随机响应技术的数据挖掘

王 泳 曾传璜 （江西理工大学信息工程学院 江西赣州 341000）

摘要：随着数据挖掘技术的发展 有关数据挖掘的个人隐私保护越来越受到关注。如何在保护隐私的情况下挖掘出有用的信息是近年来数据挖掘的研究趋势之一，为了保护个人隐私信息，我们首先对数据进行随机化的处理，然后在此基础上对数据进行分析，挖掘。本文介绍了隐私保护的发展原因，随机化处理方法及其它关于隐私保护数据挖掘的算法。

关键词：数据挖掘 随机化处理 个人隐私

1 引言

我们处在一个信息爆炸的大时代，计算机处理能力，存储技术以及互联网络的发展又极大的提高了信息的数字化处理程度，估计每隔十二个月信息量就增加一倍。如同待开垦的矿山一样，快速增长的数据背后隐藏着许多有用的消息，人们希望对积累的数据进行更高层次的分析，找出数据内部一些潜在的关系和规则，这也就导致了数据挖掘的产生。通过将未知变为可知，将数据变为真正的财富。任何事情都有其两面性，数据挖掘领域也不例外，在挖掘数据产生财富的同时，随之产生的就是隐私泄露的问题。在隐私保护问题中，人们讨论得比较多的是推论问题（Inference Problem）。推论问题是指用户提出查询，从反馈回来的结果中推理出未经授权的（即用户希望保密的）信息的过程人们提出了一些解决这些问题的途径，包括在挖掘算法中建立隐私约束规则（Privacy. Constraint Rules）、在应用挖掘算法之前对挖掘数据集应用随机化方法、对隐私建立量度评估、取样本代替真实数据、对记录进行交换等。

2 数值的随机化处理

很多公司希望建立顾客的信息集合模型。比如，一家手机店想知道最有可能买 nokia 或者 samsung 的顾客的年龄和收入；一个广告公司需要知道客人的个人爱好以便更好地、有目标地做广告；一家网上零售公司希望了解顾客的个人喜好以便更好地对网页进行排版以适应顾客的口味。以上这些情况均包括一台服务器（公司）和多个客户端（顾客）。服务器需要建立一个数据聚集模型来应用关联规则挖掘算法或分类算法。通常，结果模型不再包含个人的可识别信息，只包括建立在大量数据之上的一个平均值。一般情况下，在建立数据聚集模型之前客户已经将私有信息发布到服务器上了，但越来越多的人们开始关注自己的隐私保护，不想把和此次交易不相关的数据上传。但是，公司仍需要数据聚集模型。一个可能的方案是，在用户提交数据之前，每个客户端将数据打乱，一些真实数据被取走，而用一些假的随机产生的数据取代它，这种方法叫做随机化方法（Randomization）。在此，我们假设有 N 个客户端，分别用 Ci（ i =1,2，,N）表示，每个客户端均有要发布的属性 xi ，i =1,2，,N，设其中每个 xi 都是一个随机变量 Xi 的实例，Xi 是独立的并且可以被无差别分发的。累积的分布函数（对每个 Xi 都一样）定义为 Fx ，服务器需要知道函数 或者它的近似模拟，这些就是允许服务器事先知道的聚集模型。这样，服务器就可以通过模型获取客户端的信息，但同时我们要限制服务器知道真正的 Xi。我们考虑下面的方法，每个客户端均加上一个随机产生的偏移量 yi 到 xi 上，偏移量 yi 是用一个累积分布函数 Fy 无差别产生的独立的随机变量。是事先选定的，并且服务器知道客户端 Ci 将随机产生的值 $z_i = x_i + y_i$ 发送到服务器上，服务器的任务就是用预知的 F_x 和 $z_1, \cdots, z_n$ 的值来模拟 Fx。

Fy 的选择应该尽量满足以下原则：

（1）服务器能够合理地、很好地模拟 ；

（2）Zi 的值尽可能无法揭示 Xi 的值。这种揭示的量度由置信区间决定，给定一个置信度 c% ，对每一

个随机产生的 z，我们定义一个区间[$z-w_1, z+w_2$]。这样，对所有的未随机化的量 x，我们就有 $p[z-w_1 \leqslant x \leqslant z+w_2 \mid z = x+y, Y \sim Fy] \geqslant c\%$

