



科文图书

全美最新工商管理
权威教材系列

Mastering Data Mining

The Art and Science of
Customer Relationship Management

数据挖掘

——客户关系管理的科学与艺术

[美] 迈克尔·J·A·贝里 (Michael J.A.Berry)
戈登·S·利诺夫 (Gordon S.Linoff)

著

中国人民大学 袁卫 等译

中国财政经济出版社

全美最新工商管理权威教材系列

Mastering Data Mining

The Art and Science of
Customer Relationship Management

数据挖掘

——客户关系管理的科学与艺术

[美] 迈克尔·J·A·贝里 (Michael J. A. Berry)
戈登·S·利诺夫 (Gordon S. Linoff)

著

中国人民大学	袁 卫	王 星	戴稳胜	
	匡红波	伍业锋		译
	张尧庭	谢邦昌		校

中国财政经济出版社

著作权合同登记图字：01-2003-4151号

图书在版编目(CIP)数据

数据挖掘：客户关系管理的科学与技术/(美)贝里，(美)利诺夫著；袁卫等译。

—北京：中国财政经济出版社，2003.8

(全美最新工商管理权威教材系列)

书名原文：Mastering Data Mining: The Art and Science of Customer Relationship Management

ISBN 7-5005-6581-X

I. 数... II. ①贝... ②利... ③袁... III. 数据采集-教材
IV. TP274

中国版本图书馆 CIP 数据核字(2003)第 047406 号

Mastering Data Mining: The Art and Science of Customer Relationship Management

Copyright ©2000 by Michael J. A. Berry, Gordon S. Linoff

Original published by John Wiley & Sons, Inc.

All rights reserved.

中文简体字版版权©2003 科文(香港)出版有限公司

中国财政经济出版社出版

中国财政经济出版社 出版

(北京市海淀区阜成路甲 28 号 邮政编码：100036)

发行处电话：88190406 财经书店电话：64033436

URL: <http://www.cfeph.com.cn>

北京科文剑桥图书有限公司承销

(北京安定门外大街 208 号三利大厦 邮政编码：100011)

编辑热线：64248899-6215 购书热线：64203023

网上购书：www.dangdang.com

北京交通印务实业公司印刷装订 新华书店经销

787 毫米×1092 毫米 16 开本 25 印张 500 千字

2004 年 1 月第 1 版 2004 年 1 月第 1 次印刷

定价：38.00 元

编委会

主任：厉以宁 赵纯均

编委：徐二明 韩伯棠 魏法杰

张国华 俞 渝 彭松建

荆 新 王光远 吕一林

感谢

中国人民大学副校长	袁卫	教授
台湾政治大学商学院院长	林英峰	教授
纽约大学商学院副院长	Fred Choi	教授
《财务报表分析》杂志主编、 纽约大学会计系	Paul Brown	教授
纽约大学商学院高级经理 培训中心主任	Michael Darling	教授
纽约大学商学院策略系	Ari Ginsberg	教授

在选书上给予的支持

译者序

数据挖掘是信息领域发展最快的技术,很多不同领域的专家,比如统计学家、数据库专家等,都从中获得了发展的空间,使得数据挖掘日益成为企业界讨论的热门话题。随着信息技术的发展,人们采集数据的手段日益丰富与高明,由此积累的数据日益膨胀,数据量达到 GB 甚至 TB 级,而且高维数据也日益成为主流。这些海量数据及其高维特征使得传统的数据分析手段相形见绌。计算机性能的日益更新,使得人们能够希望计算机帮助我们分析与理解数据,帮助我们以丰富的数据为基础做出正确决策。于是数据挖掘这一融合多种分析手段,从大量数据中发现有用知识的方法就应运而生,并在使用中得以蓬勃发展。

