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
Michael J. A. Berry 著
巢文涵 张小明 王芳 译

WILEY
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
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

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

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

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
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 Petefi sh
Yvonne McCollin

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 ai ok de 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，同时在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