

Data Mining Techniques: For Marketing, Sales, and Customer
Relationship Management, Third Edition

数据挖掘技术(第3版)

——应用于市场营销、销售与客户关系管理

[美] Gordon S. Linoff 著
Michael J. A. Berry 译
巢文涵 张小明 王芳



清华大学出版社

数据挖掘技术(第3版)

——应用于市场营销、销售与客户关系管理

[美] Gordon S. Linoff 著
Michael J. A. Berry 著
巢文涵 张小明 王芳 译

清华大学出版社

北 京

Gordon S. Linoff, Michael J. A. Berry

Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management, Third Edition

EISBN: 978-0-470-65093-6

Copyright © 2011 by Wiley Publishing, Inc., Indianapolis, Indiana

All Rights Reserved. This translation published under license.

本书中文简体字版由 Wiley Publishing, Inc. 授权清华大学出版社出版。未经出版者书面许可, 不得以任何方式复制或抄袭本书内容。

北京市版权局著作权合同登记号 图字: 01-2011-3521

Copies of this book sold without a Wiley sticker on the cover are unauthorized and illegal

本书封面贴有 Wiley 公司防伪标签, 无标签者不得销售。

版权所有, 侵权必究。侵权举报电话: 010-62782989 13701121933

图书在版编目(CIP)数据

数据挖掘技术(第3版)——应用于市场营销、销售与客户关系管理/ (美)林那夫(Linoff, G. S.), (美)贝里(Berry, M.J.A.) 著; 巢文涵, 张小明, 王芳 译. —北京: 清华大学出版社, 2013.3

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

ISBN 978-7-302-31014-3

I. ①数… II. ①林… ②贝… ③巢… ④张… ⑤王… III. ①数据采集 IV. ①TP274

中国版本图书馆 CIP 数据核字(2012)第 304161 号

责任编辑: 王 军 刘伟琴

装帧设计: 康 博

责任校对: 蔡 娟

责任印制: 杨 艳

出版发行: 清华大学出版社

网 址: <http://www.tup.com.cn>, <http://www.wqbook.com>

地 址: 北京清华大学学研大厦 A 座 邮 编: 100084

社 总 机: 010-62770175 邮 购: 010-62786544

投稿与读者服务: 010-62776969, c-service@tup.tsinghua.edu.cn

质 量 反 馈: 010-62772015, zhiliang@tup.tsinghua.edu.cn

印 装 者: 清华大学印刷厂

经 销: 全国新华书店

开 本: 185mm×260mm 印 张: 40.25 字 数: 980 千字

版 次: 2013 年 3 月第 1 版 印 次: 2013 年 3 月第 1 次印刷

印 数: 1~4000

定 价: 79.80 元

产品编号: 041245-01

作者简介

Gordon S. Linoff 和 Michael J. A. Berry 在数据挖掘领域的知名度众所周知。他们是 Data Miners 公司——一家从事数据挖掘的咨询公司——的创始人，而且他们已经共同撰写了一些在该领域有影响力和得到广泛阅读的书籍。他们共同撰写的第一本书是 *Data Mining Techniques* 的第一个版本，于 1997 年出版。自那时起，他们就一直积极地挖掘各种行业的数据。持续的实践分析工作使得两位作者能够紧跟数据挖掘、预测以及预测分析领域的快速发展。Gordon 和 Michael 严格地独立于供应商。通过其咨询工作，作者接触了所有主要软件供应商(以及一些小的供应商)的数据分析软件。他们相信好的结果不在于是采用专用的还是开源的软件，命令行的还是点击的软件，而是在于创新思维和健全的方法。

Gordon 和 Michael 专注于数据挖掘在营销和客户关系管理方面的应用——例如，为交叉销售和向上销售改进推荐，预测未来的用户级别，建模客户生存期价值，根据用户行为对客户进行划分，为访问网站的客户选择最佳登录页面，确定适合列入营销活动的候选者，以及预测哪些客户处于停止使用软件包、服务或药物治疗的风险中。Gordon 和 Michael 致力于分享他们的知识、技能以及对这个主题的热情。当他们自己不挖掘数据时，他们非常喜欢通过课程、讲座、文章、现场课堂，当然还有你要读的这本书来教其他人。经常可以发现他们在会议上发言和在课堂上授课。作者还在 blog.data-miners.com 维护了一个数据挖掘的博客。

Gordon 生活在曼哈顿。在本书之前，他最近的一本书是 *Data Analysis Using SQL and Excel*，已经由 Wiley 于 2008 年出版。

Michael 生活在马萨诸塞州剑桥市。他除了在 Data Miners 从事咨询工作之外，还在波士顿大学卡罗尔管理学院讲授市场营销分析(Marketing Analytics)课程。

