差分隐私保护 k-means 聚类方法研究

李 杨[1,2] 郝志峰[2] 温 雯[2] 谢光强[1,2]
(广东工业大学自动化学院 广州 510006)[1] (广东工业大学计算机学院 广州 510006)[2]

摘 要 研究了基于差分隐私保护的 k-means 聚类隐私保护方法。首先介绍了隐私保护数据挖掘和隐私保护聚类分析的研究现状,简单介绍了差分隐私保护的基本原理和方法。为了解决差分隐私 k-means 聚类方法聚类结果可用性差的问题,提出了一个新的 IDP k-means 聚类方法,并证明了其满足 ε-差分隐私保护。最后的仿真实验表明,在相同隐私保护级别下,IDP k-means 聚类方法与差分隐私 k-means 聚类方法相比,聚类可用性得到了较大程度的提高。

关键词 差分隐私,k-均值,聚类,隐私保护
中图法分类号 TP309 **文献标识码** A

Research on Differential Privacy Preserving k-means Clustering

LI Yang[1,2] HAO Zhi-feng[2] WEN Wen[2] XIE Guang-qiang[1,2]
(School of Automation, Guangdong University of Technology, Guangzhou 510006, China)[1]
(School of Computers, Guangdong University of Technology, Guangzhou 510006, China)[2]

Abstract We studied k-means privacy preserving clustering method within the framework of differential privacy. We first introduced the research status of privacy preserve data mining and privacy preserve clustering, briefly presenting the basic principle and method of differential privacy. To improve the poor clustering availability of differential privacy k-means, we presented a new method of IDP k-means clustering and proved it satisfies ε-differential privacy. Our experiments show that at the same level of privacy preserve, IDP k-means clustering gets a much higher clustering availability than differential privacy k-means clustering method.

Keywords Differential privacy, k-means, Clustering, Privacy preserving

1 引言

随着数据库系统的日趋强大和数据存储成本的不断降低,个人数据的收集已经不再仅仅是政府和统计部门的工作,医疗机构、金融部门、搜索引擎、入侵检测系统、社交网络等各类组织都持有大量的个人数据,包括一些个人隐私数据。对这些数据进行数据挖掘能够获取大量极具价值的知识,因此,如何在挖掘这些数据的同时实现个人隐私保护是当前数据挖掘研究的一大热点问题。

数据挖掘中隐私保护的研究目的是,在保护数据中敏感属性的同时,将隐私保护方法对挖掘结果的影响程度控制在一定的范围之内。聚类分析是数据挖掘研究的重要组成部分,也是一些数据挖掘方法的基础。然而,目前数据挖掘中隐私保护方法的研究主要集中在关联规则挖掘和分类挖掘上,基于聚类分析的隐私保护方法研究相对较少。本文提出一种基于差分隐私保护的 IDP k-means 隐私保护聚类方法,差分隐私保护[1-7]是一种基于噪声添加的隐私保护方法,通过添加满足特定分布的随机噪声使数据失真,从而达到隐私保护的目的。差分隐私保护方法定义了一个极其严格的攻击模型,并对隐私泄露风险进行了严谨的数学证明和定量化表示。文章最后的仿真实验对 IDP k-means 隐私保护聚类方法的效果和可行性进行了验证。

2 相关工作

2.1 保护隐私的数据挖掘

按保护隐私所采用的技术划分,数据挖掘中的隐私保护技术可分为噪声干扰、匿名发布和数据加密等几大类。Agrawal 等人首次提出了一种在噪声干扰后的数据上构造分类树的算法,它在最大程度上保持了分类结果[8]。Sweeney 等人提出了对数据进行匿名化处理的 k-匿名算法,它保证任意一条记录与另外的 $k-1$ 条记录不可区分,从而保护了隐私数据[9]。Lindell 等人设计了一个基于数据加密的"多方安全计算协议",实现了从分布式数据源中构造分类树的方法[10]。

按隐私保护技术需对传统数据挖掘方法所做改变的程度由小到大进行排列,数据挖掘中的隐私保护方法大致分为 3 类:1)对挖掘算法透明的隐私保护技术;2)对挖掘算法不完全

到稿日期:2012-05-24 返修日期:2012-09-18 本文受国家自然科学基金(61070033),广东省自然科学基金(9251009001000005),广东省科技计划项目(2010B050400011)资助。

