软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2012,23(12):3198−3208 [doi: 10.3724/SP.J.1001.2012.04286]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

一种向量等价置换隐私保护数据干扰方法*

倪巍伟+, 张 勇, 黄茂峰, 崇志宏, 贺玉芝

(东南大学 计算机科学与工程学院,江苏 南京 210096)

Vector Equivalent Replacing Based Privacy-Preserving Perturbing Method

NI Wei-Wei+, ZHANG Yong, HUANG Mao-Feng, CHONG Zhi-Hong, HE Yu-Zhi

(School of Computer Science and Engineering, Southeast University, Nanjing 210096, China)

+ Corresponding author: E-mail: wni.seu@gmail.com, http://cse.seu.edu.cn/PersonalPage/wni/index.htm

Ni WW, Zhang Y, Huang MF, Chong ZH, He YZ. Vector equivalent replacing based privacy-preserving perturbing method. *Journal of Software*, 2012,23(12):3198−3208 (in Chinese). http://www.jos.org.cn/1000-9825/4286.htm

Abstract: Privacy-Preserving data publishing has attracted considerable research interest over the past few years. The principle difference of clustering and obfuscating burdens the trade-off between clustering utility maintaining and privacy protection. Most of existing methods such as adopting strategies of distance-preservation, or distribution-preservation, cannot accommodate both clustering utility and privacy security of the data. As a trade-off, a neighborhood-preservation based perturbing algorithm VecREP (vector equivalent replacing based perturbing method) is proposed, which realizes good clustering utility by maintaining the nearest neighborhood for each data point. The definition of a safe neighborhood is introduced to stabilize the composition of the nearest neighborhood. The equivalent replacing arc is generated to realize distribution stability of nearest neighborhood leveraging vector offset and composition. For each data point, VecREP randomly chooses a point on its equivalent replacing arc inside corresponding safe neighborhood to make substitution. The algorithm is compared with existing methods such as RBT, TDR, Camp-crest and NeNDS. Experimental results demonstrate that VecREP competes in performance with RBT on maintaining clustering quality and, outperforms the other. It can avoid a reversible attack effectively and compared to the existing solution, ARMM has a shorter handover delay and a smaller location update and delivery cost.

Key words: privacy-preserving data publishing; clustering; safe neighborhood; equivalent replacing arc; *k* nearest neighborhood

摘 要: 近年来,隐私保护数据发布得到了研究者的广泛关注,聚类与隐藏原理上的差异使得面向聚类的隐藏成为难点.针对现有保距和保分布隐藏难以有效兼顾数据聚类可用性和隐私安全的不足,提出基于保邻域隐藏的扰动算法 VecREP(vector equivalent replacing based perturbing method),通过分析数据点邻域组成结构,引入能够保持数据邻域组成稳定的安全邻域定义.进一步基于向量偏移与合成思想,提出有效保持邻域数据分布特征的等价置换弧.对任意数据点,采用随机选取位于其安全邻域内等价置换弧上点替换的策略实现隐藏.将算法与已有的 RBT,TDR,

* 基金项目: 国家自然科学基金(61003057)
收稿时间: 2010-07-25; 修改时间: 2011-11-17; 定稿时间: 2012-07-23

Camp-crest 和 NeNDS 算法进行实验比较,结果表明:VecREP 算法具有与保距隐藏算法 RBT 相近的聚类可用性,优于其余算法,能够较好地维持数据聚类的可用性.同时,具有好于其余算法的数据隐私保护安全性.

关键词: 隐私保护数据发布;聚类;安全邻域;等价置换弧;k 邻域

中图法分类号: TP309 **文献标识码:** A

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.虽然数据挖掘已经在一些深层次的研究和应用中取得较大进展,但随着人们对数据中隐私信息日益关注,对其进行挖掘也带来了隐私保护方面更迫切需要解决的问题[1−4].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.聚类是数据挖掘研究的一个重要部分,聚类结果与数据分布及数据密度特征等密切相关,而数据隐藏通过对个体数据的修改实现隐私保护,这种修改很容易引起数据分布及密度特征的改变,从而导致发布后数据聚类可用性差.已有的少数面向聚类隐藏方法[5−7]主要从保距和保分布角度实现隐藏后数据聚类可用性,这些方法存在以下问题:

① 保距隐藏需要建立任意两数据点间关于距离的强约束(距离关系不变或在某一阈值内变化),这种强约束容易引起针对原始数据的逆推猜测,导致隐藏方法的隐私保护安全性较低;

② 保分布隐藏不利于数据个体特征的维护,同时也不适用分布异常的数据集.隐藏后,数据对聚类算法的适应性较弱.

针对以上问题,提出一种介于保距与保分布隐藏之间,以保持数据邻域关系稳定为目标,基于向量等价置换的扰动方法 VecREP(vector equivalent replacing based perturbing method).实现在保护数据隐私的同时,较好地保持原始数据关于聚类的个体和分布特征信息.

1 相关工作

数据集中环境下的隐私保护微数据隐藏技术主要分为两类[3,4]:

(1) 基于数据失真(distorting)技术.通过扰动(perturbation)原始数据实现隐私保护,同时保持某些数据或数据特性(例如某些统计方面性质等)不变.包括随机化、阻塞、交换、合成数据替换等;

(2) 基于限制发布技术.主要采用数据匿名方法,一般包括抑制(suppressing)和泛化(generalization)操作,即不发布该数据项或对数据进行更概括、抽象的描述.

