隐私保护数据挖掘算法综述*)

陈晓明[1] 李军怀[1] 彭 军[2] 刘海玲[2] 张 璟[1]

(西安理工大学计算机科学与工程学院 西安 710048)[1] (重庆科技学院电子信息工程学院 重庆 400050)[2]

摘 要 如何保护私有信息或敏感知识在挖掘过程中不被泄露,同时能得到较为准确的挖掘结果,目前已经成为数据挖掘研究中的一个很有意义的研究课题。本文通过对当前隐私保护数据挖掘中具有代表性的算法按照数据分布对其中的数据更改方法、数据挖掘算法、数据或规则隐藏等进行了详细阐述,并对各自的优缺点进行了分析和比较,总结出了各种算法的特性。此外,通过对比提出了隐私保护数据挖掘算法的评价标准,即保密性、规则效能、算法复杂性、扩展性,以便在今后的研究中提出新的有效算法。

关键词 数据挖掘,隐私保护算法,数据分布

A Survey of Privacy Preserving Data Mining Algorithms

CHEN Xiao-Ming[1] LI Jun-Huai[1] PENG Jun[2] LIU Hai-Ling[2] ZHANG Jing[1]

(School of Computer Science & Enginnering, Xi'an University of Technology, Xi'an 710048)[1]

(College of Electronic Information Engineering, Chongqing University of Science and Technology, Chongqing 400050)[2]

Abstract There has been a meaningful research problem that how to protect privacy or sensitive information from leaking during data mining process, meanwhileobtain accurate result. This paper summarizes the features of privacy preserving data mining algorithms by analyzing and comparing some representative ones which include data distribution, data modification, data mining algorithms and data or rules hiding in the light of data distribution. Furthermore according to the comparison, some evaluation standards are brought forward to develop new effective algorithms for future research, such as secrecy, rules efficiency, complexity of the algorithm, expandability.

Keywords Data mining, Privacy preserving algorithms, Data distribution

1 引言

随着数据库技术和网络技术的发展,各行各业都积累了大量有用数据。如何从这些数据中提取出对决策有价值的知识,成为当务之急。数据挖掘作为一个强有力的数据分析工具,可以发现数据中潜在的模式和规律(例如一组规则、聚类、决策树、依赖网络或其他方式表示的知识),在商务决策、科学和医学研究等领域做出了巨大的贡献,具有广泛的应用前景。

与此同时,由于被挖掘的资料或数据还包含着许多敏感性的数据或知识,因此必须受到保护,即数据挖掘应该在隐私保护条件下开展。尤其是在现在,数据挖掘以及知识发现技术已经取得了进步,使用这些技术我们可以在海量的资料信息提取出隐藏的、有用的数据或知识,因而更增加了当资料公开给外界时所存在的风险,会对隐私和信息安全构成威胁。

随着越来越多的信息可以电子形式或从网上得到,人们对自己隐私的保密要求变得越来越迫切。调查显示,不管有没有隐私保护措施,17%的上网者不愿意将自身信息提供给网站,而56%的调查者只有在好的隐私保护措施下才愿意提供自身的信息给网站[1]。在这种情况下,如何在保证个人隐私的前提下进行数据挖掘,成了一个需要解决的问题。另外一个方面,在双方或多方合作进行数据挖掘时,由于某种原因,参与者往往不愿意将数据与他人共享而只愿意共享数据挖掘的结果。这种情况在科学研究、医学研究及经济和市场动态研究等方面屡见不鲜[2,3]。

因此,如何保护私有信息或敏感信息在挖掘过程中不被泄露,就成为数据挖掘研究中的一个很有意义的研究课题。

2 隐私保护数据挖掘算法

数据挖掘中的隐私保护主要关注两个方面:其一,像身份、姓名、地址和爱好等敏感的原始数据的处理,避免个人隐私信息的泄露。其二,能通过数据挖掘工具得到的敏感知识也应该被排除。隐私保护数据挖掘的主要目标是使用某种方法对原始数据进行处理,使得私有数据和知识在挖掘之后仍然是私有的。

