隐私保护数据挖掘研究进展*

张海涛，黄慧慧†，徐　亮，高莎莎
（南京邮电大学 地理与生物信息学院，南京 210003）

摘　要：近年来隐私保护数据挖掘已经成为数据挖掘的研究热点，并取得了丰富的研究成果。但是，随着移动通信、嵌入式、定位等技术的发展与物联网、位置服务、基于位置的社交网络等应用的出现，具有个人隐私的信息内容更加丰富，利用数据挖掘工具对数据进行综合分析更容易侵犯个人隐私。针对新的应用需求，对隐私保护数据挖掘方法进行深入研究具有重要的现实意义。在分析现有的隐私保护数据挖掘方法分类与技术特点的基础上，提出现有方法并应用于新型分布式系统架构应用系统、高维数据及时空数据等领域存在的挑战性问题，并指出了今后研究的方向。

关键词：隐私保护数据挖掘；新型分布式系统；高维数据；时空数据
中图分类号：TP208　　**文献标志码**：A　　**文章编号**：1001-3695(2013)12-3529-07
doi:10.3969/j.issn.1001-3695.2013.12.003

Research advances on privacy-preserving data mining

ZHANG Hai-tao, HUANG Hui-hui†, XU Liang, GAO Sha-sha
(*College of Geographic & Biologic Information, Nanjing University of Posts & Telecommunications, Nanjing* 210003, *China*)

Abstract: In recent years, the privacy-preserving data mining has become a hotspot in data mining. However, with the development of technologies (mobile communication, embedded and positioning technologies, etc), the emerging applications (the Internet of things, location-based services, social network based on location, etc) result in accumulation of abundant personal privacy information, which will easily lead to the violence of personal privacy. Therefore, it is significant to study the privacy-preserving data mining methods to meet the demands of new appplications. In view of the analyzing the characteristics and catalogs of existing privacy-preserving data mining methods, this paper proposed their challenges from the field of new distributed system, high dimensional data and spatio-temporal data, etc, as well as indicatef the future research directions.

Key words: privacy-preserving data mining; new distributed system; high dimensional data; spatio-tempral data

计算机处理能力、存储技术和网络技术的发展带来大量的数据。数据挖掘作为一个强有力的分析工具，可以从海量数据中提取隐藏的、有用的数据和知识，从而为科学、医学、商业研究等作出重要贡献[1]。然而，由于挖掘的数据中可能包含敏感数据或知识，也会对隐私和信息安全构成威胁。在 1995 年召开的第一届 KDD(knowledge discovery in datasets)会议上，隐私保护数据挖掘的概念被首次提出。1999 年 Rakesh Aggrawal 在 KDD 会议上提出将隐私保护数据挖掘作为数据挖掘领域未来研究的重点之一。目前，隐私保护数据挖掘已经取得了丰富的研究成果。

随着移动通信、嵌入式、定位等技术的发展与物联网、位置服务、基于位置的社交网络等应用的出现，使得包含个人隐私的数据量急剧增加、信息类型更加丰富。利用数据挖掘工具进行数据综合分析更容易侵犯个人隐私，从而会使得传统的隐私保护数据挖掘方法变得无效。对现有隐私保护数据挖掘方法进行深入研究，分析现有的隐私保护数据挖掘方法分类与技术特点以及其应用于当前新技术与应用存在的问题，具有重要现实意义。

1　隐私

1.1　隐私的定义

在数据挖掘领域，隐私一般被分成个人隐私和共同隐私两类[2]。个人隐私主要指不愿意被收集、发布的原始个人数据，如用户账号、密码、身份证号等；共同隐私是指不为个体所拥有，反映一类人或事物的共同信息或模式。共同隐私通常只有使用数据挖掘工具才能获取。隐私保护数据挖掘的目标是通过对原始数据进行处理，同时实现个人隐私与共同隐私的保护。

1.2　隐私和效用

隐私和效用是衡量隐私保护数据挖掘的两个重要指标，通常需要寻求两者的最佳平衡点，既要避免攻击者通过数据处理进行个体的重新识别，又要实现丢失信息的最少化。为此，需要设计相应的隐私度量方法。

1.3　隐私的度量

隐私保护效果可以通过攻击者披露隐私的程度来侧面反映。因此，隐私度量可以用披露风险来描述[3]，即攻击者通过

收稿日期：2013-04-25；**修回日期**：2013-06-08　　**基金项目**：国家自然科学基金资助项目(41201465)；江苏省自然科学基金资助项目(BK2012439)；2010 年江苏政府留学奖学金资助项目

作者简介：张海涛(1978-)，男，副教授，博士，主要研究方向为移动 GIS 理论方法与关键技术、时空数据挖掘、LBS 隐私保护等；黄慧慧(1990-)，女(通信作者)，硕士研究生，主要研究方向为空间信息系统、隐私保护的数据挖掘(St_Hillary@126.com)；徐亮(1989-)，男，硕士研究生，主要研究方向为空间信息系统、时空推理与隐私保护；高莎莎(1988-)，女，硕士研究生，主要研究方向为空间信息系统、时空数据挖掘.

对发布数据与相关背景知识的分析,披露隐私的概率。通常,相关背景知识越多,披露风险越大。具体计算如式(1)[4]所示。

$$r(s,K)=P_r(S_k) \tag{1}$$

其中:s 表示敏感数据,事件 S_k 表示攻击者基于背景知识 K 揭露敏感数据 s。若数据所有者最终发布的数据集 D 中敏感数据的披露风险小于指定的阈值 α,$\alpha\in[0,1]$,则称数据集 D 的披露风险为 α。

1.4 攻击及防御

数据性质直接决定隐私保护技术的实现方法,文献[5]基于对数据形式、性质以及攻击模型的特殊假设,提出了面向目的的隐私保护框架。具体实现过程是明确攻击者可了解到的背景知识,并基于背景知识设计攻击模型,最后确定相应的防御方法。