基于限制发布的隐藏弱化个体数据差异,适用于对个体数据差异依赖较弱的数据应用,例如计数查询等.基于数据失真的隐藏有利于对数据统计特征和个体特征的维持.近年来,面向聚类的数据隐藏开始得到研究者的关注[5−10].文献[5]提出一种基于矩阵变换的扰动方法 RBT,通过将数据属性两两分组,为每组生成满足约束的保距矩阵,实现保距(数据集为偶数维)或近似保距隐藏(数据集为奇数维).文献[6]提出 NeNDS 算法,将数据分为若干组,每组至少包含 c(分组参数)个记录,设计基于树遍历的数据交换策略在每一维分组内进行数据交换,实现数据隐私安全和聚类可用性维持.文献[7]提出一种基于 Fourier 变换的扰动方法,确保扰动前后任意两数据记录的距离差在给定范围内,实现数据基于距离分析的可用性.Camp-crest 算法[8]结合欧氏距离与敏感属性泄露概率熵定义数据点间距离,建立数据点的最小生成树,将树中边上所有近邻数据点属性值用其均值替换,实现保护数据隐私与维持数据统计信息.TDR 方法[9]采用对原数据进行预聚类生成类标签,构建满足聚簇结构约束的匿名数据集,实现数据隐私保护和聚类可用性的维持.这些方法多数采用保距或保分布隐藏思想实现聚类可用性.

隐藏操作通过映射函数 $f(t)=t'$(t 为原始记录,t'为隐藏后记录)实现对数据隐私的保护.关于映射策略的约束实例越多,对隐藏后数据进行逆推猜测的可能性越大,数据隐私安全性越低.例如:保距隐藏算法 RBT 隐藏前后任意两记录的泄漏,将导致整个隐藏数据集泄漏[10];从维持隐藏后数据聚类质量角度考虑,保距隐藏维持个体数据间距离不变,有利于数据个体特征的维护,保分布隐藏侧重对聚簇结构概要特征的维护(如图 1 所示).而聚类过程与数据的个体特征和分布特征密切相关,单纯维护数据个体特征或分布特征可能导致隐藏后数据聚类质量的偏差和聚类算法选取通用性的不足.

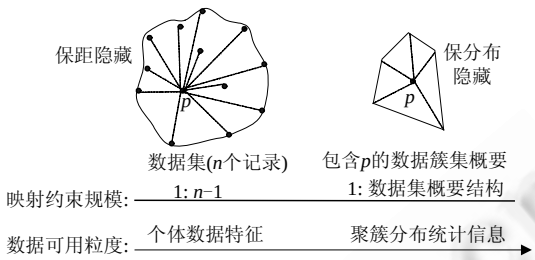


Fig.1 Illustration of distance-preservation and distribution-preservation obfuscations

图 1 保距隐藏与保分布隐藏示意

2 问题描述与分析

2.1 保邻域隐藏

邻域关系是构成聚簇的核心要素,其粒度介于数据点间距离与数据分布之间,k 邻域作为邻域关系的代表结构在聚类研究中得到了广泛应用[11,12].从映射约束规模角度考虑:维持某数据记录的 k 邻域不变,对应映射约束规模为 1:k,即需要维持该数据记录与其 k 邻域内的数据在距离及分布上隐藏前后近似不变.从维持数据聚类的可用性角度看,k 邻域是数据个体特征与数据分布的折衷,属于一种微分布(micro-distribution).

基于以上分析,提出保邻域隐藏思想——通过维持数据集中各数据点 k 邻域稳定,实现保护数据隐私和维持数据聚类质量的目的.保邻域隐藏的映射约束规模以及数据可用粒度均介于保距隐藏和保分布隐藏之间,有益于兼顾两者的优点,较好兼顾数据隐私安全性和聚类可用性.

2.2 问题描述

D 为包含 n 个数据点的 d 维数据集,对 $p,q\in D$,$dist(p,q)$表示两点的欧氏距离.p_k 表示 D 中距 p 第 k 近的数据点,p 的 k 邻域记为 $N_k(p)$:$N_k(p)=\{q|q\in D \text{ and } dist(p,q)\leqslant dist(p,p_k)\}$.需解决的问题是,如何修改 p 以保证其原始数值不泄露,同时尽量维持 $N_k(p)$稳定.$N_k(p)$稳定主要体现在组成稳定与结构稳定两个层面.

2.2.1 邻域组成稳定性

假设 p 隐藏后对应 p',这种修改可能导致 $N_k(p)$发生以下变化:

(1) $q\in N_k(p),q'\notin N_k(p')$.即原属于 $N_k(p)$的数据点隐藏后不再是 p'的 k 邻域点;

(2) $q\notin N_k(p),q'\in N_k(p')$.对应原不属于 $N_k(p)$的数据点隐藏后成为 p'的 k 邻域点.

考虑数据点 p_k 位于 $N_k(p)$的最边缘,其出现变化(1)的概率最大;数据点 p_{k+1} 是不属于 $N_k(p)$且距 p 最近的点,出现变化(2)的概率最大.因此,考虑从抑制 p_k 出现变化(1)和 p_{k+1} 出现变化(2)的角度设计隐藏策略.

2.2.2 邻域结构稳定性

聚类常通过密度来衡量数据点的聚类特性,本节从 $N_k(p)$内数据点表现出的密度趋势角度对 $N_k(p)$的结构进行分析.

定义 1(据点的邻域密度). $p\in D,p$ 的邻域密度定义为 $dens_k(p)$:$dens_k(p)=dist(p,p_k)^{-1}$.

$dist(p,p_k)$对应 p 的 k 邻域半径,半径越大,邻域内数据分布越稀疏,密度越小;反之,密度越大.

定义 2(数据点的聚类表征系数). $p\in D,p$ 的聚类表征系数定义为

$$coef_k(p)=\frac{dens_k(p)\times|N_k(p)|}{\sum_{o\in N_k(p)}dens_k(o)}.$$

基于邻域密度和聚类表征系数定义,将 $N_k(p)$内数据划分为正/负邻域点集.

