第 34 卷 第 9 期 计 算 机 工 程 2008 年 5 月
Vol.34 No.9 Computer Engineering May 2008
·软件技术与数据库· 文章编号：1000—3428(2008)09—0078—03 文献标识码：A 中图分类号：TP311

基于隐私保护的数据挖掘

马廷淮，唐美丽
(南京信息工程大学计算机与软件学院，南京 210044)

摘 要：基于隐私保护的数据挖掘(PPDM)的目标是在保护原始数据的情况下建立挖掘模型并得到理想的分析结果。该文从 PPDM 的总体需求出发，基于数据隐藏，将 PPDM 技术分为安全多方计算技术、匿名技术和数据转换技术。从准确性、隐私性和复杂性 3 个方面对 PPDM 技术进行了评估。

关键词：隐私保护；数据挖掘；分类；评价

Data Mining Based on Privacy Preserving

MA Ting-huai, TANG Mei-li
(School of Computer & Software, Nanjing University of Information Science & Technology, Nanjing 210044)

【Abstract】 The purpose of Privacy Preserving Data Mining(PPDM) is discovering accurate patterns and perfect results without precise access to the original data. The requirements of PPDM are put forward. Based on data hiding, PPDM methods are divided into three types: SMC, anonymization and data distortion, and they are evaluated under accuracy, privacy and complexity.

【Key words】 privacy preserving; data mining; classification; evaluation

1 概述

数据挖掘作为研究从海量数据中自动提取未知模式的新兴技术，在短短十几年内取得了飞速发展，随之也带来了一系列问题。通过对购物习惯、犯罪记录、健康记录、信用卡记录等数据的分析，可以得到商业信息和政府部门的决策信息，但对这些数据及其挖掘出的模式外泄而产生的不适当应用会给个人隐私带来威胁，如病人的病情信息、顾客的喜好、个人背景资料等极其隐私的信息。在数据挖掘过程中解决隐私保护问题，已经成为数据挖掘界的一个研究热点[1]。

基于非精确的原始数据挖掘出较为准确的模式与规则是隐私保护数据挖掘的出发点。基于隐私的数据挖掘是在原始数据的不准确性和数据挖掘结果的精确性之间寻求平衡[1]。

基于隐私数据保护的数据挖掘体现在如下两方面：(1)保护个人信息，就是在数据挖掘过程中不能泄漏但可以直接或间接确定用户的特征信息；(2)保护产生模式，限制挖掘中部分敏感模式的产生和泄漏。

2 基本要求

(1)数据挖掘的目标

每种数据挖掘都会有特定模式的数据挖掘结果产生。有的数据挖掘结果仅对数据本身做分析；而有的则对数据间的关系进行分析，如关联规则提取。不同的挖掘结果应采取不同的隐私保护策略，本文对前者采取个人信息保护方式，对后者须采取模式产生保护方式。

(2)数据的分布性

原始数据集中存放或分布式存放将直接影响隐私保护策略的制定。如果是分布式存放，各个分布式数据库间的数据对一个实体来说可以是垂直划分的或水平划分的。对分布式垂直划分的数据库，各个点之间要求不能进行互相串通，以免隐私的泄漏；对分布式水平划分的数据库，单点就能保存实体的所有属性数据，因此，单点的隐私性要求比较高。

(3)隐私保护的程度

正如数据挖掘没有一个普遍适应的算法一样，隐私保护的数据挖掘解决方案也是根据隐私保护的内容及数据挖掘方法差异而不同的。每一种隐私保护的技术都会限制用户信息的共享性和增强隐私的安全性。衡量一个隐私保护技术的优劣主要考虑以下几个方面：1)技术的有效性，要确定该技术的应用水平；2)隐私保护的程度，为了达到理想的挖掘结果对隐私牺牲程度。每一种隐私保护的数据挖掘只有确定了这两方面的问题，才能很好地确定隐私保护数据挖掘方案的制定策略。

3 隐私保护技术分类

基于隐私保护的数据挖掘技术可从 4 个层面进行分类：(1)数据的分布情况，可以分为原始数据集中式和分布式两大类隐私保护技术。(2)原始数据的隐藏情况，可以分为对原始数据进行扰动、替换和匿名隐藏等隐私保护技术。(3)数据挖掘技术层面，可以分为针对分类挖掘、聚类挖掘、关联规则挖掘等隐私保护技术。(4)隐藏内容层面，可以分为原始数据隐藏、模式隐藏。本文主要针对第(2)种分类进行探讨。

