WILEY

DATA MINING FOR BUSINESS ANALYTICS

CONCEPTS, TECHNIQUES, AND APPLICATIONS WITH JMP PRO

数据挖掘

商业数据分析技术与实践

[美] 盖丽特·徐茉莉（Galit Shmueli）
彼得·布鲁斯（Peter C. Bruce）
米娅·斯蒂芬斯（Mia L. Stephens）
尼廷·帕特尔（Nitin R. Patel）
著

阮敬 严雪林 周晖 译

清華大學出版社

DATA MINING FOR BUSINESS ANALYTICS

CONCEPTS, TECHNIQUES, AND APPLICATIONS WITH JMP PRO

数据挖掘

商业数据分析技术与实践

[美] 盖丽特·徐茉莉（Galit Shmueli）
彼得·布鲁斯（Peter C. Bruce）
米娅·斯蒂芬斯（Mia L. Stephens）
尼廷·帕特尔（Nitin R. Patel） 著

阮敬 严雪林 周暐 译

清华大学出版社
北京

Galit Shmueli, Peter C. Bruce, Mia L. Stephens, Nitin R.Patel
Data Mining for Business Analytic: Concepts, Techniques, and Applications with JMP Pro
EISBN: 978-1-118-87743-2

图书在版编目(CIP)数据

数据挖掘：商业数据分析技术与实践 / (美)盖丽特 • 徐茉莉(Galit Shmueli)等著; 阮敬, 严雪林, 周暐译. — 北京：清华大学出版社，2018
书名原文: Data Mining for Business Analytics: Concepts, Techniques, and Applications with JMP Pro
ISBN 978-7-302-49766-0

Ⅰ. ①数…　Ⅱ. ①盖… ②阮… ③严… ④周…　Ⅲ. ①商业信息—数据处理　Ⅳ. ①F713.51

中国版本图书馆 CIP 数据核字（2018）第 058897 号

责任编辑：刘　洋
封面设计：李召霞
版式设计：方加青
责任校对：宋玉莲
责任印制：宋　林

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：三河市铭诚印务有限公司
经　　销：全国新华书店
开　　本：187mm×235mm　　**印　　张：**26　　**字　　数：**472 千字
版　　次：2018 年 6 月第 1 版　　**印　　次：**2018 年 6 月第 1 次印刷
定　　价：118.00 元

产品编号：073707-01

序　言

无论你选择什么职业或工作地点，你的未来肯定会被数据所包围。现代世界由几十亿个键盘和数万亿个卡片刷头所发出的数据脉冲所构成，这些数据来自于电子设备和系统的各种操作，并且能够在全球范围内迅速传播。数据量是难以用数量来衡量的。但这并不在于你拥有多少数据，而是你用它做什么。把握住这个凌乱的数据世界并很好地利用它，将会成为组织运作良好和职业生涯成功的关键，它不仅仅存在于硅谷、谷歌、Facebook 这些地方，也存在于保险公司、银行、汽车制造商、航空公司、医院等地方，甚至可以说它几乎无处不在。

这就是本书（《数据挖掘：商业数据分析技术与实践》）所能给读者提供的。Shmueli 教授同她的合著者为学生们提供了这样一个非常有用的学习指南，其中涉及与复杂数据集相关的重要概念和方法。本书作者具有多年的教学经验，为了跟上本科及研究生商业分析课程中的变化，我们已经对早期的版本进行修订。最重要的是，新版本集成了 SAS 研究所用于处理和分析数据的统计工具 JMP Pro®。学习分析方法的最终目的是通过数据生成一些见解。通过强大的统计工具训练学习者敏捷的思维，是学习过程中必不可少的关键一步。

如果你把目光放在引领数字世界，那么本书将会是你为未来做准备的开始。

Michael Rappa
高级分析研究所
北卡罗来纳州立大学

译者简介

阮敬，博士，教授。现任首都经济贸易大学研究生院副院长，兼任中国统计教育学会常务理事兼高等教育分会副秘书长、中国商业统计学会常务理事、中国现场统计研究会理事、全国工业统计学教学研究会理事、北京大数据协会副秘书长等职。近年来主持国家级和省部级科研项目以及国家机关和企事业单位横向课题30余项；在国内外公开发表论文50余篇，出版专著教材6部，先后11次荣获全国统计科学研究优秀成果奖、北京市哲学社会科学优秀成果奖、北京市统计科学研究优秀成果奖等奖励。