致 谢

令我们幸运的是，在我们周围到处都是一些极有才华的数据挖掘人员，所以首先要感谢我们在 Data Miners 公司的过去的以及现在的同事 Will Potts、Dorian Pyle 和 Brij Masand，从他们身上我们学会了很多。还有一些客户，由于我们与他们的合作非常密切，所以我们同样把他们看成是同事和朋友，Harrison Sohmer、Stuart E. Ward、III 和 Michael Benigno 就属于这一类。我们的编辑 Bob Elliott 使我们能够(或多或少)如期完工，并帮助我们保持一致的风格。

SAS 研究所和数据仓库研究所在过去 12 年为我们提供了无与伦比的教学机会。我们要特别感谢 Herb Edelstein(现已退休)、Herb Kirk、Anne Milley、Bob Lucas、Hillary Kokes、Karen Washburn，以及其他许多人，是他们使得这些课程成为可能。

在过去的一年中，当我们在撰写本书时，几位朋友和同事一直都非常支持我们。我们要感谢 Diane 以及 Savvas Mavridis、Steve Mullaney、Lounette Dyer、Maciej Zworski、John Wallace、Paul Rosenblum 和 Don Wedding。

我们还要感谢多年来所有与我们一起参与数据挖掘活动的人。我们从他们每个人身上都学到了许多东西。其中许多人这些年一直在帮助我们：

Alan Parker	Gary King
Dave Waltz	Tim Manns
Craig Stanfi II	Jeremy Pollock
Dirk De Roos	Richard James
Michael Alidio	Georgia Tourasi
Michael Cavaretta	Avery Wang
Dave Duling	Eric Jiang
Jeff Hammerbacher	Bruce Rylander
Andrew Gelman	Daryl Berry
Doug Newell	Adam Schwebber
Ed Freeman	Tiha Ghyczy
Erin McCarthy	Usama Fayyad
Josh Goff	Patrick Ott
Karen Kennedy	John Muller
Ronnie Rowton	Frank Travisano
Kurt Thearling	Jim Stagnito
Mark Smith	Stephen Boyer

Nick Radcliffe	Yugo Kanazawa
Patrick Surry	Xu He
Ronny Kohavi	Kiran Nagarur
Terri Kowalchuk	Ramana Thumu
Victor Lo	Jacob Hauskens
Yasmin Namini	Jeremy Pollock
Zai Ying Huang	Lutz Hamel
Amber Batata	

当然，我们依然要感谢所有我们在第一版中感谢过的人们：

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

最后，我们要感谢家人和朋友，特别要感谢 Stephanie 和 Giuseppe，她们在我们撰写这本书期间默默地忍受和奉献。

前 言

15年前，Michael和我合写了这本书的第一版。那本书400页多一点，通过弥合技术和实践之间的差距，通过帮助商业人士了解数据挖掘技术以及帮助技术人员理解这些技术的商业应用，从而满足了我们的调查数据挖掘领域的目标。当Wiley出版社的编辑Bob Elliott让我们撰写*Data Mining Techniques*的第3版时，我们欣然同意，浑然忘记了撰写一本书给我们的个人生活所带来的牺牲。我们也知道新版本将会大幅改写以前的两个版本。

在过去的15年中，这个领域无论是在内涵上还是在字面上都已经得到了扩展，这本书中同样如此。2004年出版了第2版，这一版本增加到了600页，并引入了两个新的章节，分别介绍了生存分析和统计算法这两种新的关键技术，它们对于数据挖掘人员而言已经变得(并依然)越来越重要。现在的这个版本将再度引入新的技术领域——尤其是文本挖掘和主成分分析，同时所有章节中引入了丰富的新实例，并增强了技术描述。这些例子来自各行各业，其中包括金融服务、零售、电信、媒体、保险、保健和基于Web的服务。

作为该领域的从业人员，我们也一直在学习。我们现在大约已经有半个世纪的数据挖掘方面的经验。自1999年以来，Michael和我一直在通过SAS研究所的业务知识系列(本系列与业务的软件方面分离，引入外部专家讲授非软件特定的课程)、数据仓库研究所以及许多不同企业的现场课程进行授课。我们在这些课程中的讲师角色使我们有机会接触成千上万各种行业中的不同业务人员。其中商业数据挖掘技术这门课程就是基于这本书的第二版。这些课程提供了大量有关数据挖掘主题的反馈，比如现实世界的人们正在做什么，以及如何以最佳方式来表示这些思想，从而使它们易于理解。大部分的反馈在这个新版本中都有所反映。我们从学生那里学到的东西看起来与学生从我们这里学到的一样多。

