隐私保护的数据挖掘方法的研究

汤 琳,何 丰
(北方民族大学 计算机科学与工程学院,宁夏 银川 750021)

摘 要:介绍了隐私保护数据挖掘方法的产生背景和意义,其次概括了现阶段国内外隐私保护数据挖掘算法的研究现状,并对当前隐私保护数据挖掘领域中已提出的算法按照数据挖掘的方法、数据源分布情况、隐私保护技术和隐私保护对象以及数据挖掘应用类型等方面进行分类,然后分别详细阐述了在集中式和分布式数据分布环境下,应用在隐私保护的关联规则挖掘、分类和聚类挖掘中的一些典型的技术和算法,总结出它们的优缺点,并对这些优缺点进行剖析和对比,最后指明了隐私保护数据挖掘算法在未来的整体发展方向。

关键词:隐私保护数据挖掘;隐私保护的关联规则挖掘;分类挖掘;聚类挖掘;集中式数据;分布式数据
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2011)04-0156-04

Research on Privacy-Preserving Data Mining Method

TANG Lin, HE Feng
(School of Computer Science and Engineering, North University for Ethnics, Yinchuan 750021, China)

Abstract: Introduce the background and the significance of the privacy-preserving data mining methods. Secondly, summarized the present research status of the privacy-preserving data mining algorithm at home and abroad, and these algorithms in this area has been proposed already are classified according to the data mining methods, original data distribution, privacy-preserving techniques, privacy protection and data mining application type. Thirdly, it elaborates some typical technologies and algorithms which are used to the type such as the privacy-preserving association rule mining, and the privacy-preserving classification mining, and also the privacy-preserving cluster mining on the environment of the centralized data and the distributed data respectively. Most importantly, it also summarizes their advantages and disadvantages and then analyses and contrasts them to highlight the future direction.

Key words: privacy-preserving data mining; privacy-preserving association rule mining; privacy-preserving classification mining; cluster mining; centralized data; distributed data

0 引 言

数据挖掘是近年来十分活跃的研究领域。数据挖掘即提取或“挖掘”知识。它是从数据中抽取隐含的、未知的和潜在有用的信息[1]。

然而,随着数据挖掘过程的进行和挖掘结果的产生以及新的数据挖掘技术的不断出现,在发现知识的同时,可能导致现在备受关注的隐私权的入侵,给数据的隐私带来了威胁,严重威胁到人们的个人信息安全和机构的商业秘密安全。因此数据挖掘研究人员必须在进行数据挖掘的同时保护好数据源和挖掘结果的隐私性。

比如在医学中,为了分析某种病的发病率,几家医院可能将自己拥有的数据综合起来进行分析,但是这可能涉及病人的隐私或是病人不愿意被别人知道所患的病症而不愿意共享数据。所以必须采用某些技术手段,来控制和预防在数据挖掘过程中隐私信息的泄露问题。这些技术手段即隐私保护技术在数据挖掘中的应用,也就是在保证足够精度和准确度的前提下,使数据挖掘方在不触及实际隐私数据的同时,仍能进行有效挖掘工作,称为隐私保护的数据挖掘方法。隐私保护的有效性以及由此研究采取哪些隐私保护挖掘算法来防止敏感信息的暴露越来越显示其重要性,隐私保护数据挖掘算法在分类挖掘、聚类挖掘和关联规则挖掘方面的应用和延伸已经成为目前数据挖掘研究的公开问题,也是未来数据挖掘领域的一个富有挑战性的研究热点。

收稿日期:2010-08-06;**修回日期**:2010-11-21
基金项目:国家自然科学基金(61070131)
作者简介:汤 琳(1985-),女,河南新野人,硕士研究生,研究方向为数据挖掘;何 丰,教授,从事语义 web 和数据挖掘的研究。

1 隐私保护数据挖掘主要研究方向及国内外研究现状

1.1 主要研究方向

隐私保护在数据挖掘领域的应用可以分为三个方

向:在关联挖掘规则挖掘、分类挖掘以及聚类挖掘中的应用。隐私保护的关联数据挖掘研究已经提出了许多算法,国内的学者已经提出了大量的算法以及改进算法。以后的研究的主要方向可以转向对隐私保护的分类方法和隐私保护的聚类方法的研究。