目前,数据挖掘中的隐私保护方法研究主要有:在挖掘算法中建立隐私约束规则、在应用挖掘算法之前对挖掘数据集应用随机化方法、对隐私建立度量评估、取代本样本真实数据、对记录进行交换等,同时还有在分布式环境下的数据挖掘(数据元组水平分布和属性垂直分布)隐私保护以及通过对原始数据的混乱或扭曲进行隐私保护、敏感数据隐藏算法、规则混乱、取样法等方面[4]。这些研究的焦点主要集中在关联规则隐私保护和分类隐私保护,研究的场景可以归结为两种:个人隐私保护和分布式数据挖掘中的隐私保护。

目前在隐私保护数据挖掘领域已经有许多技术方法被采用。在实际应用中,国内外的学者已经提出了诸多算法,它们主要是集中在每一种特定情形下的算法讨论上。现有的一些

*)863项目资助(编号:2002AA414060)和2005年陕西省自然科学基金资助(编号:2005F05)。**陈晓明** 硕士研究生,主要研究方向为Web数据挖掘及Web应用;**李军怀** 副教授,主要研究方向为分布式计算、CSCW;**彭 军** 副教授,主要研究方向为计算机应用、网络安全;**张 璟** 教授,博士生导师,主要研究方向为Internet技术及应用;**刘海玲** 硕士,研究方向为分布式计算、Web服务。

隐私保护技术大体上可基于下面因素对它们分类:数据分布、隐私保护技术、数据或规则更改方法、数据挖掘算法。

表1所示的是根据数据分布方式对现有一些典型算法的一种划分。下面将着重对其进行归纳和介绍。

表1 隐私保护算法分类

数据分布方式	隐私保护技术	数据更改方法	数据挖掘算法	文献
集中式	启发式	滑动窗口法	关联规则	[5]
		随机修改部分值为1的数据为0	关联规则	[7]
		添加随机数	关联规则	[15]
	重建式	添加随机偏移量	分类	[12]
		随机修改部分数据	关联规则	[10]
		贝努里概率模型	关联规则	[11]
水平分布	加密式	加密、添加随机数	关联规则	[14]
垂直分布	加密式	加随机数	关联规则	[2]
其他				[15]、[16]、[17]、[18]、[19]

2.1 数据集中分布

2.1.1 启发式技术

针对数据挖掘中的隐私保护,很重要的一个方面都是集中在对关联规则的应用,从根本上来说都是通过各种方法来降低敏感规则的支持度或者置信度,由此诞生出了分类、聚类等方法来发现敏感规则或者敏感数据。理论上说,这些技术都是基于一种假设,即选择性的数据清理与修改是一个NP-Hard问题。而启发式算法在解决这一问题上,可以省略大量无谓的搜索路径,提高效率。

(1)基于关联规则的隐私保护

Stanley R. M. Oleveira 和 Osmar R. Zaiane 在文[5]中提出一种基于启发式的隐私保护方法,该算法通过一种单次扫描算法来实现对敏感规则的保护。

与其他的混乱式的方法不同,混乱式的方法是通过修改现有数据或者在原始数据中添加随机数据达到隐藏敏感规则的目的。而 Stanley R. M. Oleveira 提出的算法是一种非混乱式的算法,该算法通过移除部分信息的方法实现数据清理,从而隐藏敏感规则,并不对原始数据库添加任何噪声,文中称作是滑动窗口的算法。具体讲,设D是交易数据库,Rr是敏感规则,~Rr是非敏感规则,R是D中全部的关联规则,可以看出 Rr∪~Rr=R。只要交易中包含有任何敏感规则,那么它就被称作是敏感交易。为了达到隐藏敏感规则的目的,就必须对原始记录中涉及到敏感规则的敏感数据项进行处理,通过去除部分敏感项,使得规则在某种阈值范围内隐藏,而那些候选要被去除的敏感项在文中被称作是牺牲项(victim item)。

从具体实现上,数据库D被分成若干组数据交易,每组交易包含有 K 个记录,K 称作是窗口大小。数据清理的过程分为5个步骤,

第一步,从D中找出所有的R,同时区分出Rr和~Rr,并把~Rr直接复制到处理后的数据库D';

第二步,选取在当前交易敏感规则中出现频率最高的项作为牺牲项;

第三步,通过设定暴露阈值,计算要清理的交易数量;