过去两年，Michael也一直在波士顿大学的卡罗尔管理学院讲授市场营销分析课程。*Data Mining Techniques*的前两个版本在许多学院和大学的课程中也广受欢迎，包括商业课程，以及越来越多的数据挖掘课程——在过去十年中其已在各大学中出现。虽然并不打算作为教科书，但是*Data Mining Techniques*为所有类型的学生提供了一个出色的概述。多年来，我们已经在我们的网站上提供了各种可用的数据集，讲师可以在课程中使用它们。

这本书分为4个部分。第一部分讨论数据挖掘的业务上下文。第1章对数据挖掘进行了概述，并给出了如何将其用于现实世界的例子。第2章解释了数据挖掘的良性循环，以及数据挖掘如何帮助理解客户。这一章有几个例子，显示了如何在整个客户生命周期中使用数据挖掘。第3章是数据挖掘方法的概述。第5章和第12章对整体方法进行了精化，分别对应于有指导和无指导数据挖掘。第4章涉及商业统计学知识，介绍了一些贯穿整本书其余部分的关键技术思想。这一章还扩展了MyBuys的案例研究，显示了用于分析A/B营销测试结果的不同方法的长处和短处。

早期版本把所有的数据挖掘技术都放在一个单一的部分。我们现在决定把这些技术划分为两个不同的类别,因此有指导和无指导技术分别拥有它们各自的章节。有指导数据挖掘部分首先在第3章针对有指导数据挖掘对数据挖掘方法进行了精化。后续章节则介绍各种有指导数据挖掘技术,其中包括统计技术、决策树、神经网络、基于记忆的推理、生存分析以及遗传算法。

在第2版中已经覆盖了所有的有指导数据挖掘技术。然而,我们在几个重要方面对它们进行了增强,特别是包含了更多在现实世界中使用它们的例子。第7章现在包括一个关于美国银行提升建模的案例研究,同时还介绍了支持向量机。第8章讨论了径向基函数神经网络。第9章现在有两个很有趣的案例研究,一个是关于 Shazam 如何识别歌曲,另一个使用 MBR 帮助放射学家确定 X 线检查是正常还是异常。第10章介绍生存分析,其中包括了一个针对客户价值的急需的讨论。第11章介绍了遗传算法,其中还包括群体智慧——另一个来自“计算生物学”世界的相关概念,其在数据挖掘领域具有广阔的应用前景。

第三部分专门讨论了无指导数据挖掘技术。第12章解释了四种不同类型的无指导数据挖掘。聚类算法分成两章。其中第13章重点介绍了最常见的聚类技术——K-均值聚类及其三个变体:K-中位数、K-中心点和K-众数。同时它还扩展了关于群集解释的讨论,无论采用哪种技术来识别群集,解释群集都非常重要。第14章介绍了许多技术,包括层次聚类、分裂聚类、自组织网络和高斯混合模型(期望值最大化聚类),它在此版本中是新的内容。第15章的购物篮分析在例子方面进行了加强,这些例子超越了关联规则,其中还包括一个关于种族营销的案例研究。第16章是无指导数据挖掘部分的最后一章,在20世纪90年代,当我们写这本书的第1版时,它几乎还处于外围。现在,它已经处于相当中央的位置,正如这一章的三个案例研究所示。

这本书的最后一部分专注于数据挖掘这一名称中的数据。第17章介绍支持数据的计算机体系结构,例如关系数据库、数据仓库和数据集市。同时,它还介绍了 Hadoop 和分析沙箱,它们都用于处理不适合关系数据库和传统数据挖掘工具的数据。两个早期的版本也有一章介绍数据挖掘的数据准备。由于这个问题如此重要,所以这个版本将该主题分成三章。第18章是关于如何在数据中发现客户和构建客户签名,这是一种许多数据挖掘算法所使用的数据结构。第19章涉及派生变量,以及如何定义变量以帮助模型表现更好的提示和技巧。第20章侧重于如何减少变量的数量,无论是针对诸如神经网络之类的喜欢较少变量的技术,还是出于数据可视化的目的。这一章的关键技术之一——主成分,在这个版本中是新的内容。

第21章涉及的主题本身也可以是一本书,这一主题就是文本挖掘。由于分析文本是构建在本书之前所介绍的许多思想之上,所以我们认为涉及文本挖掘的章节必须放在这本书的最后。其压轴出场凸显了文本挖掘是贯穿本书所覆盖主题的高潮部分。来自 DIRECTV 的最后一个案例研究,不仅是针对业务客户服务方面的一个有趣的文本挖掘应用,同时也是一个极佳的实践中的数据挖掘例子。

