国家社科基金重大项目

GUO JIA SHE KE JI JIN ZHONG DA XIANG MU

现代统计研究丛著　主编 贺铿

数据挖掘前沿问题

吴喜之 马景义 吕晓玲 闫洁 著

SHU JU WA JUE QIAN YAN WEN TI

中国统计出版社
China Statistics Press

国家社科基金重大项目

GUO JIA SHE KE JI JIN ZHONG DA XIANG MU

现代统计研究丛著　主编 贺铿

数据挖掘前沿问题

吴喜之 马景义 吕晓玲 闫洁 著

SHU JU WA JUE QIAN YAN WEN TI

中国统计出版社
China Statistics Press

(京)新登字 041 号

图书在版编目(CIP)数据

数据挖掘前沿问题/贺铿主编. 吴喜之、马景义等著.
—北京:中国统计出版社,2009.3
(现代统计研究丛著)
ISBN 978—7—5037—5639—9

Ⅰ.数…
Ⅱ.①吴… ②马…
Ⅲ.数据采集
Ⅳ.TP274

中国版本图书馆 CIP 数据核字(2009)第 026221 号

数据挖掘前沿问题

主　　编/贺　铿
作　　者/吴喜之　马景义　吕晓玲　闫　洁
责任编辑/徐　颖
封面设计/杨燕超
出版发行/中国统计出版社
通信地址/北京市西城区月坛南街 57 号　邮政编码/100826
办公地址/北京市丰台区西三环南路甲 6 号
网　　址/www.stats.gov.cn/tjshujia
电　　话/邮购(010)63376907　书店(010)68783172
印　　刷/河北天普润印刷厂
经　　销/新华书店
开　　本/700×1000 mm　1/16
字　　数/145 千字
印　　张/9.5
印　　数/1—5000 册
版　　别/2009 年 7 月第 1 版
版　　次/2009 年 7 月第 1 次印刷
书　　号/ISBN 978—7—5037—5639—9/TP·46
定　　价/30.00 元

《现代统计研究丛著》编委会

总　序

《现代统计研究丛著》(以下简写为《丛著》)是国家社科基金重大课题的研究成果。

2004 年,“现代统计研究”在全国哲学社会科学规划办立项以后,于当年 9 月 10～12 日召开专家座谈会,研讨了课题研究思路,课题组进行了初步分工。2005 年 2 月 23 日又在北京召开课题组成员会,认真讨论了研究大纲。2006 年 5 月 6～9 日在西安召开了课题结项与《现代统计研究丛著》出版座谈会,讨论并通过了各本著作的基本内容,决定了关于出版和组建《丛著》编委会的原则。编委会由课题组成员和课题承担单位西安统计研究院的主要负责人组成,主任委员由课题组的组长担任。《丛著》编辑出版中的重大问题由编委会集体研究决定。2007 年 12 月 22 日在北京召开《丛著》主要作者座谈会,确定了各本专著的书名、署名及总序等问题。

“现代统计研究”课题立项时确定的目标是梳理统计学学术思想,阐释统计学的本质及学科边界,促进统计学的发展和应用。

究竟何谓统计学,历来学者立说纷纭。早在 1869 年,现代统计学的主要奠基人之一的 A・Quètelet 在第七次国际统计大会上报告,关于统计学的定义已有 180 种之多。近几十年来,我国还经历了“两门统计学”和“大统计思想”的大讨论。这充分说明,为了促进统计学更好发展,对统计学的学术思想和研究范围进行梳理十分必要。

“现代统计研究”课题组成员对现代统计学的本质意义有大体一致的认识。经过研讨,其基本共识是:统计学是一门方法论科学,包

括工作性方法、技术性方法和工具性方法。工作性方法系指统计工作的制度、规范;技术性方法包括统计调查方法和统计分析方法;工具性方法在现代统计学科体系中系指与统计相关的现代信息技术,缩写为SIT。