严雪林，现任SAS公司JMP事业部亚太区总经理，北京大数据协会副会长。兼任首都经济贸易大学统计学院兼职教授、全国应用统计专业学位研究生教育指导委员会“大数据硕士”兼职研究生导师、上海财经大学统计与管理学院兼职研究生导师，曾任JMP大中华地区总经理、SAS大中华地区市场及渠道总监、上海大学管理学院兼职教授、中央财经大学统计学院兼职教授等职。长期于全球信息技术及数据分析类领导企业担任高级管理职位，在数据分析、商业智能、运营、市场营销、战略、信用及风险、大数据等领域有近20年的研究与实践。曾领导多个国际团队，为财富500强企业及中国本土行业领导者提供数据分析战略、市场营销、研发优化、运营、风险等领域的服务，成功帮助客户导入大数据与分析战略，建立起基于数据分析的科学决策与运营体系，实现转型，推动变革与创新。

周暐，硕士。现任SAS公司JMP事业部数据分析专家。兼任中国质量协会六西格玛专家委员会专家委员，上海财经大学统计与管理学院应用统计硕士研究生导师（兼职），曾在西门子（Siemens）通信运营部门工作多年。精通统计学的理念与实践，擅长探索性数据分析（EDA）、实验设计（DOE）、建模预测（Modeling）、生存与可靠性分析（Reliability）、统计过程控制（SPC）、市场调查（Marketing Research）等统计分析方法在实际工作中的应用。2007年年底编著完成《六西格玛管理统计指南》（中国质量协会推荐的专业书籍），积累了统计技术在半导体、电子、医药、汽车、化工、银行、教育等行业应用的宝贵经验，曾受邀为多家著名企业和高等学府辅导统计应用及六西格玛专业知识，并获得广泛好评。

前　言

本书在2007年年初出版。随后，该教材被众多教师在授课中使用，从专用数据挖掘课程到更一般的商业分析课程（包括我们自己超过十年的在线教学和面授经验）。根据MBA、本科、管理课程的教师以及学生的反馈，在第二版中对一些现有章节进行了修订，并增加了两个新的主题：数据可视化和时间序列预测。

这本书是第一个完全集成JMP Pro®[①]的版本，而不是Microsoft Office Excel的插件XLMiner。JMP Pro®是一个运行在Mac和Windows系统上的桌面版统计工具，由SAS研究所开发。本书所有案例、特殊主题框、说明和练习都基于JMP Pro 12。专业版的JMP拥有丰富的内置工具，用于交互式数据可视化，分析和建模等[②]。

该版本还包含其他重要的变化。第一个明显的变化是标题：除了新增JMP Pro®外，我们还用“商业分析（BA）”代替“商务智能（BI）”。第二版中的这一更新反映了术语的变化：BI主要指报告和数据可视化（现在正在发生的事情），而BA还包括先进的分析方法，其中包括预测分析和数据挖掘。在这新版本中，我们更新了这些目前普遍使用的术语。

本书新增了一章——组合方法：集成算法和增量模型（第13章）。这一章在第四部分——“预测和分类方法”的最后介绍了这两种重要的方法。首先，集成是用以提高预测性能的多种模型的组合。在实际应用和数据挖掘竞赛中，集成算法通常被证明是非常有效的。其次，增量模型是一种改进方法，用于测量干预或处理影响。与其他章节类似，这一新章节包括现实世界的例子和章节末尾的练习。

其他变化包括根据实际数据添加了两个新案例（一个关于政治倾向说服问题和提升建模，另一个关于出租车订单取消），并删除“关联规则”这一章（关联规则是JMP Pro 12中不可用的功能，但将在JMP Pro 13中成为新功能）。

自从第二版问世，课程使用的教材已经得到大大扩展：最初，这本书主要用于MBA选修课程，现在用于各种商业分析和学位课程，以及本科生、研究生和高管教育项目。

---

① JMP Pro®, Version 12. SAS Institute Inc., Cary, NC 27513。关于如何获取JMP Pro®信息，请参阅第1章。

② JMP Pro 13中的相关新特性在多个章节中被提及。

这些课程开设的课时和覆盖的范围有所不同。在多数情况下，我们的书适用于多门课程。本书的设计目的是继续用于通用的预测分析或数据挖掘课程，同时也可以在一系列专门的商业分析项目课程中使用。