定义 3(正/负邻域点集). 对 $p\in D,p$ 的正/负邻域点集分别为 $N_k^+(p)$ 和 $N_k^-(p)$:

$$N_k^+(p)=\begin{cases}\{q\in N_k(p)\mid dens_k(p)\leqslant dens_k(q)\},\ coef_k(p)\geqslant 1\\ \{q\in N_k(p)\mid dens_k(p)\geqslant dens_k(q)\},\ coef_k(p)<1\end{cases},$$

$$N_k^-(p)=\{q\in N_k(p)\mid q\notin N_k^+(p)\}.$$

若 $coef_k(p)\geqslant 1$,说明 $N_k(p)$内数据相对 p 附近区域数据表现出高密度特性,$N_k(p)$内邻域密度不小于 p 点邻域密度的数据表现出与 p 一致的密度趋势;反之,$N_k(p)$内邻域密度小于 p 点邻域密度的数据点表现出与 p 一致的密度趋势.对 p 的隐藏应维持 $N_k(p)$正负邻域点集划分及正/负邻域点集内数据关于 p 的分布特性.

3 VecREP:基于向量等价置换的扰动

数据点邻域结构不仅由数据点间距离这一标量值决定,还受数据点空间位置分布影响.考虑将数据集映射到矢量空间,假设原点为 o,D 中数据点 p 对应的矢量表示为 $\overrightarrow{op}$.对数据点 p 的修改对应 $\overrightarrow{op}$ 发生偏移生成矢量 $\overrightarrow{op'}$,偏移向量为 $\overrightarrow{pp'}$,用 $|\overrightarrow{pp'}|$ 表示向量的势,$|\overrightarrow{pp'}|=\sqrt{\sum_{i=1}^{d}(p_i-q_i)^2}$.

3.1 邻域组成分析

对数据点 p 的修改可能造成 $N_k(p)$组成结构的变化,考虑对数据点 p 的修改施加某种限制,使隐藏后原 $N_k(p)$内数据点到 p'的距离仍然小于原数据集中不属于 $N_k(p)$的数据点到 p'的距离.

定理 1. 假设 p 修改为 p',若满足 $|\overrightarrow{pp'}|\leqslant 0.5(dist(p,p_{k+1})-dist(p,p_k))$,则在数据集中其余数据点不修改的情况下,用 p'替换 p 能够保证组成 $N_k(p)$的数据点不变.

证明:假设 $o\in N_k(p)$,有 $\overrightarrow{p'o}=\overrightarrow{po}+\overrightarrow{p'p}$(如图 2 所示),$|\overrightarrow{p'o}|=\sqrt{|\overrightarrow{pp'}|^2+|\overrightarrow{po}|^2-|\overrightarrow{pp'}|\cdot|\overrightarrow{po}|\cdot\cos\alpha}$.当 α=180°时,$|\overrightarrow{p'o}|$取最大值,原 $N_k(p)$中数据点到 p'的最远距离为 $|\overrightarrow{pp_k}|+|\overrightarrow{pp'}|$.

对不属于 $N_k(p)$的数据点类似分析,这些数据点隐藏后到 p'的最近距离为 $|\overrightarrow{pp_{k+1}}|-|\overrightarrow{pp'}|$.若满足 $|\overrightarrow{pp_k}|+|\overrightarrow{pp'}|\lhd|\overrightarrow{pp_{k+1}}|-|\overrightarrow{pp'}|$,隐藏前后 $N_k(p)$的组成一定不改变,推得 $|\overrightarrow{pp'}|\leqslant 0.5(dist(p,p_{k+1})-dist(p,p_k))$. □

定义 4(p 的安全邻域). $p\in D$,r 为正实数,p 的安全邻域(safe neighborhood)为 $SN_k(p)$.

$$SN_k(p)=\begin{cases}\{q\mid dist(p,q)<0.5(\overrightarrow{pp_{k+1}}-\overrightarrow{pp_k}),q\in R_d\},\ r\leqslant 0.5(\overrightarrow{pp_{k+1}}-\overrightarrow{pp_k})\\ \{q\mid dist(p,q)<r,q\in R_d\},\qquad\qquad r>0.5(\overrightarrow{pp_{k+1}}-\overrightarrow{pp_k})\end{cases}.$$

如图 3 所示,p 的安全邻域对应图中阴影区域.由定理 1,用安全邻域内数据点替换 p,能够较好地保持 $N_k(p)$邻域组成.极端情况下,数据点 p_k 与 p_{k+1} 距离可能很小,这时,安全邻域收缩于 p 点,扰动后数据易遭受近邻猜测(proximity attaching)攻击,引入安全邻域半径阈值 r 避免该问题.

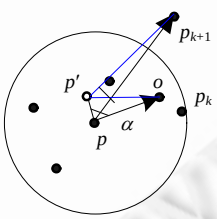


Fig.2 Change of $N_k(p)$

图 2 $N_k(p)$变化示意图

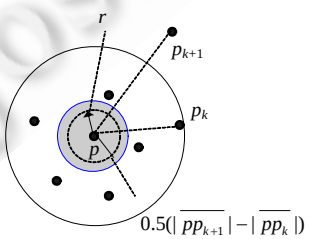


Fig.3 Illustration of the safe neighborhood

图 3 p 的安全邻域

3.2 邻域内部结构

正邻域点集内数据点表现出与 p 一致的密度特征,这些数据的分布特征可以用数据点对应矢量和描述,负邻域点集内数据对应的矢量和描述 $N_k(p)$内与 p 密度特征相反的数据分布特征.

定义 5(邻域正/负向量). $p\in D$,p 的邻域正向量为 $\overrightarrow{pp^+}$,邻域负向量为 $\overrightarrow{pp^-}$:

$$\overrightarrow{pp^+}=\sum_{q\in N_k^+(p)}\overrightarrow{pq},\quad \overrightarrow{pp^-}=\sum_{q\in N_k^-(p)}\overrightarrow{pq}.$$

如图 4 所示,邻域正/负向量表征了邻域正/负点集内数据关于 p 的分布特征:

(1) 向量的方向描述 $N_k(p)$数据分布结构特性;

(2) 向量的势表示沿向量方向分布的这种结构特性的强弱.

对 p 的修改应以尽量维持向量 $\overrightarrow{pp^+}$ 与 $\overrightarrow{pp^-}$ 所成角度以及向量势为目标.