李 杨(1980-),女,博士生,讲师,主要研究方向为隐私保护、机器学习,E-mail:kitty_llyy@163.com;郝志峰(1968-),男,教授,博士生导师,主要研究方向为机器学习、仿生算法、生物信息学;温 雯(1981-),女,博士,副教授,主要研究方向为支持向量机、核方法、模式识别;谢光强(1979-),男,博士生,讲师,主要研究方向为多智能体协调控制、人工智能。

透明的隐私保护技术;3)需重新设计挖掘算法的隐私保护技术[11]。第1)类隐私保护技术通过对原始数据进行修改来保护其中的敏感信息,因此挖掘者不用关心隐私保护的实现过程,该类技术包括数据交换[12]、数据匿名化[13]等方法;第2)类隐私保护技术要求挖掘者根据隐私保护技术的实现原理及关键参数对具体的数据挖掘过程进行一定的修改。该类技术包括数据加扰[14]、随机应答[15]等方法;第3)类方法多用于分布式挖掘中的隐私保护,需要对挖掘中的交互过程设计安全多方计算协议[16]。3类方法各有优缺点,差分隐私保护方法通过数据随机加扰的方式保护隐私,属于上述分类的第2)类。

2.2 聚类分析中的隐私保护

目前聚类分析中常见的隐私保护技术有随机扰动、数据旋转、数据交换等方法。例如,Oliveira 等人提出了一种基于数据旋转变换的 RBT 方法[17],它通过在属性之间进行旋转变换实现信息隐藏,并保证变换前后任意数据记录之间的距离近似不变,从而较好地维持了数据的聚类可用性。其不足之处在于,攻击者一旦掌握了一条记录的属性信息,就能够推算出其他记录的属性信息,从而造成隐私信息的泄露。Mukherjee 等人提出了一种基于傅里叶变换的数据扰动方法[18],其保证变换前后数据的欧氏距离不变,在维持数据基于距离的统计特性不变的同时,保护了数据的隐私信息。其不足之处在于当数据集的分布情况未知或分布不均匀时,算法的距离差阈值难以设定。Parameswaran 等人提出了局部数据交换的 NeNDS 算法[19],它通过对各维数据进行分组,再进行组内数据交换,来达到隐私保护的目的,但是仅适用于小数据集聚类的隐私保护。崇志宏等人在 NeNDS 算法的基础上,提出了 NeSDO 算法[20],其在维持聚类可用性方面做出了改进。本文提出了基于差分隐私保护的 IDP k-means 隐私保护聚类方法,其较好地解决了数据的隐私度量和可用性之间的平衡问题,在严格的隐私披露风险度量下,实现了在加入少量噪声的同时达到较高隐私保护标准的目的。

3 差分隐私保护方法

每种隐私保护方法都假设一种攻击模型(Attacking Model)。某些隐私保护方法假设攻击者对数据集中的敏感属性一无所知,一旦攻击者具有某些相关的背景知识,就会面临隐私泄露的风险。差分隐私保护方法[1-7]定义了一个极为严格的攻击模型,攻击者即使已知除一条记录之外所有记录的敏感属性,仍然无法推断出这条记录的任何敏感信息,所以,其面临的隐私披露风险极小。

差分隐私保护是基于数据失真的隐私保护技术,即通过随机噪声的添加使敏感数据失真,同时保持某些数据或数据属性不变,处理后的数据仍然保持某些统计特性,以便进行数据挖掘等操作。在差分隐私保护方法下,在数据集中添加或删除一条记录不会影响查询输出结果,因此,在上述攻击模型中,攻击者无法通过任何已知记录的敏感属性推断出其他未知记录的敏感属性,下面介绍差分隐私保护的基本原理。