第四步,针对每一个敏感规则,对前一步中的敏感交易按照其规模进行排序,将规模最小的作为被清理的对象,以达到对原始数据的影响尽可能地小;

第五步,删除每条敏感规则中的敏感项,将清理后的数据复制到不含敏感规则的数据库D'。

最后使用清理敏感规则后生成的数据库D'公布给外界,以此达到保护隐私的目的。

Stanley R. M. Oleveira 等人还在文[6]中提出了类似的方法,算法通过一个基于倒排文件索引和布尔查询的交易检索引擎来实现对数据库的清理,同样是通过减少敏感规则中的敏感数据,达到支持度的降低,从而实现对敏感规则的隐藏。同时文中针对算法提出了三条评价标准:Hiding Failure,也就是从清理后的数据库D'中发现敏感规则的百分比;Misses Cost,即在清理数据库D'中隐藏的非敏感规则所占的比例;Artifactual Pattern,即虚假的规则所占的比例。

Elena Dasseni 等在文[7]中提出了一种基于混乱的方法,通过隐藏与敏感规则相关的频繁项集,以及通过设定阈值来减少置信度,防止敏感规则的产生。在其具体实现上,是通过将频繁项集中的二进制数反转,即数值为1的变为0,为0的数值变为1的方式。这样,隐藏关联规则的方法存在一个很明显的缺点,就是会产生一些"影子规则"(ghost rules)[4],因为通过以上的反转,将会使得一些原本是非敏感的规则变为"敏感规则"。

(2)基于分类规则的隐私保护

分类是这样一个过程,它找出描述并区分数据类或概念的模型或函数,以便能够使用模型预测类标记未知的对象类。导出模型时基于对训练数据集(即其类标记已知的数据对象)的分析[3]。分类的目标就是要构造一个分类模型,从而预测未来的数据趋势。从目前来看,分类采用的方法主要有分类规则、决策树、神经网络等。而基于隐私保护的分类技术则是要在数据挖掘过程中建立一个没有隐私泄露的准确的分类模型[8]。

文[9]中提出的一种隐私保护方法就是结合了分类规则以及一种称作是吝啬降级法(parsimonious downgrading)。

吝啬降级法是这样的一种算法:针对要降级的数据中将要去掉的信息的一种形式化描述,其中所谓的降级是指从敏感级或隐私级(称之为高级别)降低到可以公布级(低级别)。算法的主要目标是要找出由于修改而导致的数据库功能性的损失是否是值得的。就其实现而言,算法通过产生一个称之为参变量基础集的方法来实现数据的降级。使用一个参数 $\theta \in [0,1]$ 来取代敏感数据(θ 表示某种属性取得一个可能的值的概率)。同时对于降级前和降级后的数值的熵进行计算,使用二者的差值同数据库变化前后置信度的降低程度比较,从而得出这种对数据库的修改是否是可以接受的,也即是否对数据库的影响是最小的。

2.1.2 重建式技术

(1)基于关联规则的隐私保护

A. Evfimievski 等在文[10]中提出了一种基于重建式的隐私保护算法,该算法使用了一种称之为统一随机化的方法(Uniform Randomization)。所谓的统一随机化是指在将一个交易发送给服务器前,客户端取走每一个项时,将之以概率 p 替换为原先在交易中没有的新项,这个过程叫做统一随机化。算法针对的是分类数据(categorical data),重点是为了发现关联规则挖掘中的频繁项集。文中对隐私漏洞(Privacy Breaches)进行了定义,提出了使用"select-a-size"和"cut-and-paste"等随机运算符来修改原始数据,控制隐私漏洞的发生。

然后根据修改后的数据来发现关联规则。以下是其中的关键性概念。

① 隐私漏洞

我们知道项集 A 产生了一个等级为 ρ 的隐私漏洞，如果对某个项目 $a\in A$ 以及某个 $i\in 1\cdots N$，有 $P[a\in t_i|A\subseteq t'_i]\geqslant\rho$，也就是说一个在随机化交易中的项集 A，如果 A 中的某个项出现在非随机化交易中的概率不小于 ρ 的话，那么就可以说项集 A 产生了等级为 ρ 的隐私漏洞。

