HZ BOOKS WILEY

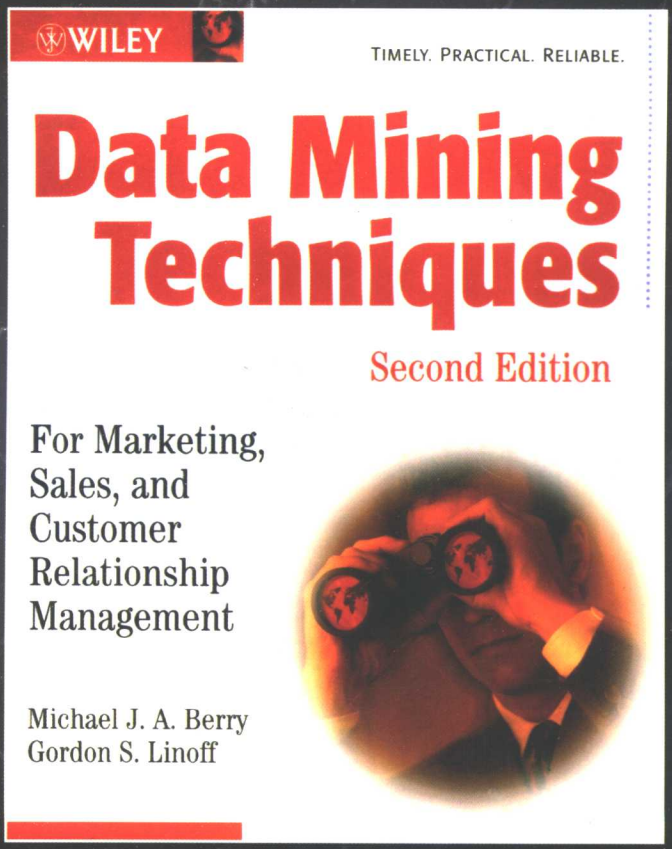
计 算 机 科 学 丛 书

原书第2版

数据挖掘技术

市场营销、销售与客户关系管理领域应用

(美) Michael J. A. Berry Gordon S. Linoff 著 别荣芳 尹静 邓六爱 译



Data Mining Techniques

For Marketing, Sales, and Customer Relationship Management

Second Edition

机械工业出版社
China Machine Press

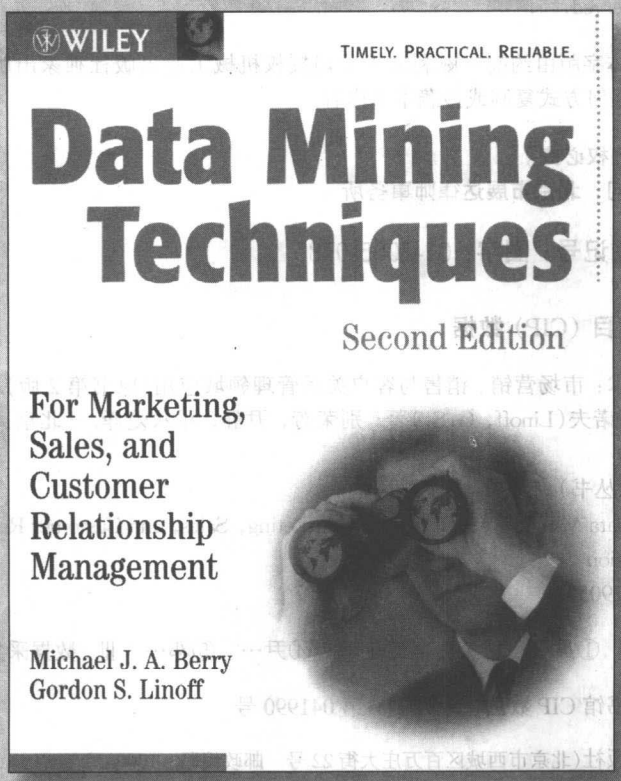
计 算 机 科 书

原书第2版

数据挖掘技术

市场营销、销售与客户关系管理领域应用

(美) Michael J. A. Berry Gordon S. Linoff 著 别荣芳 尹静 邓六爱 译



Data Mining Techniques

For Marketing, Sales, and Customer Relationship Management

Second Edition

机械工业出版社
China Machine Press

本书是一本优秀的数据挖掘教材，全面而系统地介绍了数据挖掘的商业环境、数据挖掘技术及其在商业环境中的应用。

全书共18章，内容涵盖核心的数据挖掘技术，包括：决策树、神经网络、协同过滤、关联规则、链接分析、聚类和生存分析等。此外，还提供了数据挖掘最佳实践的概观、数据挖掘的最新进展和一些极具挑战性的研究课题，极具技术深度与广度。通过学习本书，读者不仅可以精通数据挖掘的整体结构和核心技术，还可以领略数据挖掘在销售和客户关系管理等方面的成功应用，为实践数据挖掘打下坚实的基础。

本书适合作为高等院校相关专业高年级本科生或研究生的教材或参考书，也适合当前和未来的数据挖掘实践者学习和参考。

本书版权登记号：图字：01-2005-0785

图书在版编目（CIP）数据

数据挖掘技术：市场营销、销售与客户关系管理领域应用(原书第2版)/(美)贝瑞(Berry, M.J.A.),(美)莱诺夫(Linoff, G.S.)著；别荣芳，尹静，邓六爱译. －北京：机械工业出版社，2006.7

(计算机科学丛书)

书名原文：Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management, Second Edition

ISBN 7-111-19056-4

Ⅰ.数…　Ⅱ.①贝…　②莱…　③别…　④尹…　⑤邓…　Ⅲ.数据采集　Ⅳ.TP274

中国版本图书馆CIP数据核字(2006)第041990号

机械工业出版社(北京市西城区百万庄大街22号　邮政编码　100037)