2 隐私保护数据挖掘方法

Verykios 等人[3]把主流隐私保护数据挖掘方法分为五类:a)以数据的分布方式,分为集中式和分布式方法,其中分布式又分为水平分布与垂直分布;b)以数据或规则的隐藏(修改)方式,分为基于数据失真、数据匿名、数据加密等;c)以数据挖掘技术层面,分为针对关联规则挖掘、分类挖掘、聚类挖掘等;d)以隐藏的对象来看,分为原始数据隐藏、规则或模式隐藏等;e)以隐私保护技术层面,分为基于启发式、基于密码学以及基于重构技术的方法。周水庚等人[5]将隐私保护数据挖掘方法分为四个研究方向:通用的隐私保护技术、面向数据挖掘的隐私保护技术、基于隐私保护的数据发布技术与隐私保护算法,如表 1 所示。其中,通用的隐私保护技术方向研究基于概率统计的算法,且将其应用于较低的系统应用层;面向数据挖掘的隐私保护技术方向研究针对不同的数据挖掘类型(关联、分类、聚类等)的算法模型,且将其应用于较高的系统应用层;基于隐私保护的数据发布技术方向主要研究共享发布数据应用中的隐私保护方法;隐私保护算法方向主要研究优化算法性能,以及提高隐私保护数据可用性的技术。

表 1 隐私保护的研究方向[4]

研究方向	实现算法
通用的隐私保护技术	perturbation、randomization、swapping、encryption
面向数据挖掘的隐私保护技术	association rule mining、classification、clustering
基于隐私保护的数据发布技术	k-anonymity、l-diversity、m-invariance、t-closeness
隐私保护算法	anonymized publication、anonymization with high utility

本文通过对近年来相关研究成果的分析,从隐私保护数据修改的角度将现有的主要隐私保护数据挖掘方法分为基于加密技术方法、基于数据失真方法、基于数据匿名方法三类。

2.1 基于加密技术的方法

在数据挖掘过程中采用加密技术隐藏敏感数据的方法,主要应用于分布式计算环境。安全多方计算(secure multiparty computation,SMC)[6,7]是此类方法的典型代表。

2.1.1 安全多方计算

SMC 主要用于两个或多个互不信任的参与方之间进行隐私保护的协同计算。输入的独立性、计算的正确性是 SMC 的基本准则[8]。典型 SMC 模型包括参与方、安全性定义、通信网模型以及信息论安全与密码学安全等。

SMC 具体应用主要依赖于数据存储模式、站点可信度及站点行为。其中,根据所在站点的攻击者是否遵守相关计算协议,被分为半诚信攻击者和恶意攻击者。半诚信攻击者是遵守相关计算协议但仍试图进行攻击的站点;恶意攻击者是不遵守协议且试图披露隐私的站点。一般假设所有站点为准诚信攻击者。目前,SMC 方法主要应用在分布式关联规则、聚类与分类挖掘三个方面。

2.1.2 基于 SMC 的分布式关联规则挖掘

在分布式环境下,关联规则挖掘的关键是项集的全局计数,SMC 保证项集计数不会泄露隐私信息。在数据水平划分的分布式环境,项集的全局支持度是局部支持度之和,可由局部频繁项集产生全局频繁项集[9,10]。在数据垂直划分的分布式环境,项集计数的问题被简化为在隐私保护的同时,计算不同站点间的标量积[11,12]。

2.1.3 基于 SMC 的分布式聚类挖掘

该类方法的关键是安全地计算数据间的距离。由于 SMC 过程的不可逆性,将 SMC 协议运用于欧几里德距离的计算可以实现相应的隐私保护。文献[4]提出了基于 SMC 的分布式聚类挖掘的两种模型:a)Naïve 聚类模型,各个站点将加密数据安全地传递给信任的第三方,由第三方进行聚类并将聚类结果返回给各个站点;b)多次聚类模型,首先各个站点对本地数据进行聚类并发布结果,然后第三方对各个站点发布的结果进行二次处理,最终实现分布式聚类。文献[13]基于 SMC 的求平均值协议,提出了应用于垂直分布数据的 K-means 聚类方法。文献[7,14]基于 SMC 的 secure-sum 和 secure-means 技术,提出了可以同时用于与垂直分布和水平分布数据的 K-means 聚类方法。

2.1.4 基于 SMC 的分布式分类挖掘

在分布式环境下,分类挖掘的关键是生成分布式决策树,加密技术和决策树方法的结合以实现分类挖掘结果的隐私保护。ID3 和 C4.5 通常作为此类算法的原型。文献[15]基于 SMC 标量积协议,提出了应用于垂直分布数据的隐私保护 ID3 算法。文献[16,17]基于同态加密 SMC 协议,提出了应用于水平分布数据的隐私保护的 C4.5 算法。

2.2 基于数据失真的方法

该类方法要求在特定的数据或属性保持不变的前提下,对隐私敏感数据进行失真处理。这既要求失真处理的原始数据不能被重构,又要保证失真处理的数据具有原始数据的统计特性。基于数据失真的隐私保护数据挖掘方法主要包括随机化、阻塞、变形等。

2.2.1 随机化

随机扰动技术包括加性干扰和乘性干扰。目前,基于随机扰动技术的隐私保护数据挖掘主要包括关联规则与分类的挖掘两个方面。随机化扰动技术可以在不暴露原始数据的情况下进行多种数据挖掘。文献[18]通过往原始数据注入大量伪项,实现了频繁项集的隐藏,再通过在随机扰动后的数据上估计项集支持度,从而发现规则。文献[19]通过对随机干扰数据的重构,设计了高效的分类挖掘算法,利用重构数据的分布进行决策树分类器训练后,得到的决策树能很好地对数据进行分类。