定理 1 对于两个相差至多一个记录的数据集 D_1 和 D_2,$Range(K)$为一个随机函数 K 的取值范围,$\Pr[E_s]$为事件 E_s 的披露风险,若随机函数 K 提供 ε-差分隐私保护,则对于所有 $S\subseteq Range(K)$,

$$\Pr[K(D_1)\in S]\leqslant \exp(\varepsilon)\times \Pr[K(D_2)\in S] \quad (1)$$

其中披露风险取决于随机化函数 K 的值,随机函数 K 的选择与攻击者所具有的知识无关。

图1描绘了两个差别至多为一个记录的数据集 D_1 和 D_2 满足 ε-差分隐私的披露风险曲线。

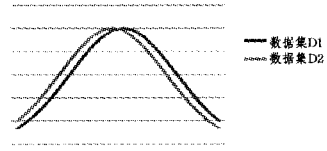

图1 数据集 D_1,D_2 的 ε-差分隐私披露风险曲线

定义 1 对于函数 $f:D\to R^k$,f 的敏感度定义为:

$$\Delta f=\max_{D_1,D_2}\| f(D_1)-f(D_2)\|_1 \quad (2)$$

其中数据集 D_1 和 D_2 相差至多一个记录。敏感度只是函数 f 的性质之一,与数据集 X 无关。

定理 2 设存在一查询函数 f、数据集 X,查询结果为 $f(X)$,通过在 $f(X)$ 上加入合适选择的随机噪声来保护隐私。函数 K 的响应值为:

$$f(X)+(Lap(\Delta f/\varepsilon))^k$$

满足 ε-差分隐私保护。

设噪声函数 $Lap(b)=\exp(-|x|/b)$ 呈标准差为 $\sqrt{2}b$ 的对称指数分布,其中,$b=\Delta f/\varepsilon$。则概率密度函数为 $p(x)=\exp(-|x|/b)/2b$,累积分布函数为 $D(x)=(1/2)(1+\mathrm{sgn}(x)(1-\exp(|x|/b)))$。拉普拉斯噪声的概率密度函数如图2所示。

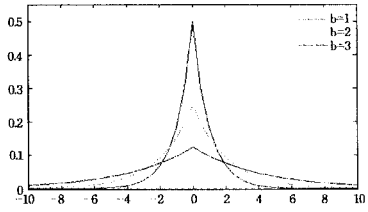

图2 拉普拉斯噪声的概率密度函数

4 差分隐私 k-means 聚类算法

差分隐私保护方法提出以来,引起了隐私保护领域研究人员的极大兴趣,决策树分类、逻辑回归、支持向量机、关联规则挖掘等数据挖掘方法的差分隐私保护成果相继涌现。聚类分析是一种重要的数据挖掘方法,差分隐私聚类方法的研究都停留在理论研究的初级阶段[1,6],对差分隐私聚类方法进行深入研究、对算法效率效果加以改进,并以仿真实验或实际应用进行验证的研究成果鲜有报道。

4.1 差分隐私 k-means 聚类方法

在 k-means 聚类算法中,计算离每个样本点最近的中心点会泄漏隐私。通过对 k-means 进行分析发现,计算中心点时需要用集合内所有点之和除以数目,因此,只要有发布点之和与数目的近似值就不会泄露隐私。

获取 ε-差分隐私保护的 k-means 算法的主要步骤如下:

首先输入 d 维空间 $[0,1]^d$ 的 n 个点 $p_1,\cdots,p_n$,从中随机

选择 k 个点 $u_1,\cdots,u_k$，返回 d 维空间$[0,1]^d$ 内 k 个添加了噪声的新点 $u_1',\cdots,u_k'$作为初始中心点，并按以下步骤更新：

1) 将每个样本点 p_i 划分到最近的中心点 u_j'，将样本集合$\{p_i\}$划分成 $S_1,\cdots,S_k$ k 个集合。

2)对于 $1\leqslant j\leqslant k$，计算集合 S_j 内点之和 $sum=\sum_{i\in S_j}p_i$ 和数目 $num=|S_j|$，分别添加噪声得到 sum' 和 num'，更新 $u_j''=sum'/num'$为集合 S_j 新的中心点。

上述过程不断迭代直至点到集合的划分不再变化或迭代次数达到上限。所添加的噪声函数为 $Lap(b)=\exp(-|x|/b)$，其中 $b=\Delta f/\varepsilon$。

4.2 IDP k-means 聚类方法

从大量的仿真实验结果来看，对于 4.1 节所述算法，当 ε 降低到一定值后，聚类的准确性急剧下降，导致误差过大的重要原因之一在于初始中心点的选择，在随机选择中心点以后，添加了噪声的新中心点往往远远偏离了原中心点，因为对于原始中心点来讲，所添加的随机噪声会使新的中心点偏离很远，所以考虑将初始中心点的选取方法加以改进。