数据挖掘是一个多学科交叉的领域。一方面,数据挖掘以计算技术的发展为首要条件,没有数据的有效组织,从一堆数据垃圾中发现有用知识是痴人说梦;没有大量计算算法的支持,即使是简单的查询也会耗时巨大,更不用说发现有用模式。另一方面,即使数据得到了有效组织,计算算法足够先进,要想发现海量数据中隐藏的有用信息,还必须综合利用统计学、模式识别、人工智能、机器学习、神经网络等学科的专业知识。比如数据挖掘使用的分析方法,有相当大比重是靠高等统计学中的多元分析来支撑的,一般将之定义为数据挖掘技术的 CART、CHAID 或模糊计算等理论方法,也都是由统计理论发展衍生的。当然,所有这些学科的发展必然会从不同角度关注数据分析技术的进展,数据挖掘也为这些学科的发展提供了新的机遇与挑战。

正因为数据挖掘的强大功能,它引起了学术界与实务界的广泛关注,吸引了众多研究与开发人员。目前国内外很多大学都开设有数据挖掘的课程,而且出版发行了若干数据挖掘方面的专业书籍。然而,长期以来,数据挖掘方面的专著往往侧重于机器学习领域,对案例的讲解所作的研究不够,而且多数专著过于理论化,与数据挖掘的本来目的——方便企业末端使用者的使用这一目标相违背,因此,选择一本合适的数据挖掘教材困难较大。

《数据挖掘——客户关系管理的科学与艺术》一书的英文原著,一直被作为我们中国人民大学统计学系数据挖掘中心,从事数据挖掘领域理论与应用研究的经典教材。在长期的探索和研究中,我们体会到从案例中学习是快速领会数据挖掘概念的最佳途径。本书通过丰富的来自于实际的案例以及大量宝贵的数据分析经验,阐述了数据挖掘的本质。它是掌握数据挖掘理论的宝典。事实上,这本书已得到了世界范围内的广泛认可。从著名的网上书店——亚马逊书店关于数据挖掘书籍的销售统计看,本书英文版长期雄居销量第一的宝座。本书的主要特点在于其清晰的理论架构、系统且易于理解的理论解说及丰富的实际案例。通过对本书的学习,读者不仅可以从中掌握数据挖掘的核心技术和理念等,而且更主要的是可以了解数据挖掘在各个业务领域的应用,从而为今后从事实际的数据挖掘工作提供攻玉之石。

本书分为三个部分共十五章。全书的翻译由袁卫和王星组织完成。参加本书初稿翻译的成员

2002/8/12



包括:袁卫(前言,第一章)、王星(第二、三、四、五、六章)、伍业锋(第七、八章)、匡红波(第九、十、十一、十二章)、戴德胜(第十三、十四、十五章)。张尧庭教授和谢邦昌教授从学术角度对本书进行了全面审校。

本书译者衷心感谢科文公司和财经出版社的编辑们,没有他们的远见和认真负责的工作,该书不可能这么快与读者见面。

由于本书涉及领域较宽,知识面广,许多领域译者并不熟悉而且有少量的专业术语目前并没有固定的译法,翻译难度很大,因此,译文中的错误和不妥之处在所难免。我们欢迎读者及数据挖掘爱好者给我们提出批评与建议。

我们的联系方式是:

地址:中国人民大学科研楼 A 座 901

电话:86-10-62511709,13683215697,13693043707

网站:<http://www.dmchina.org>

E-mail:dmcstat@sina.com

中国人民大学副校长

中国统计学会副会长

袁卫

2003 年 7 月

致 谢

在本书的写作过程中,我们从许多人手中获得了巨大的帮助。在此我们列举他们中的部分代表,献上我们最诚挚的谢意。

首先应该感谢的是我们的编辑, Bob Elliott, 是他首先倡议出版一本有关案例学习的著作, 并最终促成了这部作品的诞生。

Blue Martini 的 Ronny Kohari 和 Ralph Kimball 协会的 Ralph Kimball 两位先生, 从专业的角度审核了手稿, 并确保专业词汇使用准确无误。多亏他们的诸多建议, 才使得许多漏洞和错误得到及时纠正。