2.2.2 阻塞

不同于随机化,阻塞对原始数据的修改并不引入虚假的噪声数据,而只对原始数据进行泛化模糊处理,如将某些特定数值替换为“?”。目前,该种方法主要应用于关联规则的隐藏[18,20]。基本原理是:将数据表中的某些特定数值替换为“?”,使得项集计数变为最小、最大估计值区间范围内的一个不确定值,当区间范围的下界取值小于设定的阈值时即可实现关联规则的隐藏。示例过程如图 1 所示。其中,“1”表示在事务中出现对应的属性,“0”表示不出现。$A\rightarrow D$ 表示数据集中的敏感规则,其置信度为 0.75(3/4)。当用“?”替换了数据集中第 2 个事务的属性 D 和第 3 个事务的属性 A 后,$A\rightarrow D$ 的置信度在区间范围[0.4,0.75](2/5 ~ 3/4)进行取值,如果这一区间下界值 0.4 小于预定的阈值,即实现规则 $A\rightarrow D$ 的隐藏。

在基于阻塞技术的隐藏规则隐藏方面,最早由 Saygin 等人提出了 GIH 算法,该算法是针对生成模式的项集,通过减少其支持度范围的下边界完成规则的隐藏,选择长度最短、支持度区间下界最大的事务对其阻塞来降低对其他项集的不利影响。Saygin 等人同时还提出了 CR 算法和 CR2 算法,都是通过降低置信度隐藏规则,与 GIH 算法不同的是后两种算法更侧重规则的右件的处理。Hintoglu 等人在 Saygin 提出算法的基础上提出了改进的 MCR 和 MCR2 算法,引入了量化的估算函数来计算每一步的信息损失帮助选择损失小的阻塞方式,这种方法的特点是在减小了新旧数据库差异的同时,增加了使用评估函数计算的时间开销。

2.2.3 变形

变形技术通过对数据表中特定数值进行取反操作,实现敏感信息的隐藏。目前,该种方法也主要应用于关联规则的隐藏[21~23]。其基本原理是:a)用布尔矩阵表示事务数据库中的数据,如布尔矩阵中的“1”和“0”表示对应属性在事务数据库中是否出现;b)将敏感事务对应在布尔矩阵中的数值进行取反操作,同时修改和过滤原有事务的属性,最终使敏感规则的支持度和置信度低于设定的阈值。示例过程如图 2 所示,假设 $A\rightarrow B$ 为敏感规则,其支持度为 0.8(4/5),置信度为 1(4/4);为实现该规则的隐藏,分别选择第 2、4 事务的属性 B 对其值进行取反操作,使得规则支持度、置信度分别调整为 0.4(2/5)、0.5(2/4)。如果它们分别低于设定阈值,即可实现规则 $A\rightarrow B$ 的隐藏。

原始数据集

TID	A	B	C	D
1	1	1	0	0
2	1	0	1	1
3	0	0	1	0
4	1	1	0	1
5	1	0	1	1

阻塞算法

修改后数据集

TID	A	B	C	D
1	1	1	0	0
2	1	0	1	?
3	?	0	1	0
4	1	1	0	1
5	1	0	1	1

图 1　阻塞技术

原始数据集

TID	A	B	C	D
1	1	1	1	0
2	1	1	0	1
3	0	0	0	1
4	1	1	0	1
5	1	1	1	0

变形算法

修改后数据集

TID	A	B	C	D
1	1	1	1	0
2	1	0	0	1
3	0	0	0	1
4	1	0	0	1
5	1	1	1	0

图 2　变形技术

基于变形技术的敏感关联规则隐藏的算法最早由 Dasseni、Verykios 等人在 2001 年提出,但其设计的 Algo1a、Algo1b、Algo2a、Algo2b 及 Algo2c 算法[3,24],均存在时间开销很大、效率不高的问题。Algo1a 算法通过增加规则左件支持度的方式来降低规则置信度,在选择事务时以包含项目最多为标准;Algo1b 算法与 Algo1a 类似,不同的是对规则的右件更感兴趣,在事务的选择上更倾向于含有整条规则的事务,选择了最短的事务来减少对其他项集的影响;而 Algo2a、Algo2b 及 Algo2c 算法则是从规则的大项集入手,调整大项集中项的数量来控制支持度,过滤规则,这些算法的特点是以最小支持度和最小置信度达到阈值为标准,缺点是不能处理规则间有交集的情况[22]。这主要因为这些方法都假设规则间不存在交集项,这样每次操作只针对单一规则,要实现多条规则的隐藏需要多次扫描原始数据库。

为此,Oliveira 等人提出了 Naïve、MinFIA、MaxFIA、IGA、SWA 等算法[22,23]。Naïve 算法通过引入“冲突度”有效解决了上述敏感规则存在交集的问题,同时通过引入“公开度”,还可实现对过滤敏感事务数量的控制,该算法保护了数据集中出现次数多的项,牺牲了冲突度高的项以保持数据库中事务数目的稳定。MinFIA 和 MaxFIA 算法选用冲突度小的事务并分别选择支持度最小和最大的项作为候选。IGA 算法结合冲突度和公开度,利用敏感规则分组的方式处理规则间相交的问题,保证了非敏感规则的可用性。SWA 算法不同于以上算法,以窗口的形式对事务进行扫描,选择出现频率高的数据项和短作业,不必一次将所有数据读入内存。

阻塞和变形技术有一个无法避免的缺点,即针对不同的应用需要设计特定的算法对转换后的数据进行处理,因为所有的应用都需要重建数据的分布。基于此,文献[25]提出了凝聚技术,它将原始数据记录分成组,每一组内存储着由 k 条记录产生的统计信息,包括每个属性的均值、协方差等。这样,只要是采用凝聚技术处理的数据,都可以用通用的重构算法进行处理,并且重构后的记录并不会披露原始记录的隐私,因为同一组内的 k 条记录是两两不可区分的。

2.3 基于数据匿名的方法

匿名本意上是指无名或署名不明,隐私研究者给予了它更为明确的定义[26]。Pfitzmann 等人[27]提出,匿名是在一组对象集中单个个体无法被识别的状态,这个对象集即为匿名集。

抑制和泛化是实现数据匿名的两种主要方法。抑制方法不发布特定数据项,泛化方法对数据进行概括与抽象操作。数据匿名主要针对数据的标志属性、准标志属性或关键属性以及敏感属性。

