统计数据分析与应用丛书

Statistics

基于 SPSS Modeler 的数据挖掘

（第二版）

薛 薇 编著

中国人民大学出版社

统计数据分析与应用丛书

Statistics

基于 SPSS Modeler 的数据挖掘

（第二版）

薛　薇　编著

STATISTICS

中国人民大学出版社
· 北京 ·

图书在版编目（CIP）数据

基于 SPSS Modeler 的数据挖掘/薛薇编著. —2 版. —北京：中国人民大学出版社，2014.10
（统计数据分析与应用丛书）
ISBN 978-7-300-20069-9

Ⅰ.①基… Ⅱ.①薛… Ⅲ.①统计分析-软件包 Ⅳ.①C819

中国版本图书馆 CIP 数据核字（2014）第 225042 号

统计数据分析与应用丛书
基于 SPSS Modeler 的数据挖掘（第二版）
薛 薇 编著
Jiyu SPSS Modeler de Shuju Wajue

出版发行	中国人民大学出版社		
社　　址	北京中关村大街 31 号	**邮政编码**	100080
电　　话	010－62511242（总编室）		010－62511770（质管部）
	010－82501766（邮购部）		010－62514148（门市部）
	010－62515195（发行公司）		010－62515275（盗版举报）
网　　址	http://www.crup.com.cn		
	http://www.ttrnet.com（人大教研网）		
经　　销	新华书店		
印　　刷	北京七色印务有限公司	**版　　次**	2012 年 3 月第 1 版
规　　格	185 mm×260 mm　16 开本		2014 年 10 月第 2 版
印　　张	25.25 插页 1	**印　　次**	2014 年 10 月第 1 次印刷
字　　数	592 000	**定　　价**	49.00 元

前言
Preface

数据挖掘技术具有广阔的应用领域和发展前景，众多有识之士纷纷选择 SPSS Modeler 作为数据挖掘的工具软件，因此 SPSS Modeler 软件已经连续多年雄踞数据挖掘应用软件之首。

Modeler 的前身名为 Clementine，2009 年 IBM 公司收购了 SPSS 数据分析软件公司，并将其广受赞誉的 SPSS 统计分析软件和 Clementine 数据挖掘软件进行整合，将 Clementine 更名为 SPSS Modeler（简称 Modeler）后再次推向全球市场。

Modeler 充分利用计算机系统的运算处理能力和图形展现能力，将方法、应用与工具有机地融为一体，是解决数据挖掘问题的最理想工具。

Modeler 不但集成了诸多计算机科学中机器学习的优秀算法，同时也综合了一些行之有效的统计分析方法，成为内容最为全面、功能最为强大、使用最为方便的数据挖掘软件产品。

Modeler 继续保持了 SPSS 产品的一贯风格：界面友好且操作简捷。原因在于 Modeler 始终把自己的操作者定位于实际工作部门的一线人员，而不是数据分析专家。这种所谓“傻瓜型”软件成为 Modeler 不断开拓市场的利器。

本书作者一直从事计算机数据分析的教学与科研工作，并长期跟踪研究 SPSS 公司的数据分析系列产品，具有相当丰富的数据分析软件开发经验。因此深知，一个基础相对薄弱的读者应该从哪些方面入手，才能很快地使用 Modeler 开始数据分析工作，并逐步成长为一名有经验的多面手。

我们认为读者掌握 Modeler 软件应体现三个层面：首先是软件操作层面，读者通过实际操作，尽快掌握软件的使用方法和处理步骤；其次是结果分析层面，读者通过案例演示，基本明白软件的输出结果，从而得出正确的分析结论；最后是方法论层面，读者通过对某个算法基本思路的了解，进一步提高方法应用和分析水平，升华对数据挖掘方法的认识。所以，注重对每种方法的操作使用、结果分析和算法基本思路的讲解是本书最重要的特征。

本书适用于从事数据分析的各应用领域的读者，尤其是商业销售、财会金融、证券保险、经济管理、社会研究、人文教育等行业的相关人员。同时，也能够作为高等院校计算机类、财经类、管理类专业本科生和研究生的数据挖掘教材。

针对上述读者群，在全书的编写中我们努力体现以下特色：

1. 以数据挖掘过程为线索介绍 Modeler

目前，具备基本的计算机操作能力已经不是读者的主要障碍，数据挖掘的过程与方法才是读者关心的主题和应用的难点。所以，本书以数据挖掘的实践过程为主线，从 Modeler 数据管理入手，说明问题从浅至深，讲解方法从易到难。这样，能使读者在较短时间内掌握 Modeler 的基本功能和一般方法，并可迅速运用到实际工作中去。