最终形成的研究成果分为三类:第一类,阐释统计学的性质。着重从现代统计学发展的历史轨迹、学科体系和教育思想等方面进行梳理;第二类,介绍统计学的基本内容、前沿问题和发展趋势。主要介绍与社会经济问题相关的调查分析方法以及国民经济核算理论、指数理论、投入产出分析方法、经济计量学方法和统计信息技术(SIT)等内容;第三类,介绍国内外统计工作制度及规范。《丛著》共有8本10册专著。

《丛著》主编主要把握指导思想和研究路径。每本专著的基本内容由主编提出,经课题组成员集体研讨,作者决定。在具体学术观点上求同存异。学术观点表达及论证方法由作者(有多个作者的由第一作者)负责。我们认为,学术研究不可强求完全一致,各抒己见有利于学术发展。同时,我们在《丛著》中表达的观点在统计学界是否认同,也不是一厢情愿可以决定的。只有坚持"百花齐放,百家争鸣"的方针,才有可能真正推动现代统计学向前发展。

《丛著》付梓得到了多方面的关心和支持。在此我们对中国统计出版社、西安财经学院、九三学社机关和全国哲学社会科学规划领导小组办公室的领导及相关工作人员表示衷心感谢!

由于时间仓促和我本人水平所限,《丛著》难免存在缺点和错误。我们真诚地期待着读者的批评。

贺 铿

2008年3月

前　言

这本书所涉及的是代表统计领域发展最快的部分，也是对传统统计提出最重大挑战的部分，这就是在数据挖掘实践中发展的新方法。而本书所着重说明的前沿问题是：数据挖掘方法中最普遍应用的分类和回归中成为近年来发展热点的组合方法。

随着人类活动的不断发展，各个领域产生了不断增加的大量数据。由于许多数据的数据量很大，而且数据的结构日趋复杂，传统的统计方法无法满足分析这些数据的需要，这就产生了数据挖掘的实践，以及所产生的大量新型算法。在数据挖掘中也应用一些传统统计的方法，但这些方法的理解和检验与传统统计有所不同。更重要的是在数据挖掘中产生了许多新方法，它们从任何角度来看，都完全不同于传统统计的方法。由于这些新方法和传统统计的理念完全不同，在最初只有少数统计学家感兴趣，而大多为计算机领域的工作者所开发。因此，以数据分析为宗旨的统计学科损失了大量的机会、领域和人才。最近十多年来，不断有优秀的统计学家加入到数据挖掘的行列里来，数据挖掘也成为美国许多统计系的必修课程，数据挖掘因此进入了一个计算机算法和统计思维相结合的新时期。

由于历史原因，关于数据挖掘的文献，特别是书籍，多以计算机的术语和思维过程为特点。本书则按照统计领域的思维习惯来撰写，并

且力求通俗易懂。本书还通过一些贯穿始终的例子来增强读者对各个方法的理解。我们希望读者对本书所述内容不但能够理解，而且能够利用非商业的R软件来感受数据挖掘方法的实施。

著名统计学家C. R. Rao多次指出，统计的未来在于数据挖掘。我们希望通过本书的出版，能够使更多的统计学家参与数据挖掘的实践和理论探讨。让统计思想和计算机尽可能地完美结合，去面对世界上不断产生的新挑战！

在此，我们要诚挚感谢中国人民大学统计学院的研究生们对本书在内容校对和编写程序等方面所作的贡献。他们是：陈凯、宋捷、赵秀丽、刘苗、程晓月、魏博、斯介生、邵君舟。

吴喜之　中国人民大学
马景义　中央财经大学
吕晓玲　中国人民大学
闫　洁　首都经济贸易大学

2009年1月

目 录

第 1 章

数据挖掘概论

1.1　引言

在现代生活中，我们面对着所谓“信息爆炸”。在网络、遥感、金融、电讯、地理、商业、旅游、军事、生物医学等各个领域不断产生大量的数据。这些数据不但结构多种多样，而且数据量往往很大，数据量的大小甚至以 terabyte[①] 为单位。这些从各个领域中产生的数据远远超过了传统统计方法分析和处理它们的能力。如何能够把数据中的重要信息迅速有效地提取出来是非常重要的。在需求的拉动下，数据挖掘(data mining)就因此产生和发展了。