2.3.1 k-匿名

k-匿名最早由 Sweeney 提出[28],其基本准则是:匿名处理后的数据,其每条记录的准标志属性要与其他 $k-1$ 条记录的准标志属性不可区分。这样,数据挖掘者不能区分出隐私信息所属的个体,从而实现个体隐私的保护。将 k-匿名运用于位置隐私保护,当考虑了一个具有位置信息的对象是 k-匿名的,当且仅当该对象的位置信息与其他至少 $k-1$ 个其他对象的位置信息不可区分。

k-匿名有效地解决了链接攻击问题,但没对数据的敏感属性作任何约束。攻击者可通过对相关背景知识与 k-匿名处理后数据的关联分析,建立敏感数据与个体的联系,从而实现对个体隐私的重标志攻击。另外,攻击者也可以通过分析 k-匿名处理后数据存在的敏感信息等价类与准标志属性等价类的关联关系,实施同质性攻击[24]。

文献[24]提出了 *l*-多样性方法。*l*-多样性对 *k*-匿名进行了扩展,要求每一个等价类的敏感属性至少有 *l* 个不同值。*l*-多样性使得攻击者最多以 1/*l* 的概率识别某一个体的敏感信息,从而阻止了重标志攻击与同质性攻击。*l*-多样性也存在两方面的问题:a)当数据值过大,而 *l* 值较小时,等价类的数量急剧增加;b)如果两个等价类的敏感属性值差异过大,则很难确定敏感属性的敏感度[8]。为此,文献[29]提出了 *t*-闭合方法。*t*-闭合在 *l*-多样性的基础上进一步分析敏感属性的分布,要求所有等价类的敏感属性值其分布尽可能接近该属性的全局分布,具体的差异限值为 *t*。

k-匿名、*l*-多样性以及 *t*-闭合方法的共同特点是:*k*(*l*,*t*)值越大隐私保护效果越好,但丢失信息也越多,数据的精度和利用率也会受到影响。

2.3.2 NP 问题及解决方法

泛化和抑制是 *k*-匿名算法实现方法,对泛化空间和抑制策略的搜索直接影响到算法的性能。但是在大多数情况下,*k*-匿名最优化已经证明是 NP 问题,主要的解决方法是设计高效的近似算法。

文献[30]提出了基于逐步搜索泛化空间、同时选择局部最优,直至满足 *k*-匿名的 MinGen 算法。但是,MinGen 算法由于采用完全搜索,时间复杂度高,因此不实用。Datafly 算法在 MinGen 算法的基础上,通过引入抑制与启发式泛化指导原则,解决了时间复杂度高的问题,进一步提升了算法的效率。此外,LeFevre 等人提出了基于全域泛化技术的 Incognito 方法[31],该算法首先构建包含所有全域泛化(一种全局重编码技术)方案的泛化图,然后自底向上对原始数据进行泛化,每次选取最优泛化方案前,预先对泛化图进行修剪以缩小搜索范围。修剪的原则是如果一个节点的泛化层次满足 *k*-匿名,则它的所有泛化子集满足 *k*-匿名,因此不需要在随后的层次访问中检查它们。不断进行以上操作直到数据满足 *k*-匿名原则。文献[5]提出了可以实现多维 *k*-匿名的 Mondrian 算法,能够发布精度较高的数据,它基于多维重编码技术将原始数据映射到一个多维空间,*k*-匿名问题即转换为在空间中对多维数据进行最优化划分的问题。

实现其他匿名化原则的算法大多是基于 *k*-匿名算法,不同之处在于判断算法结束的条件,而泛化策略、对搜索空间的修剪等都是基本相同的。

2.3.3 基于 *k*-匿名的挖掘方法

基于 *k*-匿名的数据挖掘有两种实现模式[20]:匿名—挖掘(AM)和挖掘—匿名(MA),如图 3 所示。其中,PT 表示私有表;PT_k 表示对 PT 进行 *k*-匿名的结果;MD 表示直接对 PT 挖掘的结果;MD_{k1} 表示对 PT_k 挖掘的结果;MD_{k2} 表示对 MD 进行 *k*-匿名的结果;MD_{k3} 表示在 *k*-匿名约束下对 PT 的挖掘结果;图中虚线及虚线框表示数据或数据处理过程只能由数据持有者访问或执行。

1)匿名—挖掘(AM)模式

AM 模式首先对原始的私有数据表(PT)进行 *k*-匿名处理并将结果保存在数据表(PT_k)中;然后发布数据表(PT_k);最后,数据执行者或外部各方对此表进行数据挖掘,并将挖掘后的结果保存到数据表(MD_{k1})中。AM 模式基于(PT_k)而非(PT)执行数据挖掘,因此对数据挖掘结果的分析不会侵犯 *k*-匿名处理前数据(PT)中的隐私。另外,挖掘的操作并不限于数据持有者,可以提供给任何第三方,这可以大大增加匿名集数据的使用范围。AM 方法的缺点是:由于挖掘操作针对的是匿名处理后的数据,挖掘结果的效用会受到一定的影响。文献[32,33]基于 AM 方法,分别提出了自顶向下和自底向上的分类挖掘算法。

2)挖掘—匿名(MA)模式

a)MA 模式包括两种实现过程:先对原始数据表(PT)进行挖掘,并把挖掘结果保存在数据表(MD)中;然后对数据表(MD)进行匿名处理,并把处理结果保存在数据表(MD_{k2})中。

b)设计新的实现 *k*-匿名保护的数据挖掘算法,对原始数据表(PT)同时进行 *k*-匿名处理与数据挖掘操作,处理结果保存在数据表(MD_{k3})中。

两种方法的不同点是,前者的数据挖掘过程可以直接使用已有的数据挖掘工具,而后者需要设计新的方法。

MA 模式的两种方法与 AM 模式的方法不同点是,数据挖掘的过程只能由数据持有者执行。文献[20]分别给出了基于 AM、MA 模式进行关联规则与决策树挖掘的相应算法。

