数据挖掘的统计学内涵

□ 文 / 魏瑾瑞　朱建平　谢邦昌

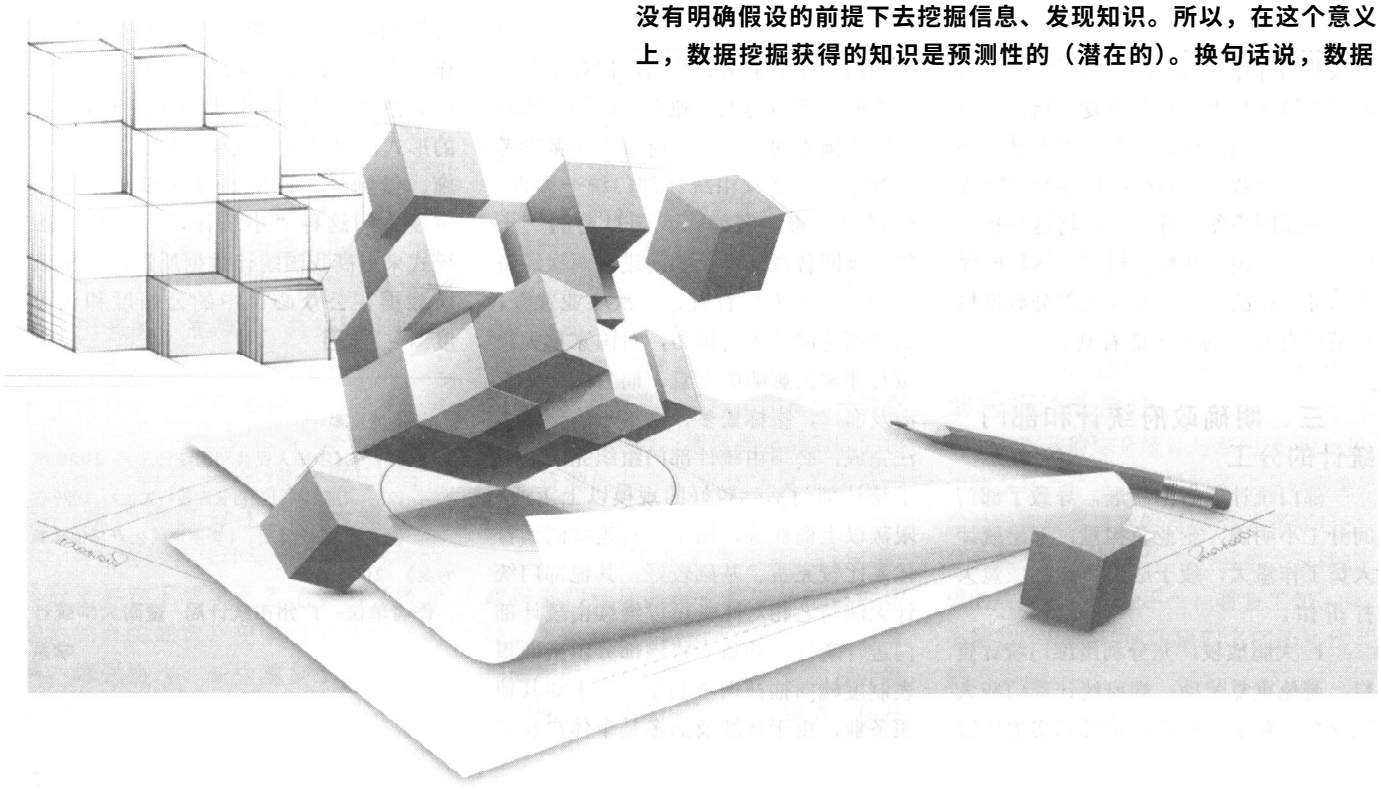
统计有培根逻辑做支撑，而数据挖掘则更多地得益于当代信息技术的飞速发展——但是相对于由此产生的大量冗余数据而言，我们似乎并没有获得多少信息。如果说数据挖掘（data mining，DM）研究的是经过清洗的全样本数据（Population），那么统计研究的则是样本数据（sample）。事实上，数据挖掘中所使用的数据也是根据特定目的抽取的，但显然它不同于统计中所谓的抽样。我们“没有时间看每一事物，尤其没有时间充分看每一事物，而且因为不看比看错了还要好些，所以他必须做出选择”（H. Poincar é，1913，The Foundations of Science）。从这个角度来讲，数据挖掘与统计之间的关系并不像我们想象中那么紧密，因为，第一，数据挖掘只是将知识发现（knowledge discovery，KD）技术应用于数据库，即数据库中的知识发现KDD（knowledge discovery in database）；也可以直接将DM看作是KDD的有效手段，或KDD的一个重要步骤：

数据准备→数据挖掘→结果的解释与评价

这类似于“统计分析”在统计流程中的角色：

数据准备→统计分析→结果的解释与评价

第二，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。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，数据