1.2 国内外研究现状

近年来,数据挖掘中的隐私保护问题得到了国内外学者的广泛研究。

首先,在隐私保护关联规则挖掘方面,国内外学者在隐私保护的关联规则挖掘的应用方面有两种主要的方法:一是运用隐藏频繁项目集这类方法实际上就是对原始数据进行隐私保护处理来防止涉及隐私及相关的重要信息的关联规则的产生;二是尽可能使涉及隐私规则或信息的置信度远远小于规定的最低置信度,也就是要运用一切可能的方法来隐藏挖掘出来的规则,以此达到需要保护或隐藏的规则不被挖掘出来的目的。

其次,在数据集中分布的隐私保护分类挖掘中,主要有两种隐私保护挖掘方法:一是使用随机响应方法,它实际上是以统计学为基础,先对数据采用随机变换方法进行变换,然后在变换后数据的基础上运用数学方法推导原始数据的取值概率,从而达到隐藏原始数据的目的,以实现隐私保护的分类挖掘。二是添加随机偏移量方法:通过加入随机偏移量对包含隐私的数据进行干扰,随后和第一种方法一样再基于干扰后的数据分布对干扰前的数据进行还原和重构,并同时对隐私保护度进行衡量。

在隐私保护聚类挖掘算法方面,目前已经主要通过几何转换等对原始数据进行转化,从而达到保护原始数据的效果。随后利用等距变换来实现原始数据隐私保护的改进算法,以及点积协议在数据垂直划分的聚类挖掘问题中的应用等。国外的研究起步较早,国内的研究相比国外的发展还是有一定的距离。

2 隐私保护数据挖掘的分类

2.1 按照基本策略进行分类

目前隐私保护的数据挖掘方法按照基本策略主要有数据扰乱法、查询限制法和混合策略。

数据扰乱法是对数据进行随机变换、数据离散化和在数据中添加噪声,从而对原始数据进行干扰,并在变换后的数据上推导原始数据的取值概率的方法。数据扰乱法的代表算法是 MASK[2](Mining Associations with Secrecy Konstraints)算法。

而查询限制的策略,是通过数据隐藏、数据抽样和数据划分等方式,从而尽量限制数据挖掘者拥有完整的原始数据,再利用概率统计的方法或者分布式计算这些数学方法来得到所需要的挖掘结果。查询限制策略存在这样一个缺陷,即所有提供的数据都是真实的原始数据,会降低隐私保护度。

2.2 按照数据源分布的情况进行分类

根据数据源的分布情况可以分为集中式数据隐私保护和分布式数据隐私保护两大类。分布式数据挖掘环境又由于数据记录的分布状态不同,又分为水平分布和垂直分布,如图1所示。

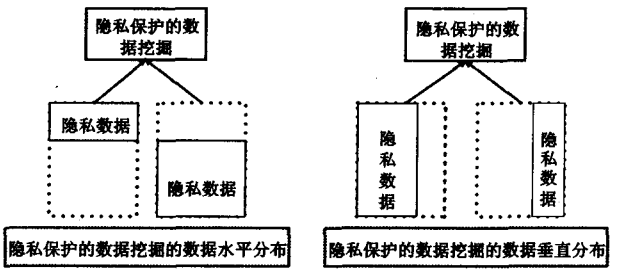


图1　隐私保护的数据挖掘的数据水平分布和数据垂直分布

2.3 按照隐私保护技术进行分类

根据隐私保护技术,可以分为启发式技术、密码技术和重构技术。

2.4 按照隐私保护的对象分类

根据隐私保护的对象,可以分为对原始数据的隐藏和对隐含的规则的隐藏。

2.5 按照隐私保护挖掘类型进行分类

根据隐私保护挖掘类型,可以分为源数据的隐私保护即输入保护和挖掘结果的隐私保护即输出保护。目前对于源数据的隐私保护主要采取随机扰动技术(Random Perturbation)和安全多方计算技术(Secure Multi-party Computation)。