2. 将数据挖掘方法、软件操作、案例分析有机结合

目前，经过消化的中文图书和资料相对短缺，Modeler 相关图书一般都比较侧重对其英文手册的翻译介绍，侧重于对计算机操作过程的描述。而对数据挖掘方法则较多地罗列数学公式，输出结果也缺少恰当的解释。本书则结合实际案例，侧重数据挖掘方法核心思想和基本原理的阐述，以使读者直观理解方法，正确掌握方法的应用范围。

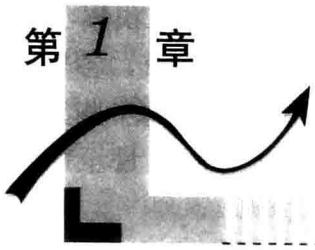
3. 数据挖掘方法讲解全面，语言通俗

本书对 Modeler 的数据挖掘算法进行了全面的分析和应用，内容力求丰富翔实。同时使用通俗的语言和示例讲述算法，尽量避免使用公式和推导堆砌算法。

请读者到人大经管图书在线（http://www.rdjg.com.cn）下载本书案例数据和数据流文件。数据流文件需使用 Modeler 14.2 以上版本打开，执行时只需修改数据源节点中的数据文件所在目录项，即可正确执行流文件。

在此特别感谢中国人民大学出版社对本书出版的大力支持和各位编辑热情细致的工作。由于水平所限，书中难免出现问题和错误，敬请各位读者批评指正。

目　录



第1章 数据挖掘和Modeler使用概述

20世纪90年代中后期以来，数据挖掘作为具有鲜明跨学科色彩的应用研究领域，已成为众多行业数据分析者瞩目的焦点。数据挖掘是一个利用各种方法，从海量数据中提取隐含和潜在的对决策有用的信息和模式的过程。因具有处理和分析海量数据的能力，注重弱化分析方法本身对数据的限制，以满足数据建模的合理性和适应性，强调与计算机技术相结合，以实现数据分析的可操作性和可实现性，数据挖掘正逐步成为数据分析应用实践的新生代和领军者。同时，随着数据挖掘方法的不断成熟及其应用的日益普及，数据挖掘软件的研发也取得了可喜的成果。目前，以SPSS Modeler为代表的数据挖掘软件，已行之有效地将束之高阁的数据挖掘理论成果解放到数据分析实践中，并普遍应用于商业、社会、经济、教育、金融、医学等领域，成为数据分析的主流工具。

§1.1 数据挖掘的产生背景

数据挖掘是在计算机数据库技术蓬勃发展、人工智能技术应用领域不断拓展、统计分析方法不断丰富发展的进程中，有效迎合数据分析的实际需求而逐步形成和发展起来的具有鲜明跨学科色彩的应用研究领域。

1.1.1 海量数据的分析需求催生数据挖掘

20世纪80年代以来，随着计算机数据库技术和产品的日益成熟以及计算机应用的普及和深化，各行业部门的数据采集能力得到了前所未有的提高。各组织通过其内部的业务处理系统、管理信息系统以及外部网络系统，获得并积累了浩如烟海的数据。

从微观管理层面看，以商业领域为例，美国某著名连锁超市的数据库中已积累 TB[①]级以上的顾客购买行为数据和其他销售数据。而随着互联网和电子商务的普及，各类网上商城、网上书店和营业厅等积累的 Web 点击流存储容量多达 GB 级。国内的一些知名电子商务平台，全国注册用户高达几亿，日交易量超过千万笔，日交易数据量至两位 TB 级。

全球著名数据挖掘咨询公司 KDnuggets 2012 年所做的调查[②]显示，被调查的 148 家公司中，2012 年大约 4%的公司处理和分析的最大数据量超过 100PB，而这个指标 2011 年为 0，如图 1—1 所示。