责任编辑：朱起飞

北京京北制版印刷厂印刷·新华书店北京发行所发行

2006年7月第1版第1次印刷

184mm×260mm·26.75印张

定价：49.00元

凡购本书，如有倒页、脱页、缺页，由本社发行部调换

本社购书热线：(010)68326294

出版者的话

文艺复兴以降，源远流长的科学精神和逐步形成的学术规范，使西方国家在自然科学的各个领域取得了垄断性的优势；也正是这样的传统，使美国在信息技术发展的六十多年间名家辈出、独领风骚。在商业化的进程中，美国的产业界与教育界越来越紧密地结合，计算机学科中的许多泰山北斗同时身处科研和教学的最前线，由此而产生的经典科学著作，不仅擘划了研究的范畴，还揭橥了学术的源变，既遵循学术规范，又自有学者个性，其价值并不会因年月的流逝而减退。

近年，在全球信息化大潮的推动下，我国的计算机产业发展迅猛，对专业人才的需求日益迫切。这对计算机教育界和出版界都既是机遇，也是挑战；而专业教材的建设在教育战略上显得举足轻重。在我国信息技术发展时间较短、从业人员较少的现状下，美国等发达国家在其计算机科学发展的几十年间积淀的经典教材仍有许多值得借鉴之处。因此，引进一批国外优秀计算机教材将对我国计算机教育事业的发展起积极的推动作用，也是与世界接轨、建设真正的世界一流大学的必由之路。

机械工业出版社华章图文信息有限公司较早意识到“出版要为教育服务”。自1998年开始，华章公司就将工作重点放在了遴选、移译国外优秀教材上。经过几年的不懈努力，我们与Prentice Hall，Addison-Wesley，McGraw-Hill，Morgan Kaufmann等世界著名出版公司建立了良好的合作关系，从它们现有的数百种教材中甄选出Tanenbaum，Stroustrup，Kernighan，Jim Gray等大师名家的一批经典作品，以“计算机科学丛书”为总称出版，供读者学习、研究及庋藏。大理石纹理的封面，也正体现了这套丛书的品位和格调。

“计算机科学丛书”的出版工作得到了国内外学者的鼎力襄助，国内的专家不仅提供了中肯的选题指导，还不辞劳苦地担任了翻译和审校的工作；而原书的作者也相当关注其作品在中国的传播，有的还专程为其书的中译本作序。迄今，“计算机科学丛书”已经出版了近百个品种，这些书籍在读者中树立了良好的口碑，并被许多高校采用为正式教材和参考书籍，为进一步推广与发展打下了坚实的基础。

随着学科建设的初步完善和教材改革的逐渐深化，教育界对国外计算机教材的需求和应用都步入一个新的阶段。为此，华章公司将加大引进教材的力度，在“华章教育”的总规划之下出版三个系列的计算机教材：除“计算机科学丛书”之外，对影印版的教材，则单独开辟出“经典原版书库”；同时，引进全美通行的教学辅导书“Schaum's Outlines”系列组成“全美经典学习指导系列”。为了保证这三套丛书的权威性，同时也为了更好地为学校和老师们服务，华章公司聘请了中国科学院、北京大学、清华大学、国防科技大学、复旦大学、上海交通大学、南京大学、浙江大学、中国科技大学、哈尔滨工业大学、西安交通大学、中国人民大学、北京航空航天大学、北京邮电大学、中山大学、解放军理工大学、郑州大学、湖北工学院、中国国家信息安全测评认证中心等国内重点大学和科研机构在计算机的各个领域的著名学者组成“专家指导委员会”，为我们提供选题意见和出版监督。

这三套丛书是响应教育部提出的使用外版教材的号召，为国内高校的计算机及相关专业

的教学度身订造的。其中许多教材均已为M. I. T.，Stanford，U.C. Berkeley，C. M. U. 等世界名牌大学所采用。不仅涵盖了程序设计、数据结构、操作系统、计算机体系结构、数据库、编译原理、软件工程、图形学、通信与网络、离散数学等国内大学计算机专业普遍开设的核心课程，而且各具特色——有的出自语言设计者之手、有的历经三十年而不衰、有的已被全世界的几百所高校采用。在这些圆熟通博的名师大作的指引之下，读者必将在计算机科学的宫殿中由登堂而入室。

权威的作者、经典的教材、一流的译者、严格的审校、精细的编辑，这些因素使我们的图书有了质量的保证，但我们的目标是尽善尽美，而反馈的意见正是我们达到这一终极目标的重要帮助。教材的出版只是我们的后续服务的起点。华章公司欢迎老师和读者对我们的工作提出建议或给予指正，我们的联系方法如下:

电子邮件: hzjsj@hzbook.com
联系电话: (010) 68995264
联系地址: 北京市西城区百万庄南街1号
邮政编码: 100037

译 者 序

随着数据库技术的应用越来越普及，人们逐渐陷入了“数据丰富，知识贫乏”的尴尬境地，因为大量数据淹没了数据中隐含的模式和有益信息。于是，致力于摆脱这一困境的数据挖掘技术从20世纪90年代起步并得到迅速发展。数据挖掘技术是数据库研究、开发和应用中最活跃的分支之一，是一种基于机器学习、统计分析等多种学科的计算机技术，能够有效地帮助人们将海量数据资源转换为有用的知识和信息，进而帮助人们科学地做出决策。

本书是数据挖掘领域的巨著，多年以来，在数据挖掘领域的地位始终无可替代，其内容也随数据挖掘技术的发展演化而不断更新。本书最早的版本是1997年出版的，补充修订后于2004年出版第2版。新版中减少了与商业相关的素材，增加了更多的技术素材，并加入了作者近年来的最新研究成果和见解，比如：关于数据挖掘在营销和客户关系管理方面的应用、基本统计学技术的使用、生存分析和为挖掘准备数据等内容。基于存储的推理增加了以最近邻技术为基础的协同过滤方法，从而在技术和应用两方面更加全面、系统地介绍了数据挖掘的商业环境、数据挖掘技术及其在商业环境中的应用。

本书共有18章，内容涵盖了核心的数据挖掘技术，包括：决策树、神经网络、协同过滤、关联规则、链接分析、聚类和生存分析等。此外，还提供了数据挖掘最佳实践的概观、数据挖掘的最新进展和一些极具挑战性的研究课题，其技术深度与广度举世公认。作者注重实效，对每类问题均提供代表性算法，以亲身经历的商业案例为实例，给出每一技术具体的应用法则。通过学习本书，读者不仅可以精通数据挖掘的整体结构和核心技术，还可以领略数据挖掘在营销、销售和客户关系管理等方面的成功应用，为实践数据挖掘打下坚实的理论和应用基础。

