一种改进的基于奇异值分解的隐私保持分类挖掘方法

李 光[1],王亚东[2]
(1.长安大学电子与控制工程学院,陕西西安 710064;2.哈尔滨工业大学计算机科学与技术学院,黑龙江哈尔滨 150001)

摘 要: 隐私保护是数据挖掘研究的重要内容之一,目前已经提出了大量隐私保持的数据挖掘算法.基于奇异值分解的方法是其中重要的一种,它是一种基于数据扰动的方法.现有的基于奇异值分解的隐私保持数据挖掘方法对所有样本和属性都进行同样强度的扰动.但不同的样本和属性可能对隐私保护有不同的要求,而且对数据挖掘的重要性也可能不同,因此最好可以对他们进行不同程度的扰动.本文对基于奇异值分解的数据扰动方法进行改进,使之可以对不同的样本和属性进行不同程度的扰动.并在此基础上提出了一种改进的隐私保持分类挖掘方法.实验表明,与原有的基于奇异值分解的方法相比,在保证数据可用性的前提下,本文方法可以对隐私数据提供更好的保护.

关键词: 隐私保持; 数据挖掘; 奇异值分解
中图分类号: TP309 **文献标识码:** A **文章编号:** 0372-2112 (2012)04-0739-06
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2012.04.019

An Improved Privacy-Preserving Classification Mining Method Based on Singular Value Decomposition

LI Guang[1], WANG Ya-dong[2]
(1. *School of Electronic and Control Engineering, Chang'an University, Xi'an, Shaanxi* 710064, *China*;
2. *School of Computer Science and Technology, Harbin Institute of Technology, Harbin, Heilongjiang* 150001, *China*)

Abstract: Privacy protection is indispensable in data mining, and many PPDM (privacy-preserving data mining) methods have been proposed. One such method based on data perturbation is SVD (singular value decomposition)-based method, which treats all samples and attributes equally. However, different samples and attributes may have different requirements for privacy protection, and may be not equally important for data mining. So, it is better to treat them differently. This paper proposed an improved SVD-based perturbation method, which can perturb different samples and attributes to different degrees. In addition, this paper proposed an improved privacy-preserving classification mining method using this improved SVD-based perturbation algorithm. The experiments showed that while maintaining data utility, the proposed privacy-preserving classification mining method can protect privacy better than the original SVD-based method.

Key words: privacy preserving; data mining; singular value decomposition

1 引言

数据挖掘是指从已有的大量数据中发现人们难以察觉却又感兴趣的模式的过程.随着数据挖掘技术的不断发展,隐私保护成为了数据挖掘应用中的一个重要问题[1,2].

为了解决这一问题,隐私保持的数据挖掘被提出[3,4],目前它已经成为数据挖掘研究的重要内容之一.传统的数据挖掘方法不考虑数据隐私保护的问题,假设所有数据都可以直接得到,这个假定在实际中是不现实的,出于隐私保护的需要,某些数据是不能公开的.隐私保持的数据挖掘克服了这一不足,它的首要研究问题是如何在无法得到精确数据的前提下得到高质量的数据挖掘结果.

现有的隐私保持的数据挖掘方法主要分为两类.第一类是基于数据扰动的方法[5~8],这类方法不公开原始数据,公开的是在原始数据上经扰动而得到的新数据,用户通过处理扰动数据来完成对原始数据的挖掘.要求从扰动数据上可以得到原始数据的模式,但得不到隐私数据的值.第二类是基于安全多方计算的方法[9~12],这类方法主要应用于分布式数据库.数据分布式存储在多个节点上,各节点希望在全体数据上进行数据挖掘却又

收稿日期:2010-11-23;修回日期:2011-05-12
基金项目:国家863高技术研究发展计划(No.2007AA02Z329)

不愿向其他节点公开自己的数据.这类方法基于安全多方计算来设计信息交流协议,在不直接分享数据的前提下,交流挖掘算法需要的信息.