与前两个版本一样,这本书的读者对象也是当前的和未来的数据挖掘从业人员和他们的经理。它不适合寻找如何实现各种数据挖掘算法详细说明书的软件开发人员,也不适合试图改进这些算法的研究人员,虽然这两组人都可以通过了解这种软件如何使用而受益。各种思想均是以非技术语言提出,其中尽量减少数学公式和神秘行话的使用。整本书的重点

既包括技术解释，也包括数据挖掘的实际应用，因此这些技术都包含了实际业务上下文的例子。

总之，我们试图写这样一本书：当我们开始自己的数据挖掘职业生涯时，也会想要阅读它。

——Gordon S. Linoff, 2011 年 1 月于纽约

目 录

第 1 章 什么是数据挖掘以及为什么要进行数据挖掘	1
1.1 什么是数据挖掘	2
1.1.1 数据挖掘是一项业务流程	2
1.1.2 大量的数据	2
1.1.3 有意义的模式和规则	3
1.1.4 数据挖掘和客户关系管理	3
1.2 为什么是现在	4
1.2.1 数据正在产生	5
1.2.2 数据正存在于数据仓库中	5
1.2.3 计算能力能够承受	5
1.2.4 对客户关系管理的兴趣非常强烈	5
1.2.5 商业的数据挖掘软件产品变得可用	6
1.3 数据挖掘人员的技能	7
1.4 数据挖掘的良性循环	7
1.5 业务数据挖掘的案例研究	8
1.5.1 识别美国银行的业务挑战	9
1.5.2 应用数据挖掘	9
1.5.3 对结果采取行动	10
1.5.4 度量数据挖掘的影响	11
1.6 良性循环的步骤	11
1.6.1 识别业务机会	12
1.6.2 将数据转换为信息	13
1.6.3 根据信息采取行动	14
1.6.4 度量结果	15
1.7 良性循环上下文中的数据挖掘	17
1.8 经验教训	19
第 2 章 数据挖掘在营销和客户关系管理中的应用	21
2.1 两个客户生存周期	21
2.1.1 客户个人生存周期	21
2.1.2 客户关系生存周期	22
2.1.3 基于订阅的关系和基于事件的关系	23
2.2 围绕客户生存周期组织业务流程	25
2.2.1 客户获取	25
2.2.2 客户激活	27
2.2.3 客户关系管理	29
2.2.4 赢回	29
2.3 数据挖掘应用于客户获取	30
2.3.1 识别好的潜在客户	30
2.3.2 选择通信渠道	30
2.3.3 挑选适当的信息	31
2.4 数据挖掘示例：选择合适的地方做广告	31
2.4.1 谁符合剖析	31
2.4.2 度量读者群的适应度	33
2.5 数据挖掘改进直接营销活动	34
2.5.1 响应建模	35
2.5.2 优化固定预算的响应	35
2.5.3 优化活动收益率	37
2.5.4 抵达最受信息影响的人	40
2.6 通过当前客户了解潜在客户	41
2.6.1 在客户成为“客户”以前开始跟踪他们	41
2.6.2 收集新的客户信息	41

2.6.3 获取时间变量可以预测将来的结果	42	3.5.2 目标数据是什么	72
2.7 数据挖掘应用于客户关系管理	42	3.5.3 输入数据是什么	72
2.7.1 匹配客户的活动	42	3.5.4 易于使用的重要性	72
2.7.2 减少信用风险	43	3.5.5 模型可解释性的重要性	72
2.7.3 确定客户价值	44	3.6 经验教训	73
2.7.4 交叉销售、追加销售和推荐	44	第4章 统计学入门：关于数据，你该了解些什么	75
2.8 保留	45	4.1 奥卡姆(Occam)剃刀	76
2.8.1 识别流失	45	4.1.1 怀疑论和辛普森悖论	77
2.8.2 为什么流失是问题	46	4.1.2 零假设(Null Hypothesis)	77
2.8.3 不同类型的流失	46	4.1.3 p-值	78
2.8.4 不同种类的流失模型	47	4.2 观察和度量数据	79
2.9 超越客户生存周期	48	4.2.1 类别值	79
2.10 经验教训	48	4.2.2 数值变量	87
第3章 数据挖掘过程	51	4.2.3 更多的统计思想	89
3.1 会出什么问题	51	4.3 度量响应	90
3.1.1 学习的东西不真实	52	4.3.1 比例标准误差	90
3.1.2 学习的东西真实但是无用	55	4.3.2 使用置信区间比较结果	91
3.2 数据挖掘类型	56	4.3.3 利用比例差异比较结果	92
3.2.1 假设检验	56	4.3.4 样本大小	93
3.2.2 有指导数据挖掘	60	4.3.5 置信区间的真正含义是什么	94
3.2.3 无指导数据挖掘	61	4.3.6 实验中检验和对照的大小	95
3.3 目标、任务和技术	61	4.4 多重比较	96
3.3.1 数据挖掘业务目标	62	4.4.1 多重比较的置信水平	96
3.3.2 数据挖掘任务	62	4.4.2 Bonferroni 修正	96
3.3.3 数据挖掘技术	66	4.5 卡方检验	97
3.4 制定数据挖掘问题：从目标到任务再到技术	66	4.5.1 期望值	97
3.4.1 选择广告的最佳位置	66	4.5.2 卡方值	98
3.4.2 确定向客户提供的最佳产品	67	4.5.3 卡方值与比例差异的比较	100
3.4.3 发现分支或商店的最佳位置	68	4.6 示例：区域和开局卡方	101
3.4.4 根据未来利润划分客户	68	4.7 案例研究：利用 A/B 检验比较两种推荐系统	103
3.4.5 减少暴露于违约的风险	69	4.7.1 第一个指标：参与会话	104
3.4.6 提高客户保留	69	4.7.2 第二个指标：每个会话的日收益	104
3.4.7 检测欺诈性索赔	70	4.7.3 第三个指标：每天谁取胜	106
3.5 不同技术对应的任务	71	4.7.4 第四个指标：每个会话的平均收益	106
3.5.1 有一个或多个目标	72		