本书的目标读者是当前和未来的数据挖掘实践者，可以作为相关专业高年级本科生的选修课教材，特别适合作为研究生的专业课教材。本书用生活实例开头，引出基本概念，同时提供大量真正的商业环境实例。因此，对于从事数据挖掘应用的读者来说，是一本必备的参考书。本书的网站还有一些推荐读物和练习，所以对于初学者来说，也是一本可读性极佳、适于循序渐进地学习数据挖掘的首选教科书。

本书主要由别荣芳、尹静和邓六爱三位翻译完成。全书由别荣芳统一审校。孙运传参与了部分审校工作。在翻译过程中，译者发现一些错误和疑似错误之处，在译文中对一般拼写错误和明显笔误均未作说明而直接进行了校正，其他错误则在相应页的脚注中给出了说明。

由于时间仓促，加上本书涉及诸多实际应用领域，原作中方言俚语和非信息技术专业词汇较多，翻译内容难免存在疏漏和不足，敬请读者谅解并批评指正。

译 者

2006.6

致　谢

非常幸运的是，我们周围有很多天才的数据挖掘专家，因此首先要感谢在 Data Miners 公司的同事，从他们那里我们学到了很多。他们是：Will Potts、Dorian Pyle 和 Brij Masand。还有许多曾经与我们密切合作的客户，我们也把他们视为同事：Harrison Sohmer 和 Stuart E. Ward，Ⅲ。编辑 Bob Elliott、编辑助理 Erica Weinstein 和责任编辑 Emilie Herman 帮助我们把握进度，并保持风格一致。毕业于麻省理工学院的 Lauren McCann，在 Data Miners 公司实习期间，准备了在很多例子中使用的人口普查数据，并创建了一些图表。

我们还要感谢过去多年来在数据挖掘方面与我们共事的所有人。我们从每个人那里学到了很多。那些数据挖掘方案对本书第 2 版有影响的人包括：

Al Fan
Alan Parker
Anne Milley
Brian Guscott
Bruce Rylander
Corina Cortes
Daryl Berry
Daryl Pregibon
Doug Newell
Ed Freeman
Erin McCarthy
Herb Edelstein
Jill Holtz
Joan Forrester
John Wallace
Josh Goff
Karen Kennedy
Kurt Thearling
Lynne Brennen
Mark Smith
Mateus Kehder
Michael Patrick
Nick Gagliardo
Nick Radcliffe
Patrick Surry
Ronny Kohavi
Sheridan Young
Susan Hunt Stevens
Ted Browne
Terri Kowalchuk
Victor Lo
Yasmin Namini
Zai Ying Huang

当然，我们仍然要感谢在第 1 版曾经感谢的人们：

Bob Flynn
Bryan McNeely
Claire Budden
David Isaac
David Waltz
Dena d' Ebin
Diana Lin
Don Peppers
Ed Horton
Edward Ewen
Fred Chapman
Gary Drescher
Gregory Lampshire
Janet Smith
Jerry Modes
Jim Flynn
Kamran Parsaye
Karen Stewart
Larry Bookman
Larry Scroggins
Lars Rohrberg
Lounette Dyer
Marc Goodman
Marc Reifeis
Marge Sherold
Mario Bourgoin
Prof. Michael Jordan
Patsy Campbell
Paul Becker
Paul Berry
Rakesh Agrawal
Ric Amari
Rich Cohen
Robert Groth
Robert Utzschnieder
Roland Pesch
Stephen Smith
Sue Osterfelt
Susan Buchanan
Syamala Srinivasan
Wei-Xing Ho
William Petefish
Yvonne McCollin

前　言

本书第1版于1997年面世。该书实际上开始于1996年，当时我和Gordon在为国家银行（NationsBank）（现在是美国银行，Bank of America）设计一天的数据挖掘研讨班。NationsBank的一位副总裁Sue Osterfelt（她还与Bill Inmon合著有一本关于数据库应用的图书）使我们深信，研讨班的材料应该整理成一本书。她把Jon Wiley & Sons公司的编辑Bob Elliott介绍给我们，在我们还没来得及仔细考虑这件事情之前，就签了一份合同。

我们两个人以前从未写过书，前面几章的草稿清楚地说明了这一点。感谢Bob的帮助，我们取得了很大的进步，最终版本仍然是相当令人骄傲的。毫不夸张地说，这一经历改变了我们的生活：第一是占用了应该散步的每一小时，甚至是应该睡觉的时间；其次，更肯定地说，提供了我们创建的Data Miners咨询公司的基础。本书第1版已经成为数据挖掘的一本标准教材，后续著作包括：*Mastering Data Mining* 和 *Mining the Web*。

那么为什么要进行修订呢？自从第1版出版以来，数据挖掘界发生了很大的变化。例如：那时候，Amazon.com才刚刚出现；美国移动电话呼叫费用平均为每分钟50美分，不超过25%的美国人拥有移动电话；KDD数据挖掘会议才举办了第二届。我们的理解也改变了很多。尽管其中的大部分核心算法仍然保持不变，但是算法嵌入的软件、应用算法的数据库以及用于解决的商业问题都有所增长和演化。

即使技术界和商业界保持不变，我们也希望更新本书第1版，因为在其间的几年，我们又学到了很多。做咨询的一大乐趣就是时刻面对新思想、新问题和新的解决方案。我们并不比当年写第1版的时候更聪明，但确实经验更丰富，而且我们的写作经验也更丰富了。稍微浏览一下本书内容目录就可以发现，我们减少了很多与商业相关的材料，而增加了更多的技术材料。另外，把一些商业材料融汇到技术章节中，因此使数据挖掘技术得以在商业环境中来讨论，希望这样可以使读者更容易领会到如何把技术应用到自己的商业问题。

我们还注意到，许多商业学校的课程使用本书作为教材。尽管我们并没有把本书写成一本教科书，在第2版中，我们努力使它可以用作教材，书中提供了大量基于公开可用的数据的实例，诸如美国的人口普查数据，在配套网站www.data-miners.com/companion中有推荐阅读材料和建议的练习。