② select-a-size 运算符

对于每个可能输入的规模为 m 的交易，一个“选择一个规模(select-a-size)”运算符 R 有如下参数：

a)某一个项默认的随机化等级 $\rho_m\in(0,1)$；

b)交易子集规模选择的概率 $p_m[0]$，$p_m[1]$，…，$p_m[m]$，且每个 $p_m[j]>=0$，$p_m[0]+p_m[1]+\cdots+p_m[m]=1$。

对于给定的交易序列 $T=(t_1,t_2,\cdots,t_N)$，运算符独立地取走每条交易 t_i，并且按以下步骤处理 $t'_i(m=|t_i|)$：

a)运算符从集合$\{0,1,\cdots,m\}$随机选择一个整数 j，这样 $P[j\text{ 被选择}]=p_m[j]$。

b)运算符同样是随机地从 t_i 中选择 j 个项目(没有替换)，这些项目(不含其他 t_i)被放入 t'_i。

c)运算符反过来考虑每个项目 $a\notin t_i$，然后以“正面”概率为 ρ_m、“反面”为 $1-\rho_m$ 抛出一枚硬币。所有结果是“正面”的项目加入到 t'_i。

③ cut-and-paste 运算符

一个“剪切粘贴(cut-and-paste)”运算符是 select-a-size 运算符中特殊的一类。对于每种可能输入规模为 m 的交易，它有两个参数：一个 $\rho_m\in(0,1)$和一个整数 $K_m>=0$，运算符独立地处理每个交易 t_i 并按以下方法得到交易 $t'_i(m=|t_i|)$：

a)选择在 0 与 K_m 间随机选择一个整数 j；如果 $j>m$，那它会设置 $j=m$。

b)随机地从 t_i 中选择 j 个项目放入 t'_i。

c)每个项(包括 t_i 剩余的部分)以概率 ρ_m 独立地放入 t'_i。

Shariq J. Rizvi 等学者在文[11]中根据朴素贝努里概型提出了一种称为 MASK 的隐私保护算法。作者使用的数据库是由固定长度的 0,1 序列形成的记录组成的。算法的具体实现是通过对所有原始数据按照贝努里概型进行变换，具体来说，设原始数据 $X=\{X_i\}$，$X_i=0$ 或 1，使用变换函数 $Y=distort(X)$，其中 $Y_i=X_i$ XOR r_i，r_i 是服从贝努里分布的一个随机变量，即取 1 的概率为 p，取 0 的概率为 $1-p$。考虑到对原始数据进行变换所耗费的时间和空间，远比在数据挖掘时对数据的重建的消耗要大很多，因而 Shariq 后来对 MASK 算法进行了一些优化。

(2)基于分类的隐私保护

数据挖掘通过对大量数据的统计计算找出趋势，而不一定要知道百分之百真实的数据。R. Agrawal 在文[12]中提出了一种基于重建式的技术，算法针对数值型数据概率分布的重建，该算法通过添加随机偏移量对原始数据进行随机化混乱，然后使用贝叶斯公式，根据原始数据的分布来重建决策树。通过重建数据分布可以建立一种准确程度接近真实数据分布的分类标记。同时，R. Agrawal 引入了一些量化方法来检验通过上述方法处理后的数据隐私泄漏状况，从置信度以及预测的准确程度上对算法进行了检验。具体地讲，如果可以以 c%的置信度预测出某个数值 x 处在某个区间内，那么这段区间的宽度就决定了带有 c%置信程度的隐私数量。但是由于贝叶斯定理的成立本身需要一个很强的独立性假设前提，而此假设在实际情况中经常是不成立的，因而其分类准确性就会下降。

2.2 数据水平分布

从表 1 中我们可以看出，如果数据的分布方式不再是集中式的，那么如果要进行隐私保护，那么大多数数据挖掘领域采用的是基于加密的技术。特别是在分布式环境下的隐私保护问题，安全多方计算(SMC)是最为常用的一个协议。安全多方计算是在一个互不信任的多用户网络中，各用户能够通过网络来协同完成可靠的计算任务，同时又能保持各自数据的安全性[13]。下面所提到的基于加密的隐私保护算法，以及后面数据垂直分布情况下的基于加密的隐私保护算法都是可以划归为一个特殊的安全多方计算协议。

