情景感知数据挖掘中隐私泄露限制方法探讨

Limiting Privacy Breaches in Context – Awareness Data Mining

李　刚　尹　涛　李海强　孟　霞　张　芸　（北京邮电大学 经济管理学院 北京 100876）

摘　要：作为未来移动电子商务的重要应用——情景感知，其通过对于海量的客户相关数据进行数据挖掘，并对用户的行为模型进行判断，进而为用户提供先知先觉的便捷服务。然而在情景感知为我们未来的生活带来无限便利的同时，也使得用户个人隐私面临巨大的泄露危险。本文对于数据挖掘中的隐私泄露给出了基于概率论的定义，并且依据该定义对如何判定和限制用户的隐私泄露进行了探讨，最后给出了"放大法"对于隐私泄露进行有效的限制。

关键词：情景感知　隐私泄露　信息揭示理论　随机化

1　引言

隐私泄露已经不是一个新的问题，但随着网络技术的发展与电子商务的兴起，"隐私泄露"问题也逐渐被放大，特别是在新的电子商务环境中，隐私泄露与保护的问题已经成为影响电子商务未来发展趋向的重要议题之一[1]。作为未来移动电子商务的重要应用之一——情景感知服务（Context – Awareness Services）是指在现有的定位服务（Location – Based Services）基础上综合考虑用户所在环境的其他因素，为用户提供更具有针对性的信息服务[2]。在情景感知为我们未来的生活带来无限便利的同时，其也使得用户对个人隐私的安全产生了担忧，使得用户个人隐私面临巨大的泄露危险。因而，在为用户提供便捷、有效服务的前提条件下，如何限制隐私泄露——将个人隐私的泄露限制在无害的范围内成为限制情景感知业务未来发展的要枢。

在本文的第二部分，我们将对于隐私泄露的概念进行讨论，并对隐私泄露给出一个具体的定义；在本文第三部分，我们基于隐私泄露的定义，通过"放大法"提出判定隐私泄露与限制的方法。

2　隐私泄露的相关概念

2.1　一些基本概念

有 n 个用户 $C_1,\cdots,C_n$ 分别与同一服务器相连，每个用户 C_i 拥有一定的隐私信息 x_i。服务器需要通过用户的数据，获取数据总体的特定（统计）属性。用户对于这种信息获取行为（即对于数据总体的数据挖掘）是可以接受的，然而用户却不愿意看见自己的隐私信息 x_i 出现泄露。为了保全每个用户的隐私，通过 $y=R(x)$ 将每个用户的个人信息加以修改，然后将修改后的 y_i 发送给服务器；服务器通过对于全体用户修改后数据的整理，进而获取其所需要的某些（统计）属性。

用户的任何隐私信息都可以通过数值形式的 x_i 进行记录，例如年龄、性别等，并且每个用户的个人隐私信息 x_i 属于同一个固定有限集合 V_x。因此该类隐私信息 x_i，例如年龄、性别等在数据整体中必然服从一个固定的概率分布，并且相互独立。这个概率分布表示为 P_x，服务器通过使用该概率分布 P_x 建立分类模型[3]。

在每个用户将其个人数据 x_i 发送给服务器前，用户通过一个随机化函数 $R(x)$ 对其个人信息进行隐藏，例如 $R(x)=x_i+r_i$ 其中的 r_i 服从正态分布或均匀分布。服务器对于 $R(x_i)$ 输出结果 y_i 进行整理，并通过期望最大化算法重建 x_i 的分布。$R(x)$ 所有可能输出结果属于一个集合 V_y，该集合为一个有限集合。对于所有的 $x\in V_x$ 和 $y\in V_y$，$R(x)$ 输出 y 的概率可以表示为，$p[x\rightarrow y]:=p[R(x)=y]$。

通过从用户 C_i 获取 y_i，服务器间接获取 x_i 的一些信息。需要指出的是，基于以上的独立性假设，所有的

$y_j(j \neq i)$不会公开任何关于x_i的信息,并且在隐私分析方面可以忽略其泄露的可能性。主要的问题在于衡量由y_i泄露的x_i相关信息的多少,以及如何通过随机化的方式显示隐私信息的泄露[4]。