定理 2. 在 D 中其余数据不修改的情况下,若$|\overrightarrow{pp^+}|\geqslant|\overrightarrow{pp^-}|$,在经过 p,p^+和 p^-的圆上,取圆弧 pp^-上的点替换 p;反之,取圆弧 pp^+上的点替换,能够维持 $N_k(p)$的邻域正/负向量角度和标量关系不变.

证明:由几何关系可知多维空间一定存在经过 p,p^+和 p^-的圆,在圆周上同一段弧对应的圆周角相等,因此,选取弧 p^+pp^-上的点 q 替换 p 一定可以保证对 p 进行替换前后 $\overrightarrow{pp^+}$ 与 $\overrightarrow{pp^-}$ 的夹角不变(如图 5 所示).

若$|\overrightarrow{pp^+}|\geqslant|\overrightarrow{pp^-}|$,圆弧 pp^-上的任意点 q 均满足$|\overrightarrow{qp^+}|\geqslant|\overrightarrow{qp^-}|$;

反之,圆弧 pp^+上的任意点 q 均满足$|\overrightarrow{qp^+}|\leqslant|\overrightarrow{qp^-}|$. □

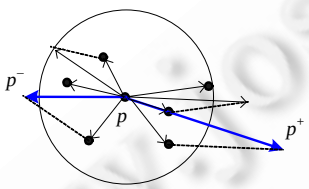


Fig.4 Illustration of positive/negative vectors

图 4 正/负向量示意

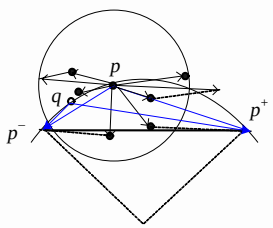


Fig.5 Illustration of the circle

图 5 p,p^+和 p^-生成圆示意图

定义 6(等价置换弧). $p\in D,p$ 的邻域正/负向量分别为 $\overrightarrow{pp^+}$ 和 $\overrightarrow{pp^-}$,经过 p,p^+和 p^-的圆弧 pp^-和 pp^+(p 点除外)构成 p 的等价置换弧.

由定理 2 可知,将 p 与其等价置换弧上的点进行置换,能够较好地维持 p 的 k 邻域内数据分布隐藏前后稳定.

定理 3. d 维空间内经过给定 3 点 p,p^+和 p^-的圆具有唯一的圆心.

证明:圆心到 3 点距离相等且距离平方和最小,由拉格朗日乘数法构建包含 d+2 个变量和 d+2 个等式的方程组,通过高斯消元法可得出方程组的唯一解. □

定理 4. 多维空间经过 3 个点的圆不唯一.

证明:假设四维空间坐标原点为 o,经过 3 点 $p_1(x_1,y_1,z_1,w_1),p_2(x_2,y_2,z_2,w_2)$和 $p_3(x_3,y_3,z_3,w_3)$所在平面的法向量为 $q(a,b,e,g)$,由法向量应垂直于向量 $\overrightarrow{op_1},\overrightarrow{op_2},\overrightarrow{op_3}$ 可以生成包含 3 个等式的四元方程组,方程要么无解或有无限多解. □

3.3 VecREP算法思想

根据定理 1、定理 2 以及 p 的安全邻域和等价置换弧概念,对任意数据点 p,取其等价置换弧上的数据点替换 p,能够有效维持 $N_k(p)$内部结构稳定;取安全邻域 $SN_k(p)$内任意数据点替换 p,能够维持 $N_k(p)$的组成稳定.VecREP 的算法思想是通过维持每个数据点隐藏前后的邻域结构,稳定实现隐藏后数据聚类可用性.具体采用生成每个数据点等价置换弧与安全邻域,在安全邻域内的等价置换弧上随机选取点替换该数据点.算法描述如下:

Input: Original dataset D, parameter k, r;

Output: Perturbed dataset D'.

(1) for each data point p in D

(2) {$NN[p]\leftarrow getNN(p,k)$;

(3) $s\text{-}radius[p]\leftarrow getSRadius(p,k,r)$;} //计算安全邻域半径

(4) $D'\leftarrow D$

(5) for each data point p in D

(6) {compute its $N_k^+(p)$ and $N_k^-(p)$;

(7) generate corresponding p^+, p^-; //生成正负邻域向量}

(8) For each data point p in D

(9) {*generateSArc*(p); //生成等价置换弧

(10) Pick a point t on the swapping arc of p, which satisfies $|\overline{tp}|$<*s-radius*[p];

(11) $p'\leftarrow t$;}

(12) return D'

对给定数据点 p,采用以下策略生成其等价置换弧:

(1) 生成 p 的正/负向量 p^+和 p^-;

(2) 根据定理 3 中拉格朗日乘数法推导求出经过 p,p^+和 p^-这 3 点圆的圆心;

(3) 根据定理 4 的推导随机选取法向量经过 3 点的一个平面;

(4) 在选定平面上,生成经过 3 点的圆,选取等价置换弧.

步骤(2)生成圆心算法描述如下:

Input: p_1, p_2, p_3, d //3 个数据点及维度;

Output: O //经过 3 点圆的圆心.

(1) For *row*=1 to d do { //初始化拉格朗日方程组对应(d+2)×(d+3)系数矩阵 *Matrix*1 的前 d 行

(2) *Matrix*1[*row*][*row*]=1; *Matrix*1[*row*][d+1]=(p_2[*row*]−p_1[*row*])/3;

(3) *Matrix*1[*row*][d+2]=(p_3[*row*]−p_1[*row*])/3; *Matrx*1[*row*][d+3]=(p_1[*row*]+p_2[*row*]+p_3[*row*])/3;}

(4) For *column*=1 to d do { //初始化 *Matrix*1 最后两行

(5) *Matrix*1[d+1][*column*]=2×(p_2[*column*]−p_1[*column*]);

(6) *Matrix*1[d+2][*column*]=2×(p_3[*column*]−p_1[*column*]);

(7) *Matrix*1[d+1][d+3]+=(p_2[*column*])^2−p_1[*column*])^2;