4.7.5	第五个指标：每个客户的 增量收益	107	5.6.4	创建一个预测模型集	130
4.8	数据挖掘与统计	107	5.6.5	创建一个剖析模型集	131
4.8.1	基本数据中没有度量误差	108	5.6.6	划分模型集	132
4.8.2	大量的数据	108	5.7	步骤 5：修复问题数据	132
4.8.3	无处不在的时间依赖性	109	5.7.1	分类变量的值太多	133
4.8.4	实验非常困难	109	5.7.2	包含偏态分布和离群点的 数值变量	133
4.8.5	数据被删截	109	5.7.3	缺失值	133
4.9	经验教训	110	5.7.4	含义随时间而变化的值	134
5	第 5 章 描述和预测：剖析与 预测建模	113	5.7.5	不一致的数据编码	134
5.1	有指导数据挖掘模型	113	5.8	步骤 6：转换数据以揭露 信息	134
5.1.1	定义模型结构和目标	114	5.9	步骤 7：构建模型	134
5.1.2	增量响应建模	115	5.10	步骤 8：评估模型	135
5.1.3	模型稳定性	116	5.10.1	评估二元响应模型和 分类器	135
5.1.4	模型集中的时间帧	117	5.10.2	利用提升评估二元 响应模型	136
5.2	有指导数据挖掘方法	119	5.10.3	利用提升图评估二元响 应模型分数	137
5.3	步骤 1：把业务问题转化为数 据挖掘问题	120	5.10.4	利用剖析模型评估二元 响应模型得分	139
5.3.1	如何使用结果	122	5.10.5	使用 ROC 图表评估二元 响应模型	139
5.3.2	如何交付结果	122	5.10.6	评估估计模型	141
5.3.3	领域专家和信息技术 的角色	123	5.10.7	利用分数排名评估估计 模型	141
5.4	步骤 2：选择合适的数据	123	5.11	步骤 9：部署模型	142
5.4.1	什么数据可用	124	5.11.1	模型部署中的实际问题	142
5.4.2	多少数据才足够	125	5.11.2	优化模型以进行部署	143
5.4.3	需要多久的历史	125	5.12	步骤 10：评估结果	143
5.4.4	多少变量	126	5.13	步骤 11：重新开始	144
5.4.5	数据必须包含什么	126	5.14	经验教训	144
5.5	步骤 3：认识数据	126	6	第 6 章 使用经典统计技术的 数据挖掘	147
5.5.1	检查分布	127	6.1	相似度模型	147
5.5.2	值与描述的比较	127	6.1.1	相似度和距离	148
5.5.3	验证假设	127			
5.5.4	询问大量问题	128			
5.6	步骤 4：创建模型集	128			
5.6.1	聚合客户签名	128			
5.6.2	创建一个平衡的样本	129			
5.6.3	包括多个时间帧	130			