3.1 SMC 技术

安全多方计算是解决分布式计算安全性的重要技术。在分布式环境中，为了保护隐私，参与数据挖掘的各个节点间互相不知道对方的原始数据，这样最能保证隐私不被泄漏。

基金项目：江苏省生产力协会基金资助项目(2006-JS-045)；江苏省青蓝工程基金资助项目；南京信息工程大学科研基金资助项目(QD 67, 2643)

作者简介：马廷淮(1974－)，男，副教授、博士，主研方向：智能计算，数据挖掘；唐美丽，讲师、硕士

收稿日期：2007-05-30 **E-mail：**thma@nuist.edu.cn

参与挖掘的各个节点互相不进行数据交流，把挖掘所需的原始数据都送到信任的第三方进行挖掘，最后取得挖掘结果。

实际上，安全多方计算是一种分布式计算协议。计算所有节点的和[2]是一种典型的安全多方计算应用。假设有1,2,⋯,s个节点，每个节点提供的值为 $u_j, j=1,2,\cdots,s$。假设所有节点的和 $U=\sum_{j=1}^{s}u_j$ 属于区间$[0,n]$。SMC下求和的执行过程如下：节点 1 随机选取一个属于$[0,n]$随机数R，然后将 $(R+u_1)\mathrm{MOD}(n)$ 传递给节点 2。节点 2 将接收到的值加上u_2取 $\mathrm{MOD}(n)$ 继续传下去，一直到节点s。对任意一节点k，其接收到 $(V=R+\sum_{j=1}^{k-1}u_j)\mathrm{MOD}(n)$，然后将 $(V+u_k)\mathrm{MOD}(n)$ 传递到节点k+1。最后节点s将结果 sum 传给节点 1，节点 1 根据 sum 减去自己确定的R，就得到真实的 $U=\sum_{j=1}^{s}u_j$。根据事先的假设 $U\in[0,n]$，U的真实值为 $sum-R+\{0\text{ or }1\}\times n$。

如上所述，安全多方计算是针对某个数据挖掘技术或挖掘目标而采取一系算法步骤的执行协议。基于安全多方计算的协议通常是根据挖掘技术来制定的。其中基于数据库横向分割和纵向分割的分类挖掘算法详见文献[3]，基于安全多方计算的聚类方法和关联规则算法详见文献[4-5]。

3.2 匿名技术

匿名技术是身份隐藏中最直接的技术。它作为隐私保护的数据挖掘技术不对数据挖掘结果进行保护，也不将原始数据进行隐藏伪装，而是公布带隐私的所有数据，但是他人拿到隐私数据却不能推导出该数据拥有者的身份。

假设一个医疗信息数据表如表 1 所示，其中将出生日期、邮编、过敏药物作为标识某个特定记录的特征属性集合，将既往病史作为隐私属性进行保护。从匿名技术进行隐私保护来讲，就是要将能够作为唯一标志的属性集合进行匿名隐藏，从而间接保护隐私。

表 1　医疗系统信息表

出生年月	邮编	过敏药物	既往病史
79-03	07030	青霉素	咽炎
57-02	07028	无过敏	中风
39-12	07030	无过敏	脊髓灰质炎
57-02	07029	磺胺	白喉
40-01	07030	无过敏	大肠炎

可以看出，标识属性值是各不相同的，根据一个标识值就能与某个特定的记录、特定的人一一对应起来，其隐私数据就跟特定的人对应起来，隐私得不到保护。而如果选用邮编、过敏药物作为标识属性，将既往病史作为隐私属性可以看到，同样是{07030，无过敏}的取值有 2 条记录，不能将其隐私属性值{脊髓灰质炎}，{大肠炎}与{07030，无过敏}为标志的记录唯一确定下来，就可以达到保护个人隐私的目的。

具体的隐私保护方法可以分为以下 2 类[6]：

(1)保护隐私属性集合

单个节点公布数据的时候，节点的标识部分不加密，将隐私属性部分进行单独加密。系统对每个节点的数据进行汇总后也不能看到每个节点的隐私数据，而只能看到该节点的标识属性数据。系统将收集到的所有节点的标识属性进行归类统计，当统计得出某个节点的标识属性在整个系统中重复次数超过K_i，系统才能根据K_i解密第i个节点的隐私属性。

(2)隐藏标志属性集合

节点参与系统的数据挖掘时候，通过一个算法要求系统给出每个节点标识规则。系统应该能测算出至少满足节点标识重复度大于临界δ的标识规则。节点根据这个规则在对外公布数据时，合理选取标识，使得系统得到的数据由于标识的重复而不能一一对应，从而保护节点的隐私。

3.3 数据转换技术

数据转换技术的主要思想是将用户的真实隐私数据进行伪装或轻微改变，通过数据挖掘，得到可以接受精度的挖掘结果。根据不同的数据挖掘技术，对原始数据的伪装方式也不尽相同。常见的数据转换技术有随机扰动方法、数据几何变换方法等。