全书仍然分为三个部分，第一部分讲述数据挖掘的商业环境。开篇章节给出了数据挖掘的简介，解释数据挖掘可以用来干什么，并且为什么需要数据挖掘。第2章介绍数据挖掘的良性循环，这是一个持续不断的过程，通过这个过程，数据挖掘将数据转变为指导行动的信息，反过来创造了更多的信息和更多的学习机会。第3章是数据挖掘的方法论和最佳实践的拓展讨论，该章比书中任何其他一章更得益于我们写第一本书以来的经历，这里介绍的方法论基于我们曾经参与的成功案例而设计。第4章在第1版中根本没有相应的部分，是关于数据挖掘在营销和客户关系管理中的应用，也正是我们现在所从事的领域。

第二部分讲解数据挖掘本身的技术内容，包含第1版描述的所有技术，但是重新进行了调整，对各种描述进行了重写，比第1版更清晰、更准确。但仍然保留了第1版的风格，即可能的地方都使用非技术语言。

除了包含第1版涵盖的7种技术：决策树、神经网络、协同过滤、关联规则、链接分析、聚类和生存分析之外，还增加了使用基本的统计学技术以及生存分析的新章节。生存分析是一项广泛应用的技术，从医学界的少量样本和连续的时间测量，到营销数据中发现的大量样本和离散时间度量，都可应用。基于存储的推理一章还包括以最近邻技术为基础的协同过滤方法，作为产生推荐的方式，已经为广大Web零售商所熟知。

第三部分讲述在商业环境中使用技术的方法，其中有一章关于在数据中发现客户，另一章关于数据挖掘和数据仓库的关系，还有一章关于数据挖掘环境（公司环境和技术环境两个方面），最后一章关于在公司中应用数据挖掘。该部分新增加了一章，介绍为数据挖掘准备数据。这是一个极其重要的话题，因为很多数据挖掘者反映，在典型的数据挖掘工程中，转换数据通常需要花费大多数的时间。

和第1版一样，本书仍然针对当前和未来的数据挖掘实践者。既不是为软件开发者提供如何实现各种数据挖掘算法的细节指导，也不是为了使研究人员改进那些算法。有关思想以非技术的语言给出，尽可能少地使用数学公式和艰涩的术语。每一种数据挖掘技术都在真实的商业环境中展示，给出大量来自商业环境的实例。简而言之，我们努力把本书写成打算开始数据挖掘生涯的技术人员喜欢读的一本书。

Michael J. A. Berry

2003年10月

目　录

译者序
致谢
前言

第 1 章　数据挖掘的缘起和内容 …… 1
1.1　分析客户关系管理系统 …… 1
1.1.1　交易处理系统的作用 …… 2
1.1.2　数据仓库的作用 …… 3
1.1.3　数据挖掘的作用 …… 3
1.1.4　客户关系管理策略的作用 …… 4
1.2　什么是数据挖掘 …… 4
1.3　数据挖掘可以完成哪些工作 …… 5
1.3.1　分类 …… 5
1.3.2　估计 …… 6
1.3.3　预测 …… 6
1.3.4　关联分组或关联规则 …… 7
1.3.5　聚类 …… 7
1.3.6　建立简档 …… 7
1.4　为什么现在研究 …… 8
1.4.1　数据正在生成 …… 8
1.4.2　数据正在形成数据仓库 …… 8
1.4.3　计算能力足以承受 …… 8
1.4.4　客户关系管理的兴趣增强 …… 9
1.4.5　商业数据挖掘软件产品已经易于使用 …… 9
1.5　目前如何使用数据挖掘 …… 10
1.5.1　超级市场成为信息经纪人 …… 10
1.5.2　基于推荐的商业 …… 10
1.5.3　交叉销售 …… 11
1.5.4　抓住好的客户 …… 11
1.5.5　淘汰差的客户 …… 11
1.5.6　变革一个行业 …… 11
1.5.7　其他 …… 12
1.6　小结 …… 12
第 2 章　数据挖掘的良性循环 …… 13
2.1　商业数据挖掘案例研究 …… 14
2.1.1　识别商务挑战 …… 14
2.1.2　应用数据挖掘 …… 14
2.1.3　按照结果采取行动 …… 15
2.1.4　测试效果 …… 16
2.2　何谓良性循环 …… 16
2.2.1　识别商业机会 …… 17
2.2.2　挖掘数据 …… 17
2.2.3　采取行动 …… 19
2.2.4　测试结果 …… 19
2.3　良性循环环境下的数据挖掘 …… 20
2.4　移动通信公司建立恰当的联系 …… 21
2.4.1　机会 …… 22
2.4.2　如何应用数据挖掘 …… 23
2.4.3　处理行动 …… 24
2.4.4　完成循环 …… 24
2.5　神经网络和决策树驱动 SUV 的销售 … 25
2.5.1　最初的挑战 …… 25
2.5.2　如何应用数据挖掘 …… 25
2.5.3　最终措施 …… 26
2.5.4　完成循环 …… 27
2.6　小结 …… 27
第 3 章　数据挖掘方法论和最佳实践 … 29
3.1　为什么需要方法论 …… 29
3.1.1　获取不真实的知识 …… 29
3.1.2　获取真实但无用的知识 …… 32
3.2　假设测试 …… 33
3.3　模型、建立简档和预测 …… 34
3.3.1　建立简档 …… 36
3.3.2　预测 …… 36
3.4　方法论 …… 36
3.4.1　第一步：将商业问题转换为数据挖掘问题 …… 37
3.4.2　第二步：选取合适数据 …… 40
3.4.3　第三步：设法理解数据 …… 43
3.4.4　第四步：创建模型集 …… 45
3.4.5　第五步：修复数据问题 …… 48
3.4.6　第六步：变换数据，获取信息 … 50