2.6 按照应用数据挖掘技术进行分类

根据应用数据挖掘技术,可以分为隐私保护的关联规则挖掘、隐私保护的分类挖掘和隐私保护的聚类挖掘等。这三种隐私保护方法将会在下面的章节中详细阐述。

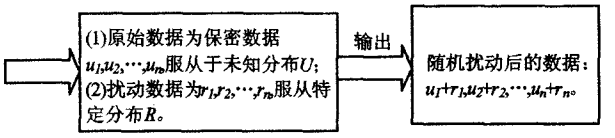
3 隐私保护的关联规则挖掘的研究

3.1 集中式数据的隐私保护的关联规则挖掘的研究

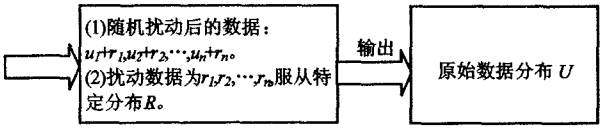
3.1.1 保护源数据

S. J. Rizvi 等在文献中提出了运用数据扰乱和分布重构的隐私保护的关联规则挖掘的代表——MASK算法,该算法通过数据扰乱和分布重构实现了隐私保护的关联规则挖掘,具体扰乱方法和分布重构过程如图2所示。但是基于扭曲数据库重构项集原始支持度呈现指数复杂度,严重影响了算法的运行效率。此算法需要对参数进行选择,它的参数必须偏离0.5;而且它经过扰乱后的数据还是和原数据息息相关,对于隐

私保护度的效果也并不很理想。后来 S. J. Rizvi 等又对 MASK 方法进行性能优化，提出高效的 EMASK(Efficient MASK)方法。EMASK 方法主要针对购物篮等布尔型和枚举型数据进行隐私保护的关联规则挖掘，对原始数据进行干扰。近年来我国的学者沈中林运用分治策略对 MASK 算法进行改进，比 MASK 算法提高了两个数量等级，从而大大降低了时间复杂度。同时朱思征等也使用数据随机化方法对原始数据进行变换，采用纵向结构组织数据与之提交变换后为"1"的数据组成数据表的方法，提出了 MASK 的改进算法 VSS-MASK 算法，通过减少提交数据量来提高算法扫描数据库的效率。随后我国学者王锐，刘杰在文献[3]中将随机相应技术与关联规则挖掘算法相结合，提出一个多参数随机扰动算法——MRD 算法，此算法通过合理参数的选择，能提高隐私保护度和挖掘准确度，性能相对于 MASK 算法大为提高。



(1) 随机扰动过程



(2) 重构过程

图 2 随机扰动过程和重构过程

3.1.2 规则隐藏

随着隐私保护数据挖掘方法的进行，关联规则中的规则隐藏也成为其中的研究热点。

Algo[4] 系列的算法，Naive[5]，MinFIA[5]，MaxFIA[5]，IGA[5]，RRA[6]，RA[6]，SWA[6] 等隐藏规则的算法先后被提出。文献[4]前提是所要隐藏的不同的敏感规则是相互独立的，按照最传统的方法，即降低支持度和置信度来隐藏规则。而后面的几个文献中所提出的算法隐藏的规则可以是不相互独立的，即是有交集项的。其中上述算法中 SWA 算法在对非敏感规则支持度的影响相对来说较小一些。

学者吴方对数据进行代替，从而降低所要隐藏规则中项目的支持度和置信度，同时尽可能少地影响留下的非敏感性规则，以便尽可能高地保持数据质量。戴智丽等把相关系数的概念引入隐私保护的数据挖掘领域，通过对规则左右件相关系数的调整，使敏感规则无法被发现，最后达到隐藏规则的目的。李霞等通过删除项和增加项两种操作相结合的方法来改变规则的支持度，用这种方法清洗数据所产出的规则丢失率和相异度均有所下降。文献[7]中提出了用于隐藏那些包含敏感项目的关联规则的处理算法 PPARM。该算法通过对原始事务库中的事务进行更新操作，达到了隐藏敏感的关联规则的目的，但是该算法磁盘读写操作次数过高，需要减少执行时间。

3.2 分布式数据的隐私保护的关联规则挖掘的研究

3.2.1 运用在分布式环境中的主要技术

在分布式环境中，计算全局频繁项集是关联规则挖掘的要点，在计算全局频繁项集的同时，恰当运用加密技术成为保证隐私信息不会泄露的重要方面。对于分布式数据的隐私保护数据挖掘，原始数据的保护方法主要基于密码学的方法来达到效果[8]。任何一种普通的计算都可以转化为 SMC 的框架这句话已经被证明。