(1)随机扰动技术[7]

把单个节点的原始值 $x_1,x_2,\cdots,x_n$ 看作是 n 个具有相同分布的独立随机变量 $X_1,X_2,\cdots,X_n$ 的值，随机变量 $X_1,X_2,\cdots,X_n$ 具有相同的分布，密度函数是 F_X。真实提供给系统的数据是 $x_1+y_1,x_2+y_2,\cdots,x_n+y_n$，$y_i$ 是加入的噪声数据，对应随机变量 Y_i 值，Y 的密度函数是 F_Y (均值为 0 的正态分布或者均匀分布)。对于挖掘算法，已知 x_i+y_i 和 F_y，要推断出 X_i 的取值，才能进行挖掘计算。重构 X_i 的主要思路是利用贝叶斯规则迭代进行近似估算 F_X。利用随机扰动技术进行隐私保护的数据挖掘方法有决策数构造法、关联规则发现。

(2)数据的几何变换法

利用计算机图形学中的几何变换思想来对数据进行变换达到保护原始数据的目的。经过几何变换的数据与原始数据相差较大，对部分挖掘方法的挖掘结果影响较大。常见的几何变化方法有数据平移、缩放、旋转等。该类数据转换方法在聚类挖掘技术中应用较好[8]。聚类技术的核心是考虑数据间的距离，此距离可以化为一个无量纲的相对距离。原始数据的平移、缩放、旋转等都不会改变数据间的相对距离的大小，实践证明其对聚类方法的挖掘结果影响较小[8]。

4 算法的评价

对一个基于隐私保护的数据挖掘方法总体上来说应该从算法挖掘结果的准确程度、隐私保护程度、算法性能 3 个方面来度量。由于每种数据挖掘技术的不同，每个隐私保护方案的不同，对隐私的度量和对挖掘结果的影响度量也不同。

4.1 准确度

基于隐私保护的数据挖掘算法的准确性包含 2 个方面：数据挖掘算法本身的准确性，在隐私保护下该算法相对于非隐私情况下的准确性。这里只讨论第 2 种情况。

(1)对于安全多方计算，其隐私保护前后算法的准确性是能得到保证的，一般靠牺牲算法的性能来满足算法的准确性。

(2)对于匿名技术，采用的实质是一种属性归纳和泛化的方法，是针对隐私保护而设计的，不存在隐私保护前后算法的准确性比较。因此，匿名技术无准确性度量。

(3)对于数据转换技术，其准确性的分析根据挖掘算法的不同分为以下几个方面：

1)基于数据扰动的隐私保护关联规则挖掘

关联规则的数据挖掘分为 2 步：根据支持度进行频繁集的选择，根据信任度生成关联规则。关联规则的生成基于频繁集，而不针对具体的操作型数据库，因此，数据扰动的原始数据不直接影响关联规则的生成。

数据扰动对频繁集的影响可以考虑相对非扰动下的频繁集的增加或减少。文献[9]中用应用 $\sigma^+=\frac{F-R}{F},\sigma^-=\frac{R-F}{F}$ 来评价，σ^+,σ^- 分别是新增加的频繁集比例和减少的频繁集比

例，F 是扰动前的频繁集数目，R 是采用数据扰动后的频繁集数目。

2)基于数据扰动的隐私保护分类挖掘

在数据扰动下的分类算法考虑最多的是将原始数据随机扰动后采用 Bayes 公式对原始数据的分布密度函数进行估算。因此，分类算法的精度直接受原始数据估算分布函数的估算精度影响。一般说来，经过数据扰动后的分类算法的精度小于非扰动前的分类算法的精度。

3)基于数据变换的聚类

聚类算法的准确度描述主要是看是否能正确的形成聚集点。文献[8]中采用误分类百分率M_E来比较分析转换前后各个聚类中数据点的分布变化。

4.2 隐私保护程度

理论上设计的每一个算法和协议都是完全保护隐私的。但由于算法协议的复杂性，对数据的扭曲导致算法的不精确性等限制，实际案例中的隐私保护不得不在隐私的损失和算法精度和性能的提高上进行协调。

隐私的保护在安全多方计算中，保证参与计算的各节点不能了解其他节点的原始数据，只能知道最后的挖掘结果。

匿名技术中隐私的保护主要是根据系统设定同时混淆的个体数目来决定。如果选择的混淆程度越大，隐私就保护得越好，根据已有的信息推断出确定个体的可能性就越小。

数据变化技术中的隐私保护是通过扭曲原始数据来实现的。衡量隐私保护的程度也就是原始数据的安全程度，包含 3 个方面的内容：(1)原始数据的改变程度；(2)原始数据经数据修改后被推测出正确值的程度；(3)隐含在原始数据中的规则模式被暴露的程度。

