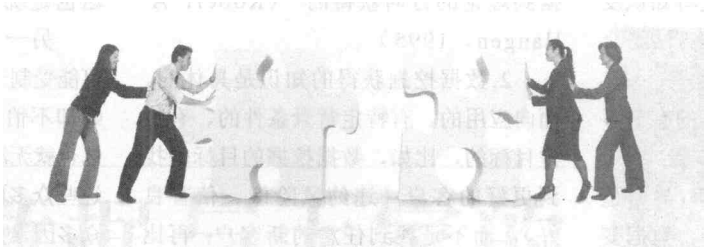

# 数据挖掘与隐私

□ 文/魏瑾瑞 谢邦昌 朱建平

## 引言：数据挖掘并不侵犯隐私

网络抓取个人浏览信息、购物清单卖给商家，这不是数据挖掘。数据挖掘的目的不是挖掘个人的隐私，它针对的是“典型人”的特征，考察的是大多数人的行为，是统计意义上的，不是个体的。比如，“买黄瓜的客户中，20%也买了西红柿”。而不会给出像 “小明买黄瓜的时候，可能（20%）也会买西红柿” 这样具体到某个人的结论。统计部门统计企业信息的时候也是一样的，它只会给出一个汇总的结果，而不会公布单个企业的任何信息。

集合中的任何一个点都包含有共性和特殊性的成分，统计抽离的是共性。用统计的术语来讲，统计主要关注的是平均、大概的情形，即3 σ以内。它指向的是一个常规的、典型的状态，是“大众行为”，不是极端值，也不是个别值。换句话说，统计只为沉默的大多数代言。

对于具体微观数据，除了访问权限和保密合约的限制之外，特别是向第三方（potential intruder）提供的时候，一般都会对敏感数据实施保密处理。值得注意的是，去掉身份标识（de-identified）的数据仍然是不安全的，我们需要对机密属性（confidential）做进一步保统计性质的变换，即数据伪装（data masking），包括去除某些变量、某些记录局部隐藏、全局重新编码、截断、数据互换、增加噪音等。所以，从技术本身而言，统计（包括数据挖掘方法）并不侵犯个人隐私。

## 什么是数据挖掘

经典的回归和分类 方法是建立在独立重复观测基础上的，并且通常附有诸如正态分布之类的假定，模型也大都有显性表达式；而现代回归和分类方法则是建立在一些训练样本基础之上的数值算法，进而转换成计算机程序来实现的。比如数据挖掘中使用的决策树、随机森林、助推法、神经网络、最近邻估计、核估计……事实上，从某种程度上而言，数据挖掘可以看做是现代统计学，或至少是现代统计学的一部分。正如伍德里奇所言，现代的实证研究越来越与经典范式格格不入。与手工计算时代相比，计算机时代的统计学正处在蜕变之中。

简言之，数据挖掘（Data Mining，DM），就像是矿工采矿、考古学家挖掘有价值的文物一样，它结合知识发现的各种算法、统计学的数据分析方法、可视化等各种技术，从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中 ，提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识 的过程，并且这个过程不是一蹴而就的，而是多重往复进行的。

事实上，数据挖掘可以被看做是在寻找模式或规则的过程；也可以被看做是一个数据压缩的过程，即，将大量细致的数据映射到一个简单的、容易理解的集合，比如，分类实际上就是将数据集映射到一个所谓的类的过程，其中关键是寻找这个映射规则；再比如描述统计、概念描述等。

（一）数据挖掘与知识发现

1.数据挖掘可以看做是将知识发现（knowledge discovery，KD） 技术应用于数据库，即KDD（knowledge discovery in database）。

值得注意的是，尽管数据挖掘的对象可以是传统事务性数据库、数据仓库、关系数据库、空间数据库、文本数据库、web数据库 等，但数据仓库可能是更好的挖掘环境，效率可能也更高一些，因为数据仓库中的数据已经完成了预处理和整合。另一方面，如果把数据挖掘放在机器学习（人工智能）的范畴里面来看，机器学习强调的是算法（遗传算法，人工神经网络）；如果把数据挖掘放在统计学的范畴里面来看，统计学强调的是模型（描述变量之间的

关联)、推断（由样本归纳出未知总体的特征)、分类、概括等。所以，综合以上两个角度，数据挖掘是将知识发现的技术（算法）和统计学的方法结合应用于数据（仓）库。

2.DM也可以看做是KD的有效手段，或是KDD的一个重要步骤：

（1）数据准备：数据清理(异常值处理)、数据集成、数据选择、数据变换(比例缩放)等。

（2）数据挖掘（选择有效的算法来找到感兴趣的模式)。

（3）模式生成与评估。

（4）知识表示。

在论文集《知识发现与数据进展》(1996）中，Fayyd、Piatetsky-Shapiro和Smyth将KDD与DM做了明确的区分：KDD是从数据中辨别有效的、新颖的、潜在的、有用的、最终可理解的模式的过程；而DM是KDD中通过特定的算法（在可接受的计算效率的限制内）生成特定模式的一个步骤。