2.2 隐私泄露的定义

隐私泄露是指,某一用户C_i,对于其随机化的信息y_i的公开将导致C_i的特定隐私信息特征的泄露。例如,我们将年龄属性x_i,通过与服从[-50,50]均匀分布的r_i求和的方式进行随机化。假设,服务器接收到的一个用户随机化年龄为120,服务器可以确定的推断出该用户的真实年龄不会小于70岁,即否则$R(x)=x_i+r_i<70+50=120$。服务器得到了了一条关于用户具有潜在价值的信息,并且该信息的正确可能性为100%。[5]

假设:C_i为任意用户,x_i为其隐私信息。

在随机化之前,对于服务器,每个C_i个人信息的可能取值x都有一个$p_x(x)$。定义随机变量X,$P[X=x]:=p_x(x)$。随机变量X是服务器对于x_i预先了解的最好描述。现在,假设用户通过$y_i=R(x_i)$,随机化x_i,并将随机化后的y_i发送给服务器。从服务器的方面来看,已经随机化的y_i是随机变量Y的一个实例;Y可以表示为,

$$P[Y=y]:=\sum_{x \in V_X} P[X=x] \cdot P[x \to y]。$$

随机变量X和Y是非独立的,它们的联合分布如下:

$$P[X=x,Y=y]P_x(x) \cdot P[x \to y]。$$

通过给出的y_i,服务器可以更好的预测C_i个人信息可能取值的概率。使用贝叶斯方程并计算后验概率:

$$P[X=x|Y=y_i]:=\frac{P[X=x] \cdot P[x \to y_i]}{P[Y \to y_i]}。$$

我们同样也可以计算出任何特征的先验概率,其中$Q:V_x \to \{true, false\}$:

$$P[Q(x)|Y=y_i]=\sum_{Q(x).x \in V_x} P[X=x|Y=y_i]。$$

通俗地讲,隐私泄露就是指对于某一特征$Q(x_i)$,由于y_i对服务器的公开而引起的该特征函数$Q(x_i)$概率的明显提高。如果该隐私信息中的特征$Q(x_i)$的保密,对于用户非常重要;那么这一概率的明显上升将会对用户隐私造成侵犯。

在此,我们给出隐私泄露的正式定义:

定义1:对于特征函数$Q(x)$,如果有某一$y \in V_Y$,$P[Q(x)] \leqslant \rho_1$且$P[Q(x)|Y=y] \geqslant \rho_2$,其中$0<\rho_1<\rho_2<1$且$P[Y=y]>0$,则称该状态为关于特征$Q(x)$的$\rho_1-\rho_2$隐私泄露。

依据定义1,本节开始处的隐私泄露例子则可以被称为关于“大于或等于70岁”年龄特征的30% -100%的隐私泄露。

下面再让我们看一个关于隐私泄露的例子。

假设隐私信息x是一个在0至1000之间的自然数。这个数字被作为一个随机变量来选择,其中0的概率为1%,其他非零数字的概率为0.099%,即

$$P[X=0]=0.01, P[X=k]=0.00099, k \in [1,1000]。$$

如果我们要通过新的随机数字$y=R(x)$替代它方式,使这个数字随机化;值得注意的是$y=R(x)$中仍保留关于原数字的部分信息。这里我们给出三种可能的处理方式:

(1)给定x,设定$R_1(x)$等于x的概率为20%,等于其他数字的概率为80%(随机进行选择);

(2)给定x,设定$R_2(x)$等于$x+\xi(\mathrm{mod}1001)$,其中ξ在[-100,100]随机选取;

(3)给定x,设定$R_3(x)$等于$R_2(x)$的概率为50%,等于其他数字的概率为50%(随即进行选择)。

表1 上例特征的先验概率与后验概率

给定值	$X=0$	$X \notin (200,800)$
空	1%	≈40.5%
$R_1(x)=0$	≈71.6%	≈83%
$R_2(x)=0$	≈4.8%	100%
$R_3(x)=0$	≈2.9%	≈70.8%

在表1中,我们计算了X两个特征的先验概率与后验概率,其特征分别为特征$Q_1(X) \equiv$"$X=0$"与$Q_2(X) \equiv$"$X \notin (200,800)$"。由此我们发现,当$R_1(X)$恰巧等于0的时候,随机化函数R_1给出了很多关于X的信息。在不需要知道$R_1(X)=0$的情况下,服务器认为$X=0$的概率仅仅是1%;但是当$R_1(X)=0$被给出以后,$X=0$的概率提升为70%左右。当$R_2(X)=0$被给出的时候,就不会发生上述情况,$X=0$的概率仅仅为4.8%。不论怎样,另一种个人隐私被泄露了——服务器有百分之百的把握确定不是在200与800之间。这