数据挖掘是利用各种复杂的数据分析工具来发现大型数据集中的各种未知模式和数据的内在关系，并据此进行预测的一个过程。数据挖掘的工具包括传统统计模型、数学方法及机器学习方法(包括诸如神经网络、决策树或其组合算法等通过对训练集的学习来改进预测性能的方法)。因此数据挖掘不仅包括收集和管理数据，而且包括了描述、分析和预测。数据挖掘的基础是数据，因此，收集、探索、选择合适的数据是至关重要的。其最简单、最初等的内容是描述数据，即通过考察所产生的描述性统计量及各种图表来发现潜在的数据内部的联系。但是描述数据本身不能产生指导性的结论，必须建立可预测的模型。这些模型不是精确的，而是现实世界的一个近似。一个好的模型可以帮助我

① 1 terabyte=10^{12}字节；1 gigabyte=10^{9} 字节；1 megabyte=10^{6} 字节。

们理解世界。模型建立之后需要在实践中进行验证，所有由模型产生的结果必须符合现实世界的规律。显然，这一切也是传统统计的目标和做法。

和传统统计一样，进行数据挖掘，必须要理解其目的，理解相关的领域，理解你所用的数据中各个变量的意义。所有的步骤都不应该盲目进行。数据挖掘不是一个全自动的过程，而是基于对数据和领域理解基础上的许多决策的结果。使用数据挖掘软件和使用一般“傻瓜”统计软件一样，可以使你避免纠缠具体方法的细节，但你必须理解各种方法的意义，能够在使用软件每一步时进行合理的决策或选择，并且能够解释各种输出。

数据挖掘的产品可以是功能强大的工具，但它们并不是自给自足的。要想成功地进行数据挖掘，需要熟练的技术和分析专家；这些专家必须能够设计分析和解释输出的结果。这样数据挖掘主要被与数据或与人有关的因素所局限。数据挖掘可以发掘模式和关系，但它并不告诉用户这些模式是否显著。这都依赖于用户的理解。另外，数据挖掘产生的模式必须和现实世界比较。这并不是数据挖掘本身可以解决的。此外，和传统统计分析一样，数据挖掘可以识别变量之间的关系，但这不意味着任何因果关系。

看上去，数据挖掘和统计没有什么区别。下面就此做一些讨论。

1.2 统计学家和计算机学家从不同角度看数据挖掘

数据挖掘实际上和统计的目标是没有什么区别的。按照不列颠百科全书，统计可定义为**收集、分析、展示、解释数据的科学**（Encyclopædia Britannica 2008）。这是历史相对悠久的统计在其发展中慢慢形成的统计界认可的定义，它包含了一系列固有的概念、理论和方法，有一个比较稳定的知识结构和基础。数据挖掘也完全符合这个定义，但由于它的发展经历较短，初期主要由计算机科学家开创，而且是脱离统计的传统体系而发展的，因此，它有其自己的特点。**根据计算机科学家的理解**，数据挖掘也被称为数据库的知识发现（Knowledge Discovery in Databases，KDD）；但严格说来，数据挖掘仅仅是 KDD 过程的一个组成部分。当大量数据存放在数据库等地方，需要把其中的信息和知识提取出来。这不是一步就可以完成的过程。它基本上可以分成几个部分：

1. 数据清洗（data cleaning 或 data cleansing）：把噪声和无关的数据去掉；

2. 数据整合(data integration):把许多常常是非常不同的数据源整合成一个共同的数据源;

3. 数据选择(data selection):决定并取出需要分析的数据;