除了我们二人的工作, 本书还汇集了其他数据挖掘工作者的卓越成果和宝贵经验。其中有 Apower Solutions 的 Alan Parker 以及 Geoff Parker。SAS 研究所的 Anne Milley 的工作, 以及 R.R. Donnelly & Sons 公司的 Bob Evans 等人的工作。另外, 在与许多专家学者的探讨中, 我们受益匪浅, 其中包括 Dorian Pyle, Herb Edelstein, Bill Inmon 和 Erik Thomsen。我们对数据挖掘在社会中地位的诸多理解取之于和 Esther Dyson 和 Nolan Bowie 等人的谈话。

我们特别感谢数据挖掘工具的几位馈赠者。他们为我们提供了软件和相关培训, 甚至使我们有机会接触到软件开发者和设计者。我们要向 SGI 公司的 Cliff Brunk, Brad Fiery, Peter Maysek, Marc Ondrechen, Lorna Lusic, Manuel Hoffman, Aydin Senkut 及 Mario Schkolnick 等人表示感谢, 是他们使我们如愿以偿成功地将 Mineset 数据开发工具应用于本书的各个项目中。除此之外, 在 SAS 研究所, 我们要感谢 Will Potts, 是他教会我们有关类神经网络的使用方法, 并和其他人员 Herbert Kirk, Anne Milley, Austin Tripensee, Padraic Neville, Mark Brown 一起使我们成功地使用了 SAS Enterprise Miner 软件。Thinking Machines 公司的 Charlie Berger 提供了有关 Darwin 软件的资料和使用权。SPSS 公司的 Debra Daily 和 Ken Elliott 提供了有关 Clementine 软件的资料和使用机会。Angoss 的 Ken Ono 和 Eric Apps 提供了有关数据流的软件资料, 并使我们有机会接触了 Knowledge Studio 软件的使用。Ab Initio 公司的 Cliff Lasser, Craig Stanfill, Marie Campbell, Sheryl Handler 以及 Paul Bay 等人提供了有关数据流软件的资料和使用机会。Tessera Enterprise Systems 的 Beth Maerz, Brett Mc Tammany, Susan Eliis 以及 Tatyana Kofman 提供了有关 Rapid Modeling Environment 的资料和使用机会。Torent Systems 的 Rob Utzschneider, Paul Guerin 和 Katherine 提供了有关 Orchestrate 的资料。Peter Adams, David Sabel 以及 Matt Grosso 提供有关电子商务以及他们所在 Primary Knowledge 公司的资料。本书的部分项目是在 DST/Condor 的 DSS 实验室完成, 我们向 Erik Thomsen, George Spofford 以及 Barry Grushkin 等人的支持表示感谢。

当然, 我们要感谢我们的客户, 因为没有他们那些诱人的商业问题, 我们只能纸上谈兵。在过去的几年中, 我们和许多蓬勃发展的企业结下了深厚的友情, 获得了许多帮助, 其中有 NCC, Wells Fargo, Banco Velox, Huntington Bank, American Express, Customer Analytics, Cogit Corporation, Blue Marti-



ni, SK Telecom, MCI/WorldCom, SkyTel, BellSouth, Vermont Country Stores, WAM! Net, Liberty Mutual, John Hancock 等单位,——数量之多不胜枚举。

最后,我们谨向以下人员表示感谢:

Abhay Mehta

Al Fan

Anjana Pursnani

Anne M. Grim

Anthony Cote

Ben Swanson

Bill Walsh

Brain Guscott

Clark Abrahams

Dave Trevorrow

David Anderson

Ed Zyszkowski

Elly Pieper

Eric Williams

Erin McCarthy

Evangelos Simoudis

Greg James

Injung Hwang

Jason Mallen

Jim Bell

Jim Stitzlein

Joan

Judy barden

Larry Shaw

Lounette Dyer

Mary Bohannon

Mayank Prakash

Mike Wade

Miles Adams

Rebekah Peterson

Ruy Cardoso

Sabine Shea

Sandra Batra

Sheridan Young

Steve Mohan

Sue Osterfeld

Su Yoon

Ted Browne

Thulasiram Kesavan

- Vandy Johnson

Youngho Oh

序 言

迈克尔和我的第一本书《市场营销、销售及客户服务的数据挖掘技术》(*Data Mining Techniques for Marketing, Sales, and Customer Support*, John Wiley & Sons 出版, 1997, 以下简称《数据挖掘技术》)发行以来,读者反应强烈、好评如潮,令我们深受感动。我们编写那本书的初衷是对数据挖掘作个综合且又易于理解的介绍,看来我们的目的是达到了。