6.1.2 示例: 产品普及率的相似 度模型.....	148	第7章 决策树	179
6.2 表查询模型	153	7.1 决策树是什么以及如何使用 ...	180
6.2.1 选择维度.....	153	7.1.1 一棵典型的决策树.....	180
6.2.2 维度的划分.....	154	7.1.2 使用决策树学习客户流失 ...	181
6.2.3 从训练数据到得分.....	154	7.1.3 使用决策树来了解数据和 选择变量.....	182
6.2.4 通过删除维度处理稀疏和 缺失数据.....	155	7.1.4 使用决策树生成排名.....	183
6.3 RFM: 一种广泛使用的 查询模型	155	7.1.5 使用决策树估计类别概率... ..	183
6.3.1 RFM 单元格迁移.....	156	7.1.6 使用决策树分类记录.....	184
6.3.2 RFM 与测试和度量 (Test-and-Measure)方法论 ..	156	7.1.7 使用决策树估计数值.....	184
6.3.3 RFM 和增量响应建模.....	157	7.2 决策树是局部模型	184
6.4 朴素贝叶斯模型	158	7.3 决策树的生长	187
6.4.1 概率论的一些思想.....	158	7.3.1 发现初始划分.....	187
6.4.2 朴素贝叶斯计算.....	160	7.3.2 生成整棵决策树.....	189
6.4.3 与表查询模型的比较.....	160	7.4 寻找最佳划分	190
6.5 线性回归	161	7.4.1 Gini(总体多样性)作为划 分标准.....	191
6.5.1 最佳拟合曲线.....	162	7.4.2 熵减少或信息增益作为划 分标准.....	192
6.5.2 拟合的优点.....	164	7.4.3 信息增益率.....	193
6.5.3 全局效应.....	166	7.4.4 卡方检验作为划分标准.....	194
6.6 多元回归	166	7.4.5 增量响应作为划分标准.....	195
6.6.1 等式.....	166	7.4.6 减小方差作为数值型目标 的划分标准.....	196
6.6.2 目标变量的范围.....	166	7.4.7 F 检验.....	198
6.6.3 解释线性回归方程的系数... ..	167	7.5 剪枝	198
6.6.4 用线性回归捕捉局部影响... ..	168	7.5.1 CART 剪枝算法.....	198
6.6.5 使用多元回归的其他 注意事项.....	169	7.5.2 悲观修剪: C5.0 剪枝算法 ...	202
6.6.6 多元回归的变量选择.....	170	7.5.3 基于稳定性的修剪.....	202
6.7 逻辑回归分析	171	7.6 从决策树中提取规则	203
6.7.1 建模二元输出.....	171	7.7 决策树变种	204
6.7.2 逻辑函数.....	172	7.7.1 多路划分.....	204
6.8 固定效应和分层效应	174	7.7.2 一次在多个字段上进行 划分.....	205
6.8.1 分层效应.....	175	7.7.3 创建非矩形框.....	205
6.8.2 内部效应与之间效应.....	175	7.8 评估决策树的质量	209
6.8.3 固定效应.....	175	7.9 什么时候使用决策树才合适 ...	209
6.9 经验教训	177	7.10 案例研究: 咖啡烘焙厂的 过程控制	210

7.10.1	模拟器的目标	210	8.12	神经网络模型是否能解释	241
7.10.2	构建烘焙机模拟器	210	8.12.1	灵敏度分析	241
7.10.3	评价烘焙机模拟器	211	8.12.2	使用规则来描述得分	242
7.11	经验教训	211	8.13	经验教训	242
第 8 章	人工神经网络	213	第 9 章	最近邻方法：基于记忆的推理和协同过滤	245
8.1	历史回顾	214	9.1	基于记忆的推理	246
8.2	生物学模型	215	9.1.1	类众模型	247
8.2.1	生物神经元	216	9.1.2	实例：使用 MBR 估计纽约州 Tuxedo 镇的房租价格	248
8.2.2	生物输入层	217	9.2	MBR 面临的挑战	250
8.2.3	生物输出层	217	9.2.1	选择一个平衡的历史记录集	250
8.2.4	神经网络与人工智能	217	9.2.2	训练数据表示	250
8.3	人工神经网络	218	9.2.3	确定距离函数、组合函数和邻居数	253
8.3.1	人工神经元	218	9.3	案例研究：使用 MBR 分类乳房 X 线照片异常	253
8.3.2	多层感知器	220	9.3.1	业务问题：识别 X 射线异常	253
8.3.3	神经网络的一个例子	221	9.3.2	使用 MBR 应对这一问题	253
8.3.4	神经网络拓扑结构	223	9.3.3	总体解决方案	255
8.4	应用实例：房地产估价	224	9.4	距离和相似度计算	255
8.5	神经网络的训练	227	9.4.1	距离函数是什么	256
8.5.1	神经网络如何使用反向传播算法学习	227	9.4.2	“一次一个字段”地建立距离函数	257
8.5.2	神经网络的修剪	228	9.4.3	其他数据类型的距离函数	259
8.6	径向基函数网络	230	9.4.4	当存在一个距离度量指标时	260
8.6.1	RBF 神经网络概述	230	9.5	组合函数：向邻居征求建议	260
8.6.2	选择径向基函数的位置	231	9.5.1	最简单的方法：一个邻居	260
8.6.3	万能逼近器	232	9.5.2	针对类别目标的基本方法：民主	261
8.7	神经网络的应用	233	9.5.3	针对类别目标的加权投票	262
8.8	选择训练集	235	9.5.4	数值目标	262
8.8.1	覆盖特征的所有值	235	9.6	案例研究：Shazam——发现音频文件的最近邻居	263
8.8.2	特征数	235	9.6.1	为何这一技能存在挑战	264
8.8.3	训练集大小	235			
8.8.4	输出的数目和值域	235			
8.8.5	使用 MLP 的经验规则	235			
8.9	数据准备	236			
8.10	神经网络输出结果的解释	238			
8.11	时间序列神经网络	239			
8.11.1	时间序列建模	239			
8.11.2	时间序列神经网络的示例	240			