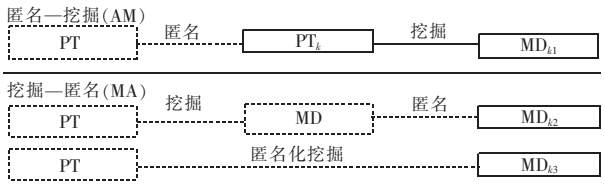

图 3 基于 *k*-匿名的挖掘方法

2.4 算法性能评估

综上分析,基于加密技术、数据失真以及数据匿名的三类隐私保护数据挖掘方法均是针对特定的应用目的而设计,因此很难设计统一的性能评估方法。现有的评估方法主要从以下四个方面开展:a)隐私的保护程度,通常使用“披露风险”计算[4],披露风险越小,隐私保护程度越高;b)数据的效用,用于度量发布的隐私保护数据的质量,数据缺损越高,信息丢失越多,数据效用越低;c)算法性能,数据计算开销(时间复杂度)与数据通信开销是对算法性能度量的重要指标;d)算法的适应性,是描述算法适应不同应用环境的重要指标。

2.5 隐私保护数据挖掘方法对比

三类隐私保护数据挖掘方法具有不同的特点,在不同应用需求下,它们的适用范围、性能表现等不尽相同。从表 2 可以看出,基于加密技术的方法具有高隐私保护程度及数据效用,数据缺损最少、真实可用,当用户更关注于数据挖掘时的隐私保护甚至要求达到完美保护时,应考虑基于加密技术的方法。但是,该方法的代价是计算开销较高,即具有较高的时间复杂度,且在分布式环境下通信开销也随之增加。当针对特定数据实现隐私保护数据挖掘且要求计算开销较低时,基于数据失真的方法更加适合;而基于数据匿名的方法能以较低的计算开销和信息缺损实现对数据较高的隐私保护程度。

表 2 隐私保护数据挖掘方法的性能评估

方法	隐私保护程度	数据效用	计算开销	通信开销
加密技术	高	高	高	高
数据失真	中	低	低	低
数据匿名	高	中	中	低

表 3 从典型的隐私保护技术、主要应用以及优缺点对隐私保护数据挖掘方法进行了进一步的对比分析。基于加密技术的方法主要应用在分布式环境中,利用 SMC 技术实现敏感数

据的隐藏;但该方法在实际施行中,部署难度较大,应用难度较高。基于数据失真的方法可以应用于各类数据挖掘中,典型应用有关联规则挖掘和分类挖掘,该方法容易实现,但严重依赖于数据,需要为特定的数据设计特定的算法,且数据缺损较大。基于数据匿名的方法能够实现数据匿名化,在匿名化数据的基础上进行各种数据挖掘,相较于其他两种方法,各方面表现较为平衡。

表 3　隐私保护数据挖掘方法对比分析

方法分类	隐私保护技术	典型应用	优点	缺点
加密技术	SMC	分布式关联规则挖掘,分布式分类挖掘,分布式聚类挖掘	隐私保护程度高、数据效用高	计算开销及通信开销高、应用难度大
数据失真	随机化阻塞变形	各类隐私保护数据挖掘:关联规则挖掘,决策树分类器构建	容易实现、计算开下及通信开销低	数据效用低、严重依赖于数据
数据匿名	k-匿名 l-多样性 t-闭合	实现数据匿名化、在匿名化数据的基础上进行各种数据挖掘:关联规则挖掘,决策树分类器构建,聚类挖掘	算法通用性较高、通信开销低、能保证数据的真实性	存在一定的数据缺损、存在一定的隐私保护漏洞、实现最优化的计算开销大

3　隐私保护数据挖掘新的挑战

3.1　新型分布式系统架构的隐私保护

物联网、位置服务、基于位置的社交网络等应用均构建在移动互联网、无线自组织网络(Ad hoc)、无线传感网(WSN)等新型的分布式网络架构之上[34~37],而这些应用对协同、交互以及信息传输等方面提出了更高的要求。因此,研究新型分布式系统架构的隐私保护方法既具有重要的现实意义,又是一项极具挑战性的课题。

文献[38]对无线传感器网络的隐私保护进行综述,主要包括数据查询隐私保护、数据聚合隐私保护和位置隐私保护等几个方面,分别介绍了基于加密技术和路由协议技术的隐私保护方法。基于数据加密的保护方法中,通过加密机制实现了他方对原始数据的不可见性以及数据的无损失性,既保证了数据的机密性,又保证了数据的隐私性。路由协议方法主要用于无线传感网中的节点位置隐私保护,无线传感网的无线传输和自组织特性使得传感器节点的位置隐私保护尤为重要[35]。

现有的分布式隐私保护方法大多需要引入一个集中式的第三方服务器,所有信息都由第三方服务器处理,易导致信息阻塞、系统崩溃,并且第三方的可信度也很难保证;如果第三方服务器被攻击,则会泄露用户的隐私。因此文献[39]引入了P2P 的思想,取消了第三方的服务器,由系统中的对等点来充当第三方服务器,从而解决了集中式第三方的处理瓶颈问题。然而,P2P 方式产生大量的网络流量且移动端的计算量加大,会影响用户的使用效果。

此外,新型分布式系统架构的多源异构性使其安全面临巨大的挑战,因此如何建立有效的多网融合的隐私保护模型是今后研究的一个重要方向。

3.2　高维数据的隐私保护

随着移动通信、嵌入式、定位等技术的发展,数据获取能力得到很大提升,数据量以及维数都急剧增加[40]。由于隐私保护者并不能准确了解攻击者掌握的属性信息,其采用的隐私保护方法就需要涵盖所有属性信息。但是目前加密技术、失真技术以及匿名技术还只适合有限的属性信息,处理高维数据会引起所谓的“维数灾难”问题。维数灾难通常是指在涉及到向量的计算问题中,随着维数的增加,计算量呈指数倍增长的一种现象。从理论上看来,当维数较高时,无法实现对高维数据的隐私保护,但由于真实世界中,高维数据往往是稀疏的,因此研究人员主要对稀疏高维数据的隐私保护进行研究。