在基于数据扰动的方法中,有一类重要的方法是基于矩阵分解的.包括奇异值分解[13,14]以及非负矩阵分解[15].这类方法通过矩阵分解来分析数据,提取并保留对于数据挖掘来讲是重要的信息以保持数据的可用性,去除对数据挖掘不重要的信息来实现数据扰动以保护隐私.

现有的基于矩阵分解的方法对所有的样本和属性都进行同样强度的扰动.但在实际中,不同的样本及属性对于隐私保护的要求是不同的,而且对于数据挖掘的重要性也是不同的,因此应该进行差别化处理,对于不同的样本和属性应该进行不同程度的扰动.

本文分析并改进了基于奇异值分解的隐私保持的数据挖掘方法,使之可以对不同的样本和属性进行程度不同的扰动,并在此基础上结合样本选择和属性选择,提出一种改进的基于奇异值分解的隐私保持分类挖掘方法.实验表明,与原有的基于奇异值分解的方法相比,本文方法可以在保持数据可用性的前提下对隐私数据提供更好的保护.

2 基于奇异值分解的隐私保持的数据挖掘方法

在基于奇异值分解的隐私保持的数据挖掘方法中,数据被排列为矩阵形式.假设原始数据包括 n 个元组,m 个属性,则他们可以组成一个 $n \times m$ 的矩阵 A,A 的行代表元组,列代表属性.矩阵 A 的奇异值分解为

$$A = USV^T$$

其中 U 是一个 $n \times n$ 的正交阵;S 是一个 $n \times m$ 的矩阵,假设 A 的秩为 R,则 S 的左上角 $R \times R$ 子阵是一个对角阵,对角线元素都大于零且按照降序排列,S 其他位置上的元素都是零;V^T 代表矩阵 V 的转置,是一个 $m \times m$ 的正交阵.

考虑到 S 中,非零元素仅出现在左上角 $R \times R$ 子阵的主对角线上,矩阵 A 的奇异值分解可以简写为截尾奇异值分解,公式如下:

$$A = U_R S_R V_R^T$$

其中 U_R 是一个 $n \times R$ 的矩阵,由 U 的前 R 个列组成;S_R 是 S 的左上角 $R \times R$ 子阵;V_R^T 是一个 $R \times m$ 的矩阵,由 V^T 的前 R 个行组成.

基于奇异值分解的隐私保持的数据挖掘方法主要有两种形式:BSVD(basic singular value decomposition)算法以及 SSVD(sparsified singular value decomposition)算法.

BSVD 算法有一个参数 k,$0 \leqslant k \leqslant R$. BSVD 算法的扰动数据为

$$A_k = U_k S_k V_k^T$$

其中 U_k 是一个 $n \times k$ 的矩阵,由 U 的前 k 个列组成;S_k 是 S 的左上角 $k \times k$ 子阵;V_k^T 是一个 $k \times m$ 的矩阵,由 V^T 的前 k 个行组成.在 BSVD 算法中,参数 k 表征了扰动的强度.k 的值越大,扰动的强度越小,扰动数据具有更好的可用性,但安全性下降.

SSVD 算法有两个参数 k 和 d,$0 \leqslant k \leqslant R$,$d \geqslant 0$. SSVD 算法的扰动数据为

$$\overline{A_k} = \overline{U_k} S_k \overline{V_k^T}$$

其中 $\overline{U_k}$和$\overline{V_k^T}$分别是在 U_k 和 V_k^T 的基础上得到的,假设 u_{ij} 和 v_{ij}分别是 U_k 和 V_k^T 的第 i 行第 j 列的元素.如果 $|u_{ij}| < d$,$\overline{U_k}$的第 i 行第 j 列的元素为 0,否则,该元素等于 u_{ij}.同样的,如果 $|v_{ij}| < d$,$\overline{V_k^T}$的第 i 行第 j 列的元素为 0,否则,该元素等于 v_{ij}.