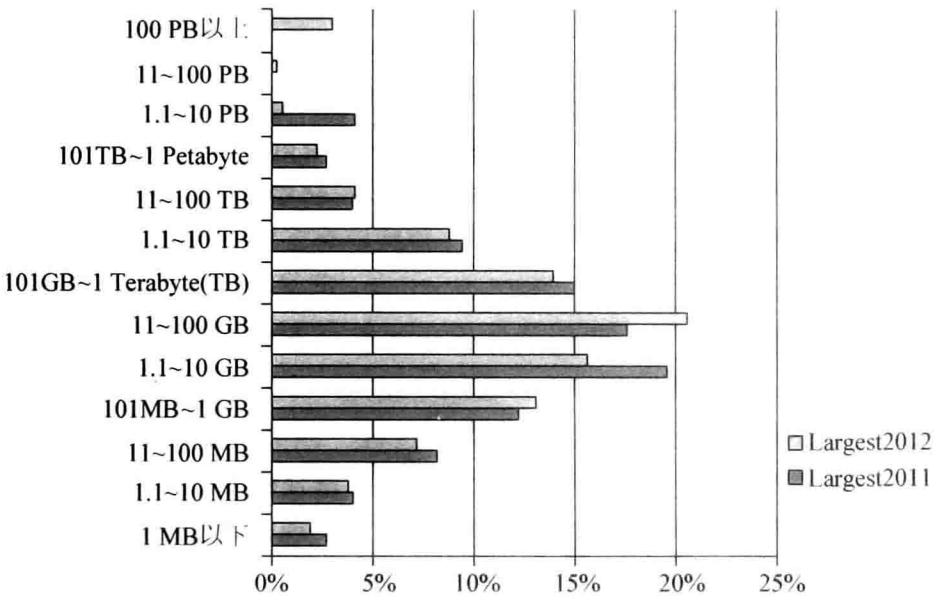


图 1—1　KDnuggets 的 2012 年数据量调查结果

在严酷的市场竞争压力下，为更客观地把握自身和市场状况，提升内部管理和决策水平，企业管理者面对如此丰富的海量数据，分析需求愈发强烈，希望借助有效的数据分析工具，更多、更有效地挖掘出“隐藏”在数据中的有助于管理和决策的有价值的信息。

例如，制造业已从过去的粗放式生产经营模式过渡到精细化的生产管理。决策者需要通过对已有数据的挖掘，了解客户偏好，设计最受市场欢迎的产品；制定合适的价格，确保企业的利润；了解市场需求，调整产销计划，优化库存结构；评估供应商质量、供应合同和订单违约率，提高产品合格率以及风险控制能力等。

再例如，电子商务盛行的今天，电子商务平台的管理者更需通过对交易数据的分析挖掘，计算各类指数（如买家指数、卖家指数、产品指数等），了解买卖双方的特征规律，掌握产品交易的地区和行业分布特点及趋势，预测热销品牌，制定商业广告投放策略，进而实现业务优化分析及促销活动效果评估等。此外，还可为国家宏观经济决策提供依据。

---

① 1TB=1 024GB，1PB=1 024TB。

② 参见 http://www.kdnuggets.com/polls/2012/largest-dataset-analyzed-data-mined.html。

从宏观管理层面看，国家政府部门所积累的数据量也令人瞠目。例如，一次全国经济普查或人口普查所采集和处理的数据条目均在千万级以上。另外，在互联网的大背景下，微博、博客等新媒体已成为传递民声的不可忽视的重要渠道。目前，国内大型网站的微博注册用户数过亿。海量信息的互联网已经成为各阶层利益表达、情感宣泄、思想碰撞的舆论渠道，成为反映社会舆情的主要载体，也是政府治国理政、了解民意的新平台。因此，为保证出台政策的科学性和全面性，需要以互联网络数据为分析对象，借助有效的数据挖掘方法，剖析社会网络结构，研究热点话题的形成脉络，分析社会情绪倾向，进而为政府管理和政策制定等提供参考。

总之，引用数据仓库领域的革新者、作者、教育家和顾问，世界知名的数据仓库专家拉尔夫·金博尔（Ralph Kimball）[①] 在其 1998 年的著作 *The Data Warehouse Lifecycle Toolkit: Expert Methods for Designing, Developing, and Deploying Data Warehouses* 中的一句话："我们花了 20 多年的时间将数据放入数据库，如今是该将它们拿出来的时候了。"

然而，令人棘手的问题接踵而来。以往人们得不到想要的数据，是因为数据库中没有数据，而现在仍然无法快捷地得到想要的数据，其原因是数据库里的数据太多了。由于缺少获取数据库中利于决策的有价值数据的有效方法和操作工具，人们对规模庞大、纷繁复杂的数据显得束手无策，原本极为宝贵的数据资源反成了数据使用者的负担。于是，所谓"信息爆炸"、"数据多但知识少"成为一种普遍的怪现象。

究其原因，一方面，辅助决策的数据大多来自企业的多个业务处理系统。业务处理系统是面向特定业务问题，按照特定业务的流程和规范开发的。企业内部不同业务处理系统往往是分散独立研制的，加上规划和技术等诸多原因，使得各系统基本处于"封闭"状态，系统之间的数据交换需求不多，数据交换渠道也多不畅通，"信息（数据）孤岛"现象比比皆是。这样的客观条件必然给数据整合带来极大障碍。而如果无法有效快捷地将各系统中的数据整合到一起，就无法及时得到全面准确的数据，更无法进行分析进而做出正确决策。

另一方面，数据的定量分析是科学决策的前提。但实施定量分析需要深厚的专业知识，更需要有效的分析工具。一般业务处理系统中的数据分析功能相对简单，通常只能制作各种数据汇总报表，无法实现对数据的深层次分析，不能很好地满足决策者的定量分析需求。