改进方法将 $p_1,\cdots,p_n$ 平均分成 k 个子集，按 4.1 节所述算法的第 2)步分别计算 k 个子集加扰后的中心点，将其作为初始中心点。实验证明改进方法在维持噪声添加量和隐私保护级别 ε 不变的前提下大大提高了聚类的准确度。

设 D_1 和 D_2 代表相互之间仅相差一条记录的两个数据集，计算 k 个中心点的过程等价于对空间$[0,1]^d$ 进行划分的直方图查询，分母 num 的 Δf 为 1。对于分子 sum，由于 $p_i\in[0,1]^d$，因此分子的 Δf 最大为 d。在 d 维空间$[0,1]^d$ 的点集中添加或删除一个点，对于每一维的和，敏感度为 1。因此整个查询序列的敏感度为 $d+1$。设 $Par(D_1)$和 $Par(D_2)$分别对应 D_1 和 D_2 添加噪声之后的聚类结果，$Partition$ 表示任意一种聚类划分，根据定理 2 可知，$\frac{\Pr[Par(D_1)=Partition]}{\Pr[Par(D_2)=Partition]}\leqslant e^{\varepsilon}$ 证明了 IDP k-means 算法满足 ε-差分隐私。

根据上述分析，假设迭代次数固定为 N，可以通过添加分布 $Lap((d+1)N/\varepsilon)$来获取 ε-差分隐私保护。如果迭代次数未知，可以在计算过程中逐步调高参数，例如，第一次迭代参数为$(d+1)/(\varepsilon/2)$，下次迭代参数为$(d+1)/(\varepsilon/4)$，每次将消耗剩余“隐私预算”的一半。

5 实验分析

文中涉及的算法均基于 Weka 开源机器学习框架[21]，使用 JAVA 语言实现，编程环境为 MyEclipse6.0，实验环境为 Windows XP 3.6GHz，2.00GB，算法所涉及的数据均来自 UCI Knowledge Discovery Archive database (http://archive.ics.uci.edu/ml/datasets.html)，数据基本信息如表 1 所列。

表 1 实验数据信息

数据集名称	属性数	属性数据类型	记录数	数据集别名
Blood Transfusion Service Center Data Set	5	real	748	DS1
MAGIC gamma telescope data 2004	11	real	19020	DS2

5.1 实验评价指标

聚类可用性采用 F-measure 评价指标[22]来衡量。对数据集用 k-means 算法和差分隐私 k-means 算法分别进行聚类，使用这两个聚类结果计算 F-measure，F-measure 的结果越大表示两个聚类结果相似度越大，即差分隐私保护算法所添加的噪声对聚类可用性影响越小。当两个聚类结果相同时，F-measure 的结果取最大值 1。

下面简单介绍 F-measure 的计算方法。

用 C 表示对数据集进行 k-means 聚类得到的结果，C_p 表示差分隐私 k-means 算法的聚类结果。其中 X_i 表示 C 中的任意聚簇，Y_j 为 C_p 中的任意聚簇，$n_{ij}=|X_i\cap Y_j|$，$|T|$为数据集样本数，按下列公式计算 $F(C_p)$即得到 F-measure 的结果。

$$Recall(X_i,Y_j)=\frac{n_{ij}}{|X_i|}$$

$$Precision(X_i,Y_j)=\frac{n_{ij}}{|Y_j|}$$

$$F(X_i,Y_j)=\frac{2\times Recall(X_i,Y_j)\times Precision(X_i,Y_j)}{Recall(X_i,Y_j)+Precision(X_i,Y_j)}$$

$$F(C_p)=\sum_{X_i\in C}\frac{|X_i|}{|T|}\max_{Y_j\in C_p}\{F(X_i,Y_j)\}$$

隐私泄露风险评价用 ε 指示，在满足 ε-差分隐私保护条件下，ε 越小，加入的噪声越多，隐私保护的级别越高，隐私泄露风险越低。因此可以通过 ε 值来划分隐私保护的等级。