在 SSVD 算法中,参数 k 和 d 共同表征了扰动的强度.k 的值越大,d 的值越小,扰动的强度越小,扰动数据具有更好的可用性,但安全性下降.

容易看出,基于奇异值分解的扰动方法对所有数据进行同等程度的扰动.无论 BSVD 算法还是 SSVD 算法,对所有数据都使用同样的参数值进行扰动.

考虑到某些元组和属性并不是隐私数据,不需要保护,Jie Wang 等人[14]将原始数据矩阵划分为分别由隐私数据以及非隐私数据构成的子阵,仅对由隐私数据构成的子阵进行基于奇异值分解的扰动.在这一方法中,对所有隐私数据仍然都使用同样的参数值进行扰动.仍然没有突破基于奇异值分解的方法的局限.

3 本文方法

本文对基于奇异值分解的隐私保持的数据挖掘方法进行了分析和改造,使之可以对不同样本和不同属性进行不同程度的扰动.

将矩阵 A 的截尾奇异值分解 $A = U_R S_R V_R^T$ 展开.假设 a_{ij},u_{ij}和v_{ij}分别是 A,U_R 和 V_R^T 的第 i 行第 j 列的元素,令 s_i 为 S_R 的第 i 行第 i 列的元素,则有

$$a_{ij} = \sum_{p=1}^{R} u_{ip} s_p v_{pj}$$

假设矩阵 A 经过 BSVD 算法扰动后得到的矩阵为 A_k,令 b_{ij}为 A_k 的第 i 行第 j 列的元素,展开 A_k 的计算式 $A_k = U_k S_k V_k^T$,可得

$$b_{ij} = \sum_{p=1}^{k} u_{ip} s_p v_{pj}$$

由这两个展开式容易看出:

(1)奇异值分解后,原始数据矩阵 A 中的每一个元素被分解为 R 项的和,每项对应着一个奇异值.参数为 k 的 BSVD 算法实质上是对 A 中的每个元素,仅保留奇异值较大的 k 项,而把奇异值较小的 $R-k$ 项舍弃.

(2)矩阵 A 中的每个元素奇异值分解后得到的 R 个项中,每一项均由三个分别来自于 U_R, S_R 和 V_R^T 的元素相乘而得.只要将这三个元素中的任意一项置零便可将此项舍弃.

(3)矩阵 A 的第 i 行,即原始数据中的第 i 个样本,仅与矩阵 U_R 的第 i 行元素有关,与矩阵 U_R 的其他行元素无关.如果将矩阵 U_R 的第 i 行元素中的后 $R-k$ 个置零,其他不变,便可仅对 A 的第 i 行元素进行参数为 k 的 BSVD 扰动.另一方面,矩阵 A 的第 i 列,即原始数据中的第 i 个属性,仅与矩阵 V_R^T 的第 i 列元素有关,与矩阵 V_R^T 的其他列元素无关.如果将矩阵 V_R^T 的第 i 列元素中的后 $R-k$ 个置零,其他不变,便可仅对 A 的第 i 列元素进行参数为 k 的 BSVD 扰动.

根据以上分析结果可以看出,如果要对 A 中第 i 行第 j 列的元素 a_{ij} 使用参数为 k_{ij} 的 BSVD 方法进行扰动,那么只需首先对 A 进行截尾奇异值分解,然后将 a_{ij} 展开式中的后 $R-k_{ij}$ 项舍弃便可以了.这样做非常灵活,对每个样本在每个属性上的值都可以进行程度不同的扰动,但也正因如此,该方法包含了大量未定参数,所有的 k_{ij} 都是待定的.这种方法在某些场合下可能会很有价值,但在一般情况下不方便使用.