3.4.7 第七步：建立模型 …… 52
3.4.8 第八步：评估模型 …… 52
3.4.9 第九步：部署模型 …… 57
3.4.10 第十步：评估结果 …… 57
3.9.11 第十一步：重新开始 …… 57
3.5 小结 …… 58
第 4 章 数据挖掘在市场营销和客户关系管理中的应用 …… 59
4.1 寻找潜在客户 …… 59
4.1.1 识别好的潜在客户 …… 59
4.1.2 选择沟通渠道 …… 60
4.1.3 遴选适当的信息 …… 60
4.2 为选择正确的广告场所进行数据挖掘 …… 61
4.2.1 谁匹配简档 …… 61
4.2.2 测量读者群组的匹配度 …… 62
4.3 通过数据挖掘改进定向市场营销活动 …… 64
4.3.1 响应建模 …… 65
4.3.2 优化固定预算的响应率 …… 65
4.3.3 优化营销活动收益 …… 67
4.3.4 接触那些受相关信息影响最大的人们 …… 71
4.3.5 差别响应分析 …… 72
4.4 使用当前客户来了解潜在客户 …… 73
4.4.1 在他们成为客户前就开始跟踪客户 …… 73
4.4.2 从新客户那里收集信息 …… 74
4.4.3 获取时间变量可预测未来结果 … 74
4.5 客户关系管理数据挖掘 …… 74
4.5.1 按客户需求策划营销活动 …… 75
4.5.2 划分客户群体 …… 75
4.5.3 减少信用风险 …… 77
4.5.4 决定客户价值 …… 77
4.5.5 交叉销售、提升销售和销售推荐 …… 78
4.6 保持和流失 …… 78
4.6.1 识别流失 …… 78
4.6.2 流失为什么重要 …… 79
4.6.3 不同类型的流失 …… 80
4.6.4 不同类型的流失模型 …… 80
4.7 小结 …… 81
第 5 章 统计学的魅力：数据挖掘常用的工具 …… 83
5.1 Occam 的剃刀 …… 84
5.1.1 原假设 …… 84
5.1.2 p 值 …… 85
5.2 观察数据 …… 85
5.2.1 观察离散数值 …… 85
5.2.2 观察连续变量 …… 92
5.2.3 另一对统计概念 …… 93
5.3 测定响应 …… 94
5.3.1 比例标准误差 …… 94
5.3.2 使用置信界限比较结果 …… 95
5.3.3 使用比例差值比较结果 …… 96
5.3.4 样本大小 …… 97
5.3.5 置信区间的真正含义 …… 97
5.3.6 实验的测试群组和对照群组大小 …… 98
5.4 多重比较 …… 99
5.4.1 多重比较下的置信层次 …… 99
5.4.2 Bonferroni 修正 …… 100
5.5 卡方检验 …… 100
5.5.1 期望值 …… 100
5.5.2 卡方值 …… 101
5.5.3 卡方与比例差值的比较 …… 103
5.6 示例：区域和起点的卡方 …… 103
5.7 数据挖掘和统计学异同 …… 106
5.7.1 原始数据中没有测量误差 …… 106
5.7.2 有大量的数据 …… 106
5.7.3 时间从属性随处出现 …… 107
5.7.4 试验是艰难的 …… 107
5.7.5 数据审查和截取 …… 107
5.8 小结 …… 108
第 6 章 决策树 …… 111
6.1 什么是决策树 …… 111
6.1.1 分类 …… 112
6.1.2 评分 …… 112
6.1.3 估计 …… 114
6.1.4 树以多种形态生长 …… 114
6.2 决策树是如何长成的 …… 115
6.2.1 发现拆分 …… 115
6.2.2 生成完全树 …… 118
6.2.3 度量决策树的有效性 …… 118

6.3 选择最佳拆分的测试 …… 119
6.3.1 纯度和发散性 …… 119
6.3.2 基尼或总体发散性 …… 120
6.3.3 熵归约或信息增益 …… 121
6.3.4 信息增益比率 …… 121
6.3.5 卡方检验 …… 122
6.3.6 方差归约 …… 124
6.3.7 F 测试 …… 124
6.4 修剪 …… 124
6.4.1 CART 修剪算法 …… 125
6.4.2 C5 修剪算法 …… 128
6.4.3 基于稳定性的修剪 …… 129
6.5 从树中提炼规则 …… 130
6.6 考虑成本 …… 131
6.7 决策树方法的进一步修正 …… 132
6.7.1 每次使用多于一个字段 …… 132
6.7.2 倾斜超平面 …… 133
6.7.3 神经树 …… 134
6.7.4 使用树分段回归 …… 135
6.8 决策树的替代表示法 …… 135
6.8.1 方格图 …… 135
6.8.2 树年轮图 …… 137
6.9 实际应用中的决策树 …… 138
6.9.1 决策树作为数据探查工具 …… 138
6.9.2 把决策树方法应用于顺序事件 …… 139
6.9.3 模拟未来 …… 140
6.10 小结 …… 142
第 7 章 人工神经网络 …… 143
7.1 历史回眸 …… 143
7.2 房地产评估 …… 144
7.3 用于定向数据挖掘的神经网络 …… 148
7.4 神经网络是什么 …… 149
7.4.1 神经网络的单元是什么 …… 150
7.4.2 前馈神经网络 …… 153
7.4.3 神经网络如何使用反向传播学习 …… 154
7.4.4 前馈网络和反向传播网络的启发 …… 156
7.5 选择训练集 …… 157
7.5.1 覆盖所有特征值 …… 157
7.5.2 特征数目 …… 157
7.5.3 训练集的大小 …… 158
7.5.4 输出数目 …… 158
7.6 准备数据 …… 159
7.6.1 具有连续数值的特征 …… 159
7.6.2 具有有序、离散(整数)数值的特征 …… 161
7.6.3 具有分类数值的特征 …… 162
7.6.4 其他类型的特征 …… 163
7.7 解释结果 …… 163
7.8 时间序列神经网络 …… 165
7.9 如何了解在神经网络内部正在运行的事情 …… 167
7.10 自组织映像 …… 168
7.10.1 什么是自组织映像 …… 168
7.10.2 实例：发现簇 …… 171
7.11 小结 …… 172
第 8 章 最近邻方法：基于存储的推理和协同过滤 …… 175
8.1 基于存储的推理 …… 175
8.2 MBR 面临的挑战 …… 178
8.2.1 选择一组平衡的历史记录 …… 179
8.2.2 表示训练数据 …… 179
8.2.3 确定距离函数、组合函数和邻居的数目 …… 180
8.3 案例研究：分类新闻报导 …… 181
8.3.1 什么是代码 …… 181
8.3.2 应用 MBR …… 181
8.3.3 结果 …… 183
8.4 测量距离 …… 184
8.4.1 什么是距离函数 …… 184
8.4.2 每次每个字段只建立一个距离函数 …… 186
8.4.3 其他数据类型的距离函数 …… 189
8.4.4 当距离度量已经存在时 …… 189
8.5 组合函数：向邻居求答案 …… 190
8.5.1 基本的方法：民主 …… 190
8.5.2 加权投票 …… 191
8.6 协同过滤：可以做出推荐的最近邻方法 …… 192
8.6.1 建立简档 …… 192
8.6.2 比较简档 …… 193
8.6.3 做出预测 …… 193
8.7 小结 …… 194