5.2 实验结果

首先对 DS1 和 DS2 进行预处理，将每个属性的取值归一化到$[0,1]$范围，对两个数据集分别进行 ε-差分隐私 k-means 聚类和 IDP k-means 聚类，逐步将 ε 的值从 6 调低到 0.05，观察 F-measure 的结果随 ε 的变化情况。对应每个 ε 值，对 DS1 和 DS2 分别调用两个算法聚类 10 次，取 F-measure 结果的平均值，在数据集 DS1 和 DS2 上运行的 F-measure 结果分别如图 3、图 4 所示。

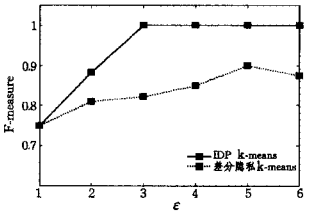

图 3 DS1 上运行的 F-measure 结果比较

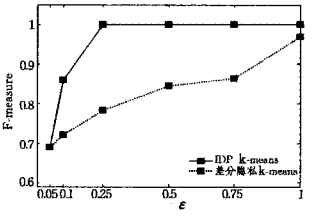

图 4 DS2 上运行的 F-measure 结果比较

对比图 3、图 4 发现，对于数据集 DS1 和 DS2 来说，在相同的 ε 下，本文提出的 IDP k-means 算法与差分隐私 k-means 聚类算法相比，F-measure 的结果值得到了较大程度的提高，这说明我们的算法在相同隐私保护级别下获得的聚类可用性更高。另一方面，实验结果也显示，对于 DS2 这种大数据集，可以在较高的隐私保护级别下，取得较好的聚类可用性，这是因为算法具有加入的噪音量与数据集大小无关的性质，所以对于小数据集，在同等的隐私保护级别下，聚类可用性低于大数据集。

结束语 ε-差分隐私保护下的 IDP k-means 算法添加的噪声与数据集大小无关，这意味着对于相等的隐私保护级别 ε 而言，数据集越大，隐私保护下的聚类可用性越高。与现有

的隐私保护 k-means 聚类方法相比，IDP k-means 聚类方法在聚类可用性和隐私保护级别两方面，取得了更好的平衡。下一步将深入研究在相同的隐私保护级别下，如何进一步提高聚类的可用性，特别是如何提高小数据集差分隐私保护下的聚类可用性，并将继续研究其它聚类算法的差分隐私保护方法。

参 考 文 献

[1] Blum A, Dwork C, McSherry F, et al. Practical Privacy: The SuLQ Framework[C]//24th ACM SIGMOD International Conference on Management of Data / Principles of Database Systems, Baltimore (PODS 2005). Baltimore, Maryland, USA, June 2005

[2] Dwork C. Differential Privacy[C]//33rd International Colloquium on Automata, Languages and Programming, part II (ICALP 2006). Venice, Italy, Springer Verlag, July 2006

[3] Dwork C. Differential Privacy: A Survey of Results[C]//Theory and Applications of Models of Computation (TAMC2008). Xi'an, China, Springer Verlag, April 2008

[4] Dwork C. The Differential Privacy Frontier[C]// 6th Theory of Cryptography Conference (TCC 2009). San Francisco, CA, Springer Verlag, March 2009

[5] Dwork C. Differential Privacy in New Settings[C]//Symposium on Discrete Algorithms (SODA), Society for Industrial and Applied Mathematics. Austin, TX, January 2010

[6] Dwork C. A Firm Foundation for Private Data Analysis [J]. Communications of the ACM, 2011, 54(1): 86-95

[7] Dwork C. The Promise of Differential Privacy. A Tutorial on Algorithmic Techniques[C]//52nd Annual IEEE Symposium on Foundations of Computer Science. Palm Springs, CA, October 2011

[8] Agrawal R, Strikant R. Privacy-preserving data mining [C]// Proceedings of the 2000 ACM SIGMOD International Conference on Managementof Data. Dallas, Texas, May 2000: 439-450

[9] Sweeney L. K-anonymity: A Model for Protecting Privacy[J]. International Journal on Uncertainty[J]. Fuzziness and Knowledge-based Systems, 2002, 10(5): 557-570

[10] Lindell Y, Pinkas B. Privacy preserving data mining[C]// Proceedings of the 20th Annual International Cryptology Conference on Advances in Cryptology. Santa Barbara, California, August 2000: 36-54