一般来说，“商业分析”“预测分析”和数据挖掘课程常见于MBA和本科专业选修课中，包括书中前三部分的全部内容和第四、五部分中的部分内容。教师可以选择用案例作为团队作业、课堂讨论对象或项目。对于两个学期的课程，可以将第六部分考虑在内。在一系列专门的商业分析课程中，可以使用本书第二版。

预测性分析：监督学习 在专门的商业分析项目中，预测性分析的主题通常是基于很多课程的。第一个课程将涵盖第一到第四部分，并且教师会根据课程的长度选择部分内容进行讲授。我们建议在这样的课程中增加第 13 章的内容。

预测性分析: 无监督学习 本课程介绍数据探索和可视化、降维、挖掘关系和聚类（第三和五部分）。如果本课程遵循预测性分析：监督学习课程，探索那些继承了有监督和无监督学习的例子和方法是极其有用的。

预测分析：关于时间序列预测的专门课程将在第六部分讲授。

在所有课程中，我们强烈建议组建一个项目，其中的数据要么根据学生的兴趣收集，要么由老师提供（例如从许多数据挖掘竞赛中获取数据集）。根据我们以及其他教师的经验，这些项目能够增强学生的学习效果，并且可以为他们提供一个很好的机会去了解数据挖掘的优势，以及处理数据和解决实际问题所面临的挑战。

致　谢

笔者感谢许多帮助我们改进第一版并在第二版中帮助进一步改进的人，在现在这个JMP版本中，Anthony Babinec为我们提供了详细和专业的纠正，他在Statistics.com上的数据挖掘课程中一直在使用这本书的书稿。Statistics.com网站团队同样也提供了有价值的校对、纠正和批判，包括教学操作主管Kuber Deokar，助理教师Shweta Jadhav和Dhanashree Vishwasrao。我们也要感谢在Statistics.com上使用本书早期版本并给出评价的学生们。

同样，Dan Toy和John Elder IV对我们的项目充满热情，并就早期版本提出了详细且有用的意见。Boaz Shmueli和Raquelle Azran在前两版中给出了详细的编辑意见和建议；Noa Shmueli对新版本进行了编辑；Bruce McCullough和Adam Hughes在第一版中做了同样的工作。Ravi Bapna在印度商学院的数据挖掘课程中使用了本书的早期版本，并给出了宝贵的建议和意见。还有一些有用的评价和反馈来自于许多在课堂上使用过这本书的老师。

马里兰大学史密斯商学院的Shrivardhan Lele，Wolfgang Jank和Paul Zantek提供了实用的建议和意见。我们感谢Robert Windle、MBA学生Timothy Roach，Pablo Macouzet和Nathan Birckhead所提供的宝贵的数据集。我们也感谢MBA学生Rob Whitener和Daniel Curtis提供的热力图和地图图表。我们感谢许多MBA学生进行了富有成效的讨论和有趣的数据挖掘项目，这帮助塑造并完善了本书。

如果没有麻省理工学院斯隆管理学院教师的支持，本书不会有今天的成果。我们特别感谢Dimitris Bertsimas，James Orlin，Robert Freund，Roy Welsch，Gordon Kaufmann和Gabriel Bitran。作为斯隆数据挖掘课程的教学助理，Adam Mersereau对本书笔记和案例作了详细的评论，Romy Shioda帮助编写了本书所使用的案例和练习，Mahesh Kumar提供了聚类分析部分的相关资料。我们非常感谢斯隆商学院的MBA学生，因为他们在课堂上的讨论，促成了这些笔记的完善。

Chris Albright，Gregory Piatetsky-Shapiro，Wayne Winston和Uday Karmarkar提供给我们有用的建议。Anand Bodapati提供了数据和建议，Suresh Ankolekar和Mayank Shah

帮助完善了几个案例，并提供了宝贵的教学评论，Vinni Bhandari 帮助撰写了 Charles Book Club 案例。

我们要感谢 Marvin Zelen，哈佛大学的 L.J.Wei 和 Cyrus Mehta 以及普纳大学的 Anil Gore，他们就统计学和数据挖掘之间的关系进行了发人深省的讨论。感谢麻省理工学院工程系统部的 Richard Larson，关于在复杂系统建模中数据挖掘的作用，他提出了许多激动人心的想法。他们帮助我们在新兴的数据挖掘领域建立一个平衡的哲学观点。