第 9 章　购物篮分析和关联规则 …… 195
9.1　定义购物篮分析 …… 196
9.1.1　购物篮数据的三个层次 …… 196
9.1.2　订单特征 …… 197
9.1.3　项流行性 …… 199
9.1.4　跟踪市场干预 …… 199
9.1.5　按用途聚类产品 …… 200
9.2　关联规则 …… 201
9.2.1　可操作的规则 …… 201
9.2.2　平凡的规则 …… 201
9.2.3　费解的规则 …… 202
9.3　一个关联规则有多好 …… 203
9.4　建立关联规则 …… 205
9.4.1　选择恰当的项集 …… 206
9.4.2　从所有这些数据中生成规则 …… 209
9.4.3　克服实际局限 …… 211
9.4.4　大数据的问题 …… 213
9.5　扩展思想 …… 213
9.5.1　使用关联规则比较店铺 …… 213
9.5.2　无关规则 …… 214
9.6　使用关联规则的顺序分析 …… 215
9.7　小结 …… 215
第 10 章　链接分析 …… 217
10.1　图论基础 …… 217
10.1.1　哥尼斯堡七桥问题 …… 219
10.1.2　旅行推销员问题 …… 221
10.1.3　有向图 …… 222
10.1.4　检测图中的环 …… 223
10.2　链接分析的一个熟悉的应用 …… 223
10.2.1　Kleinberg 算法 …… 224
10.2.2　细节：查找网络中心和权威 … 225
10.2.3　实践中的网络中心和权威 …… 226
10.3　案例研究：谁在家中使用传真机 … 227
10.3.1　为什么发现传真机是有用的 … 227
10.3.2　用数据画图 …… 227
10.3.3　方法 …… 228
10.3.4　一些结果 …… 229
10.4　案例研究：分段移动电话客户 …… 232
10.4.1　数据 …… 232
10.4.2　不使用图论的分析 …… 232
10.4.3　两位客户的对比 …… 232
10.4.4　链接分析的力量 …… 234
10.5　小结 …… 234
第 11 章　自动聚类探测 …… 235
11.1　搜索单纯岛状片段 …… 235
11.1.1　星光与星的亮度 …… 236
11.1.2　适应多维情况 …… 237
11.2　K 平均聚类 …… 238
11.2.1　K 平均算法的三个步骤 …… 238
11.2.2　K 的意义 …… 240
11.3　相似性和距离 …… 241
11.3.1　相似性度量与变量类型 …… 242
11.3.2　相似性的常规度量 …… 242
11.4　聚类过程的数据准备 …… 244
11.4.1　利用比例缩放使变量相对一致 …… 245
11.4.2　使用权重编码外部信息 …… 245
11.5　聚类探测的其他途径 …… 246
11.5.1　高斯混合模型 …… 246
11.5.2　凝聚聚类 …… 247
11.5.3　分裂聚类 …… 249
11.5.4　自组织映像 …… 250
11.6　评价簇 …… 250
11.6.1　在簇内部 …… 251
11.6.2　在簇之外 …… 251
11.7　案例研究：聚类城镇 …… 251
11.7.1　创造城镇特征 …… 252
11.7.2　创建簇 …… 253
11.7.3　利用主题簇调整区域边界 …… 256
11.8　小结 …… 256
第 12 章　市场营销中的风险函数和生存分析 …… 259
12.1　客户保持 …… 260
12.1.1　计算保持 …… 260
12.1.2　保持曲线揭示的内容 …… 261
12.1.3　从保持曲线找出平均保有期 … 262
12.1.4　把客户保持看做衰变 …… 263
12.2　风险 …… 266
12.2.1　基本思想 …… 266
12.2.2　风险函数示例 …… 268
12.2.3　审查 …… 270
12.2.4　其他类型的审查 …… 271
12.3　从风险到生存 …… 273
12.3.1　保持 …… 273
12.3.2　生存 …… 274