一般地，针对一个置信区间，我们用它在 c% 置信度级别上的最短宽度 $w = w_1 + w_2$ 来衡量隐私的数量。

一旦分发函数 Fy 已定义并且数据经过随机化处理之后，服务器将面对重构问题。问题可以这样描述，给定及 F_y 及 $z_1, \cdots, z_n$。我们用基于 Bayes 定律的循环算法来处理。定义 Xi（Fx 的引申值）的分布密度为 f_x，定义 Yi（y 的引申值）的分布密度为 f_y，算法描述如下：

（1） f_x^0 = 统一分布；

（2） j: =0；//计数器

（3） loop

① $$f_x^{j+1}(a) = \frac{1}{N}\sum_{i=1}^{N}\frac{f_Y(z_i-a)f_x^j(a)}{\int_{-\infty}^{\infty} f_Y(z_i-z)f_x^j(x)\,dz}$$

② j: =j+1；

until（不满足准则）为了让算法更实用，我们可以把属性域分成 k 个区间，分别为 $I_1, \cdots, I_K$，用分段的常数函数来代替密度函数 f_x^j，其中 $m(I_t)$ 是 I_t 的中点，则上面的公式变为：

$$f_x^{j+1}(I_p) := \frac{1}{N}\sum_{i=1}^{N}\frac{f_Y(m(Z_i)-m(I_p))f_x^j(I_p)}{\sum_{t=1}^{k} f_Y(m(Z_i)-m(I_t))f_x^j(I_t)\,|I_t|}$$

当然，这个公式也可以用积聚分布函数的形式来改写，具体应该根据实际情况去选择。

下面是我们在一个公司客户档案管理数据库之上应用随机化处理方法之后，再应用上述算法进行重构后结果分布的对照。下图说明我们的重构算法很好地模拟出了原始的年龄分布函数，基本上达到了我们的目的。

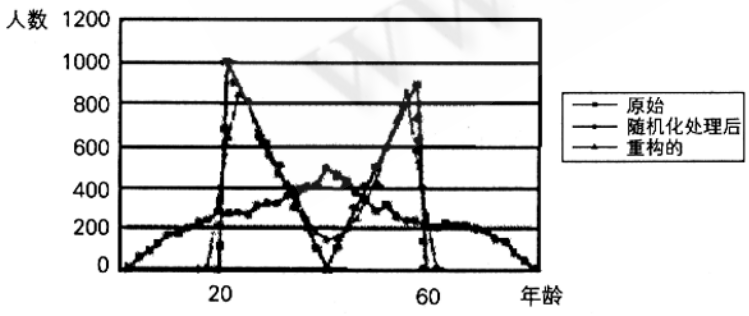

图 1

3 项集随机化处理

在隐私保护关联规则挖掘算法中，提出了“select – a – size”和“cut – and – paste”随机运算符来变换原始数据，然后再从变换后的数据中来推算项集的支持计数，以找出频繁项集。

一个长度为 m 的交易的“select – a – size”随机运算符有如下参数：

- 单个项的缺省随机概率 $\rho_m \in E(0,1)$；
- 交易子集长度选择概率 Pm[0]，Pm[1]，…Pm[m]，且 $Pm[j] \geqslant 0$ $(0 \leqslant j \leqslant m)$ 和 $Pm[0] + Pm[1] + \cdots + Pm[m] = 1$；

对交易集 $T = (t_1, , t_z, \ldots, t_N)$，“select – a – size”随机运算符采用如下方法独立地处理每一条交易 t_i（$m = |t_i|$）为 t_i'

（1）从{0，1，…，m}中随机选择一个整数 j，且 P[选择 j] = Pm[j]；

（2）从 t_i 中随机均匀地选择 j 个项，放入 t_i' 中；

（3）对每个项 $a \notin t_i$，随机地以正面为 ρ_m、反面为 $1-\rho_m$ 的概率抛硬币，将那些结果为正面的项加入 t_i'，中；

一个长度为 m 的交易的“cut – and – paste”随机运算符有如下参数：

单个项的缺省随机概率 $\rho_m \in E(0,1)$；

整数 $Km \geqslant 0$；

对交易集 $T = (t_1, , t_z, \ldots, t_N)$，“select – a – size”随机运算符采用如下方法独立地处理每一条交易 t_i（$m = |t_i|$）为 t_i'