(8) *Matrix*1[d+2][d+3]+=(p_3[*column*])^2−(p_1[*column*])^2;}

(9) For i=1 to d+1 do //将系数矩阵化为上三角矩阵;利用前 d 行对矩阵第 d+1 行降维

(10) For *column*=d+2 to i do

(11) *Matrix*1[d][*column*]−=*Matrix*1[d][i]*Matrix*1[i][*column*];}

(12) For *column*=d+2 to d do

(13) *Matrix*1[d][*column*]/=*Matrix*1[d][d];

(14) For i=1 to d+2 do { //利用前 d+1 行对矩阵第 d+2 行降维

(15) For *column*=d+1 to i do

(16) *Matrix*1[d+1][*column*]−=*Matrix*1[d+1][i]*Matrix*1[i][*column*];

(17) For *column*=d+3 to d+1 do {*Matrix*1[d+1][*column*]/=*Matrix*1[d+1][d+1]}

(18) For *row*=1 to d do //将系数矩阵化为对角线矩阵,利用第 d+1 行对矩阵前 d 行降维

(19) For *column*=d+3 to d+1 do

(20) *Matrix*1[*row*][*column*]−=*Matrix*1[*row*][d+1]*Matrix*1[d+1][*column*];

(21) For *row*=1 to d+1 do //利用第 d+2 行对矩阵前 d+1 行降维

(22) *Matrix*1[*row*][*column*]−=*Matrix*1[*row*][d+2]*Matrix*1[d+2][*column*];

(23) For *row*=1 to d do //圆心 O 的各维坐标分别为矩阵前 d 行的第 d+3 列值

(24) O[*row*]=*Matrix*1[*row*][d+3];

(25) Return O;

假设过 p,p^+和 p^-这 3 点的平面方程为 $t_1x_1+t_2x_2+\ldots+t_dx_d+e=0$,平面的法向量对应($t_1$,$t_2$,…,$t_d$).选取法向量的

d–2 维并赋予随机值(要求 p,p^+和 p^-在剩余两维上不共线),由法向量与平面上非平行两向量内积为零原理得到关于剩余两维的二元一次方程组,求解方程组确定法向量,步骤(3)算法描述如下:

Input: p_1, p_2, p_3, d //3 个数据点及数据集维度;

Output: T, e //平面法向量 T 与参数 e.

(1) do {i=*random*(1,d); j=*random*(1,d) //从 d 维中随机选取两个维度 i,j;满足 i<j

(2) a=($p_3[i]-p_1[i]$)*($p_2[j]-p_1[j]$)−($p_3[j]-p_1[j]$)*($p_2[i]-p_1[i]$);

(3) b=($p_2[i]-p_1[i]$)*($p_3[j]-p_1[j]$)−($p_2[j]-p_1[j]$)*($p_3[i]-p_1[i]$);

(4) Until a!=0 and b!=0 //向量 p_3p_2,p_2p_1 在维度(i,j)上不共线,即矩阵秩 r(*Matrix*)=2

(5) *Matrix*2[1][1]=$p_3[i]-p_1[i]$; //初始化二元一次方程组对应的 2×3 系数矩阵 *Matrix*2

(6) *Matrix*2[1][2]=$p_3[j]-p_1[j]$;

(7) *Matrix*2[2][1]=$p_2[i]-p_1[i]$;

(8) *Matrix*2[2][2]=$p_2[j]-p_1[j]$;

(9) For k=1 to d do {

(10) If k!=i and k!=j {$T[k]$=*random*(−∞,+∞); //对法向量 T 剩余 k–2 维随机取值

(11) *Matrix*2[1][3]+=($p_3[k]-p_1[k]$)*$T[k]$; *Matrix*2[2][3]+=($p_2[k]-p_1[k]$)*$T[k]$;}

(12) ($T[i]$,$T[j]$)=*GetSolution*(*Matrix*2[2][3]) //求解二元一次方程组

(13) For k=1 to d do{e−=$T[k]\times p_1[k]$}

(14) Return T and e;

4 实验分析

这部分对算法隐藏效果进行实验分析,算法采用 VC++6.0 实现,实验环境为 Windows XP 1.8GHz 1.00GB.实验数据来源于 UCI Knowledge Discovery Archive database(http://archive.ics.uci.edu/ml/datasets.html),具体见表 1.采用 F-measure[9]指标衡量隐藏后数据的聚类可用性,F-measure 将同一聚类算法作用于隐藏前后数据集,其值越大,表明隐藏操作对数据集的聚类可用性维持效果越好.

Table 1 Data set information

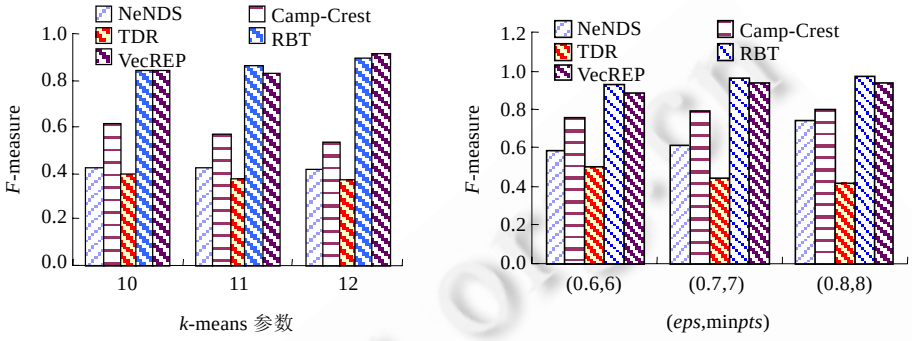
表 1 实验数据信息

数据集名称	别名	属性数目	记录数目	数据类型
Transfusion	DS_1	4	748	Real
Gamma telescope	DS_2	10	1 822	Real
Letter recognition	DS_3	16	4 356	Real
Poker_Hand_Testing	DS_4	5	11 200	Real

分别用 NeNDS,RBT,Camp-crest,TDR 和 VecREP 算法对各个数据集进行隐藏,对隐藏前后数据集分别采用 k-means 和 DBScan 算法[11]聚类,对比所得 F-measure 值,结果如图 6~图 9 所示.图中横坐标对应算法 k-means 和 DBScan 参数取值,其中,k-means 的参数为设定的聚簇数目,DBScan 的参数为邻域半径 *eps* 和核心点阈值 *MinPts*(参数通过采样分析设置).VecREP 算法参数 k 与 r,NeNDS 算法参数 c,Camp-crest 算法参数 b 见图中标注;RBT 算法角度参数采用在给定安全对阈值约束下,随机选取满足条件角度方法实现,TDR 算法中采用 k-means 聚类的类标签数是 2,准标示符匿名阈值取 10.