12.4 比例风险 …… 275
12.4.1 比例风险实例 …… 276
12.4.2 分层：测量生存的初始结果 … 276
12.4.3 Cox 比例风险 …… 277
12.4.4 比例风险的局限性 …… 277
12.5 生存分析实践 …… 278
12.5.1 处理不同的流失类型 …… 278
12.5.2 客户何时会回来 …… 279
12.5.3 预测 …… 280
12.5.4 风险随时间变化 …… 281
12.6 小结 …… 282
第 13 章 遗传算法 …… 283
13.1 遗传算法如何工作 …… 284
13.1.1 计算机上的遗传学 …… 284
13.1.2 表示数据 …… 290
13.2 案例研究：使用遗传算法进行资源优化 …… 290
13.3 模式：遗传算法为什么起作用 …… 291
13.4 遗传算法的更多应用 …… 294
13.4.1 在神经网络方面的应用 …… 294
13.4.2 案例研究：为响应建模完善一个解决方案 …… 295
13.5 超越简单算法 …… 298
13.6 小结 …… 299
第 14 章 数据挖掘贯穿客户生存周期 …… 301
14.1 客户关系层次 …… 301
14.1.1 深度亲密 …… 302
14.1.2 大众亲密 …… 303
14.1.3 中间关系 …… 304
14.1.4 间接关系 …… 304
14.2 客户生存周期 …… 305
14.2.1 客户生存周期：生存阶段 …… 306
14.2.2 客户生存周期 …… 306
14.2.3 基于订阅关系和基于事件关系的比较 …… 307
14.3 围绕客户生存周期组织商业过程 … 309
14.3.1 客户获取 …… 310
14.3.2 客户激活 …… 312
14.3.3 关系管理 …… 313
14.3.4 保持 …… 314
14.3.5 赢回 …… 315
14.4 小结 …… 315
第 15 章 数据仓库、OLAP 和数据挖掘 …… 317
15.1 数据结构 …… 318
15.1.1 交易数据——基础层 …… 318
15.1.2 操作汇总数据 …… 319
15.1.3 决策支持汇总数据 …… 319
15.1.4 数据库模式 …… 320
15.1.5 元数据 …… 323
15.1.6 商业规则 …… 323
15.2 数据仓库的大致结构 …… 324
15.2.1 源系统 …… 325
15.2.2 提取、转化和加载 …… 325
15.2.3 中央储存库 …… 326
15.2.4 元数据储存库 …… 328
15.2.5 数据集市 …… 329
15.2.6 操作反馈 …… 329
15.2.7 最终用户和桌面工具 …… 329
15.3 OLAP 适用于何处 …… 331
15.3.1 立方体中的内容 …… 332
15.3.2 星形模式 …… 337
15.3.3 OLAP 和数据挖掘 …… 339
15.4 数据挖掘在哪里切入数据仓库 …… 340
15.4.1 大量数据 …… 340
15.4.2 一致的、清洁的数据 …… 340
15.4.3 假设测试和测量 …… 341
15.4.4 可升级硬件及 RDBMS 支持 … 341
15.5 小结 …… 342
第 16 章 构造数据挖掘环境 …… 343
16.1 以客户为中心的组织 …… 343
16.2 理想的数据挖掘环境 …… 344
16.2.1 确定什么数据可用的能力 …… 344
16.2.2 将数据转化为可操作信息的技巧 …… 345
16.2.3 所有必需的工具 …… 345
16.3 返回现实世界 …… 345
16.3.1 建立以客户为中心的组织 …… 345
16.3.2 创建单个客户视图 …… 346
16.3.3 定义以客户为中心的度量标准 …… 346
16.3.4 收集正确的数据 …… 347
16.3.5 从客户交互到学习机会 …… 348

16.3.6 挖掘客户数据 …… 348
16.4 数据挖掘组 …… 348
16.4.1 外包数据挖掘 …… 349
16.4.2 内部数据挖掘 …… 350
16.4.3 数据挖掘组成员需要具备的条件 …… 351
16.5 数据挖掘基础设施 …… 351
16.5.1 挖掘平台 …… 352
16.5.2 评分平台 …… 352
16.5.3 一个产品数据挖掘结构实例 … 352
16.6 数据挖掘软件 …… 355
16.6.1 所应用的技术范围 …… 355
16.6.2 可扩展性 …… 356
16.6.3 评分支持 …… 357
16.6.4 用户界面的多种层次 …… 357
16.6.5 可理解的输出 …… 358
16.6.6 处理各种数据类型的能力 …… 358
16.6.7 文档及简单使用 …… 358
16.6.8 对新手和高级用户的培训、咨询和支持 …… 358
16.6.9 卖方可信度 …… 359
16.7 小结 …… 359
第 17 章 为挖掘准备数据 …… 361
17.1 数据应该像什么 …… 361
17.1.1 客户特征标识 …… 362
17.1.2 列 …… 363
17.1.3 模型在建模中的角色 …… 366
17.1.4 变量度量 …… 368
17.1.5 用于数据挖掘的数据 …… 373
17.2 构建客户特征标识 …… 373
17.2.1 编写数据目录 …… 374
17.2.2 识别客户 …… 374
17.2.3 第一次尝试 …… 376
17.2.4 取得进展 …… 377
17.2.5 实际的问题 …… 378
17.3 探查变量 …… 378
17.3.1 直方图分布 …… 378
17.3.2 随时间变化 …… 378
17.3.3 交叉表 …… 380
17.4 衍生变量 …… 380
17.4.1 提取来自单个数值的特征 …… 380
17.4.2 在记录内合并数值 …… 381
17.4.3 查找辅助信息 …… 381
17.4.4 转轴正则时间序列 …… 383
17.4.5 汇总交易记录 …… 384
17.4.6 汇总跨越模型集的字段 …… 385
17.5 基于行为变量的例子 …… 385
17.5.1 购买频率 …… 386
17.5.2 衰减使用 …… 387
17.5.3 旋转者、交易商和便利用户：定义客户行为 …… 388
17.6 数据的黑暗面 …… 393
17.6.1 缺失值 …… 394
17.6.2 脏数据 …… 395
17.6.3 不一致数值 …… 396
17.7 计算问题 …… 396
17.7.1 源系统 …… 397
17.7.2 提取工具 …… 397
17.7.3 专用代码 …… 397
17.7.4 数据挖掘工具 …… 397
17.8 小结 …… 398
第 18 章 应用数据挖掘 …… 399
18.1 开始 …… 399
18.1.1 从概念验证方案中能期待什么 …… 400
18.1.2 识别概念验证方案 …… 400
18.1.3 实现概念验证方案 …… 401
18.2 选择数据挖掘技术 …… 404
18.2.1 将商务目标转换为数据挖掘任务 …… 404
18.2.2 决定数据的相关特性 …… 404
18.2.3 考虑混合方法 …… 405
18.3 公司如何开展数据挖掘 …… 406
18.3.1 保持的对照实验 …… 406
18.3.2 数据 …… 408
18.3.3 一些发现 …… 409
18.3.4 实践出真知 …… 409
18.4 小结 …… 410

第 1 章　数据挖掘的缘起和内容

在本书第 1 版中，第 1 章的第一句就写到："马萨诸塞州萨默维尔市，本书作者之一的故乡……"，接着讲述了那个镇上的两个小店和他们如何与客户形成学习关系（learning relationship）的故事。该章描述了梳小辫的小女孩和给她梳辫子的人的关系，在其间的几年中，这个小女孩已经长大成人，离开小镇，也不再梳着小辫，她的父亲也搬到附近的剑桥居住。但是有一件事情没变，作者仍然是 Wine Cask 商店的忠实客户。正是在这个小店，同样忠诚的一些客户在 1978 年将便宜的阿尔及利亚红酒介绍给他，后来介绍给他法国的葡萄酒产区，现在正帮他开发意大利和德国的酒源。