一特征的先验概率大约为 40%。至此,仅仅 R_3 看起来是一个对隐私保护较好的随机化方法。

正如上例中展示的一样,一些随机化的函数(或方法)对于隐私保护来说可能不是安全的,有时在知晓一个随机化值的情况下,对于某一特征的原隐私值的先验概率有很明显的影响。为了避免上述问题,我们要么必须去确定所涉及的特征对服务器的公开都是无害的(即使出现隐私泄露的结果,对于用户的损害也是可以接受的),或者确保没有特征的先验概率被明显地改变了。

在此,我们选择后一种方法,依据我们对于隐私泄露的定义 1,对于 $R_1(x)$,我们得到关于特征 $Q_1(x)$ 的 1% -70% 的隐私泄露;对于 $R_2(x)$,我们得到关于特征 $Q_2(x)$ 的 40% -100% 的隐私泄露。

在上面例子中,我们将两类概率的改变归为"明显"的改变。

(1) 某一特征 $Q_1(x)$ 的先验概率很小,而在知晓 $R(x)=y$ 后变得较大;在上例中,当知晓 $R_1(x)=0$ 以后,特征 $X=0$ 的概率从 1% 扩大为 70%;

(2) 某一特征 $Q_2(x)$ 的概率远小于 100%,即不确定,但在知晓 $R(x)=y$ 后,变得接近 100%;在上例中,当知晓 $R_2(x)=0$,特征"$X \notin (200,800)$"的概率从 40% 增至 100%,亦即"$200 \leqslant X \leqslant 800$"的概率从 60% 减为 0。

这个观测表明对于隐私泄露,我们可以将其归为两个主要的子类。下面让我们给出关于这两个子类的正式定义。

定义 2:ρ_1、ρ_2 为特征 $Q(x)$ 的相关概率,其中 $\rho_1<\rho_2$;

对于特征函数 $Q_1(x)$,如果有某一 $y \in V_Y$,$P[Q_1(x)] \leqslant \rho_1$ 且 $P[Q_1(x) | R(x)=y] \geqslant \rho_2$,其中 $0<\rho_1<\rho_2<1$ 且 $P[R(x)=y]>0$,则称该状态为关于特征 $Q_1(x)$ 的 $\rho_1-\rho_2$ 正隐私泄露;

对于特征函数 $Q_2(x)$,如果有某一 $y \in V_Y$,$P[Q_2(x)] \geqslant \rho_2$ 且 $P[Q_2(x) | R(x)=y] \leqslant \rho_1$,其中 $0<\rho_1<\rho_2<1$ 且 $P[R(x)=y]>0$,则称该状态为关于特征 $Q_2(x)$ 的 $\rho_2-\rho_1$ 负隐私泄露;

至此,我们已经对于隐私泄露给出了完备的定义。基于我们对于隐私泄露的定义,我们将会从机制设计与随机化技术两个方面,对于限制隐私泄露的方法进行更深地探讨。

3 隐私泄露的限制方法

如果我们试图通过关于隐私泄露的定义 1 直接检查一个给定的随机化函数是否会造成隐私泄露,那么将会发现两个巨大的困难:

(1) 可能的属性有 $2^{|V_x|}$ 种,其数量多到对其全部进行检验没有任何可行性;

(2) 如果我们不知道 X 的先验概率分布 P_x,就无法使用定义 1。然而在实际情况中,随机化函数可以在知晓 P_x 之前进行选定。

事实上,存在没有上述两个缺陷的充分验证集;同时在实际情况当中,也存在满足该验证集的有效的随机函数。这个验证集是基于对比同一个 $y \in V_y$ 而非不同的 $x \in V_x$ 的随机化函数的转移概率(transitional probabilities)$P[x \to y]$ 的方式得到的。直观上说,似乎所有 x 的值通过随机化成为一个给定的 y 都是合理的,因此对于 $R(x)=y$ 的公开不会对 x 造成什么隐私泄露。由于我们使用该方法来限制特定 $P[x \to y]$ 相对于其他的转移概率其可以被放大的程度,因此我们称这种方式为放大法[6]。

定义 3:当

$$\forall x_1, x_2 \in V_x : \frac{p[x_1 \to y]}{p[x_2 \to y]} \le \gamma \qquad \text{(a)}$$