在基于加密技术的隐私保护方面，对于隐私的定义，Benny Pinkas 在文[8]是这样给出的：限制通过分布式计算所能获得的满足其要求的所泄漏的信息称之为隐私。Benny Pinkas 对安全多方计算在基于加密技术的隐私保护方面的应用作了一定的探讨，认为安全两方计算的构建要易于安全多方计算，同时影响协议的一个重要方面是用于计算评价函数的最佳组合电路的规模，而影响计算的因素则来自健忘传输协议以及协议的改进程度。

Murat Kantarcioglu 在文[14]中提出了一种数据水平分布下的针对关联规则的隐私保护数据挖掘算法。数据水平分布是指数据按照记录分布在各个站点，在此条件下，各个站点不必知道其他站点的具体记录信息，就可以计算出全局的关联规则。算法提出的目标就是各个站点除了知道全局的结果之外，对其他各站点的信息一无所知。

算法通过两个步骤找出全局频繁项集：

第一步使用交换加密的方法发现候选集。各个站点加密各自的频繁项集，然后将结果传递到下一个站点。传递的同时去掉重复的集合，整个过程一直持续到所有的站点加密完所有的项集，然后各个站点使用自己的密钥对得到的结果进行解密，最后得到一个公共的项集。

第二步找出满足条件的全局频繁集。首先第一个站点计算由第一步得到的项集在本地的支持度与最小支持度阈值之间的差，然后加上一个随机数 R，将结果传给下一个站点；第二个站点做与第一个站点相同的工作，同时加上第一个站点传来的值，接着将结果传递至下一个站点；依次直至传递完所有站点，最后的值传递回第一个站点。最后在第一个站点将结果与 R 比较，如果该值不小于 R，则说明该项集为全局频繁项集。

最后通过计算开销和通讯开销对上述算法进行了评价，同时 Murat Kantarcioglu 还提出了一些改进的方向。

2.3 数据垂直分布

Jaideep Vaidya 在文[2]中提出了一种数据垂直分布条件下的基于关联规则的隐私保护算法。垂直分布是指数据按照属性分布在各个站点，在这种条件下，可以通过发现项集的支持计数来进行数据挖掘。如果某个项集的支持计数可以被安全地计算，那么通过检查计数和预先设定的阈值比较，就可以知道该项集是否是频繁项集。Jaideep Vaidya 通过安全地计算代表子项集的标量积的方法来得到项集的支持计数。在这里，作者同样将目标数据库视作是由固定长度的 0、1 序列构

成的，同时考虑的是数据垂直分布在两个站点上的情况。

设数据库是由 n 条记录构成，对于其中的一个 k-项集，站点 A 拥有其中的 p 个属性 $a_1, a_2, \cdots, a_p, a_{i1}, a_{i2}, \cdots, a_{ip}$，表示第 i 条记录对应在这些属性上的值，$\vec{X}$ 表示一个 n 维矢量，第 i 维的值 $x_i = \prod_{j=1}^{p} a_{ij}$；站点 B 拥有剩余的 q 个属性 $b_1, \cdots, b_q$，$b_{i1}, \cdots, b_{ip}$ 表示第 i 条记录对应在这些属性上的值。对于 B，类似的有：一个 n 维矢量 $\vec{Y}$，其第 i 维的值 $y_i = \prod_{j=1}^{q} b_{ij}$，于是有 $k = p+q$。这样，通过计算 $\vec{X} \cdot \vec{Y} = \sum_{i=1}^{n} xi * yi$，可以得出 k-项集的支持计数，从而得出全局频繁集以及关联规则。

但是，如果按照上面的方法，要计算支持计数，那么站点 A 或 B 都必须公布各自的私有信息，暴露了自己的隐私。针对这样的情况，Jaideep Vaidya 提出的算法就是一种不向对方公布自己的向量的情况下计算标量积的方法。他的根据就是解一个 n 元线性方程组，而方程的个数小于 n，其结果是不确定的。通过这样的方法达到保护隐私的目的，同时还能保证各方只能得到全局的频繁项集和关联规则。对各站点将其拥有的属性构成一个 n 维系数矩阵，通过产生随机的 n 个数 $R_1, R_2, \cdots, R_n$，使之与其拥有的属性线性组合，通过交换计算结果得到规则。