> 大规模海量数据的整合处理和深层次量化分析的实际需求，直接孕育了 20 世纪 90 年代初期的两项重大技术：数据仓库和数据挖掘。

数据仓库和数据挖掘的产生和发展，使得当今的计算机网络应用体系从业务管理层逐

---

① 自 1982 年以来，拉尔夫·金博尔博士一直是数据仓库行业最主要的开拓者，目前是最知名的演讲人、咨询师与培训员之一，《智能企业》（*Intelligent Enterprise*）杂志"数据仓库设计者"（Data Warehouse Designer）专栏的撰稿人，也是最畅销的《数据仓库生命周期工具箱》（*The Data Warehouse Lifecycle Toolkit*）与《数据仓库工具箱》（*The Data Warehouse Toolkit*）两部著作的作者，被列入数据库名人堂（Database Hall of Fame）。

步跃升到决策支持层。同时，两者在技术和产品上互相补充、互相促进，逐渐形成了融合发展的可喜局面，为最终形成具有一定通用意义的决策支持系统奠定了良好的基础。

1.1.2 应用对理论的挑战催生数据挖掘

应用需求对理论研究的牵引力是巨大的，没有应用背景的理论研究是没有价值的。在海量数据管理和分析应用呼声不断的同时，相关理论研究和应用实践的脚步从未停止。

数据库与数据仓库、人工智能与机器学习、统计学等理论的应用是数据挖掘诞生发展的坚实理论基础。

1. 数据库和数据仓库

计算机应用从其刚刚诞生时的以数值计算为主发展到当今的以数据管理为主，数据库的理论实践起到了巨大的推动作用。从最初的文件系统研究，到后来的层次模型、网状模型，直至 1969 年科德（E. F. Codd）提出关系数据模型，可以说数据库理论开创了数据管理的新时代。数据库因其卓越的数据存储能力和数据管理能力，得到了极为广泛的应用。但随着数据库中数据的不断积累以及人们对海量数据分析的强烈需求，针对数据库的理论实践，人们开始思考这样的问题：

“是否存在更有效的存储模式以实现高维海量数据的存储管理?”

“数据库中的数据处理能力仅仅局限在简单的查询和汇总层面吗?”

应用呼唤理论的发展和理论的再实践。通过数据库研究者的不懈努力，在数据库基础上逐渐发展完善起来的数据仓库（Data Warehouse，DW）技术，已经成为一种有效的面向分析主题的数据整合、数据清洗和数据存储管理集成工具。

同时，在机器学习和统计学等领域研究成果的基础上，数据仓库也在不断吸纳经典数据分析方法的精髓。众多知名的数据库厂商已经行动起来，通过调整企业发展策略，拓展和强化产品自身的数据分析能力，将数据挖掘的理论成果完美融入到产品研发和推广中。如微软公司的 Sql Server 提供了多种典型的数据挖掘算法；甲骨文公司的 Oracle 产品包含了包括关联规则和贝叶斯算法在内的众多数据挖掘算法；IBM 公司更是斥资 12 亿美元收购了享誉业界、麾下拥有 SPSS 统计分析软件和 Clementine 数据挖掘产品的 SPSS 公司，极大扩展了 IBM 的“信息随需应变”软件组合和商业分析能力。IBM 表示，收购 SPSS 将增强公司“信息议程战略”（Information Agenda Initiative）的业务实力，帮助客户更有效地将信息转化为战略资产。

与此同时，研究者也在为实现数据仓库中数据和分析模型的“无缝”连接、不同数据仓库产品间数据挖掘算法及分析结果的共享而不懈努力。例如，1999 年，微软公司提出了 OLE（Object Linking and Embedding）DB（DataBase）for DM（Data Mining）规范，研发了模型建立、模型训练和模型预测的数据挖掘语言。其核心思想是利用 SQL 和 OLE DB 将数据库中的关系概念映射到数据挖掘中。包括 IBM，微软，甲骨文，SAS，SPSS 等

大公司在内的数据挖掘协会，提出了预测模型标记语言 PMML（Predictive Model Markup Language），使常见数据挖掘算法的模型内容标准化，并以 XML 格式存储，使不同软件之间的模型交换和共享成为可能。以微软公司的 Sql Server 产品和 IBM 公司的 Clementine 产品为例，当用户在计算机中安装了 Sql Server，在 Clementine 中建立和执行数据挖掘流时，Clementine 会自动将数据挖掘流提交给数据库，并利用数据库系统所提供的各种数据管理和优化机制，直接读取数据库中的数据而不必下载到 Clementine 中，且模型结果可存储于数据库中。

数据库和数据仓库技术发展的直接应用成果是企业决策支持系统（Decision Support System，DSS）的盛行一时以及商业智能（Business Intelligence，BI）的大行其道。决策支持系统是辅助决策者通过数据、模型和知识，以人机交互方式进行半结构化或非结构化决策的计算机应用系统，是管理信息系统（Management Information System，MIS）向更高阶段发展的必然产物，能够为决策者提供分析问题、建立模型、模拟决策过程和方案的环境，通过调用各种信息资源和分析工具，帮助决策者提高决策的水平和质量。如果说早期的决策支持系统强调与专家系统相结合，那么随着数据仓库和联机分析处理（On-Line Analysis Processing，OLAP）等新技术的出现，多维报表分析则是新型决策支持系统的亮点。