自从那本《数据挖掘技术》问世后,世界发生了很大变化。我们办了自己的公司——数据挖掘者(Data Miners),专门致力于数据挖掘工作。数据挖掘者的使命是像强调模型的结果一样强调对方法的理解,像重视数据挖掘技术一样重视数据挖掘的过程。

对于毫无经验的创业者来说,放着定期领取工资的稳定工作不做,而去独立闯荡肯定是一件既刺激又冒险的行动。但它确实给我们提供了接触企业并向实践学习的机会,使我们学会了如何面对客户、如何分配资金以及如何选择客户等。它还为我们创造了与数据挖掘主要客户以及该领域顶尖人物合作的机会。

我们的朋友和家人不止一次地问过我们,为什么还要花费那么多时间写第二本书?简单地说是《数据挖掘——客户关系管理的科学与艺术》一书需要写。数据挖掘的领域最近几年发展得非常迅速,商界和技术领域的实践需要我们写,希望这本书能够满足读者的要求。

要了解数据挖掘市场发展得有多快,我们只需看一看商业专栏的记载就够了。在本书写作的过程中,我们见证了由于数据挖掘在客户关系管理和电子商务领域的杰出作用使得大公司发生并购或者扩展的许多范例。

- 世界最大的私人控股软件公司——SAS 公司,推出了公司历史上最成功的产品——数据挖掘软件包。我们习惯开玩笑地说:“在 SAS 系统中,统计就是一切。那么数据挖掘就只能是这一切之外的了。”若按照这一定义,数据挖掘就没有存在的必要了!
- 在分析软件市场上,SAS 最强的竞争对手 SPSS 通过开发研制国际领先的数据挖掘软件,使得其竞争力得到提高,而在几年前 SPSS 还无力与 SAS 竞争。
- DoubleClick, 这一互联网最大的广告销售商购买了 Abacus, 一个善于用目录数据建立预测模型的公司。

种种迹象表明,数据挖掘已经成为各种问题的解决方法之一。如果做得好的话,数据挖掘确实可以解决许多难题,但这需要理解数据挖掘的整个过程。这一理解来自于过去所做项目成功与失败两个方面的经验和教训。过去的每一个项目都积累了宝贵的经验。在这本书中,我们将带领你们去重温这些项目,以分享我们的经验。

《数据挖掘——客户关系管理的科学与艺术》一书将对数据挖掘作全面的介绍。第一部分从商业领域开始。本部分的四章将回答如下的问题:数据挖掘为什么重要?数据挖掘有哪些诀窍?



第四章将特别介绍客户及客户关系管理。尽管数据挖掘在许多领域都有成功的应用,但在客户关系管理领域引起了人们格外的兴趣和重视。

第二部分从技术层面上介绍数据挖掘。第五章回顾数据挖掘技术(这些技术和方法在我们的第一本书里有详细的介绍),第六章讨论数据,接下来的第七章研究如何建立好模型。这一章十分重要,因为它包含了我们多年的经验和教训。

第三部分是本书最长的部分,也是最重要的部分。这部分是数据挖掘的案例研究。尽管这些案例都是商业领域的,但涉及的范围还是很广的,从对数十万兆字节数据的探索(hundreds of gigabytes of data)到为网络银行客户服务预报下一个标题广告,从而改进印刷过程。所有这些案例讨论的都是实际的问题,同时都给出了其中所用的方法、数据、所得到的结果以及经验和教训。