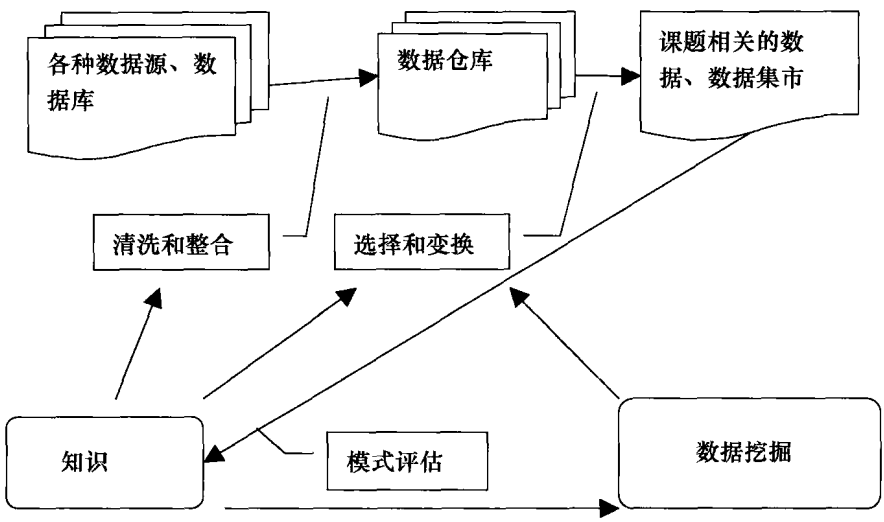
4. 数据变换(data transformation 或 data consolidation):把数据变换成适合于数据挖掘的形式;

5. 数据挖掘:应用各种技术把潜在有用的模式提取出来;

6. 模式评估(pattern evaluation):基于确定的标准,把确实感兴趣的模式识别出来;

7. 知识展示(knowledge representation):所有发现的知识被直观地通过可视化技术展示给用户,以帮助用户理解和解释数据挖掘的结果。

整个过程如下图所示。图中的流向并不是单向的,经常需要反馈。



显然,**按照统计学家的观点**,其中的1、2、3部分是分析前的数据准备部分,而第4、5、6、7部分是统计学家拿到数据后进行建模所必然要做的部分。按照计算机学家的知识和训练结构,他们在前面数据准备部分强调的较多也较细致,同时由于计算机学科通常不讲那么多的统计,因此把统计的建模过程分成了后面的4个部分来详细说明。

由于知识结构不同,由计算机学家撰写的文献,对于统计学家而言有些不

习惯。这首先表现在思维方式上，例如，有些统计学家很明白的概念，他们介绍得很详细，而另外一些诸如数据库、计算机科学的概念，他们一带而过。更令统计学家不习惯的是，在术语上计算机学家的用语远远不如具有严格数学训练的统计学家那么确定。比如，在一个文献中，同一术语在同一文献的不同地方可能以多种不同的名字出现。比如，在统计中具有深厚背景的**变量(variable)**一词，在统计文献中基本上没有同义词，而在计算机学家的文献中该术语除了变量之外，也随意地称为属性(attribute)、特征(feature)、特性(characteristic)、字段(field)等，而**数量变量(quantitative variable)**也叫“指标(index)”，**定性变量或分类变量(qualitative variable** or **categorical variable)**也叫“维度(dimension)”。此外统计中的**观测值(observation)**在计算机学科中可称为记录(record)、对象(object)、点(point)、向量(vector)、模式(pattern)、事件(event)、例(case or instance)、样本(sample)、项、实体(entity)等。由于观测值的术语在统计中的背景不那么强，由此统计中也常用点(point)、向量(vector)、例(case)等术语，但绝对不能用事件(event)、样本(sample)等另有确切数学含义的名词。值得注意的是，数据挖掘软件多半采用统计学家不那么熟悉的术语，各软件也并不统一。

