足 e -差分隐私保护模型的要求。对于隐私保护水平，本文给出了较为详细的数学证明过程。实验结果显示，所提算法作用后的数据可用性得到了很大提升。但是聚类过程本身也会造成信息量损失，因此，当 k 值较小时，聚类造成的数据损失得不到补偿，也会使数据的可用性低于直接添加拉普拉斯噪声。所以，实验中，k 值的选择十分关键。

本文给出了一种高可用性的差分隐私保护数据发布方法，然而考虑的场景仅仅是数值型数据，并没有对类别型数据和混合型数据的隐私保护数据发布进行研究。类别型数据和混合型数据的差分隐私保护需要考虑随机概率选举及加噪操作混合的问题，处理上较为复杂，本文中未有涉及。因此，后续的工作将是设计更为通用的聚类匿名化差分隐私保护数据发布方法，以满足不同类型数据的差分隐私保护需求。

参考文献：

[1] ADAM N R, WORTHMANN J C. Security-control methods for statistical databases: a comparative study [J]. ACM Computing Surveys (CSUR), 1989, 21(4): 515-556.

[2] LATANYA S. *k*-anonymity: a model for protecting privacy [J]. International Journal on Uncertainty Fuzziness Knowledge-based Systems, 2002, 10(5): 557-570.

[3] DWORK C. Differential privacy[M]. Encyclopedia of Cryptography and Security. Springer US, 2011: 338-340.

[4] VERYKIOS V S, BERTINO E, FOVINO I N, et al. State-of-the-art in privacy preserving data mining[J]. ACM Sigmod Record, 2004, 33(1): 50-57.

[5] DWORK C, NISSIM K. Privacy-preserving datamining on vertically partitioned databases[C]//Advances in Cryptology–CRYPTO 2004. California, USA: Springer Berlin Heidelberg, c2004: 528-544.

[6] FRIEDMAN A, SCHUSTER A. Data mining with differential ivacy [C]//The 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington DC, USA: ACM, c2010: 493-502.

[7] CHAUDHURI K, MONTELEONI C, SARWATE A D. Differentially private empirical risk minimization[J]. The Journal of Machine Learning Research, 2011, 12: 1069-1109.

[8] OLIVEIRA S R M, ZAIANE O R. Privacy preserving clustering by data transformation [J]. Journal of Information and Data Management, 2010, 1(1): 37.

[9] **崇志宏，倪巍伟，刘腾腾，等. 一种面向聚类的隐私保护数据发布方法**[J]. **计算机研究与发展**, 2010, 47(12): 2083-2089.
CHONG Z H, NI W W, LIU T T, et al. A privacy-preserving data publishing algorithm for clustering application[J]. Journal of Computer Research and Development, 2010, 47(12): 2083-2089.

[10] **李杨，郝志峰，温雯，等. 差分隐私保护 *k*-means 聚类方法研究**[J]. **计算机科学**,2013,40(3): 287-290.
LI Y, HAO Z F, WEN W, et al. Research on differential privacy preserving *k*-means clustering [J]. Computer Science, 2013, 40(3): 287-290.

[11] KUMAR K A, RANGAN C P. Privacy preserving DBSCAN algorithm for clustering[C]//Advanced Data Mining and Applications. Harbin, China, c2007: 57-68.

[12] AMIRBEKYAN A, ESTIVILL-CASTRO V. Privacy preserving DBSCAN for vertically partitioned data[C]//Intelligence and Security Informatics. San Diego, CA, USA: Springer Berlin Heidelberg, c2006: 141-153.

[13] XU W, HUANG L, LUO Y, et al. Protocols for privacy-preserving DBSCAN clustering [J]. Int J Secur, 2007, 1(1): 45-56.

[14] HAEBERLEN A, PIERCE B C, NARAYAN A. Differential privacy under fire[C]//USENIX Security Symposium. San Francisco, CA, USA: c2011.

[15] DWORK C, MCSHERRY F, NISSIM K, et al. Calibrating noise to sensitivity in private data analysis[M]. Theory of Cryptography. Springer Berlin Heidelberg, 2006: 265-284.

[16] MCSHERRY F D. Privacy integrated queries: an extensibl platform for privacy-preserving data analysis[C]//The 2009 ACM SIGMOD International Conference on Management of Data. Providence, Rhode Island: ACM, c2009: 19-30.

作者简介：

刘晓迁（1989-），女，河北保定人，南京理工大学博士生，主要研究方向为数据挖掘、机器学习与隐私保护等。

李千目（1979-），男，江苏南京人，博士，南京理工大学教授、博士生导师，主要研究方向为信息安全、传感网技术、智能决策与数据挖掘等。