1996 年，高德纳咨询公司（Gartner Group）[①] 提出了商业智能的概念。商业智能能够提供使企业迅速分析数据的技术和方法，包括收集、管理和分析数据，并将数据转化为有用的信息。数据挖掘的兴起和商业化发展，完成了商业智能中智能层次的飞跃，即从多维报表分析到问题的解决和数据的预测。

2. 人工智能和机器学习

人工智能和机器学习的理论研究一开始就具有浓厚的神秘色彩。针对如何利用计算机模拟人脑的部分思维，如何利用计算机进行实际问题的求解和优化等，人工智能和机器学习的理论研究可以说成果丰硕。然而，其理论实践的进程中却遇到了许多无法逾越的鸿沟。

以专家系统为例，作为人工智能和机器学习应用研究成果之一的专家系统，在某种意义上能够代替专家给病人看病，能够帮助人们识别矿藏，但却很难解决那些看似简单的问题。例如，专家系统建立过程中的知识获取问题。其中涉及诸如人脑是如何思维的，计算机技术人员以怎样的方式与领域专家交流，才能克服知识传递过程中的随意性和跳跃性等问题，以实现对专家领域知识的全面系统的获取。又例如，专家系统的知识表示问题。由于计算机的知识表示通常采用简单机械的“如果……那么……”方式，而专家领域的知识形式却是丰富多彩的，并不是所有知识都能够概括成“如果……那么……”的模式。再例如，专家系统存储的知识绝大部分是领域的专业知识，常识性知识很少。而没有常识的专家系统有时比傻子还傻。有“专家系统之父”之称的人工智能学家爱德华·艾伯特·费根

① 高德纳咨询公司，全球最具权威的 IT 研究与顾问咨询公司，成立于 1979 年，研究范围覆盖全部 IT 产业，就 IT 的研究、发展、评估、应用、市场等问题，为客户提供客观、公正的论证报告及市场调研报告。

堡姆（Edward Albert Feigenbaum）[①] 曾估计，一般人拥有的常识存入计算机大约有 100 万条事实和抽象经验，而将如此庞大的事实和抽象经验整理、表示并存储在计算机中，难度极大。

以计算机博弈为例，计算机博弈是人工智能和机器学习的另一项重大的应用研究成果。从 20 世纪 70 年代开始，世界各地的人工智能研究学者投入大量心血对国际象棋、中国象棋、五子棋、围棋等进行研究。各种计算机开始和人类下国际象棋，其间互有胜负。1997 年 5 月，IBM 研制的“深蓝”超级智能计算机与国际象棋大师卡斯帕罗夫进行的 6 局制比赛成为计算机博弈的巅峰之战，结果计算机以两胜三平一负的成绩获胜。“深蓝”出神入化的棋艺依赖于它的评估功能，即评估每一种可能走法的利弊。而评估功能的背后除了高性能的计算机硬件系统之外，还需要拥有数千种经典对局和残局的数据库，以及由国际象棋大师乔约尔·本杰明等人组成的参谋团队。计算机博弈的最大“死穴”是不按“套路出牌”所导致的低级失败。

“深蓝”赢了卡斯帕罗夫之后，计算机下棋的热点渐渐退去，人类自然语言的理解成为人工智能研究的新焦点。IBM 制造的超级智能计算机沃森（Watson）就是其中的最高成就之一。2011 年 4 月 1 日，借助美国著名的问答节目《危险边缘》，沃森与人类的“情人节人机大战”展开。《危险边缘》是一个综合性问答节目，题目涵盖时事、历史、艺术、流行文化、哲学、体育、科学、生活常识等几乎所有已知的人类知识。与沃森同场竞技的两位人类选手绝非泛泛之辈，他们是该节目有史以来成绩最好的人类参赛者。选手肯·詹宁斯（Ken Jennings）曾经在 2004—2005 年赛季中连续赢了 74 场，创造了该节目的纪录，赢得超过 250 万美元。另一位选手布拉德·拉特（Brad Rutter）则创造了节目最高个人奖金的纪录，奖金数达到 325 万美元。然而，比赛的最终结果是沃森以近 8 万分的得分，将两位得分均在 2 万左右的人类选手远远地甩在了后面。

尽管如此，智能计算机面临的自然语言理解的挑战仍是严峻的。例如，如何克服“机械性”理解的弱点。沃森曾经错答这样一道题目：“这个被信赖的朋友是一种非奶制的奶沫。”正确答案是咖啡伴侣。因为咖啡伴侣多是植物制的奶精，并非奶制品，且人类做这道题时会很快想到“朋友”对应“伴侣”。但沃森却需要在数据库里寻找“朋友”、“非奶制”、“奶沫”这几个词的关联，结果关联最多的是牛奶。当然，如何领悟双关、反讽之类的语言修辞，分析比语言理解本身更复杂的情感问题等，都是智能计算机需克服的困难。