本文提出的基于奇异值分解的扰动算法有两个参数 k_1 和 k_2.首先将样本划分为两个样本集 O_1 和 O_2,将属性也划分为两个属性集 P_1 和 P_2.扰动数据可以通过下式计算

$$\overline{A} = \overline{U}S\overline{V}^T$$

其中 $\overline{U}$ 和 $\overline{V}^T$ 是由 U 和 V^T 扰动得来的.如果第 i 个样本在 $O_1(O_2)$ 中,则保持 U 中第 i 行的前 $k_1(k_2)$ 个元素不变,其他元素置零.如果第 i 个属性在 $P_1(P_2)$ 中,则保持 V^T 中第 i 列的前 $k_1(k_2)$ 个元素不变,其他元素置零.

在这种方法中,对 O_1 在 P_1 上的值进行了参数为 k_1 的 BSVD 扰动,对其它元素进行了参数为 k_2 的 BSVD 扰动.

本文基于这种可以对不同的样本和属性进行差异化扰动的方法,结合样本选择和属性选择,提出了一种改进的基于奇异值分解的隐私保持分类挖掘方法.

本文方法使用样本选择来划分样本集,被选中的样本组成 O_1,表示对数据挖掘来讲是重要的样本,其他样本组成 O_2,表示对数据挖掘来讲是不重要的样本.类似的,用属性选择来划分属性集,被选中的属性组成 P_1,表示重要的属性,其他属性组成 P_2,表示不重要的属性.本文方法中 $k_1 > k_2$,对重要样本在重要属性上的值进行强度较小的扰动,对其它数据使用高强度扰动.

本文中,样本选择使用 WCNN[16] 算法(Weighted Condensed Nearest Neighbor).属性选择使用 WEKA[17] 软件(Waikato Environment for Knowledge Analysis,本文使用的是 3.6.0 版)中的属性选择模块 CfsSubsetEval.

WCNN 算法可以由用户来指定选择的样本的个数,因此本文方法有三个参数:k_1, k_2 以及 p(重要样本所占的比例).

实验表明,与原有的基于奇异值分解的方法相比,本文方法可以在保持数据可用性的前提下对隐私数据提供更好的保护.

4　实验

4.1　可用性度量

如果可以从扰动数据上得到高质量的数据挖掘结果,特别是,如果从扰动数据上和原始数据上得到的挖掘结果相近,就称扰动数据具有高可用性.假设在原始数据和扰动数据上训练出的分类器的分类正确率分别为 R_o 和 R_p,则使用 $r = (R_o - R_p)/R_o$ 来衡量扰动数据的可用性.显然,r 越小,数据可用性越好.本文使用最近邻分类器以及 WEKA 中的 J48 决策树来计算可用性度量.最近邻以及决策树都是非常著名和常用的分类方法,WEKA 中的 J48 决策树实质上是用 C4.5 算法进行决策树训练,C4.5 算法是经典的决策树训练算法.假设 r_n 和 r_j 分别表示最近邻分类器和 J48 决策树对应的 r 值,使用 $\max(r) = \max\{r_n, r_j\}$ 来度量可用性.本文假定当 $\max(r) < 0.02$ 时,数据的可用性可以接受.

4.2　隐私性度量

隐私性度量表征隐私数据是否得到了很好的保护.本文使用基于矩阵分解的隐私保持的数据挖掘算法中使用的隐私性度量指标[13~15].该指标一共包含五项,分别为 *VD*, *RP*, *RK*, *CP* 和 *CK*.简单来讲,对隐私数据的保护越好,*VD*, *RP* 和 *CP* 的值越大,而 *RK* 和 *CK* 的值越小.

假设 A 为原始数据矩阵,*MA* 为扰动后的数据矩阵,A 和 *MA* 都是 $n \times m$ 的矩阵,则 *VD*, *RP*, *RK*, *CP* 和 *CK* 的定义如下.