由图 6~图 9 可以发现:当数据集维度为偶数时,VecREP 算法具有与 RBT 相近的 F-measure 值,能够较好地保持数据聚类可用性;图 9 对应数据维度为奇数情况,此时,RBT 提供近似保距隐藏,VecREP 算法对应的 F-measure 值明显高于 RBT 算法.VecREP 算法在保持聚类可用性方面明显优于 NeNDS,Camp-crest 与 TDR 算法.NeNDS 算法维持各个属性维分组内数据分布,容易割裂属性间关联,缺少对数据集多维属性上分布特征的维持.Camp-crest 算法着眼于数据子集分布的维护,隐藏过程缺少对个体数据聚类特征的有效维持.TDR 算法对聚类可用性的维持依赖于预聚类算法对数据集的适用性,对不同数据集,聚类质量存在较大波动.通过对比各算

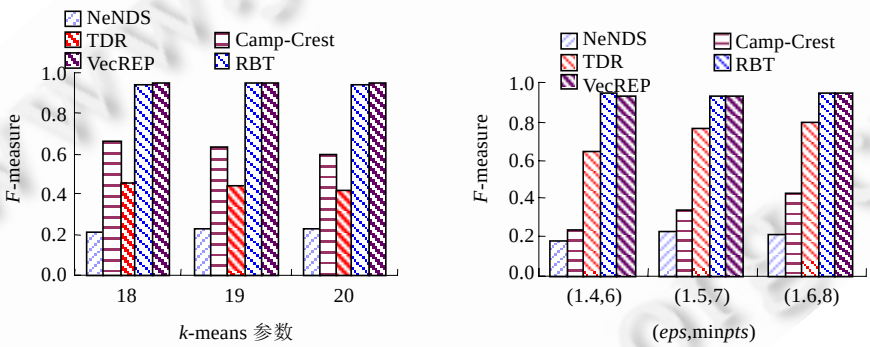
法对不同类型聚类算法的适应性可以发现:不管采用基于划分的聚类算法*k*-means,还是采用基于密度的聚类算法 DBScan,VecREP 算法的 *F*-measure 值均能保持稳定;RBT,Camp-crest,TDR 和 NeNDS 算法均存在显著的波动,验证了 VecREP 算法对不同类型聚类算法具有良好的适应性.



(a) *k*-means 聚类算法 (b) DBScan 聚类算法

Fig.6 *F*-measure evaluation on DS_1 (*k*=7,*r*=0.06,*c*=4,*b*=5)

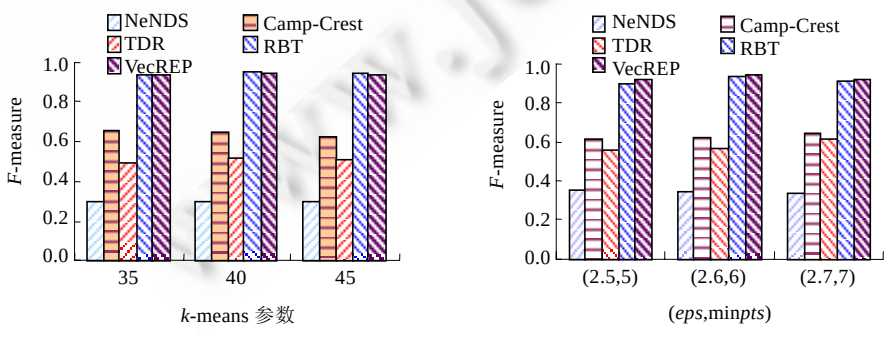
图 6 DS_1 上 *F*-measure 对比(*k*=7,*r*=0.06,*c*=4,*b*=5)



(a) *k*-means 聚类算法 (b) DBScan 聚类算法

Fig.7 *F*-measure evaluation on DS_2 (*k*=5,*r*=0.24,*c*=4,*b*=6)

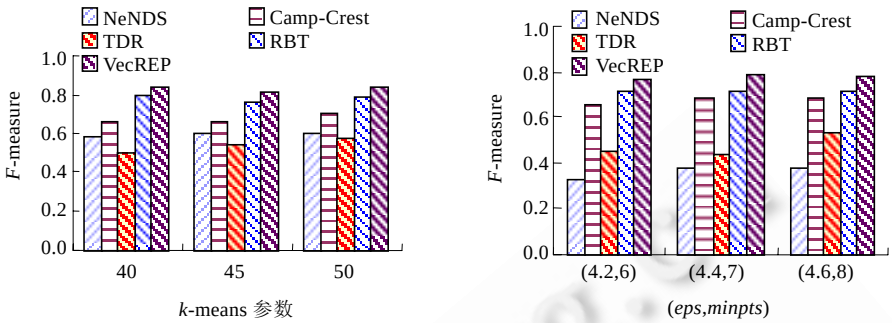
图 7 DS_2 上 *F*-measure 对比(*k*=5,*r*=0.24,*c*=4,*b*=6)



(a) *k*-means 聚类算法 (b) DBScan 聚类算法

Fig.8 *F*-measure evaluation on DS_3 (*k*=9,*r*=1.4,*c*=5,*b*=8)

图 8 DS_3 上 *F*-measure 对比(*k*=9,*r*=1.4,*c*=5,*b*=8)