正是这样，目前，人工智能和机器学习的应用实践开始从专家系统、博弈、自然语言理解等向更具现实意义的数据分析领域拓展。机器学习方法，如决策树、神经网络、推理规则等，能够模拟人类的学习方式，向数据案例学习，并通过学习实现对新事物所具模式的识别和判断。这种学习方式恰恰为数据挖掘提供了绝妙的研究思路。

3. 统计学

从 17 世纪中叶起源于英国古典政治经济学开创者威廉·佩蒂（William Petty，

① 爱德华·艾伯特·费根堡姆，计算机人工智能领域的科学家，被誉为“专家系统之父”，1994 年获得计算机科学领域最高声望奖。

1623—1687）的《政治算术》、英国人约翰·格朗特（John Graunt，1620—1674）的《关于死亡表的自然观察与政治观察》和古典概率，到 19 世纪末的古典统计学（Classical Statistics）；从 20 世纪初伴随大工业发展孕育而生的现代统计学雏形，到后续包含极大似然估计、方差分析、置信区间和假设检验等在内的现代统计学基本框架；从参数模型（Parametric Model）假设中的联合分布、因素独立（Factor Independency）、线性叠加（Linear Additivity）、数值连续数据，到非线性模型研究的丰硕成果；从古典统计与贝叶斯统计两大流派的争议，到多元统计分析、现代时间序列分析，乃至机器学习中贝叶斯分类和贝叶斯网络、神经网络和决策树模型的热捧以及统计学习理论的大发展，可以说，统计学为数据收集、整理、展现和分析过程提供了完整的理论框架和实践依据，与其他学科融合发展的轨迹，将现代数据分析的特色和需求展现得淋漓尽致。

在信息技术迅猛发展、数据量高速膨胀、数据类型日益丰富、数据管理和分析需求不断提升的过程中，统计学的理论研究和应用实践一直面临着诸多挑战。

例如，通过样本推断总体特征，经典推断统计具有极高的应用价值。但在数据采集能力极强的今天，有时摆在人们面前的不再是样本，而是海量的高维总体。此时推断不再有意义，原本较小的参数差异在大样本条件下都表现出了“显著”。再例如，经典统计分析方法往往是模型驱动式的演绎推理，是验证驱动（verification-driven）型分析。以统计学中应用极为广泛的线性回归分析方法为例，是首先确定模型，然后利用数据建立模型、验证模型和应用模型。这样的研究模式是建立在对模型的“先知先见”基础上的。但在数量庞大、结构复杂的海量数据面前，这种“先知先见”几乎不再可能。于是，数据驱动式的归纳分析，发现发现驱动（discovery-driven）型思路似乎更为现实。

为克服统计分析方法应用过程中的诸多问题，20 世纪 60 年代，稳健统计开始盛行。其通过敏感性分析、异常值诊断等手段，开创性地解决了数据与理论分布假设有偏差的分析问题。20 世纪 70 年代中期，约翰·怀尔德·图基[①]（John Wilder Tukey，1915—2000）提出的探索性数据分析（Exploratory Data Analysis，EDA）方法，打破了统计方法中分布假设的古典框架，注重从数据的特征出发研究和发现数据中有价值的信息。在此后至今的几十年发展历程中，统计方法在与数据相结合的道路上硕果累累，许多新的统计技术应运而生。在摆脱古典框架约束方面，通过马尔可夫链蒙特卡罗（Markov Chain Monte Carlo，MCMC）模拟以及贝叶斯统计等方法，着力解决复杂模型识别和分析问题。利用 Jack-Knife 刀切法、Cross-Validation 交叉验证、Bootstrap 等方法解决模型评价和选择问题。此外，在分析结果展示方面，除用传统的数学语言表示之外，统计也力图更多地借助现代计算机技术，实现高维数据分布特征以及分析结果的图形化展示，数据的可视化技术已成为统计和计算机界共同的热门话题。

同时，在应用实践中，一方面，数据整理是统计分析必不可少的重要环节。在数据量相对较少的过去，数据整理可以通过手工或借助简单工具实现。但随着数据量的快速膨胀，该问题已从量变转化成质变。从工作量看，数据整理的工作量已经占到整个统计分析工作量的 70%～80%或更高；从工作方式看，手工或借助电子表格软件整理数据的方式已

① 约翰·怀尔德·图基，美国著名的统计学家、信息科学家、数据分析师、化学家、拓扑学家、教育家。

显得无能为力。

表面上，上述问题源于数据整理手段和工具效率不高，但本质上却源于数据的存储组织模式。数据整理的高效率是建立在良好的数据组织模式基础上的，只有好的数据组织模式才可能支撑高效率的数据整理。因此，过去在统计应用视野之外的数据存储和组织问题，今天成为统计应用实践的瓶颈。统计应用与计算机数据库技术相结合已是大势所趋。