VD 是 *F*-范数下的 *MA* 和 A 的相对误差.

$$VD = \| A - MA \|_F / \| A \|_F$$

其中,对于一个 $n \times m$ 的矩阵 A,假设 a_{ij} 表示矩阵 A 的第 i 行第 j 列的元素,矩阵 A 的 *F*-范数定义如下.

$$\| A \|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m} |a_{ij}|^2}$$

假设在 A 和 *MA* 中,第 i 行第 j 列的元素按升序顺序,分别是第 j 列元素中的第 $Rank_j^i$ 个和第 $MRank_j^i$ 个,则 *RP* 定义为

$$RP = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} |Rank_j^i - MRank_j^i|}{nm}$$

RK 用来衡量扰动前后,在自己所在的列中保持大小顺位不变的元素的比例. RK 的计算式为

$$RK = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} Rk_j^i}{nm}$$

其中, $Rk_j^i = \begin{cases} 1 & Rank_j^i = MRank_j^i \\ 0 & otherwise \end{cases}$

指标 CP 用于衡量扰动前后各属性平均值大小顺位的变化情况.假设按照升序顺序,在 A 和 MA 中,第 i 列元素的平均值分别是所有属性平均值中的第 $RankV_i$ 个和第 $MRankV_i$ 个,则 CP 定义为

$$CP = \frac{\sum_{i=1}^{m} \left| RankV_i - MRankV_i \right|}{m}$$

类似于 RK, CK 表示扰动前后在各属性平均值中保持大小顺位不变的属性的比例. CK 定义为

$$CK = \frac{\sum_{i=1}^{m} Ck_i}{m}$$

其中, $Ck_i = \begin{cases} 1 & RankV_i = MRankV_i \\ 0 & otherwise \end{cases}$

4.3 实验数据

本文实验使用两个实际数据集:WBC(the original Breast Cancer Wisconsin Data Set)和 PID(the Pima Indians Diabetes Data Set).他们都来自于 UCI(University of California at Irvine)机器学习数据库.UCI 机器学习数据库是著名的免费数据库,研究者们经常使用该数据库的数据进行实验.WBC 数据包含 9 个属性和 699 个样本,本文只使用其中完整的不重复样本,共 449 个.PID 数据包含 8 个属性和 768 个样本.

4.4 实验结果

本文实验中,首先通过实验为各算法选定参数,选择参数的原则是在保持数据可用性的前提下,最大化数据隐私保护程度;随后,通过比较此时的数据隐私保护程度来比较各算法的优劣.

图 1 表示当参数 k 取不同值时,经 BSVD 算法扰动后数据的可用性. k 的最小值为 1,最大值为原始数据的属性个数,步长为 1.如前所述, k 的值越大,扰动数据的可用性越好,但安全性下降.因此,在保持数据可用性的前提下, k 的值应该尽可能的大.对于 WBC 数据,本文取 $k=7$;对于 PID 数据,本文取 $k=6$.

SSVD 算法有两个参数 k 和 d.从本质上来讲,SSVD 算法有两个步骤.首先使用参数为 k 的 BSVD 算法来对数据进行扰动,然后再在扰动后数据上进行参数为 d 的附加扰动. k 的值越大, d 的值越小,扰动数据可用性越好,但安全性下降.本文使用贪心策略来为 SSVD 算法选择参数值.首先令 SSVD 算法的参数 k 等于 BSVD 算法中参数 k 的优化值,即对于 WBC, $k=7$;对于 PID, $k=6$.使得在保持数据可用性的前提下最大化第一步的扰动强度.然后在保持数据可用性的前提下,选择尽可能大的 d.本文实验中,参数 d 的值由另一个参数 e 来确定, e 表示在第二步的附加扰动中被收缩为零的元素的比例,即绝对值小于 d 的元素的比例.图 2 表示当参数 k 取优化值,参数 e 取不同值时,经 SSVD 算法扰动后的数据的可用性. e 的最小值为 0.05,最大值为 0.95,步长为 0.05.对于 WBC 数据,本文取 $e=0.45$;对于 PID 数据,本文取 $e=0.15$.