（二）从数据中挖掘什么：未知的、有用的、具体的知识

1．现代科学基于“首要原则模型”，然后用实验数据来验证之，即先是逻辑模型，尔后是实证模型。但是，对于很多复杂系统而言，“首要原则模型”往往是未知的，且这样的系统生成了大量的数据，“数据过剩”、“信息混沌空间”、“丰富的数据贫乏的知识”……我们的电脑的容量越来越不够用、书籍的出版和我们的藏书也有泛滥之势，然而，其中有利用价值的部分却少之又少或尘封窖藏难以被发现。在这种情况下，开采有用的知识等同于抛弃无用的信息。

数据挖掘与传统的数据分析的本质区别是，数据挖掘是在没有明确假设的前提下去挖掘信息、发现知识。DM与OLAP（联机分析处理）的本质区别是，DM不是用来验证假设的，而是产生假设。

所以，在这个意义上，数据挖掘获得的知识是预测性的（潜在的)。换句话说，DM不是验证性的，而是探索性的。数据挖掘获得的知识是新奇的、人们事先不知道的，倘若挖掘到的是常识，比如，情人节鲜花热卖，则相当于证明了糖是甜的，这是没有价值的。“我们大部分最伟大的胜利是按照从数据到理论的方向获得的”(Robert A. Haugen，1995)

2.数据挖掘获得的知识是具体的、面向应用的、有特定背景条件的、有特定目标的，比如，数据挖掘的目标是找到更好的客户（违约风险低、信誉良好)，而不是找到任意的新客户；再比如，有目标的营销，在限制营销成本的同时，增加了营销的收益。再比如，使用数据挖掘来分析客户的生命期信息。另一方面，数据挖掘也应该容易被用户所理解和接受，所以，结合可视化、与用户交互是必要的。“为了提出一个有意义的问题的陈述，拥有领域内详尽的知识和经验是必不可少的。不幸的是，许多应用研究往往以牺牲对问题的清晰描述为代价而集中在数据挖掘技术上。”

（三）数据挖掘是一个过程

数据挖掘是一个过程，并且，这个过程不是线性的流程，而是反复的迭代。因为很可能在分析阶段会对数据处理有新的要求。

“实际上，现实中所发生的是：数据挖掘变成了一个反复的过程。一个人对数据进行研究，利用一些分析工具对数据进行检查，决定从另外一个角度来看它，可能会对数据进行修改，然后又回到开始，应用别的数据分析工具，得到一个更好的或不同的结果。这个过程可能循环许多次，每一种技术都被用到，以便查明数据的细微的不同的父母……”

## 结 语

诚然，在犯罪侦查中，我们完全可以根据个人的历史数据来确定嫌疑人。然而，即便是对个人信息的挖掘，它也不会将巨细靡遗的记录公之于众，或断章取义，散布其中的一条或多条记录，而是总结这些记录的特征。这在统计学中，就相当于概率分布的期望、方差、偏度等概率分布的特征。

为什么统计不关心一条具体的记录？因为一条记录并不说明什么，一次体检报告并不能说明健康状况，一次考试也并不能说明个人的能力，这里面有扰动。所以，我们需要更多的数据来看这个过程的平均趋势。勿要因噎废食。这也是统计学所蕴含的道理。

另一方面，不同个体的经济行为，可能受制于不同的因素，比如有的人怕蟑螂却不怕老鼠、而有的人恰恰相反，所以也就无法统一控制这些因素。如何驾驭这些众多的影响因素？统计学告诉我们，众多因素只是表象，我们可以找到其中的共同因素（影响结果的主要因素)；或者异质性特别大的时候，我们需要按照不同的类别来找共同因素。总之，个性——平均之外——在统计中常常是被忽略的（放入扰动项)，也即，隐私一般是在统计学视野之外的。 CS

参考文献：

[1]Mehmed Kantardzic. Data Mining: Concepts, Models, Methods and Algorithms[M]. IEEE Press(2002). 清华大学出版社（2003).

[2]高济等编. 人工智能基础[M]. 高等教育出版社. 2002.

[3]苏新宁等著. 数据仓库和数据挖掘[M]. 清华大学出版社. 2006.

[4] 毛国君等编著. 数据挖掘原理与算法[M]. 清华大学出版社. 2005.

[5] 段云峰等编著. 数据仓库及其在电信领域中的应用[M]. 电子工业出版社. 2003.

[6] David Hand, Heikki Mannila, Padhraic Smyth. 数据挖掘原理[M]. 机械工业出版社. 2003.

[7] Jiawei Han, Micheline Kamber. 数据挖掘:概念与技术[M]. 机械工业出版社. 2001.

[8] 朱世武，崔嵬，张尧庭，谢邦昌. 数据挖掘运用的理论与技术[J]. 统计研究. 2003年第8期.

[9] 朱建平著. 数据挖掘的统计方法及实践[M]. 中国统计出版社. 2005.

[10]William A. Building the Data Warehouse [M]. 机械工业出版社. 第四版.

[11]刘丽，白雪梅，刘永久. 国外微观数据发布的进展与启示[J]. 统计研究. 2010年第8期.

[12]Krishnamurty Muralidhar and Rathindra Sarathy. Data Shuffling: A New Masking Approach for Numerical Data. Management Science.Vol. 52, No. 5 (May, 2006), pp. 658-670 .

作者单位：厦门大学经济学院
台湾辅仁大学