其中 $\gamma \geqslant 1$,并且 $\exists x : p[x \to y]>0$。

那么对于 $y \in V_y$,一个随机化函数 $R(x)$ 是最 γ-amplifying 大。如果随机化函数 $R(x)$ 对于所有适合的 $y \in V_y$ 最大放大 γ,那么随机化函数 $R(x)$ 是最大 γ-amplifying。

条件:R 为一个随机化函数,其中 $y \in V_y$ 为一个随机化值,且 $\exists x : p[x \to y]$,概率 $0<\rho_1<\rho_2<1$(参照定义 2)。假设 R 是 y 的最大 γ-amplifying。在满足下述条件时,

$$\frac{\rho_2}{\rho_1} \cdot \frac{1-\rho_1}{1-\rho_2} > \gamma \qquad \text{(b)}$$

公开 $R(x)y$,对于任何特征 $Q(x)$ 既不会导致 $\rho_1-\rho_2$ 正隐私泄露,也不会导致 $\rho_2-\rho_1$ 负隐私泄露。

证明:$\exists x \in V_x : P[x \to y]>0$,否则 $\gamma \Rightarrow \infty$;将 $Y=R(X)$ 作为一个随机变量。对于任何分布 P_x,因为其至少存在一个 $x \in V_x$ 使得其不为 0,因此有

$$P[Y=y] \ge P[X=x] \cdot p[x \to y] > 0 \text{。}$$

矛盾之处在于，如果我们假设对于特征 $Q(x)$，存在一个 $\rho_1-\rho_2$ 隐私泄露，特征 $Q(x)$ 对于所有 $x\in V_x$ 都不为真，因为根据隐私泄露的定义，$P[Q(x)]\leqslant\rho_1<1$。相似地，特征 $Q(x)$ 对于所有 $x\in V_x$ 亦不能为假，因为 $P[Q(x)|Y=y]\geqslant\rho_2>0$。因此，存在以下表达：

$$x_1\in\{x\in V_x\left|Q(x),p[x\to y]=\max_{Q(x')}p[x'\to y]\right.\}\;;$$

$$x_2\in\{x\in V_x\left|\neg Q(x),p[x\to y]=\max_{\neg Q(x')}p[x'\to y]\right.\}\;。$$

表面上，x_1 是一个拥有特征 $Q(x)$ 的隐私值，且很有可能通过随机化成为 y；x_2 是不满足 $Q(x)$ 的隐私值，且接近不可能被随机化成为 y。依据条件概率的定义，

$$P[Q(X)|Y=y]=\sum_{Q(x)}P[X=x|Y=y]$$

$$=\sum_{Q(x)}\frac{P[X=x]\cdot p[x\to y]}{P[Y=y]}$$

$$\leq\frac{p[x_1\to y]}{P[Y=y]}\cdot\sum_{Q(x)}P[X=x]=p[x_1\to y]\cdot\frac{P[Q(X)]}{P[Y=y]}\;;$$

同样地，

$$P[\neg Q(X)|Y=y]=\sum_{\neg Q(x)}P[X=x|Y=y]$$

$$=\sum_{\neg Q(x)}\frac{P[X=x]\cdot p[x\to y]}{P[Y=y]}$$

$$\geq\frac{p[x_2\to y]}{P[Y=y]}\cdot\sum_{\neg Q(x)}P[X=x]=p[x_2\to y]\cdot\frac{P[\neg Q(X)]}{P[Y=y]}\;。$$

我们知道 $P[Q(x)|Y=y]\geqslant\rho_2>0$，且 $P[Q(x)]>0$。通过上式（b）我们得到不等式，

$$\frac{P[\neg Q(X)|Y=y]}{P[Q(X)|Y=y]}\geq\frac{p[x_2\to y]}{p[x_1\to y]}\cdot\frac{P[\neg Q(x)]}{P[Q(x)]}\;,$$

因为 $R(x)$ 是对于 y 的最大 γ – amplifying：

$$\frac{1-P[Q(X)|Y=y]}{P[Q(X)|Y=y]}\geq\frac{1}{\gamma}\cdot\frac{1-P[Q(X)]}{P[Q(X)]}$$