文献[41]阐述了三种对用于多维数据的隐私保护方法:多关系 k-匿名[42]、多维数据的 l-多样性[43]和 k^m-匿名。多关系 k-匿名将传统 k-匿名扩展到多关系层面的应用,即将高维数据分散到多个关系表中,每个表函数依赖于特定的主体,将复杂的高维数据匿名转换为多个关系表的简单匿名。文献[42]提出用于应对高维问题的基于聚类的 MiRaCle 匿名算法,尽管该算法基于对多关系 k-匿名数据库的严格假定,它仍具有重要意义。该算法的匿名化过程相较于传统的 k-匿名更为高效;其缺点在于将 k-匿名扩展到多关系模式的多维空间需要对原始数据进行大量的转变,这个操作将导致信息丢失。文献[43]提出多维数据的 l-多样性的匿名算法能够保证每个事物具有不同的准标志属性和敏感值。该算法结合了泛化和扰动技术的优点,既防止了泛化在高维数据隐私保护时可能出现的信息丢失,同时它能够保持准标志属性和敏感值间的关系;不足之处在于,该算法的实行需要较大的内存支持。由于上述两种算法并不适用于所有稀疏多维数据,文献[41]提出 k^m-匿名方法。k^m-匿名将所有项作为准标志符同时也作为敏感值,对数据库中的记录,任何掌握了多达记录的 m 个项的攻击者都不能使用这些项从数据库中识别少于 k 个元组。也就是说,由攻击者提出的 m 或者小于 m 的子集查询,应返回多于 k 个记录或者什么都不返回。k^m-匿名对数据作出与上述两种方法截然不同的假设,它不涉及知道所有的准标志符的攻击者以及它并未考虑否定的知识和 l-多样性而言,它提供了一个较弱的隐私保护形式。k^m-匿名适用于敏感和非敏感值区别不明显的数据,它能够实现从具备部分敏感值的攻击者中保护数据,并且信息丢失较少。

这些方法提供了一些基本的解决方法,但是还不能够覆盖所有数据类型以及满足可用性要求。如何设计针对高维数据,更加有效的隐私保护数据挖掘方法也是今后研究的一个主要方向。

3.3　时空数据的隐私保护

移动通信和移动定位技术的快速发展,使得 LBS(location based service)成为移动通信、地理信息等领域近年来一个新兴的研究方向[44~47]。随着 LBS 的深入发展与广泛应用,位置隐私泄露或者被非法使用等隐私安全问题逐渐成为公众关注的焦点,位置隐私保护也成为 LBS 进一步深入发展亟待解决的一项关键问题[48,49]。

早期的关于位置隐私保护的研究主要集中在标准规范与法律法规的制定,但由于存在不够灵活和滞后于技术发展的问题,国内外学者近年来主要侧重于对位置隐私保护技术的研究。主要的技术方法包括直接去标志、使用假名、标志与查询内容相分离等方法,但这些方法都存在一定的隐私保护安全性不足的问题。基于 LBS 查询数据的时空特性,通过对用户运动轨迹的跟踪连接、分类、聚类分析,可以进行用户位置隐私的标志攻击[50]。此后,学术界又出现了基于密钥技术的安全多路计算方法、基于假位置[51]的方法、基于 Hilbert 曲线的空间转换[52]以及 PIR 技术[53]等方法。但这些方法又都存在匿名

数据可用性不足的问题,基于假位置过多地引入了虚假的噪声数据;基于 Hilbert 曲线的空间转换会出现转换前后时空邻近性不一致的错误;PIR 技术采用复杂的加密协议算法,不允许对隐私保护数据进行任何的分析利用。

2003 年由 Gruteser 等人[54]提出的基于时空 *k*-匿名的 LBS 隐私保护方法(简称时空 *k*-匿名),以匿名数据的真实可用、方法实现简洁灵活以及更适合 LBS 移动计算环境等特点,成为近年来研究的主流方向[55]。时空 *k*-匿名方法的基本思想是:提交给 LBS 的查询请求不再包括精确的时空信息,而替换为由时空近邻的 *k* 个移动对象形成的匿名集。时空 *k*-匿名方法实现了 LBS 位置隐私以及标志隐私的同时保护。时空 *k*-匿名方法在快照查询的扩展主要包括查询标志隐私的增强性保护[56]、隐私保护级别的灵活设定[57]、多模式查询的隐私保护[58]、空间网络[59]与分布式传感网络[45]的隐私保护等。而其在连续查询的扩展研究主要包括 memorization、plain KAA、advance KAA[60]、连续查询发送模型[61]以及基于运动模型预测的方法[62]。

具有时空特性是位置信息的一大特点。相对于传统的属性数据、空间数据等,时空数据具有更加复杂的拓扑关系,且具有数据量大、更新频繁、非连续、信息不完整等特性[63]。研究时空数据的隐私保护挖掘算法[64],不仅要保证信息提供者的隐私安全,更要实现服务提供商对数据挖掘分析,以实现提供个性化服务、优化资源管理等辅助决策功能。这也是隐私保护数据挖掘今后研究的又一热点。

4 结束语

隐私保护数据挖掘是近年来学术界新兴的研究课题,国内外大量的文献对此进行了研究,提出了很多有用的方法和算法。本文对隐私保护数据挖掘已取得的研究成果进行分析总结,阐明了现有隐私保护数据挖掘方法的分类,并从加密技术、数据失真和数据匿名三个方面分析各方法及其特点并进行性能对比,最后结合当前的新兴技术与应用,分析了目前存在的主要问题及今后研究的热点领域方向。

参考文献:

[1] HAN Jia-wei, KAMBER M. Data mining: concepts and techniques [M]. 2nd ed. San Francisco: Morgan Kaufmann Publishers, 2006.