目前在分布式关联规则隐私保护挖掘方面的应用大多使用安全多方计算方法，安全和运算(Secure Sun)，安全交集大小运算(Secure Size of Set Intersection)，安全求并集求法(Secure Set Union)，安全比较和标量积协议运算(Scalar Product)这几种技术来实现分布式数据隐私保护挖掘。

文献[9]中打破常规，在不使用当前流行的安全多方计算的条件下，用较简单的方法进行隐私保护关联规则挖掘，降低了运算量。同时也较好地保持了各个站点的数据和信息。

3.2.2 数据水平分布环境的隐私保护关联规则挖掘算法

在分布式隐私保护数据挖掘中，设计一个好的保护隐私的挖掘算法需要考虑以下几个方面：正确的挖掘结果，计算开销，通信代价和安全强度。

黄高琴在文献[10]中从数据到规则的方法提出一种新的有效的数据水平分布环境下的隐私保护关联规则挖掘算法，首先运用数学方法即项集转移概率矩阵对各个分布站点的数据进行变换，接着对全局计数项集支持度进行恢复，进而找出挖掘过程中所需的全局频繁集，最后找出全局关联规则。此算法能很好地完成数据挖掘的同时，也有效保护了各用户的私有数据。

3.2.3 数据垂直分布环境的隐私保护关联规则挖掘算法

在数据垂直分布环境中，同时利用各个站点的数据计算所有项集的计数，找出支持度大于阀值的全局频繁项集已成为解决问题的关键。目前有的学者已经用计算标量积的方法来解决此问题。在文献中 Jaidepe vaidya 提出从垂直分布数据中挖掘全局关联规则的隐私保护算法，其中项集的支持计数通过密码学中的安全计算代表子项集的标量积的方法得到，有效地解

决了这个难题。

目前我国学者对这方面的研究也已经探讨了在数据垂直分布环境中,如何在保护各方隐私数据的前提下挖掘全局频繁项集。并在商品服务器模型的研究基础上,提出了一种基于可逆方阵的两方和多方安全加密协议,该协议在垂直划分的分布式数据库中具有较好的隐藏性、高效性和准确性。

4　隐私保护的分类挖掘研究

4.1　集中式数据隐私保护的分类挖掘

Rkaesh Agrawal 在文献中提出了集中式数据的隐私保护的分类挖掘算法。该算法使用添加随机偏移量的方法达到隐私保护的目的,随后利用贝叶斯公式来重建判定树对原始数据的支持度进行重构来完成集中式数据隐私保护的分类挖掘,在此过程中,通过给值域空间分区来加快计算速度。但是原数据要明确给出,带来隐私泄露的隐患是此方法的缺陷,同时此算法中推出原始数据的密度函数使用的迭代法计算量特别大,并且要求原数据均匀分布,只适合均匀分布的数据类型,所以此算法仍然有进一步改进的余地。

外国学者 warner 在文献中成功使用随机响应技术来解决涉及隐私的问题调查问卷,先选择一个群体填写与自身隐私有关的特性调查问卷来估计群体中拥有这种特性比例,由于被调查人群可能会拒绝填写或者直接填写错误答案,于是 warner 提出了两种经典模型并圆满解决了这个问题。Wenliang Du 等在文献中提出了基于随机响应技术的隐私保护分类挖掘算法,将该技术扩展到了需要处理多个布尔属性的分类挖掘上,不过此算法隐私保护度差,并只能适用于布尔类型的数据。

4.2　分布式数据的隐私保护分类挖掘

分布式数据的隐私保护分类挖掘利用 SMC 协议已在金融领域中应用,已取得明显的效果,文献中是将隐私保护方法和分类中的决策树方法结合在一起,在保护金融体系中的隐私信息的前提下来识别洗钱交易行为,是应用成功的先例。