9.6.2	音频签名	264	10.6	经验教训	299
9.6.3	相似度计算	265	第 11 章	遗传算法与群体智能	301
9.7	协同过滤: 一种用于推荐的最近邻方法	267	11.1	优化	302
9.7.1	构建个人信息	268	11.1.1	优化问题是什么	302
9.7.2	比较个人信息	268	11.1.2	蚁群世界的优化问题	302
9.7.3	预测	269	11.1.3	合众为一(E Pluribus Unum)	303
9.8	经验教训	270	11.1.4	聪明的蚂蚁	304
第 10 章	了解何时应担忧: 使用生存分析了解客户	271	11.2	遗传算法	306
10.1	客户生存	273	11.2.1	一点历史	306
10.1.1	生存曲线揭示的含义	273	11.2.2	计算机中的遗传学	306
10.1.2	从生存曲线中寻找平均持续期	274	11.2.3	基因组的表示	312
10.1.3	使用生存分析保留客户	276	11.2.4	模式: 遗传算法的构造模块	313
10.1.4	将生存视为衰变	277	11.2.5	超越简单算法	315
10.2	风险概率	279	11.3	旅行商问题	316
10.2.1	基本思想	279	11.3.1	穷举搜索	316
10.2.2	风险函数例子	280	11.3.2	简单的贪婪算法	317
10.2.3	删截	282	11.3.3	遗传算法的方法	317
10.2.4	风险计算	283	11.3.4	群体智慧的方法	317
10.2.5	其他类型的删截	284	11.4	案例研究: 使用遗传算法优化资源	319
10.3	从风险到生存	285	11.5	案例研究: 进化出分类投诉的解	320
10.3.1	保留	285	11.5.1	业务上下文	320
10.3.2	生存	286	11.5.2	数据	321
10.3.3	比较保留和生存	287	11.5.3	评论签名	321
10.4	比例风险	288	11.5.4	基因组	322
10.4.1	比例风险的示例	288	11.5.5	适应度函数	323
10.4.2	分层: 度量生存的初始影响	289	11.5.6	结果	323
10.4.3	Cox 比例风险	290	11.6	经验教训	323
10.5	生存分析实践	292	第 12 章	一些新知识: 模式识别与数据挖掘	325
10.5.1	处理不同的客户流失类型	292	12.1	无指导技术和无指导数据挖掘	326
10.5.2	客户何时还会返回	293	12.1.1	无指导技术与无指导技术的对比	326
10.5.3	理解客户价值	295			
10.5.4	预测	297			
10.5.5	风险随时间变化	298			