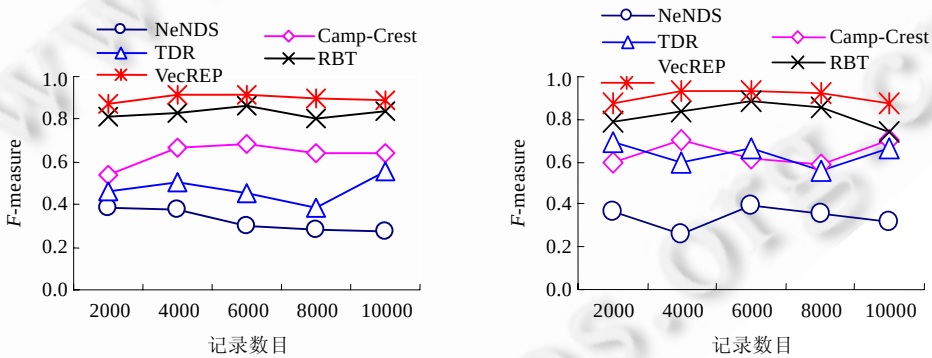


(a) *k*-means 聚类算法 (b) DBScan 聚类算法

Fig.9 *F*-measure evaluation on DS_4 (k=9,r=1.4,c=5,b=10)

图 9 DS_4 上 *F*-measure 对比(k=9,r=1.4,c=5,b=10)

进一步分析算法对数据集规模的可扩展性,采用 Poker_hand_testing 数据,采样生成包含 2 000,4 000,6 000,8 000,10 000 个数据点的数据子集,测试各个算法的隐藏效果.如图 10 所示,随着数据规模的增大,相对于其余算法,VecREP 算法的 *F*-measure 值保持稳定,表现出良好的可扩展性.其原因在于:数据集规模较大时,数据分布变得不规范和复杂,保持数据子集分布稳定变得困难;数据量的增加也使得数据间的距离差异变得细微,使得保距隐藏效果容易受参数选取的影响.VecREP 算法隐藏和维持的对象是邻域集合,其粒度介于距离和数据分布之间,能够较好地平抑数据规模增加对隐藏效果的影响.



(a) *k*-means 聚类算法 (b) DBScan 聚类算法

Fig.10 Scalability performance

图 10 算法扩展性

5 问题讨论

5.1 数据隐私安全性

能否有效防止各种可能的逆推猜测,是隐藏算法隐私安全性的重要表现.逆推猜测的常见模式是假设隐藏算法是公开的,攻击者除了获取隐藏后的所有数据,还有可能获取一些关于原始数据的背景知识(例如,获知少量数据的原始数值等).从对逆推猜测的防御角度分析,RBT 算法的隐私保护安全性最低,任意两条记录隐藏前后数值的泄露,将导致整个隐藏数据集的泄露;NeNDS 算法采用对各维数据进行分组,选择代价最小的交换策略进行置换隐藏,c 个数值有 $c!$种排列方式,猜测出某个分组所选排列的概率为 $1/c!$;Camp-Crest 算法在不改变敏感属性条件下,将 b 条邻近记录的属性值用均值来替换,若 b 个记录中敏感属性值的种类数为 m,则敏感属性泄露的概率为 m/b;TDR 算法的隐私安全与各准标示符的匿名阈值有关,优化往往需要设置较低的阈值,而较低的

阈值容易导致扰动后属性值接近原始数据造成隐私泄露.

VecREP 算法在多维空间选取圆弧上的某一点替换数据集中原始数据点,逆推概率近似为 0;且定理 4 已证明,多维空间存在多个满足条件的同心圆,进一步保证扰动数据的不可逆推.VecREP 算法在隐私保护安全性上优于 RBT,NeNDS,Camp-crest 与 TDR 算法.

5.2 参数的选取

本节对算法 VecREP 中参数的选取进行分析,参数 k 用于定义数据点邻域的范围.取值太小难以体现数据点的局部聚簇分布特征,太大则 k 邻域介于数据点间距离与数据分布之间的微分布优势无法体现.如图 11 所示,可以发现,无论采用 k-means 或 DBScan 算法,k 的取值与 F-measure 值的关系都不是线性的.取得太小或太大均可能造成 F-measure 值的下降.但 k 的变化曲线均存在一段相对稳定的高 F-measure 值区间,例如,对测试数据集 DS_1,k 位于区间[6,9]时,F-measure 值较高且保持稳定;对数据集 DS_2,相应区间为[6,8].参数 r 设置的太大将导致扰动后数据的聚类可用性较低,太小则达不到防止逆推猜测与近似攻击猜测的效果.这两个参数的类型和作用与算法 DBScan 的参数 minPts 和半径参数 eps 相似,类似地,可以采用对原始数据集进行 k-dist 分析设置.