2.4 其他

除了以上提到的一些典型方法，其他的学者也提出了另外的一些隐私保护算法。例如：Y. Saygin 等在文[15]中通过在记录中添加"?"的方法，来对敏感规则进行保护。文[16]中利用安全标量积协议提出了一个在数据垂直分布情况下，通过生成随机向量的方法进行隐私保护；文[17]中在数据水平分布方式下，利用加密方法建立判定树保护隐私。Zhangqiang Yang 等在文[18]中所提到的是一种基于分类规则的隐私保护方法，算法针对全分布环境下的数据，使用一种简单的加密方法，各个站点间仅仅向挖掘者进行一次数据传输，在保护隐私的同时最大限度地保证挖掘的准确性。Lin Xiaodong 等在数据水平分布环境下，使用聚类理论提出一种期望最大化(EM)混合模型[19]。

3 算法的评价标准

目前国内外已经产生出了针对不同环境下的很多的隐私保护算法。通过上面对一些目前典型的算法的介绍可以看出，很明显，在进行隐私保护数据挖掘研究中，对算法作出适当的评价。或者说，在进行应用开发的时候，采用哪一种隐私保护算法是非常重要的。隐私保护算法可以从下列方面进行评价和比较：

(1)保密性。也就是说站在隐私保护的角度，如何能够最大限度地防止入侵者非法获取隐私数据，对隐私进行有效的保护。

在现有的算法中，保密是一个最基本的方面，各个算法都从不同的角度进行了实现。但是不同的算法都设定了一个特定的数据模型，而且更重要的是这些算法针对非法入侵者都进行了一个基本假定，即所有的非法入侵者都是采用同样的入侵手段来获得数据的。而实际中，这显然是理想化的。综合来看，前面提到的不同算法，所能做到的保密性都是有限的。

(2)规则效能。规则效能是指在使用隐私保护算法处理数据的时候，对原始信息的修改使得挖掘结果，也即最终得出的全局关联规则，与原始数据之间关系的匹配程度。

规则效能其实反映的是挖掘结果的有效性、可用性。很多的隐私保护算法是用了混乱或者相似的技术对原有数据进行了"净化"，主要是针对其中的隐私数据进行了处理。这样，处理后的数据如果经过挖掘得出的是错误的，或者说不能反映真实状况的规则，那么原有的数据也就失去了价值，而这样处理数据的算法也同样失去了效用。因而在考虑保护个人隐私的同时，算法还要能在整体上反映出规则联系。

例如，对于基于关联规则的隐私保护算法，可以从经过挖掘算法处理后的数据库所得到的规则数目与原始规则的数目相比较，来得出算法的规则效能；针对分类规则，也可以使用类似的方法。

(3)算法复杂性。具体指算法的时间复杂性和空间复杂性，也即算法的执行时间以及在进行数据处理时使用处理资源的消耗程度上，可以说这是直接与计算效率直接相关的一条标准。

算法复杂性的高低体现在该算法所需要的计算机资源的多少上。所需资源越多，该算法的复杂性就越高；反之，所需资源越少，该算法的复杂性就越低。具体来说需要时间资源的量称为时间复杂性，需要空间资源的量称为空间复杂性。特别地，在分布式环境下，通讯复杂性也是一个主要因素。无疑，复杂性尽可能低的算法设计算法时所追求的一个重要目标。

(4)扩展性。指对算法在处理海量数据集时的能力，或者是在数据量增加时，其处理效率的变化趋势。

算法的扩展性的好坏直接反映在当所处理的数据量急剧增大的时候，算法的处理效率是否下降得很剧烈。很明显，一个扩展好的算法在数据量增大的同时，其效率的变化是相对缓慢的。

算法的扩展性在一定程度上是与其复杂性相关的。例如基于混乱技术的算法在处理数据时，算法从时间复杂度上讲是相对较低的，但是从空间复杂度方面，由于其要遍历整个数据库，计算其中的频繁集，对内存资源的消耗是很大的。特别是数据库中的数据量急剧增大的时候，其处理效率会显著降低，扩展性不好。