另一方面，整体解决方案已成为统计应用实践的大趋势。过去，人们的统计应用实践往往呈现出“片段性”的特点，原本完整的统计应用却呈“割裂”状。以企事业统计为例，统计应用实践应包括建立指标体系、采集数据、存储和管理数据、分析数据和制定决策等多个相互影响和制约的环节，而将其割裂开，必然会出现各自为政、各行其是的局面。统计人员脑子中“我只负责指标框架设计而不考虑具体实施”、“你给我数据，我给你分析”的工作模式不足为奇，数据上报是基层统计人员的额外负担并非罕事。

没有从系统和工程的角度提供统计应用整体解决方案，是导致以上问题的根本原因。事实上，首先，企事业统计更需要的是服务于企事业决策的统计指标体系。其理论框架固然重要，但更应建立在对业务充分理解、广泛调研和可行性深入分析的基础上。指标体系的建立不仅涉及统计制度的建设，还必须考虑其可操作性，并体现在业务处理系统或信息管理系统中。其次，统计数据的采集应纳入企事业的日常管理流程中，应能够通过业务处理系统或信息管理系统自动生成所需的统计数据，并以面向主题的方式存储于统计数据库中。统计数据库支持数据不同表式和格式数据的自由转化，支持通过统计数据库的灵活查询提取分析所需的数据。最后，分析方法可以“无缝”嵌入决策支持系统中，统计建模过程可以不透明，分析结果可以用业务人员熟悉的语言陈述，且可随数据的不断更新而动态调整。

所以，现代统计应用实践需要依托数据库和网络技术，实现从海量数据的收集、存储管理到有效分析的整体解决方案，是统计与计算机相结合的产物。

总之，在海量复杂数据的存储和分析需求，数据库和数据仓库技术、人工智能和机器学习、统计学三者的理论发展和应用实践，以及各学科领域融合发展态势的大背景下，数据挖掘这个新兴的应用研究领域诞生了。

§1.2 什么是数据挖掘

海量数据的分析需求，理论研究的拓展和相互渗透，利用数据库、数据仓库技术存储管理数据，利用机器学习和统计方法分析数据，这种多学科交叉融合发展和实践的思想，催生了备受人们关注的新兴领域——数据挖掘。

> 数据挖掘是一个利用各种方法，从海量的有噪声的凌乱数据中，提取隐含和潜在的对决策有用的信息和模式的过程。

1.2.1　数据挖掘和数据库中的知识发现

1995 年，在加拿大蒙特利尔召开了第一届“知识发现和数据挖掘”国际学术会议。从此，数据挖掘（Data Mining，DM）一词很快流传开来。人们将存储在数据库中的数据比喻为“矿石”，数据挖掘则是一个从数据“矿石”中开采知识“黄金”的过程。

数据库中的知识发现（Knowledge Discovery in Database，KDD）概念是由计算机科学界提出的。顾名思义，KDD 的目的是发现数据库中的知识。完整的 KDD 过程包括数据源的建立和管理、从数据源中提取数据、数据预处理、模型建立、模型评估、模型可视化以及模型应用等一系列步骤。

早期的数据挖掘是作为 KDD 的一个重要环节提出的，特指模型建立。由于数据源通常以数据库和数据仓库的形式存在，业界大多认为数据挖掘离不开数据库和数据仓库的支撑。因此，正像巴瓦尼·图拉辛加姆（Bhavani Thuraisingham）[①] 在她 1998 年的著作 *Data Mining: Technologies, Techniques, Tools and Trends* 中指出的，数据挖掘是对存储于数据库中的海量数据，通过查询和抽取方式获得以前未知的有用信息、模式和规则的过程。

随着对数据挖掘认识和应用实践的不断深入，人们发现，模型建立仅依赖对数据库的简单查询和抽取是不够的，还需要更多的建模理论和量化分析方法。KDD 不仅需要数据库和数据仓库的研究者，也离不开机器学、统计学等其他学科领域学者的参与。同时，人们还认识到，KDD 中的模型建立（数据挖掘）环节不可能脱离数据准备和模型评价等阶段而独立存在，有效的数据准备和合理的模型评价是数据挖掘成功的基础。

为此，数据挖掘的内涵不再局限于数据建模，还囊括了模型建立过程中必不可缺的环节，即数据抽取、数据预处理以及模型评价等，如图 1—2 所示。因此，正如迈克尔·J·A·贝里（Michael J. A. Berry）[②] 和戈登·林诺夫（Gordon Linoff）[③] 在其 1997 年所著的 *Data Mining Techniques for Marketing, Sales and Customer Support* 和 1999 年所著的 *Mastering Data Mining: The Art and Science of Customer Relationship Management* 中指出的，数据挖掘是一种通过自动或半自动方式，探索和分析海量数据，以发现其中有意义的模式和规则的过程。