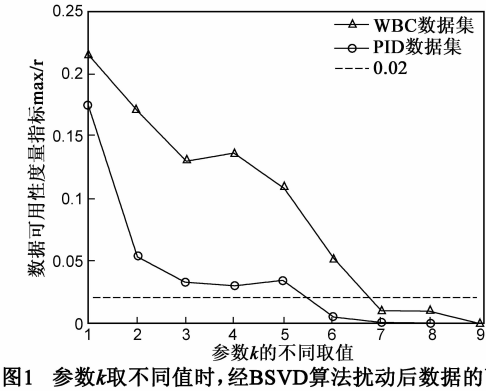

图1 参数k取不同值时,经BSVD算法扰动后数据的可用

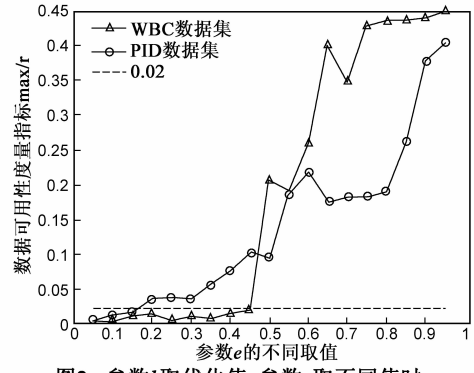

图2 参数k取优化值,参数e取不同值时,经SSVD算法扰动后的数据的可用性

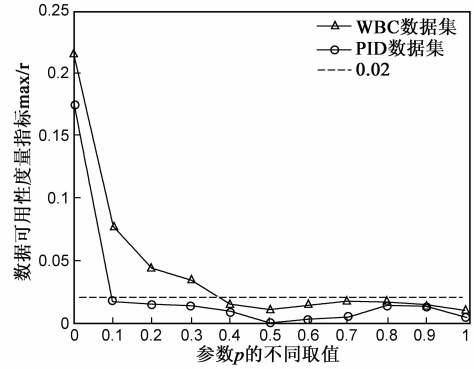

图3 参数k_1和k_2取优化值,参数p取不同值时,经本文算法扰动后的数据的可用性

在本文算法中,有三个参数: k_1, k_2 和 p.考虑到保持数据的可用性,令 k_1 等于 BSVD 算法中参数 k 的优化取值,即对于 WBC 数据, $k_1=7$;对于 PID 数据, $k_1=6$.为了尽量加大对于数据的扰动,令 $k_2=1$.图 3 表示

了 k_1 和 k_2 取选定值,参数 p 取不同值时经本文算法扰动后数据的可用性. p 的最小值为 0,最大值为 1,步长为 0.1. 当 $p=1$ 时,相当于对数据进行 $k=k_1$ 的 BSVD 扰动;当 $p=0$ 时,相当于对数据进行 $k=k_2$ 的 BSVD 扰动.可以看出,当 p 较大时(对于 WBC, $p\geqslant 0.4$;对于 PID, $p\geqslant 0.1$),扰动数据都保持了不错的可用性,当 $p=0.5$ 时,扰动数据的可用性最好.本文为参数 p 设置了两套取值:取 $p=0.5$ 时,可以得到较好的数据可用性;对 WBC,取 $p=0.4$,对 PID,取 $p=0.1$ 时,可以对隐私数据提供较好的保护.

表 1 表示本文算法和原有的基于奇异值分解的方法的可用性及隐私性度量. WSVD1 和 WSVD2 都表示本文算法,而且参数 k_1 和 k_2 都取优化值($k_2=1$;对 WBC, $k_1=7$,对 PID, $k_1=6$). 在 WSVD1 中,参数 p 取第一组优化值($p=0.5$);WSVD2 中,参数 p 取第二组优化值(对 WBC, $p=0.4$,对 PID, $p=0.1$). 参数 p 的第一组优化值注重可用性,第二组优化值注重隐私性. BSVD1 和 BSVD2 分别表示当 k 取优化值(即 $k=k_1$)时以及 $k=1$ (即 $k=k_2$)时的 BSVD 算法. SSVD 表示参数取优化值时(对 WBC, $k=7$, $e=0.45$;对 PID, $k=6$, $e=0.15$),用 SSVD 算法对数据进行扰动.