[11] 杨维嘉. 在数据挖掘中保护保护隐私信息的研究[D]. 上海：上海交通大学，2009：13-29

[12] Fienberg S E, Mclntyre J. Data swapping: Variations on a theme by Dalenius and Reiss[C]// Proceedings of the Privacy in Statistical Databases(PSD). Barcelona, Spain, 2004: 14-29

[13] Kifer D, Gehrke J. Injecting utility into anonymized data-sets[C]// Proceedings of the ACM SIGMOD Conference on Management of Data (SIGMOD). Atlanta, Georgia, USA, 2006: 217-228

[14] Agrawal R, Srikant R. Privacy preserving data mining[C]//Proceedings of the ACM SIGMOD Conference on Management of Data(SIGMOD). Dalias, Texas, 2000: 439-450

[15] Du W, Zhan Z. Using randomized response techniques for privacy-preserving data mining[C] // Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington D C, USA, August 2003: 505-510

[16] Clifton C, Kantarcioglou M, Lin X, et al. Tools for privacy preserving distributed data mining[J]. ACM SIGKDD Explorations, 2002, 4(2): 28-34

[17] Oliveira S R M, Zaiane O R. Achieving privacy preservation when sharing data for clustering[C]// Secure Data Management Proceedings. Toronto, Canada, Berlin: Springer, 2004: 67-82

[18] Mukherjee S, Chen Zhi-yuan, Gangopadhyay A. A privacy preserving technique for Euclidean distance-based mining algorithms using Fourier-related transforms [J]. The International Journal on Very Large Data Bases, 2006, 15(4): 293-315

[19] Parameswaran R, Blough D M. Privacy preserving data obfuscation for inherently clustered data[J]. International Journal of Information and Computer Security, 2008, 2(1): 1744-1765

[20] 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

[21] Witten I H, Frank E. Data Mining: Practical Machine Learning Tools and Techniques[M]. Morgan Kaufmann, 2005

[22] van Rijsbergen C J. Information Retrieval (2nd edition)[M]. London: Butterworths, 1979

(上接第 270 页)

[15] Liu Man-hua, Zhang Dao-qiang, Shen Ding-gang. Ensemble sparse classification of Alzheimer's disease[J]. Neurolmage, 2012, 2(60): 1106-1116

[16] 王晓丹，高晓峰，姚旭，等. SVM 集成研究与应用[J]. 空军工程大学学报：自然科学版，2012，13(2)：84-89

[17] Alonso-Gonzalez C J, Moro-Sancho Q I, Ramos-Munoz I, et al. Rotation Forest on Microarray Domain: PCA versus ICA[C]// IEA/AIE 2010, Part II. LNAI 6097, 2010: 96-105

[18] Maudes J, Rodrlguez J J, Garcla-Osorio C, et al. Random Projections for SVM Ensembles[C]//Garcla-Pedrajas N, et al., eds. IEA/AIE 2010, Part II. LNAI 6097, 2010: 87-95

[19] Kuncheva L I, Rodrlguez J J. An Experimental Study on Rotation Forest Ensembles[C]// Haindl M, Kittler J, Roli F, eds. MCS 2007. LNCS 4472, 2007: 459-468

[20] Zhang Chun-xia, Zhang Jiang-she. RotBoost: A technique for combining Rotation Forest and AdaBoost [J]. Pattern Recognition Letters, 2008, 29: 1524-1536

[21] Yang Fei-hu, Cheng Wei-qing, Dou Ren-fu, et al. An Improved Feature Selection Approach Based on ReliefF and Mutual Information[C] // International Conference on Information Science and Technology. Nanjing, China, 2011: 246-250

[22] Valentini G, Dietterich T G. Bias-variance Analysis of Support Vector Machines for the Development of SVM-Based Ensemble Methods [J]. Journal of Machine Learning Research, 2004, 5: 725-775

[23] 王晓丹，孙东延，郑春颖. 一种基于 AdaBoost 的 SVM 分类器[J]. 空军工程大学学报：自然科学版，2006，7(6)：54-57

[24] Demsar J. Statistical Comparisons of Classifiers over Multiple Data Sets[C]//J. Machine Learning Research. 2006, 7: 1-30