挖掘不是验证性的，而是探索性的，且其所获得的知识是新奇的、人们事先不知道的。即，数据挖掘的目的是要找到“例外”，“以规则的事实开始是合适的，但是规则一经确立，与它完全一致的事实不久以后也就没有意义了，因为它们不能再告诉我们任何新的东西，这时候，例外变得重要起来。”（H．Poincaré，1913，The Foundations of Science）

从统计学的观点来看，例外至少有两种，一种是异常值，它被认为是个别现象、偶然的，故而不予考虑或剔除；还有一种是珍贵值，它被认为是系统变异，其中潜藏着重要的质变信息或“扩展信息”，需要特别重视 。这时候，例外并不是规律以外的偶然事件，而很可能是未经确认的“新规律”。“偶然性并不是理由，它必定能在某个意想不到的定律中找到”（H． Poincaré，1913，The Foundations of Science）。比如一个非整数的阶乘的出现开启了一个新的领域——Γ（x），非整数次幂导致了对数的产生。因此从这个意义上来讲，科学的前进来自于对规律失效的关注 。

然而如果将传统的统计学看作是手工计算时代的统计学，那么数据挖掘完全可以看作是计算机时代的统计学（现代统计学），或至少是现代统计学的一部分。比如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。在这种情况下，要对评价结果进行评价，我们很难从逻辑上做出证明（如统计量的无偏、有效等性质），而只能通过模拟来实现。注意，对结果的评价除了机器（模拟）评价之外，更重要的是用户评价。这实际上对数据挖掘提出一个要求，即，数据挖掘的结果应该容易被用户所理解和接受。所以，结合可视化、与用户进行沟通是必要的，这不仅体现在结果的评价上面，而且贯穿于整个挖掘过程。“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 。”

注意：数据经过处理之后仍是数据，处理的目的仅仅是为了便于解释，倘若经过处理后的数据具有了某种特定的意义，我们才说它是信息。换句话说，数据只有经过解释（获得了某种意义）才成为信息，比如17.37是数据，而“原油平均每桶17.37美元”则是信息；战争年代为防敌军截获信息而将其编译为密码，就是将信息还原为数据进行传递，而密码破译则可看作是数据挖掘；有些人背诵圆周率这个无理数也大都将其转换为有意义的信息才能记得住。因此，数据是信息的载体（表现形式）；信息是数据的含义（解释），而数据挖掘的目标是要从数据中获得信息 。不过，我们从数据中挖掘到的规律也并不都是规律，这取决于能否做出合理的解释。比如根据数据，我们发现火山爆发的次数与冰激凌销量有很大的相关关系，但这只是数据上表现出来的关系，并存在逻辑关系，而啤酒与尿布的相关关系是可以被解释的，因此后者成为一条规则。这里值得注意的是，（1）这里的“解释”与前文提到的“评价”是两个不同的概念。（2）解释分为内部解释（interpretation）和外部解释（explanation），前者是人文社会科学所强调的，而自然科学则往往倾向于外部解释，比如光为什么是红的？外部解释是，这取决于它的波长。谈论人，不从吃饭而是从心理谈起，这是内部解释。（3）解释往往都是回顾性的，比如进化论，再如1900年召开于巴黎的第二届国际数学家大会，人数寥寥，但现在我们知晓其意义深远。正如Hilbert所言，“很难且常常不可能提前判断一个问题的价值”。可能伟人或伟大的事件都需要痛苦来滋养。

事实上，数据挖掘之所以称之为数据挖掘，很大程度上是因为数据挖掘的对象是海量数据，而不是因为所使用的工具。比如海量数据加一般传统的统计方法，要比小容量样本加数据挖掘工具（关联规则、神经网络、支持向量机），来得更“数据挖掘”。事实上，对于海量数据而言，传统的统计方法若不做适当的修正则很难直接发挥作用，比如一个很大的数X，由于舍入误差，计算机运行就会出现：X+10-X=0，X-X+10=10，这样的尴尬现象。此外，海量数据一个不容忽视的重要特点是，因数据量庞大而无法保证数据的完整性、准确性和一致性，比如与低频数据相比，实际上高频数据的质量并不高（交易数据会因种种原因而缺失，某些交易的确切时间也不见得准确，再有就是微结构噪音等因素干扰）。 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]吕晓玲. 谢邦昌编著. 数据挖掘方法与应用[M]. 中国人民大学出版社. 2009

[10]谢邦昌编著. 数据挖掘Clementine应用实务[M]. 机械工业出版社. 2008

[11]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

[12]谢邦昌编著. 商务智能与数据挖掘Microsoft SQL Server应用[M]. 机械工业出版社. 2008

[13]谢邦昌．朱建平．来生强．Excel 2007数据挖掘完全手册[M]. 清华大学出版社. 2008

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

[15]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

[16]吴国盛著. 技术哲学讲演录[M]. 中国人民大学出版社. 2009

作者单位：厦门大学经济学院计划统计系