容易看出, $k=1$ 时的 BSVD 算法在隐私保护方面表现最佳,但数据可用性很差. 当参数 k 取优化值时, BSVD 算法可以保持很好的数据可用性,但隐私性表现最差.本文方法在隐私性度量指标上仅略差于 $k=1$ 时的 BSVD 算法,同时保持了很好的数据可用性. 这是因为本文算法将样本和属性分成了重要的和不重要的,对重要样本在重要属性上的值进行强度受限的扰动,对其它数据进行高强度扰动. 而在 BSVD 算法中不存在这种差别化扰动. BSVD1 对所有样本和属性都进行强度受限的扰动,因此可用性较好而隐私性不佳,而 BSVD2 对所有数据都进行高强度扰动,因此隐私性很好而可用性很差. 同为对 BSVD 算法的改进,本文算法无论隐私性还是可用性都优于 SSVD 算法.

表 1　经本文算法和原有的基于奇异值分解的方法扰动后的数据的可用性及隐私性度量

数据	方法	max(r)	*VD*	*RP*	*RK*	*CP*	*CK*
WBC	BSVD1	0.0102	0.12	31.87	0.019	0.26	0.78
WBC	BSVD2	0.2139	0.37	63.91	0.006	0.79	0.33
WBC	SSVD	0.0180	0.25	36.95	0.014	0.27	0.76
WBC	WSVD1	0.0101	0.28	47.07	0.008	0.57	0.54
WBC	WSVD2	0.0143	0.30	50.26	0.008	0.58	0.53
PID	BSVD1	0.0051	0.01	48.35	0.126	0	1
PID	BSVD2	0.1747	0.46	180.13	0.004	0.25	0.75
PID	SSVD	0.0172	0.02	54.76	0.071	0	1
PID	WSVD1	0.0002	0.33	134.44	0.005	0.01	0.99
PID	WSVD2	0.0170	0.44	176.89	0.004	0.22	0.78

5　结论

在现有的基于奇异值分解的隐私保持数据挖掘算法中,所有的样本和属性都被看作是平等的,并被施以相同程度的扰动.但在实际中,不同的样本和属性对隐私保护可能有不同的要求,而且他们可能对于数据挖掘有不同的重要性.因此对样本和属性进行差别化的扰动是必要的.

正是出于这一目的,本文分析并改进了原有的基于奇异值分解的扰动方法,使之可以对不同的样本和属性进行不同程度的扰动.并在此基础上,结合样本选择和属性选择,提出了一种改进的基于奇异值分解的隐私保持分类挖掘方法.实验表明,与原有的基于奇异值分解的方法相比,在保持数据可用性的前提下,本文方法可以为隐私数据提供更好的保护.

参考文献

[1] Herman T Tavani. Information privacy, data mining, and the internet [J]. Ethics and Information Technology, 1999, 1(2): 137 – 145.

[2] A Cavoukian. Data mining: staking a claim on your privacy [OL]. http://www. ipc. on. ca/english/Resources/Discussion-Papers/Discussion-Papers-Summary/? id = 342. 1998-01-01/2010-10-15.

[3] Vassilios S Verykios, Elisa Bertino, Igor Nai Fovino, Loredana P Provenza, Yucel Saygin, Yannis Theodoridis. State-of-the-art in privacy preserving data mining [J]. ACM SIGMOD Record, 2004, 33(1): 50 – 57.

[4] Elisa Bertino, Igor Nai Fovino, Loredana Parasiliti Provenza. A framework for evaluating privacy preserving data mining algorithms [J]. Data Mining and Knowledge Discovery, 2005, 11(2): 121 – 154.

[5] Rakesh Agrawal, Ramakrishnan Srikant. Privacy-preserving data mining [J]. ACM SIGMOD Record, 2000, 29(2): 439 – 450.