当数据挖掘内涵得到扩展后，KDD 的提法在一定程度上受到了影响。

---

① 巴瓦尼·图拉辛加姆，得克萨斯大学工程与计算机科学学院教授。

② 迈克尔·J·A·贝里，数据挖掘顾问，数据挖掘创始人之一，波士顿大学 Carroll 学院教授。与同事戈登·林诺夫所著的数据挖掘著作，被翻译成法语、意大利语、日语和汉语等多国语言，成为最受读者欢迎的数据挖掘图书。

③ 戈登·林诺夫，数据挖掘顾问，数据挖掘创始人之一。独立或合作出版了多部数据挖掘著作。如 *Data Analysis Using SQL and Excel*，2008；*Data Mining Techniques for Marketing, Sales, and Customer Relationship Management*，Second Edition，2004；*Mining the Web: Transforming Customer Data into Customer Value*，2001；*Mastering Data Mining: The Art and Science of Customer Relationship Management*，1999。

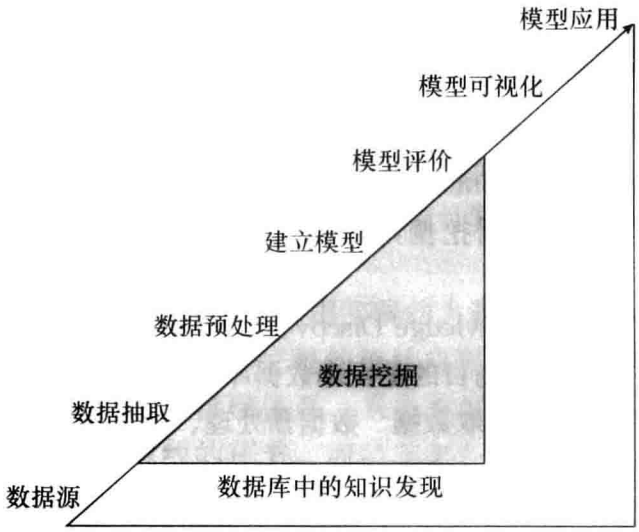


图 1—2　数据挖掘和 KDD 的关系

国内外众多学科领域的学者参与到了数据挖掘的研究中，如图 1—3 所示，涌现出了大批学术论文、著作以及商业应用成功案例。不同学科领域学者对数据挖掘的研究有不同的出发点和侧重。例如，从数据库和数据仓库技术角度，侧重拓展数据挖掘过程中的数据管理理论和技术，以及数据挖掘产品的商业化实现；从机器学习和统计学角度，侧重探讨各种算法的精度和效率改进策略，关注建模过程的模型搜索和参数优化，以及评价函数和模型选择等问题；从可视化角度，侧重研究低维空间中高维数据的展示问题；从计算性能角度，侧重并行算法研究以提高海量数据的计算效率等。

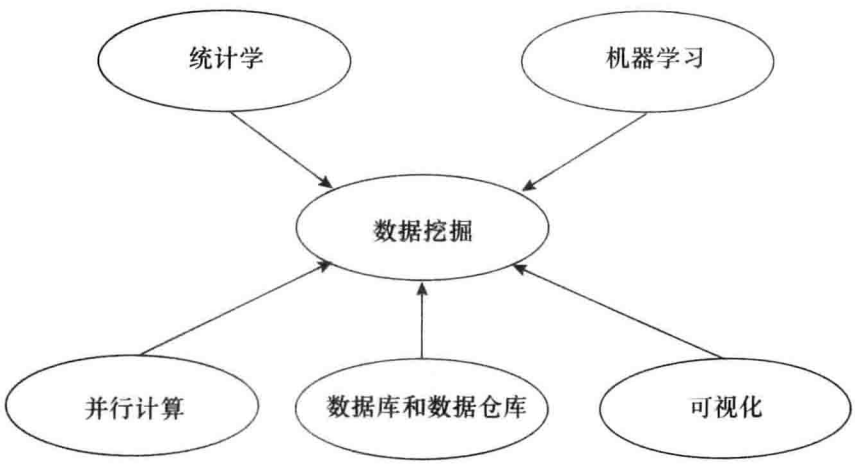


图 1—3　数据挖掘涉及的研究领域

本书并非从理论角度探讨上述方面，而是从数据挖掘的应用角度，重点讲解数据挖掘模型的基本原理和案例的软件操作实现。

基于上述认识，从理论研究角度看，数据库、数据仓库与数据挖掘有这样的内在联系。首先，数据库和数据仓库是数据挖掘诞生发展的重要原因。数据仓库能够有效实现数据的集成、清洗，保证数据的完整性和一致性，为数据挖掘奠定了良好的数据基础。同时，数据仓库并不是数据挖掘的先决条件，但与数据仓库协作会大大提高数据挖掘的效率。