4.3 算法性能

算法的性能一般是指算法的时间复杂度。基于隐私保护的数据挖掘技术与常规的数据挖掘技术相比，无论是采用安全多方的协议控制还是进行数据转换，都会带来时间的开销。每种数据挖掘算法的时间复杂度不一样，在每种隐私保护解决方案下所带来的时间开销也不一样，因此，对基于隐私保护的数据挖掘的时间复杂度没有一个统一的描述。

除算法本身时间开销外，通信开销在分布节点的数据挖掘中也显得比较重要。特别是基于 SMC 的解决方案，由于计算协议的复杂导致数据的传输量大，占用的时间增加，通信开销十分可观。

5 结束语

受数据挖掘技术多样性的影响，隐私保护的数据挖掘方法呈现多样性。在常见的分类、聚类、关联分析方面都出现了隐私保护技术，但是隐私保护数据的一些基本理论和应用还需进一步细化。

(1)隐私保护的量化。不同的隐私保护方法对隐私的保护程度不同，应该对隐私保护数据挖掘建立一个评价体系和量化标准。

(2)隐私保护与普适网络环境相结合。在普适网络环境下，建立隐私保护下的用户数据挖掘架构，实现优质的服务推荐系统(recommand system)。

(3)设计基于隐私保护的分布式数据节点的协同挖掘算法，提高数据挖掘算法的性能。

参考文献

[1] Clifton C, Kantarcioglu M, Vaidya J. Defining Privacy for Data Mining[C]//Proceedings of the National Science Foundation Workshop on Next Generation Data Mining. Baltimore, MD, USA: [s. n.], 2002.

[2] Clifton C, Kantarcioglou M, Lin Xiadong, et al. Tools for Privacy Preserving Distributed Data Mining[J]. SIGKDD Explorations, 2002, 4(2): 28-34.

[3] Lindell Y, Pinkas B. Privacy Preserving Data Mining[M]. [S. l.]: Springer-Verlag, 2000-08.

[4] Lin Xiaodong, Clifton C. Privacy Preserving Clustering with Distributed EM Mixture Modeling[J]. Knowledge and Information Systems, 2005, 8(1): 68-81.

[5] Kantarcoglu M, Clifton C. Privacy Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data[C]//Proc. of DMKD'02. Madison, WI, USA: [s. n.], 2002.

[6] Zhong Sheng, Yang Zhiqiang, Wright R N. Privacy-enhancing K-anonymization of Customer Data[C]//Proceedings of the 24th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems. Baltimore, MD,USA: [s. n.], 2005.

[7] Agrawal R, Srikant R. Privacy Preserving Data Mining[C]//Proc. of ACM SIGMOD Conf. on Management of Data. Dallas, Texas, USA: [s. n.], 2000.

[8] 黄伟伟, 柏文阳. 聚类挖掘中隐私保护的几何数据转换方法[J]. 计算机应用研究, 2006, 23(6): 180-181.

[9] Rizvi S, Haritsa J R. Maintaining Data Privacy in Association Rule Mining[C]//Proc. of the 28th Int'l Conf. on Very Large Databases. Hong Kong, China: [s. n.], 2002.

（上接第 77 页）

参考文献

[1] Agrwal R, Imielinsk T, Swami A. Mining Associations Between Sets of Items in Massive Database[C]//Proc. of the ACM SIGMOD Int'l Conference on Management of Data. Washington D. C., USA: [s. n.], 1993.

[2] Agrawal R, Srikant R. Fast Algorithms for Mining Association Rules in Large Databases[C]//Proceedings of the 20th International Conference on Very Large Data Bases Table of Contents. Santiago, Chile: [s. n.], 1994.

[3] Bayardo R J. Efficiently Mining Long Patterns from Databases[C]//Proceedings of International Conference on Management of Data. New York, USA: ACM Press, 1998.

[4] Han Jiawei, Kamber M. 数据挖掘——概念与技术[M]. 范 明, 孟小峰, 译. 北京: 机械工业出版社, 2001.

[5] Han Jiawei, Pei Jian, Yin Yiwen. Mining Frequent Patterns Without Candidate Generation[C]//Proceedings of the ACM SIGMOD International Conference on Management of Data. Dallas, Texas, USA: [s. n.], 2000.

[6] Li Jiuyong, Shen Hong, Topor R. Mining the Smallest Association Rule Set for Prediction[C]//Proc. of IEEE Int'l Conf. on Data Mining. San Jose, California, USA: [s. n.], 2001.