容易得到， $\dfrac{1-\rho_2}{\rho_2}\geq\dfrac{1-P[Q(X)|Y=y]}{P[Q(X)|Y=y]}$；

$$\frac{1-P[Q(X)]}{P[Q(X)]}\geq\frac{1-\rho_1}{\rho_1}\;。$$

因此在条件（a）下，推理出现了矛盾，即在条件（b）下不存在 $\rho_1-\rho_2$ 正隐私泄露。

为了证明对于 $\rho_2-\rho_1$ 负隐私泄露的情况，我们用 $\rho_1'=1-\rho_2$ 与 $\rho_2'=1-\rho_1$ 与分别替换 ρ_2 与 ρ_1，从而得到 $\rho_1'-\rho_2'$ 正隐私泄露，同时其仍然满足条件（b）：

$$\frac{\rho'_2}{\rho'_1}\cdot\frac{1-\rho'_1}{1-\rho'_2}=\frac{1-\rho_1}{1-\rho_2}\cdot\frac{\rho_2}{\rho_1}>\gamma$$

显然得证，在此不再赘述。

我们可以称不等式（a）为对于给定 $y\in V_y$ 的放大条件。如果在不考虑随机化的值 $R(x)=y$ 时我们不希望出现隐私泄露，我们需要对于全部 $y\in V_y$ 遵守这一条件。

在 2.2 所举出的例子中，随机化函数满 $R_3(x)$ 足放大条件（a），且 $\gamma<6$。事实上，对于这个随机化函数，其转移概率可表示如下：

$$p[x\to y]=\begin{cases}\dfrac{1}{2}(\dfrac{1}{201}+\dfrac{1}{1001}),\text{当}y\in[x-100,x+100];\\ \dfrac{1}{2}(0+\dfrac{1}{1001}),\ \text{其他};\end{cases}$$

其分数差分为 $1+1001/201<6$。借助上述条件，我们可以确定其不存在 $\rho_1=1/7\approx14\%$ 至 $\rho_2=1/2\approx50\%$ 的正隐私泄露，也不存在相反的负隐私泄露。对于这个结论，我们甚至不需要知道 P_x 的分布。

在存在关于用户的某些特定背景信息时，放大条件（a）可以限制隐私泄露。假设用户 C_i 拥有个人信息 x_i，并且服务器掌握一些函数 $f(x_i)$ 的值，抑或说某一依赖于 x_i 的变量 Z。从服务器的角度，对于 x_i 的可能值的概率分布，其先验概率与后验概率，成为有条件的：

* 先验概率：$P[X=x]\to P[X=x|Z=z]$
* 后验概率：

$$P[X=x|R(x)=y]\to P[X=x|R(X)=y,Z=z]$$

如果背景信息对于随机化函数是独立的，那么所有的转移概率都是同样的，因此放大条件没有受到影响，并且上述条件仍然适用。然而隐私泄露的定义 1 在背景信息存在的情况下，却发生了改变，亦即

$$P[Q(X)|Z=z]\leq\rho_1,P[Q(X)|Y=y,Z=z]\geq\rho_2$$

（下转第 20 页）

(上接第 38 页)

4 结论

在本文中,我们对于情景感知数据挖掘中隐私泄露的方法进行了探讨。我们对于数据挖掘中的隐私泄露给出了具体的定义;而且在后文中,依据这个定义,我们提出通过放大法对隐私信息的泄露进行有效的控制。

在本文中,虽然对于可以限制隐私泄露的随机化函数 R(X) 进行了定义,但缺少对给定条件下如何构建随机化函数 R(X) 的方法进行探讨。

希望在未来的研究中能得到进一步的完善与修正。

参考文献

1 严中华,关士继,米加宁. 电子商务隐私保护的重要性及其经济分析. 物流科技,2005(4).

2 Huebscher M C, McCann J A. A Learning Model for Trustworthiness of Context – awareness Services //Proceedings of the 3rd Int'l Conf. on Pervasive Computing and Communications Workshops, 2005.

3 Agrawal D, Aggrawal CC. On the design and quantification of privacy preserving data mining algorithms, //Proceedings of the 20th Symposium on Principles of Database System, California, USA, 2001.

4 Evfimievski. A. Randomization in Privacy Preserving Data Mining, SIGKDD Exploration, 2002,4(2):43–48.

5 Agrawal D,Srikant R. Privacy preserving data mining, //Proceedings of the 19th ACM SIGMOD Conference on Management of Data, Texas, USA, 2000.

6 Evfimievski A, Gehrke J, Srikant R. Limiting Privacy Breaches in Privacy Preseving Data, PODS 2003,2003.