5　隐私保护的聚类挖掘研究

对数据间的距离进行安全地计算是基于隐私保护的分布式聚类的重中之重,目前研究的基于距离的隐私保护聚类挖掘方法居多,在这类挖掘中,早已有文献将 SMC 协议运用到欧几里得距离,可以精确地得到欧几里得距离,并且由于 SMC 的不可逆,从而可以有效地保护隐私。Stanley R. M. Olvieria 在文献中提出了一种变换原始数据的方法即旋转变换方法来保持变换后数据间的相对距离,来进行隐私保护的聚类挖掘。

隐私保护的聚类挖掘方法有几种常用的模型[11]:Naive 隐私保护聚类模型:所有站点的信息必须无条件地交给第三方,当然第三方至少需要是半诚实模型的,与此同时运用加密算法进行聚类并返回结果。隐私保护的多次聚类模型:各个站点的信息不用交由第三方,而是首先对本站点的数据进行隐私保护聚类并返回结果,随后对这些结果要进行二次处理来实现隐私保护分布式聚类[12]。有的文献很早就已经实现了两次聚类,在以后的研究中可以多在多次聚类方面进行研究,不过要注意算法效率问题。除了上述常用的两种方法外,还有经典的任意划分数据的环境下的 k-mean、期望最大化隐私保护聚类算法,一维空间以及后来逐渐发展的多维空间的数据等距变换等隐私保护聚类算法,都能有效地解决隐私保护聚类挖掘中的数据类型适用性不强等缺点。

6　结束语

文中首先介绍了现阶段隐私保护数据挖掘算法的研究现状和算法中存在的问题,有的算法是以牺牲空间效率为代价来提升隐私保护度的。上述章节中明确了目前一些算法的优缺点,为以后算法的改进明确了方向。特别是基于隐私保护的分类挖掘算法和聚类算法的研究,随着社会中的信息技术越来越发达,个人和集体隐私问题也越来越受重视,这时隐私保护挖掘算法就具有了非常大的应用前景,也是我们以后的研究热点。

一个算法只有真正在现实生活中被应用,才会具有旺盛的生命力,而每一个算法的提出特别是验证都需要一个严谨的过程,所以提出一个算法其实并不是真正的难点,重要的是在提出之后需要先用理论验证它的性能、隐私保护度及实用性,并对其进行合理的评价,同时还要在现实中有广泛的应用才能体现它的价值。正所谓实践是检验真理的唯一标准,只有经过实践检验的算法才是真正的好算法。

参考文献:

[1] 王爱平,王占凤,陶嗣干,等. 数据挖掘中常用关联规则挖掘算法[J]. 计算机技术与发展,2010,20(4):105-108.

[2] 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: 50-57.

[3] 王　锐,刘　杰. 隐私保护关联规则挖掘算法的研究[J]. 计算机工程与应用,2009,45(26): 126-130.

[4] Verykios V S, Elmagarmid A, Bertino E, et al. Association Rule Hiding[J]. IEEE Trans. on Knowledge and Data Engineering, 2006,16(4): 434-447.

(下转第226页)

CPLD 成功地解决了高速 DSP 与低速 LCD 之间通信接口问题,该方法充分利用了 CPLD 的片上资源,硬件结构简单,CPLD 编程容易,为高速处理器与低速外设之间实现高速、大数据量传输问题提供了一种有效的解决方案。文中的设计方案已经成功应用于某型便携式数字超声波探伤系统,并取得了良好的使用效果。

参考文献:

[1] 王金友. 用 CPLD 实现 DSP 与外设芯片的速度匹配[J]. 电子测量技术,2006,29(4):73-75.

[2] 彭志刚,董金明. CPLD 在 DSP 设计中的应用[J]. 电子测量技术,2004(4):55-56.

[3] 周 博,杨 超,司锡才. 用 FPGA 实现 DSP 与液晶显示器的快速接口[J]. 电子技术应用,2003 (4):73-75.

[4] 杨旭光,丁铁夫,刘维亚,等. 基于 FPGA 的高速 DSP 与液晶模块接口的设计[J]. 液晶与显示, 2007 ,22(3):342-345.

[5] XC95144XL High Performance CPLD[M]. [s. l.]: Xilinx Inc,2001.

[6] 薛红娟,江海河,张飞军. 基于 DSP 与 CPLD 的液晶模块的设计[J]. 微计算机信息,2009(23):114-116.