最后一章是以更广阔的视野来观察数据挖掘,即从社会的角度。数据挖掘一方面要反对专制,提倡数据公开,同时要重视保护个人的隐私。迈克尔和我试图充分展示数据挖掘的功能,但在分析数据时要特别小心,以保护个人隐私。

在编写这本书时,我们特别强调数据挖掘的实际应用。我们希望这本书能够帮助读者掌握这门艺术。

目 录

第 1 部分 本书焦点

第 1 章 数据挖掘概述	(3)
1. 什么是数据挖掘?	(4)
2. 数据挖掘能做什么?	(5)
分类	(5)
估计	(5)
预测	(6)
组合或关联法则	(6)
聚类	(6)
描述与可视化	(7)
3. 商业领域的数据挖掘	(7)
作为研究工具的数据挖掘	(7)
改进生产过程的数据挖掘	(8)
市场营销中的数据挖掘	(8)
客户关系管理中的数据挖掘	(9)
4. 技术层面的数据挖掘	(9)
数据挖掘与机器学习	(9)
数据挖掘与统计学	(10)
数据挖掘与决策支持	(10)
数据挖掘与计算机技术	(12)
5. 数据挖掘的社会背景	(12)

第 2 章 为什么要精通数据挖掘这门艺术	(14)
1. 数据挖掘的四种方法	(15)
购买评分	(16)
购买软件	(16)
聘请编外专家	(22)
培养企业内部骨干	(24)



本章小结	(25)
第 3 章 数据挖掘方法论:互动循环系统	(26)
1. 数据挖掘的两种类型	(27)
有监督的数据挖掘	(27)
无监督的数据挖掘	(28)
2. 数据挖掘的互动循环过程	(29)
3. 正确识别业务问题	(30)
实施数据挖掘是否必要?	(31)
是否存在最让人感兴趣的客户子群或客户细分?	(31)
相应的行业规范有哪些?	(31)
关于数据	(31)
印证业内专家的观点	(32)
4. 将数据转换成可操作的决策	(32)
确认和获取数据	(32)
生成有效数据、探索数据以及清洁数据	(34)
将数据转换成具有合适的粒度的数据	(35)
加入衍生变量	(36)
准备建模数据集	(36)
选择建模技术和训练模型	(37)
检测模型的执行效率	(37)
5. 将结果生成决策	(39)
6. 评测模型的有效性	(40)
7. 成功建立预测模型的要点	(41)
预测模型的时间范围	(41)
模型的使用有效期	(42)
假定 1:过去是将来的预言家	(42)
假定 2:数据是可以获得的	(43)
假定 3:数据中应包括我们的预期目标	(44)
本章小结	(44)
第 4 章 客户和他们的生命周期	(46)
1. 谁是企业的客户	(46)
消费者	(47)
企业客户	(48)
客户市场细分	(50)



2. 客户的生命周期	(51)
客户生命周期的不同阶段	(52)
客户生命周期中的重要事件	(53)
客户生命周期中不同的时段所产生的资料	(55)
3. 客户的生理生命周期	(56)
4. 选择最佳时机, 锁定最佳客户	(57)
预算最优化	(57)
促销活动最优化	(59)
客户最优化	(61)
本章小结	(64)

第 2 部分 数据挖掘的三大支柱

第 5 章 数据挖掘技术与算法	(71)
1. 不同的目标要求不同的技术	(71)
不同的数据类型要求不同的方法	(73)
2. 三种数据挖掘技术	(73)
3. 自动类别侦测	(74)
K—均值类别侦测的工作原理	(74)
选择聚类所产生的后果	(77)
4. 决策树	(79)
决策树的工作原理	(79)
决策树的建立过程	(81)
选择决策树所产生的后果	(85)
5. 神经网络	(85)
神经网络的训练	(89)
选择神经网络所产生的后果	(90)
本章小结	(91)

第 6 章 无所不在的数据	(92)
1. 数据结构	(93)
行	(93)
列	(95)
数据挖掘中列的作用	(97)
数据挖掘中的数据	(99)
2. 数据看起来究竟像什么?	(100)