结束语 本文对目前比较典型的各种算法进行了一定的介绍和分析，综合提出了评价隐私保护算法好坏的 4 条标准。实际上，上述这些算法在保密的准确度方面差别不大，在当今数据量急剧增长的时代，算法的复杂性、扩展性以及规则的效能等特性更为重要。此外，由于算法针对的目标不同，效果一般和数据的特点有关，目前还不存在能适合各种不同数据的最佳通用方法。一种各方面特性都很好的算法，仍有待进一步研究。

参 考 文 献

1 Cranor L F, Reagle J, Ackerman M S. Beyond concern: Understanding net users' attitudes about online privacy: [Technical Report TR 99. 4. 3.]. AT&T Labs-Research, 1999. http://www.ressearch. att. com/ library/

2 Vaidya J, Clifton C. Privacy preserving association rule mining in vertically partitioned data. In: the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2002. 639~644

3 Han Jiawei, Kamber M. 数据挖掘概念与技术. 范明, 孟小峰, 等译. 北京：机械工业出版社, 2001. 8

4 Verykios V S, Bertino E, Nai Fovino I. State-of-the-art in Privacy-preserving Data Mining, SIGMOD Record, 2004, 33(1)

5 Oliveira S R M, Zaiane O R. Protecting Sensetive Knowledge by Data Sanitization. In: Proceedings of the Third IEEE International Conference on Data Mining, 2003

(下转第 199 页)

表 1 孤立点挖掘结果列表

孤立点序号	会员编号	X1	X2	X3
1	464	51	48484	50774
2	708	11	56984	41891
3	535	14	56284	32381
4	978	15	52322	43462
5	989	11	44453	34373
6	1071	17	56191	40252
7	302	26	36946	55740
8	938	31	55304	45005

表 2 三个指标的平均值与标准差

	X1	X2	X3
平均值	13.392	13391.4	12999.33
标准差	3.285	7611.83	5831.29

表 3 各指标相关系数

	X1	X2	X3
X1	1.0000	0.3980	0.5390
X2	0.3980	1.0000	0.9238
X3	0.5390	0.9238	1.0000

从表 3 可以看出，各指标存在一定的相关性，有必要用独立成分分析的方法对原指标数据进行处理。从表 1 与表 2 可以看出，孤立点 1,7,8 的三个指标都很明显偏离中心值，孤立点 2,3,4,5,6,7 的指标 X2,X3 偏离中心值，与实际情况基本相符合。如果孤立点数据被确定，应当对其内容进行细查，根据其特征和挖掘目的而确定其是否为真正的孤立点数据。我们从该航空公司的会员信息表中查询，确认对表 1 中列出的八个会员为该航空公司的金卡会员，是重点客户，这也说明了 IMVOM 模型的合理性。

小结 一般高维数据的内部十分复杂，很难被探测，而人眼视觉对二维的几何空间分布却有着绝佳的分析能力，如果能够将高维数据资料转成人眼可以观察到的图形的话，对于高维数据的孤立点挖掘将是很有帮助的。本节在 ICA 与 MViSOM 的基础上提出了一个 IMVOM 孤立点挖掘模型，IMVOM 模型先通过 ICA 去除了数据之间的相关性，并通过 MViSOM 算法保持数据之间一定的拓扑结构不变性，并取得数据的可视化效果。最后，在孤立点挖掘的过程中，加入了“人为因素”。结合了“人类擅长于模式识别的能力”与“电脑擅长于大量地记忆、快速地计算的能力”的双方优点，IMVOM 方法避免了对高维数据内部结构的复杂探测，最终为克服高维数据集孤立点挖掘过程中的困难提供了基础。实验结果也验证了该模型的合理性，从而为数据挖掘提供了一种有效的方法。

参 考 文 献

1 Liu X. Strategies for outlier analysis. Birkbeck College University of London, 2000

2 Johanna H, Rocke D. Outlier detection in the multiple cluster setting using the minimum covariance determinant estimator. Computational Statistics & Data Analysis, 2004, 44: 625～638

3 Bayarri M J, Morales J. Bayesian measures of surprise for outlier detection. Journal of Statistical Planning and Inference, 2003, 111: 3～22