[7] 康 岭,朱齐丹. 基于 CPLD 的 DSP 与 LCD 接口的设计与实现[J]. 应用科学,2008,35(5):60-64.

[8] TMS320C6713 Data Sheet[M]. [s. l.]:TI, 2004.

[9] 陈亚萍,陈 明. 基于 DSP 和 CPLD 的液晶显示控制器的设计[J]. 计算机测量与控制,2007,15(4):482-484.

[10] 李方慧,王 飞,何佩琨. TMS320C6000 系列 DSPs 原理与应用[M]. 第 2 版. 北京:电子工业出版社,2003.

[11] TMS320C6000 DSP McBSP Reference Guide[M]. [s. l.]: TI,2004.

[12] NHC_34 彩色液晶显示控制模块使用说明[M]. 北京:北京宁和电子科技发展有限公司,2004.

[13] 云创工作室,詹仙宁,田 耕. VHDL 开发精解与实例剖析[M]. 北京:电子工业出版社,2009.

(上接第 159 页)

[5] Oliveira S R M, Zaiane O R. Privacy Preserving Frequent Itemset Mining[C]//Proc. of the IEEE ICDM Workshop on Privacy, Security and Data Mining. Maeoasm, Australian: IEEE Computer Society, 2004.

[6] Oliveira S R M, Zaiane O R. Algorithms for Balancing Privacy and Knowledge Discovery in Association Rule Mining [C]//Proc. of the 7th Int'l Database Engineering and Application Symp. Hong Kong, China: IEEE Computer Society, 2005.

[7] 张 瑞,郑 诚. 基于隐私保护的关联规则挖掘算法[J]. 软件技术与数据库,2009,35(4): 78-82.

[8] 周志纯. 隐私保护数据挖掘研究[D]. 合肥:合肥工业大学,2008.

[9] 汪晓刚,惠 蕙,孙志挥. 基于共享的隐私保护关联规则挖掘[J]. 软件导刊,2009,8(9):150-153.

[10] 黄高琴. 基于隐私保护的分布式关联规则数据挖掘[J]. 微计算机信息,2009,25(9):99-100.

[11] 周水庚,李 丰,陶宇飞,等. 面向数据库应用的隐私保护研究综述[J]. 计算机学报,2009,32(5):847-858.

[12] Jagannathan G, Pillaipakkamnatt K, Wright R N. A new privacy-preserving distributed k-clustering algorithm[C]//Proceedings of the 2006 SIAM International Conference on Data Mining(SDM). Bethesda, Maryland:[s. n.], 2006: 492-496.

(上接第 222 页)

[3] Hsiao C H, Lin C T, Cassidy M. Application of Fuzzy Logic and Neural Networks to Automatically Detect Freeway Traffic Incidents[J]. Journal of Transportation Engineering, 1994, 120(5):753-772.

[4] Magistretti E, Lee U, Gerla M, et al. Smart mobs for urban monitoring with a vehicular sensor network[J]. IEEE Wireless Comm,2006, 13(5):158-162.

[5] 贾元华. 高速公路交通事件自动检测系统结构框架[J]. 佳木斯大学学报(自然科学版), 2004, 22(2):242-246.

[6] 罗晓辉,王忠仁. 遥感微波检测器(RTMS)简介[J]. 公路交通科技,1997, 14(4):62-64.

[7] 朱 茵,王军利,周彤梅. 智能交通系统导论[M]. 北京:中国人民公安大学出版社, 2007.

[8] 陈德望,高海军,陈 龙,等. 城市高速道路微波检测器 RTMS 的检测精度分析[J]. 公路交通科技,2002,19(5): 122-124.

[9] 刘廷新. 高速公路监控通信管理[M]. 北京:人民交通出版社,2005.

[10] Tang S M, Gao H J. Traffic-incident detection-algorithm based on nonparametric regression[J]. IEEE Transactions on Intelligent Transportation Systems,2005(6):38-42.

[11] 刘伟铭. 高速公路系统控制方法[M]. 北京:人民交通出版社,1998.

[12] 史新宏,蔡伯根. 高速公路自动事件检测算法[J]. 交通运输系统工程与信息, 2001, 1(4):306-310.

[13] 邓毅萍. 高速公路路段运行状况评价与分析研究[D]. 南京:东南大学, 2005.