[6] Li Liu, M Kantarcioglu, B Thuraisingham. The applicability of the perturbation based privacy preserving data mining for real-world data[J]. Data & Knowledge Engineering, 2008, 65(1): 5 – 21.

[7] 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(10): 2021 – 2029.
HAN Jian-min, CEN Ting-ting, YU Hui-qun. Research in microaggregation algorithms for k-anonymization [J]. Acta Electronica Sinica, 2008, 36(10): 2021 – 2029. (in Chinese)

[8] S Kisilevich, L Rokach, Y Elovici, B Shapira. Efficient Multidimensional Suppression for K-Anonymity [J]. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(3): 334 – 347.

[9] Yehuda Lindell, Benny Pinkas. Privacy preserving data mining

[J]. Journal of Cryptology, 2002, 15(3): 177 – 206.
[10] Benny Pinkas. Cryptographic techniques for privacy-preserving data mining [J]. ACM SIGKDD Explorations Newsletter, 2002, 4(2): 12 – 19.
[11] 张锋,孙雪冬,常会友,赵淦森.两方参与的隐私保护协同过滤推荐研究[J].电子学报,2009,37(1):84 – 89.
ZHANG Feng, SUN Xue-dong, CHANG Hui-you, ZHAO Gan-sen. Research on privacy-preserving two-party collaborative filtering recommendation [J]. Acta Electronica Sinica, 2009, 37(1): 84 – 89. (in Chinese)
[12] F Emekci, O D Sahin, D Agrawal, A El Abbadi. Privacy preserving decision tree learning over multiple parties [J]. Data & Knowledge Engineering, 2007, 63(2): 348 – 361.
[13] Shuting Xu, Jun Zhang, Dianwei Han, Jie Wang. Singular value decomposition based data distortion strategy for privacy protection [J]. Knowledge and Information Systems, 2006, 10(3): 383 – 397.
[14] Jie Wang, Jun Zhang, Shuting Xu, Weijun Zhong. A novel data distortion approach via selective SSVD for privacy protection [J]. International Journal of Information and Computer Security, 2008, 2(1): 48 – 70.
[15] Jie Wang, Weijun Zhong, Jun Zhang. NNMF-Based Factorization Techniques for High-Accuracy Privacy Protection on Non-negative-valued Datasets [A]. In Proceedings of ICDMW'06 [C]. Washington, DC: IEEE Computer Society, 2006. 513 – 517.
[16] 郝红卫,蒋蓉蓉.基于最近邻规则的神经网络训练样本选择方法[J].自动化学报,2007,33(12):1247 – 1251.
HAO Hong-Wei, JIANG Rong-Rong. Training Sample Selection Method for Neural Networks Based on Nearest Neighbor Rule [J]. ACTA AUTOMATICA SINICA, 2007, 33(12): 1247 – 1251. (in Chinese)
[17] Ian H Witten, Eibe Frank. Data Mining: Practical Machine Learning Tools and Techniques (Second Edition) [M]. Morgan Kaufmann, 2006.

作者简介

李 光 男.1982 年 2 月生于陕西富平.分别于 2004、2006 和 2011 年在哈尔滨工业大学计算机科学与技术学院获工学学士、硕士和博士学位.现任教于长安大学,从事隐私保持数据挖掘,生物信息学等方面的有关研究.
E-mail: hit6006@126.com

王亚东 男.1964 年 6 月生于黑龙江.教授,博士生导师,哈尔滨工业大学计算机科学与技术学院院长,国家十五"863"计划生物信息技术主题第一届、第二届专家组专家,黑龙江省生物医学信息技术与系统工程研究中心主任,黑龙江省生物信息技术重点实验室主任,中国人工智能学会理事.主要研究方向:分布式人工智能、专家系统、机器学习、知识工程、生物信息技术等,主持完成国家自然科学基金、国家 863 计划项目、国际合作项目等二十余项,在国内外重要期刊发表论文五十余篇.
E-mail: ydwang@hit.edu.cn