4 Kantardzic M. Data Mining Concepts, Models, Methods, and Algorithms. Tsing hua University Press, 2003

5 De Groot P J, Postma G J, et al. Application of principal component analysis to detect outliers and spectral deviations in near-field surface-enhanced Raman spectra. Analytica Chimica Acta, 2001, 446: 71～83

6 Jutten C, Herault J. Independent component analysis versus PCA. In: Proceeding of European Signal Processing Conf. 1988. 287～314

7 Kocsor A, Csirik J. Fast Independent Component Analysis in Kernel Feature Spaces. LNCS, 2001, 2234: 271～281

8 Kohonen T. Self-organizing maps. 3rd ed. Berlin Heidelberg New York: Springer, 2001

9 Yin H. ViSOM——a novel method for multivariate data projection and structure visualization. IEEE Transaction on Neural Networks, 2002, 1: 237～243

10 Yin H. Data visualization and manifold mapping using the ViSOM. Neural Networks, 2002, 15: 1005～1016

11 Sarvesvaran S, Yin H. Visualisation of Distributions and Clusters Using ViSOMs on Gene Expresssion Data. Lecture Notes in Computer Science, 2004, 3177: 78～84

12 Wu S, et al. PRSOM: A New Visualization Method by Hybridizing Multidimensional Scaling and Self-Organizing Map. IEEE Transactions on Neural Networks, 2005, 5: 1362～1380

13 彭红毅，蒋春福，朱思铭. 一种改进的高维数据可视化模型. 计算机科学(已录用)

14 彭红毅，蒋春福，朱思铭. 基于 ICA 与 SVM 的孤立点挖掘模型. 计算机科学，2006, 33(9)

(上接第 186 页)

6 Oliveria S R M, Zaiane O R. Privacy Preserving Frequent Itemset Mining. In: Workshop on Privacy, Security, and Data Mining at The 2002 IEEE International Conference on Data Mining (ICDM'02), Maebashi City, Japan, December 2002

7 Dasseni E, Verykios V S, Elmagarmid A K, et al. Hiding Association Rules by Using Confidence and Support. In: Procedings of the 4th Information Hiding Workshop, 2001. 369～383

8 Pinkas B. Cryptographic techniques for privacy-preserving data mining. SIGKDD Explorations, 2002, 4(2)

9 Chang Li Wu, Moskowitz I S. Parsimonious downgrading and decisions trees applied to the inference problem. In: Proceedings of the 1998 New Security Paradigms Workshop, 1998. 82～89

10 Evfimievski A, Srikant R, Agrawal R, et al. Privacy preserving mining of association rules. In: Proc. of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM Press, 2002. 217～228

11 Rizvi S J, Haritsa J R. Maintaining data privacy in association rule mining. In: Proceedings of the 28th International Conference on Very Large Data Bases, Hong Kong, China, August 2002

12 Agrawal R, Srikant R. Privacy- preserving data mining. In: Proceedings of the 2000 ACM SIGMOD Conference on Management of Data, 2000

13 秦静，张振峰，冯登国，等. 无信息泄露的比较协议. 软件学报，2004, 15(3)

14 Kantarcioglu M, Clifton C. Privacy- preserving Distributed Mining of Association Rules on Horizontally Partitioned Data. In: ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, 2002

15 Saygin Y, Verykios S, Elmagarmid A K. Privacy Preservin, Association Rule Mining. In: Procedings of the 12th Int'l Workshop on Research Issues in Data Engineering: Engineering e-Commerce/e-Bussiness Systems (RIDE'02)

16 Du Wenliang, Zhan Zhijun. Building decision tree classifier on private data. In: Procedings of the IEEE ICDM Workshop on Privacy, Security and Data Mining, 2002

17 Lindell Y, Pinkas B. Privacy preserving data mining. In: Advances in Cryptology-Crypto, 2000. 36～54

18 Yang Zhangqiang, Zhong Zhong, Wright R N. Privacy-preserving Classification of Customer Data Without Loss of Accuracy. In: Proceedings of the 5th SIAM International Conference on Data Mining, 2005. 21～23

19 Lin Xiaodong, Clifton C, Zhu M. Privacy-preserving clustering with distributed EM mixture modeling. Knowledge and Information Systems, 2004