12.1.2	无指导数据挖掘与有指 导数据挖掘的对比	327	13.4.3	使用决策树描述群集	361
12.1.3	案例研究: 使用有指导 技术的无指导数据挖掘	327	13.5	评价聚类	362
12.2	什么是无指导数据挖掘	329	13.5.1	群集的度量和术语	362
12.2.1	数据探索	329	13.5.2	群集轮廓	363
12.2.2	划分和聚类	330	13.5.3	为打分限制群集直径	365
12.2.3	当目标不明确时目标 变量的定义	332	13.6	案例研究: 城镇聚类	366
12.2.4	模拟、预测和基于智能 体的建模	335	13.6.1	创建城镇签名	366
12.3	无指导数据挖掘的方法论	344	13.6.2	创建群集	367
12.3.1	不存在方法论	345	13.6.3	确定合适的群集数目	367
12.3.2	需要谨记的事情	345	13.6.4	评价群集	368
12.4	经验教训	345	13.6.5	使用人口统计学群集 调整区域边界	370
12.4.6	商业成功	370	13.7	K-均值算法的变种算法	371
13.7.1	K-中位数、K-中心点和 K-众数	371	13.7.1	K-中位数、K-中心点和 K-众数	371
13.7.2	K-均值的软层面	374	13.7.2	K-均值的软层面	374
13.8	聚类的数据准备	375	13.8	聚类的数据准备	375
13.8.1	一致性缩放	375	13.8.1	一致性缩放	375
13.8.2	使用权重编码外部信息	375	13.8.2	使用权重编码外部信息	375
13.8.3	选择聚类变量	376	13.8.3	选择聚类变量	376
13.9	经验教训	376	13.9	经验教训	376
第 13 章	发现相似的岛屿: 自动群集 检测	347	第 14 章	其他的群集检测方法	379
13.1	搜索简化的岛屿	348	14.1	K-均值聚类的缺点	379
13.2	客户细分和聚类	349	14.1.1	合理性	380
13.2.1	相似性聚类	350	14.1.2	一个直观的例子	380
13.2.2	基于群集划分的跟踪 活动	351	14.1.3	通过改变度量范围来 修正问题	382
13.2.3	聚类揭示被忽视的细分 市场	352	14.1.4	这在实际中意味着什么	383
13.2.4	适应军队需求	353	14.2	混合高斯模型	383
13.3	K-均值聚类算法	353	14.2.1	把高斯过程引入 K-均值 聚类	384
13.3.1	K-均值算法的两个步骤	354	14.2.2	回到混合高斯模型	386
13.3.2	Voronoi 图和 K-均值 群集	355	14.2.3	混合高斯模型的打分	388
13.3.3	选择群集种子点	357	14.2.4	混合高斯模型的应用	388
13.3.4	选择 K 值	357	14.3	分裂聚类	389
13.3.5	使用 K-均值检测 离群点	358	14.3.1	一种类决策树的聚类 算法	390
13.3.6	半指导聚类	359			
13.4	解释群集	359			
13.4.1	使用质心表征群集	359			
13.4.2	使用群集之间的差异表 征群集	360			

14.3.2	分裂聚类的打分	391	15.4.4	大数据问题	432
14.3.3	群集和树	391	15.5	思想扩展	432
14.4	凝聚(层次化)聚类	392	15.5.1	左右两侧包含不同的项目	432
14.4.1	凝聚聚类方法的综述	392	15.5.2	利用关联规则比较商店	433
14.4.2	凝聚聚类算法	395	15.6	关联规则和交叉销售	434
14.4.3	为凝聚群集打分	397	15.6.1	一个经典的交叉销售模型	435
14.4.4	凝聚聚类的局限性	398	15.6.2	更可信的倾向度产生方法	435
14.4.5	凝聚聚类的实际应用	399	15.6.3	使用置信度所产生的结果	436
14.5	自组织映射	400	15.7	序列模式分析	436
14.5.1	什么是自组织映射	401	15.7.1	序列的发现	436
14.5.2	SOM 的训练	403	15.7.2	序列关联规则	439
14.5.3	SOM 的打分	404	15.7.3	利用其他数据挖掘技术的序列分析	440
14.6	继续搜索简化的岛屿	404	15.8	经验教训	440
14.7	经验教训	405	第 16 章	链接分析	443
第 15 章	购物篮分析和关联规则	407	16.1	图论基础	444
15.1	购物篮分析的定义	408	16.1.1	图是什么	444
15.1.1	购物篮数据的四个级别	408	16.1.2	有向图	445
15.1.2	购物篮分析的基础: 基本度量	409	16.1.3	加权图	446
15.1.3	订单特征	410	16.1.4	哥尼斯堡的七桥问题	447
15.1.4	项目(产品)人气	411	16.1.5	图中的回路检测	449
15.1.5	跟踪市场干预	412	16.1.6	旅行商问题的反思	449
15.2	案例研究: 西班牙语或英语	413	16.2	社交网络分析	452
15.2.1	业务问题	413	16.2.1	六度分割理论	453
15.2.2	数据	414	16.2.2	你朋友说了关于你的什么事情	454
15.2.3	“西班牙裔城市”偏好的定义	414	16.2.3	发现托儿福利欺诈	454
15.2.4	解决方案	415	16.2.4	交友网站中谁响应了谁	455
15.3	关联分析	416	16.2.5	社会营销	456
15.3.1	规则不是万能的	416	16.3	呼叫图挖掘	456
15.3.2	关联规则中的项目集	418	16.4	案例研究: 追踪领袖	458
15.3.3	关联规则的益处	420	16.4.1	业务目标	458
15.4	构建关联规则	421	16.4.2	数据处理面临的挑战	459
15.4.1	选择正确的项目集	422			
15.4.2	从所有这些数据中生成规则	426			
15.4.3	克服实际限制	429			