25 年后，他们仍然有一位忠实的客户，这并非偶然。在 Wine Cask 商店的 Dan 和 Steve 了解他们的客户的口味和可承受价位，当有客户询问时，他们的回答除了基于本店库存外，还有因日积月累而得到的有关该顾客口味和经济能力方面的信息。

Wine Cask 商店的人掌握很多有关葡萄酒的知识，尽管这种知识是很多人来这里买酒而不是去大的折扣酒店的原因之一，但是他们对每个客户的详细了解才是客户持续购买的主要原因。也许可以在大街对面开另一个酒店，同样雇用一批品酒专家，但是要达到对客户了解程度具有同样水平至少需要几个月甚至数年时间。

经营好的小商店自然与他们的客户形成学习关系。久而久之，他们对客户的了解越来越多，然后用这种了解更好地为客户服务，结果不仅获得忠实的客户，还盈利颇丰。拥有数十万乃至上百万客户的大公司，难以形成与每个客户的密切关系，这些公司必须依赖其他方法形成与客户的学习关系。特别是，他们必须充分利用自己拥有的大量东西，那就是几乎每笔客户交易所产生的数据。本书将要讲述的就是如何把客户数据转换为客户知识的分析技术。

1.1　分析客户关系管理系统

人们普遍认为，任何规模的公司都需要学会效仿那些以服务为本的小企业的成功之处——与客户建立一对一的关系。客户关系管理（customer relationship management，CRM）系统是很多书和会议中广泛讨论的主题，从引导追踪软件到调用中心软件的外围管理软件都被称为客户关系管理工具。本书主要关注的是数据挖掘（data mining）在提高公司与客户形成学习关系的能力，进而改善客户关系管理中所起的作用。

在任何行业，有远见的公司正在向着下面的目标努力：努力了解每个客户个体，并且利用这种了解使客户选择与他们进行商业活动，而不是选择他们的竞争对手。这些公司也正在学习认识每个客户的价值，进而知道哪些人值得投入资金和精力来保持联系，哪些人可以放弃。从重视广泛的市场到重视客户个体的这种转变，需要整个企业在市场、销售和客户支持等方面适应这种转变。

对大多数公司来说，围绕客户关系建立商业活动是一种全新的变革。银行一贯关注如何保持存款应付利息和贷款应收利息的差额，电信公司关注网络内通话连接，保险公司关注处理理赔和投资管理。仅使用数据挖掘并不足以把一个注重产品的组织转变为以客户为中心的组织。如果管理者的奖金基于新物品的季度销售数量而不是小部件的销售数量，一个建议给

某个客户提供一个小部件而不是一件新物品的数据挖掘结果极容易被忽略，尽管也许后者盈利更多。

狭义地讲，数据挖掘是一系列工具和技术的集合，是支持以客户为中心的组织需要的多项技术之一；广义地讲，数据挖掘是一种态度，它表明商业活动应该基于认知，分析获得的决策比没有任何分析所得的决策好得多，经过测算的结果更利于商业盈利。数据挖掘还是应用这些工具和技术的过程和方法论。为进行有效的挖掘，分析客户关系管理系统的其他要求也必须到位。为了与其客户形成学习关系，公司必须做到：

- 注意客户正在做什么
- 记住公司及其客户曾经做过什么
- 从记住的信息学习
- 按照获得的知识进行商业活动使顾客更加受益

本书的目标是上述第三个方面，也即从过去发生的事情中学习，这种学习不可能凭空进行。必须依靠交易处理（transaction processing）系统收集客户数据，用数据仓库存储客户历史行为信息，使用数据挖掘把历史数据转变成未来行动计划，然后通过某种客户关系策略将这一计划付诸实施。

1.1.1 交易处理系统的作用

小企业通过注意客户的需求，记住客户的喜好，从过去的交流中学习如何在未来更好地服务他们，由此建立与客户的关系。但是对于大多数雇员从来不与客户交流的大公司来说，如何完成类似的事情呢？在这种公司中即使有一些客户交流，也可能仅仅是与销售职员或不知名字的客服中心的员工进行交流，那么，公司怎么可能注意到或记住这些信息，并且从这种交流中获取信息呢？什么东西能够替代可以识别客户姓名、面孔、声音，记住客户的习惯和喜好的独特的创造性直觉呢？

一句话，没有东西可以代替它！但这不代表我们不可以做尝试。通过灵活运用信息技术，即使是最大的公司也可获得惊人的相近结果！在大商业公司，注意客户的行为这一步已经高度自动化，交易处理系统无处不在，收集几乎所有的数据：自动售货机、电话交换机、网络服务器和售点扫描仪等生成的数据，都是数据挖掘的主要素材。

目前，每天的生活都可产生一系列的交易记录。当拿起电话从 L.L.Bean 预订一只皮划艇桨或者从 Victoria's Secret 定制一个缎纹文胸，市话公司就生成详细电话记录，显示呼叫时间、呼叫电话号码以及被叫长途电话公司等。在长途电话公司，也会生成类似的记录，包括持续通话时间和使用的交换机中的具体路由线路。这些数据连同个人账号信息、姓名和地址等其他记录产生一个账单。订购公司也会记录你的呼叫，连同预订项信息以及对一些推销商品的反应。当接听电话的销售服务代表询问你的信用卡号码以及交付期限时，信息很快转入转账的信用卡验证系统，这样又生成了一条记录。然后转账业务抵达发行信用卡的银行，出现在下个月的银行账单中。当订单连同商品号码、型号和颜色进入订单系统，在付账系统和库存控制系统中将产生另外的记录。几小时后，你的订单又会在 UPS 或者 FedEx 的计算机系统中产生交易记录，它们在你家和公司仓库之间进行多次扫描，可以使你通过检查邮递公司的网页来方便地追踪所订购的物品。

这些交易记录不是专门为数据挖掘生成的，而是公司的运作需要。然而所有记录均包含