[2] CLIFTON C, KANTARCIOUGLU M, VAIDYA J. Defining privacy for data mining[C]//Proc of National Science Foundation Workshop on Next Generation Data Mining. 2002: 126-133.

[3] VERYKIOS V S, BERTINO E, FOVINO I N, *et al*. State-of-the-art in privacy persevering data mining[J]. ACM SIGMOD Record, 2004, 33(1): 50-57.

[4] 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.

[5] BONCHI F, FERRARI E. Privacy-aware knowledge discovery novel application and new techniques[M]. Boca Raton: CRC Press, 2011.

[6] YAO A C. How to generate and exchange secrets[C]// Proc of the 27th IEEE Symposium on Foundations of Computer Science. Washington DC: IEEE Computer Society, 1986: 162-167.

[7] CLIFTON C, KANTARCIOGLU M, VAIDYA J, *et al*. Tools for privacy preserving distributed data mining [J]. ACM SIGKDD Explorations, 2002, 4(2): 28-34.

[8] 刘英华, 杨炳儒, 马楠, 等. 分布式隐私保护数据挖掘研究[J]. 计算机应用研究, 2011, 28(10): 3606-3610.

[9] KANTARCIOGLU M, CLIFTON C. Privacy-preserving distributed mining of association rules on horizontally partitioned data[J]. IEEE Trans on Knowledge and Data Engineering, 2004, 16(9): 1026-1037.

[10] CHEUNG D W, HAN Jia-wei, NG V T, *et al*. A fast distributed algorithm for mining association rules[C]//Proc of the 4th International Conference on Parallel and Distributed Information Systems. 1996: 31-44.

[11] VAIDYA J, CLIFTON C. Privacy preserving association rules mining in vertically partitioned data[C]//Proc of the 8th ACM SIGMOD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2002: 639-644.

[12] 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, 9(8): 150-153.

[13] VAIDYA J, CLIFTON C. Privacy-preserving K-means clustering over vertically partitioned data[C]//Proc of the 9th ACM SIGMOD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2003: 206-215.

[14] JAGANNATHAN G, WRIGHT R N. Privacy preserving distributed K-means clustering over arbitrarily partitioned data [C]//Proc of the 11th ACM SIGMOD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2005: 593-599.

[15] DU Wen-liang, ZHAN Zhi-jun. Building decision tree classifier on private data [C]//Proc of IEEE International Conference on Privacy, Security and Data Mining. 2002: 1-8.

[16] XIAO Ming-jun, HUANG Liu-sheng, LUO Yong-long, *et al*. Privacy preserving ID3 algorithm over horizontally partitioned data[C]//Proc of the 6th International Conference on Parallel and Distributed Computing, Applications and Technologies. 2005: 239-243.

[17] XIAO Ming-jun, HAN Kai, HUANG Liu-sheng, *et al*. Privacy preserving C4.5 algorithm over horizontally partitioned data[C]// Proc of the 5th International Conference on Grid and Cooperative Computing. Washington DC: IEEE Computer Society, 2006: 78-85.

[18] SAYGIN Y, VERYKIOS V S, ELMAGARMID A K. Privacy preserving association rule mining[C]//Proc of the 12th International Workshop on Research Issues in Data Engineering. 2002: 151-158.

[19] AGRAWAL R, SRIKANT R. Privacy preserving data mining[J]. ACM SIGMOD Record, 2000, 29(2): 439-450.

[20] AGGARWAL C C, YU P S. Privacy-preserving data mining: models and algorithms [M]. New York: Springer-Verlag, 2008.

[21] 魏晓辉. 敏感规则隐藏算法的研究[D]. 哈尔滨: 哈尔滨工程大学, 2010.

[22] ZHANG Nan, ZHAO Wei. Distributed privacy preserving information sharing[C]//Proc of the 31st International Conference on Very Large Data Bases (VLDB). 2005: 889-900.

[23] MACHANAVAJJHALA A, GEHRKE J, KIFER D, *et al*. *l*-diversity: privacy beyond *k*-anonymity[C]//Proc of the 22nd International Conference on Data Engineering. 2006: 24-35.

[24] AGRAWAL D, AGGARWAL C C. On the design and quantification of privacy preserving data mining algorithms [C]//Proc of the 20th ACM Symposium on Principles of Database Systems. New York: ACM Press, 2001: 247- 255.

[25] AGGARWAL C C, YU P S. A condensation approach to privacy preserving data mining[C]//Proc of the 9th International Conference on Extending Database Technology. Berlin: Springer-Verlag, 2004: 183-199.

[26] SIMPSONAND J A, WEINER E S C. Oxford English dictionary[M]. 2nd ed. [S. l.]: Clarendon Press, 1989.

[27] PFITZMANN A, KOEHNTOPP M. Anonymity, unobservability, and pseudonymity:a proposal for terminology[C]//Proc of International Workshop on Design Issues in Anonymity and Unobservability. Berlin:Springer-Verlag,2000:1-9.

[28] SWEENEY L. *k*-anonymity:a model for protecting privacy[J]. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems,2002,10(5):557-570.

[29] LI Ning-hui, LI Tian-cheng, VENKATA-SUBRAMANIAN S. *t*-closeness: privacy beyond *k*-anonymity and *l*-diversity [C]//Proc of the 23rd International Conference on Data Engineering. New York: ACM Press,2007:106-115.

[30] SWEENEY L. Achieving *k*-anonymity privacy protection using generalization and suppression[J]. International Journal on Uncertainty,2002,10(5):571-588.

[31] LeFEVRE K, DeWITT D J, RAMAKRISHNAN R. Incognito:efficient full domain *k*-anonymity[C]//Proc of ACM SIGMOD Conference on Management of Data. New York:ACM Press,2005:49-60.

[32] FUNG B C M, WANG Ke, YU P S. Anonymizing classification data for privacy preservation[J]. IEEE Trans on Knowledge and Data Engineering,2007,19(5):711-725.

[33] WANG Ke, YU P S, CHAKRABORTY S. Bottom-up generalization: A data mining solution to privacy protection[C]//Proc of the 4th IEEE International Conference on Data Mining. Washington DC:IEEE Computer Society,2004:249-256.