（1）从 0 和 KM 中随机均匀地选择一个整数 j，如果 j > m，则设置 j = m；

（2）从 t_i 中随机均匀地选择 j 个项，放入 t_i'，中；

（3）对其它项（包括。剩余部分），随机地以正面为 ρ_m、反面为 $1-\rho_m$ 的概率抛硬币，将那些结果为正面的项加入 t_i' 中；

设 T 是总数为 N 的交易集合，A 是一个项集，则定义 A 的交集为 l 的局部支持度 $\sup p_l^T$

$$(A) = \frac{\#\{t \in T \backslash \#(A \frown t) = L\}}{N}$$

假设随机运算符是保持交易和项不变的，设一个

长度为 m 的交易 t，以及一个长度为 k 的项集 A，t 被随机化为 t'，定义 $p_m^k(l\to l') = p[l\to l'] = p[\#(t'\frown A) = l'] \mid \#(t\frown A) = l]]$ 这里 l，l' 都是{0，1，...，k}中的整数。

再假设 T 中所有的 N 条交易长度都为 m，则对一个长度为 k 的项集 A，随机向量 $N^*(s_0', si', \ldots, sx')$ 是一个(k+l)个独立的多项式分布的随机向量的和，这里 $S_l' = \sup p_l^T(A)$。

这样就得到了根据随机化后的局部支持度计算出来的原始局部支持度的无偏差估计值 $\bar{S}:est = Q * \bar{S}'$。

故：$COV\,\bar{S}est = Cov(Q * \bar{S}') = Q(COV\,\bar{S}')\}Q^T = \frac{1}{N} * \sum_{l=0}^{k} s_l QD[l]Q^T$。

用 $\bar{S}$ 来表示 $\bar{S}est$ 的第 k 个坐标，用 $\bar{S}est$ 来估计 s，用 q[l←l'] 来表示 Qll'，则 $\bar{S} = \sum_{l=0}^{k} S'_l * q[k\leftarrow l']$；此公式可以用来恢复项集的支持计数。

4 隐私保护技术的评估

关于隐私保护技术的评估，最重要的工作是建立合适的评估标准与相关的参考标准。本人在阅读多篇文献中得出，在实际的应用中，不可能用一个标准衡量所用的隐私保护技术，也就是说没有一种的隐私保护技术可以比其它方法好，只是在特定的应用下比其它方法好。通常隐私保护技术的技术有：性能，算法精度，隐私保护程度和通用程度。

评价性能也就是计算开销，包括估计算法的时间复杂度与空间复杂度。算法精度是指与非隐私保护的同类算法相比，其判断精度，丢失比例和误差生成比例。隐私保护程度是指防止原始数据或隐含知识被导出的程度。隐私保护的最终目的是反对信息的未授权者获取该信息。这些侵犯者往往会利用各种数据挖掘技术危害隐私，因此一个针对具体的挖掘技术而研制的隐私保护算法是不可能用于其它所有的挖掘算法。所以通用程度指的是某一隐私保护算法能运用到不同的数据挖掘技术。

5 总结

结果表明，用随机化处理过的隐私保护已经显示了它的好处，并且导致了很好的效果，这种基于安全和隐私上的统计数据压缩，也提供了联系了基于加密系统上的香农原理的基础工作，也让我们从不同与加密技术的角度去看待隐私，也介绍了相关的隐私保护算法评价标准。这种随机化不依赖复杂的假设，而且也不需要复杂的操作和完善的协议。在未来的工作可能我们会联合统计近似利用复杂和安全的多维计算去保护隐私数据，目前，我国数据挖掘中的隐私保护技术研究才刚刚开始。随着社会的发展，公民的隐私保护会受到越来越多的重视。希望本文能够起到多米诺效用，让更多的人重视和研究此项技术。

参考文献

1 D. Agrawal and C. C. Aggarwal. On the design and quantification of privacy preserving data mining algorithms. In Proceedings of the 20th Symposium on Principles of Database Systems, Santa Barbara, California, USA, May 2001.

2 C. Clifton, M. Kantarcioglu, X. Lin, J. Vaidya, and M. Y. Zhu. Tools for privacy preserving distributed data mining. SIGKDD Explorations, Newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining, 2003.

3 G. T. Duncan and S. Mukherjee. Optimal disclosure limitation strategy in statistical databases: Deterring tracker attacks through additive noise. Journal of the American Statistical Association, 95 (451):720 - 729, 2000.

4 A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke. Privacy preserving mining of association rules. In Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery in Databases and Data Mining, pages 217 - 228, Edmonton, Alberta, Canada, July 23 - 26 2002.

5 李蒙－等，数据挖掘中隐私保护的随机化处理方法。计算机工程与技术 2005，27：58－59.

6 Jiawei Han, Micheline Kamber. 范明、孟小峰译，数据挖掘概念与技术 I－M. 北京：机械工业出版社，2001.