数据从哪里来·····	(100)
粒度的合适水平·····	(105)
度量数据取值的不同方法·····	(107)
3. 多少数据才足够呢? ·····	(110)
4. 衍生变量 ·····	(111)
使用衍生变量时应该注意的问题 ·····	(112)
离群点的处理 ·····	(112)
列变量的组合 ·····	(113)
分类汇总·····	(114)
从某一列中提取信息·····	(115)
时间序列·····	(117)
5. 案例:客户行为的界定 ·····	(119)
6. 受污染的数据 ·····	(125)
缺失数据·····	(125)
定义模糊·····	(126)
谬误值·····	(126)
本章小结·····	(127)

第7章 建立有效的预测模型 ·····(129)

1. 建立好的预测模型 ·····	(129)
预测模型的建立过程·····	(130)
对模型效果的衡量·····	(131)
模型稳定性·····	(135)
保持模型稳定性所面临的挑战·····	(136)
2. 对模型集进行处理 ·····	(136)
分割与掌握:训练集、测试集与评价集·····	(137)
模型集规模对模型效果的影响 ·····	(137)
模型集密度对模型效果的影响 ·····	(138)
抽样 ·····	(138)
何谓过抽样? ·····	(139)
利用时间相关资料来建立模型 ·····	(142)
模型输入和模型输出 ·····	(144)
执行时间:考虑模型的建立时间 ·····	(145)
时间和遗漏数据 ·····	(148)
建立时间上易于转换的模型 ·····	(149)
字段命名 ·····	(151)
3. 使用多个模型 ·····	(151)

多个模型的表决	(152)
将输入分段	(155)
对模型进行组合的其他原因	(156)
4. 做试验!	(157)
模型集	(158)
不同类型的模型以及模型参数	(158)
时间范围	(159)
本章小结	(160)
 第 8 章 实施控制:建立数据挖掘环境	(161)
1. 起步	(161)
何谓数据挖掘环境?	(162)
四个案例研究	(162)
数据挖掘环境得以成功的要素	(162)
2. 案例 1:建造公司内部核心竞争力	(163)
保险行业的数据挖掘	(163)
开端	(164)
3. 案例 2:创造新的商机	(166)
向网上发展	(166)
环境	(167)
潜在客户的数据仓库	(167)
下一个步骤	(169)
4. 案例 3:在数据仓库工作中培养数据挖掘技能	(169)
特殊类型的数据仓库	(169)
数据挖掘的计划	(171)
信息技术部门内部的数据挖掘	(171)
5. 案例 4:利用特斯拉快速建模环境法(RME)进行数据挖掘	(172)
建立高级数据挖掘环境所需的条件	(172)
什么是 RME?	(173)
RME 如何运作?	(173)
RME 如何协助数据准备	(175)
RME 如何支持抽样	(176)
RME 如何协助建立模型	(177)
RME 如何协助模型评估和管理	(177)
本章小结	(179)



第3部分 案例研究

第9章 数据挖掘在目录直销业中的应用

——有谁会需要香油袋和长裤拉伸器？	(186)
1. 佛蒙特乡村小店	(186)
VCS的发家史	(187)
预测模型	(189)
2. 商业问题	(189)
3. 数据	(191)
4. 技术路线	(192)
数据挖掘软件的选择	(192)
RFM与细分的基础	(192)
挑战者——神经网络、决策树和回归分析	(194)
决定可能已经发生的事	(196)
计算投资回报率	(196)
5. 未来	(196)
期望收益	(197)
本章小结	(197)

第10章 数据挖掘在在线银行业中的应用

——顾客垂青的下一个产品是什么？	(198)
1. 获取利润	(198)
2. 商业问题	(199)
3. 数据	(200)
从账户到客户	(202)
推出产品	(204)
4. 解决问题的方法	(206)
标准分数	(206)
如果走起来像只鸭	(207)
这个方法的陷阱	(208)
5. 建模	(209)
决策树模型	(211)
建立其他模型	(218)
得到交叉销售模型	(218)
6. 更完美的世界	(219)