[34] 钱萍,吴蒙. 物联网隐私保护研究与方法综述[J]. 计算机应用研究,2013,30(1):13-20.

[35] 贾金营,张凤荔. 位置隐私保护技术综述[J]. 计算机应用研究,2013,30(3):641-646.

[36] 谈啸,顾君忠,林欣,等. 基于用户隐私保护的区域多对象聚集问题[J]. 计算机应用,2011,31(9):2389-2394.

[37] 杨煜尧,赵方,罗海勇,等. 一种基于地理位置信息的移动互联网社交模型[J]. 计算机研究与发展,2011,48(S1):307-313.

[38] LI Na, ZHANG Nan, DAS S K, *et al*. Privacy preservation in wireless sensor networks: a state-of-the-art survey[J]. Ad hoc Networks, 2009,7(8):1501-1514.

[39] CHOW C Y, MOKBEL M F, LIU Xuan. A peer-to-peer spatial cloaking algorithm for anonymous location-based services[C]//Pro of the 14th ACM International Symposium on Advances in Geographic Information Systems. New York:ACM Press,2006:171-178.

[40] 孟小峰,慈祥. 大数据管理:概念、技术与挑战[J]. 计算机研究与发展,2013,50(1):146-169.

[41] TERROVITIS M, MAMOULIS N, KALNIS P. Privacy preservation in the Publicatin of spare multidimensional data[M]. London: Taylor and Francis Group,2011:35-56.

[42] NERGIZ M E, CLIFTON C, ERGIZ A E. Multirelational *k*-anonymity [C]//Proc of the 23rd IEEE International Conferenc on Data Engineering. 2007:1417-1421.

[43] GHINITA G, TAO Yu-fei, KALNIS P. On the anonymization of spare high-dimensional data[C]//Proc of the 24th International Conference on Data Engineering. 2008:715-724.

[44] ALI S, TORABI T, ALI H. Location aware business process deployment [C]//Proc of International Conference on Computational Science and Its Applications. Berlin:Springer-Verlag,2006:217-225.

[45] BALDAUF M, DUSTDAR S, ROSENBERG F. A survey on context-aware systems[J]. International Journal of Ad hoc Ubiquitous Computing,2007,2(4):263-277.

[46] 赵文斌,张登荣. 移动计算环境中的地理信息系统[J]. 地理与地理信息科学,2003,19(2):19-23.

[47] 张海涛,闾国年,张书亮,等. 移动 GIS 中 GML 数据压缩技术研究[J]. 地理与地理信息科学,2008,24(5):21-24.

[48] 彭志宇,李善平. 移动环境下 LBS 位置隐私保护[J]. 电子与信息学报,2011,33(5):1211-1216.

[49] KULIK L. Privacy for real-time location-based services[J]. SIGSPATIAL Special,2009,1(2):9-14.

[50] HOH B, GRUTESER M, XIONG Hui, *et al*. Enhancing security and privacy in traffic-monitoring systems[J]. IEEE Pervasive Computing,2006,5(4):38-46.

[51] KIDO H, YANAGISAWA Y, SATOH T. Protection of location privacy using dummies for location-based services[C]//Proc of the 21st International Data Engineering Workshops. Washington DC:IEEE Computer Society,2005.

[52] UM J H, KIM H D, CHANG J W. An advanced cloaking algorithm using hilbert curves for anonymous location based service[C]//Proc of the 2nd International Conference on Social Computing. Washington DC:IEEE Computer Society,2010:1093-1098.

[53] SHANG Ning, GHINITA G, ZHOU Yong-bin, *et al*. Controlling data disclosure in computational PIR protocols[C]//Proc of the 5th ACM Symposium on Information, Computer and Communications Security, 2010:310-313.

[54] GRUTESER M, GRUNWALD D. Anonymous usage of location-based services trough spatial and tmporal cloaking[C]//Proc of the 1st International Conference on Mobile Systems, Applications and Sevices. New York:ACM Press,2003:31-42.

[55] KRUMM J. A survey of computational location privacy[J]. Personal and Ubiquitous Computing,2009,13(6):391-399.

[56] GEDIK B, LIU Ling. Location pivacy in mobile systems:a personalized anonymization model[C]//Proc of the 25th International Conference on Distributed Computing Systems. Washington DC: IEEE Computer Society,2005:620-629.

[57] XU T, CAI Ying. Feeling-based location privacy protection for location-based services[C]//Proc of the 16th ACM Conference on Computer and Communications Security. New York:ACM Pres,2009:348-357.

[58] CHOW C Y, MOKBEL M F, AREF W G. Query processing for location services without compromising privacy[J]. ACM Trans on Database Systems,2009,34(4):1-45.

[59] KU W S, ZIMMERMANN R, PENG W C, *et al*. Privacy protected query processing on spatial networks[C]//Proc of the 23rd International Data Engineering Workshops. Washington DC: IEEE Computer Society,2007:215-220.

[60] CHOW C Y, MOKBEL M F. Enabling private continuous queries for revealed user locations[C]//Proc of the 10th International Conference on Advances in Spatial and Temporal Databases. Berlin: Springer-Verlag,2007:258-275.

[61] 林欣,李善平,杨朝晖. LBS 中连续查询攻击算法及匿名性度量[J]. 软件学报,2009,20(4):1058-1068.

[62] PAN Xiao, MENG Xiao-feng, XU Jian-liang. Distortion-based anonymity for continuous queries in location-based mobile services[C]//Proc of the 17th ACM International Conference on Advances in Geographic Information Systems. New York:ACM Press,2009:256-265.

[63] 刘大有,陈慧灵,齐红,等. 时空数据挖掘研究进展[J]. 计算机研究与发展,2013,50(2):225-239.

[64] 李军怀,高苗,陈晓明,等. 时空特性约束下的数据挖掘隐私保护方法[J]. 计算机工程与应用,2008,44(9):139-142.