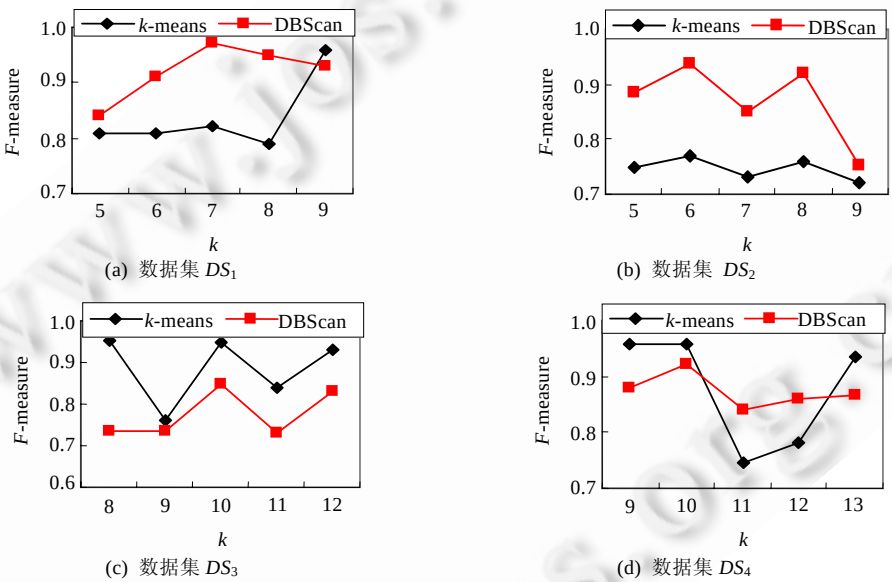


Fig.11 Effect of parameter k of VecREP

图 11 VecREP 中参数 k 的选取

6 论文总结与工作展望

针对面向聚类的隐私保护微数据发布问题,提出基于保邻域隐藏思想的扰动算法 VecREP,引入能保持数据点邻域结构稳定的安全邻域和等价置换弧定义.对任意数据点,采用随机选取位于其安全邻域内等价置换弧上数据点进行替换的策略实现隐藏.算法能有效兼顾隐藏后数据的聚类可用性和隐私安全性.

随着数据维度增加,算法 VecREP 中安全邻域半径可能越来越小,直至趋于 0.这时,r 将难以设定.维数灾难的影响困扰已有的几乎所有隐私保护微数据发布算法,后续将考虑解决扰动算法对维度的扩展性问题.

References:

[1] Kantarcioglu M, Jin JS, Clifton C. When do data mining results violate privacy? In: Kim W, Kohavi R, Gehrke J, DuMouchel W, eds. Proc. of the 10th ACM SIGKDD on Int'l Conf. Knowledge Discovery and Data Mining. New York: ACM Press, 2004. 599–604. [doi: 10.1145/1014052.1014126]

[2] Chen BC, Kifer D, LeFevre K, Machanavajjhala A. Privacy-Preserving data publishing. Foundations and Trends in Databases, 2009,2(1-2):1−167. [doi: 10.1561/1900000008]

[3] Zhou SG, Li F, Tao YF, Xiao XK. Privacy preservation in database applications: A survey. Chinese Journal of Computers, 2009, 32(5):847−858 (in Chinese with English abstract). [doi: 10.3724/SP.J.1016.2009.00847]

[4] Yang XC, Wang YZ, Wang B, Yu G. Privacy preserving approaches for multiple sensitive attributes in data publishing. Chinese Journal of Computers, 2008,31(4):574−587 (in Chinese with English abstract).

[5] Oliveira SRM, Zaïane OR. Achieving privacy preservation when sharing data for clustering. In: Jonker W, Petkovic M, eds. Proc. of the Int'l Workshop on Secure Data Management in a Connected World. Berlin: Springer-Verlag, 2004. 67−82. [doi: 10.1007/978-3-540-30073-1_6]

[6] Parameswaran R, Blough DM. Privacy preserving data obfuscation for inherently clustered data. Int'l Journal of Information and Computer Security, 2008,2(1):1744−1765. [doi: 10.1504/IJICS.2008.016819]

[7] Mukherjee S, Chen ZY, Gangopadhyay A. A privacy-preserving technique for Euclidean distance-based mining algorithms using Fourier- related transforms. The Int'l Journal on Very Large Data Bases, 2006,15(4):293−315. [doi: 10.1007/s00778-006-0010-5]

[8] Li XB, Sarkar S. Data clustering and micro-perturbation for privacy-preserving data sharing and analysis. In: Sabherwal R, Sumner M, eds. Proc. of the Int'l Conf. on Information Systems. Saint Louis: Association for Information Systems, 2010. 58−58. http://aisel.aisnet.org/icis2010_submissions/58

[9] Fung BCM, Wang K, Wang LY, Hung PCK. Privacy-Preserving data publishing for cluster analysis. Data & Knowledge Engineering, 2009,68(6):552−575. [doi: 10.1016/j.datak.2008.12.001]

[10] Ni WW, Chen G, Chong ZH, Wu YJ. Privacy-Preserving data publishing for clustering. Journal of Computer Research and Development, 2012,49(5):1095−1104 (in Chinese with English abstract).

[11] Ester M, Kriegel HP, Sander J, Xu XW. A density based algorithm for discovering clusters in large spatial databases with noise. In: Simoudis E, Han JW, Fayyad UM, eds. Proc. of the 2nd Int'l Conf. on Knowledge Discovery and Data Mining (KDD'96). Menlo: AAAI Press, 1996. 226−231.

[12] Ni WW, Chen G, Wu YJ, Sun ZH. Local density based distributed clustering algorithm. Journal of Software, 2008,19(9): 2339−2348 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/19/2339.htm [doi: 10.3724/SP.J.1001.2008.02339]

附中文参考文献:

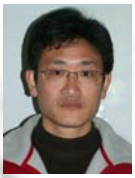
[3] 周水庚,李丰,陶宇飞,肖小奎.面向数据库应用的隐私保护研究综述.计算机学报,2009,32(5):847−858.

[4] 杨晓春,王雅哲,王斌,于戈.数据发布中面向多敏感属性的隐私保护方法.计算机学报,2008,31(4):574−587.

[10] 倪巍伟,陈耿,崇志宏,吴英杰.面向聚类的数据隐藏发布研究.计算机研究与发展,2012,49(5):1095−1104.

[12] 倪巍伟,陈耿,吴英杰,孙志挥.一种基于局部密度的分布式聚类挖掘算法.软件学报,2008,19(9):2339−2348. http://www.jos.org.cn/1000-9825/19/2339.htm [doi: 10.3724/SP.J.1001.2008.02339]

倪巍伟(1979－),男,江苏淮安人,博士,副教授,CCF 会员,主要研究领域为数据挖掘,数据隐私安全保护.



崇志宏(1969－),男,博士,副教授,主要研究领域为数据流,语义数据管理.

张勇(1987－),男,硕士生,主要研究领域为隐私保护数据发布,数据挖掘.

贺玉芝(1987－),女,硕士生,主要研究领域为隐私保护数据发布,数据挖掘.

黄茂峰(1987－),男,硕士生,主要研究领域为隐私保护数据发布,数据挖掘.