由于术语不同，思维方式不同，再加上由数据挖掘软件的供应商支持的介绍数据挖掘的各种演示(ppt)的华而不实的宣传，给人们以数据挖掘神乎其神，以及买了数据挖掘软件就可以解决一切问题的感觉。难怪，这种对数据挖掘的宣传被人们认为目的是“挖掘挖掘者口袋(mining miners' pockets)”。

显然，数据挖掘的目的和传统统计的目的是完全一样的，即分类(判别)、聚类、回归、模式和规则的识别与发现等。但是，数据挖掘发展了许多对应于各种形式数据和大型数据的方法，特别是许多建模方法是基于算法而不是基于有限数学公式的。

而统计学家所关心的主要是数据挖掘的建模方法。特别是那些与传统统计根本不同的算法建模方法。这也是本书所关注的焦点。

1.3 数据源

由于数据挖掘最初是由计算机学家发展的，所以，数据源是不能回避的内

容。这部分工作多由数据库工作者承担，而发展建模方法的人，除了在线方法，通常不那么强调这部分知识的细节。

数据挖掘的数据源可以是大的数据(data)，或者数据库(database)，或者数据仓库(data warehouse)，或者数据集市(data mart)。当然，这些都是从外部世界收集的原始数据被存储的地方。数据挖掘需要保证可靠和方便的数据源，这意味着有大量的涉及数据库的工作要做。对数据进行分析之前，还必须根据需要来筛选，预处理和净化数据。

特别设计的为了用计算机快速搜寻和提取的数据或信息的集合称为数据库。数据库的构造使得在各种数据过程操作中可以很方便地存入、提取、修改和删除数据。数据库能够存储在磁盘、磁带、光盘或其他二级存储设备中。比如，电讯部门所有的通讯记录、银行所有业务的记录都形成了不断变化的数据库。而数据仓库则是一个面向目标的、整合的、只能够读的数据集合，是为了管理决策而建立的。对于一个企业来说，数据仓库把整个企业的各种不同的数据库整合起来，易于查询，易于对和既定目标有关的数据作出分析。而数据集市则是数据仓库的子集。比如，数据仓库是针对整个企业或总体战略的，而数据集市则可能是针对具体部门或者某一项目标的。

1.4 数据挖掘的应用

数据挖掘可应用于各个科学领域。比如核实验室统计粒子数目、动物保护部门对野熊的无线电项圈发出的信号进行解释等；地球周围无数的卫星传回的各种遥感数据、图片及其他信息的数量非常巨大，其中许多在收到后马上就公布了，以希望有人能够分析和利用。

许多企业早已运用数据挖掘来进行客户全程管理：吸引新客户，从现有客户得到最大回报，以及尽量保住有价值的客户。根据各种(有价值的和已经流失的)客户的特征，尽量避免流失好客户。电讯和信用卡领域是最先及最充分应用数据挖掘的领域之一，它们用数据挖掘来发现对它们服务的欺诈和滥用；保险和股票也使用数据挖掘来减少欺诈行为；医药行业用数据挖掘来预测手术结果、进行医学医药试验；公司在金融市场中用数据挖掘来了解市场和工业特征，并且预测公司及相关股票的业绩；零售商用数据挖掘来确定什么商品出

现在什么商店，甚至什么商品应该摆在商店内部的什么位置，还可以确定促销手段是否有效；制药业要挖掘大量化合物和遗传物质，以发现和开发有效的药物。

目前，传播领域已经基本上数据化了，大量的录相、录音及其他多媒体信息需要数据挖掘方法来管理；各种工程及其他领域所用的计算机辅助系统和软件工程都产生大量数据，包括设计图、电路、软件代码、函数库等，需要管理和维持；大量的文字和诸如电子邮件之类的信息、互联网及其产生的海量数据更是数据挖掘系统的大用户。可以说，没有人能够穷举数据挖掘可能产生的用武之地，对数据挖掘新的无法预料的需求每天都在产生。