我们感谢 Ajay Sathe 和他的 Cytel 同事帮助启动这个项目，他们是 Suresh Ankolekar，Poonam Baviskar，Kuber Deokar，Rupali Desai，Yogesh Gajjar，Ajit Ghanekar，Ayan Khare，Bharat Lande，Dipankar Mukhopadhyay，SVSabnis，Usha Sathe，Anurag Srivastava，V. Subramaniam，Ramesh Raman 和 Sanhita Yeolkar。

Wiley 的 Steve Quigley 从一开始就表现出对本书的信心，帮助本书以极快的速度顺利出版，同时由于 Curt Hinrichs 的鼓励和支持使 JMP Pro 版本成为可能。Wiley 的 Jon Gurstelle，Allison McGinniss，Sari Friedman 和 Katrina Maceda，以及 Thomson Digital 的 Shikha Pahuja，在我们完成新版本的 JMP Pro 时也给出了帮助和响应。

我们还要感谢马里兰大学人类 - 计算机交互实验室的 Catherine Plaisant，他给出了数据可视化章节的练习和插图；在时间序列章节给出评论和想法的得克萨斯农工大学的 Marietta Tretter；对数据可视化章节和总体设计提供反馈和建议的 Stephen Few 和 Ben Shneiderman。

KDNuggets.com 的创始人 Gregory Piatetsky-Shapiro 多年以来花费大量的时间对该项目提供宝贵的建议。Ken Strasma 是微目标公司 HaystaqDNA 的创始人，同样也是 2004 年克里竞选和 2008 年奥巴马竞选的主管，他为增量模型部分提供了场景和数据。

最后，我们要感谢审稿人对第一个 JMP Pro® 版本的反馈和建议，以及 JMP 文档教育和开发团队的成员，感谢他们的支持和耐心，以及对我们所提问题和请求的回应。还要感谢 L. Allison Jones-Farmer，Maria Weese，Ian Cox，Di Michelson，Marie Gaudard，Curt Hinrichs，Rob Carver，Jim Grayson，Brady Brady，Jian Cao，Chris Gotwalt 和 Fang Chen。最重要的是感谢 John Sall，他的创新、灵感和无私的奉献，为友好的桌面统计软件 JMP 以及本书的出版提供了可能。

译者简介

阮敬，博士，教授。现任首都经济贸易大学研究生院副院长，兼任中国现场统计研究会经济与金融统计分会副理事长、中国统计教育学会常务理事兼高等教育分会副秘书长、中国商业统计学会常务理事、全国工业统计学教学研究会理事、北京大数据协会副秘书长等职。近年来主持国家级和省部级科研项目以及国家机关和企事业单位横向课题 40 余项；在国内外公开发表论文 50 余篇，出版专著教材 7 部，先后 11 次荣获全国统计科学研究优秀成果奖、北京市哲学社会科学优秀成果奖、北京市统计科学研究优秀成果奖等奖励。

严雪林，现任 SAS 公司 JMP 亚太区总经理，北京大数据协会副会长，兼任首都经济贸易大学统计学院兼职教授、全国应用统计专业学位研究生教育指导委员会“大数据硕士”兼职研究生导师、上海财经大学统计与管理学院兼职研究生导师。曾任 JMP 大中华区总经理、SAS 大中华区市场及渠道总监、上海大学管理学院兼职教授、中央财经大学统计学院兼职教授等职。长期于全球信息技术及数据分析类领军企业担任高级管理职位，在数据分析、商业智能、运营、市场营销、战略、信用及风险、大数据等领域有 19 年以上的研究与实践。曾领导多个国际团队，为财富 500 强企业及中国本土行业领导者提供数据分析战略、市场营销、研发优化、运营、风险等领域的服务，成功帮助客户导入大数据与分析战略，建立起基于数据分析的科学决策与运营体系，实现转型，推动变革与创新。

周暐，硕士。现任赛仕软件（北京）有限公司 JMP 事业部咨询经理，兼任中国质量协会六西格玛专家委员会专家委员、上海财经大学统计与管理学院应用统计硕士研究生导师（兼职），曾在西门子（Siemens）通信运营部门工作多年。精通统计学的理念与实践，擅长探索性数据分析 EDA、实验设计 DOE、建模预测 Modeling、生存与可靠性分析 Reliability、统计过程控制 SPC、市场调查 Marketing Research 等统计分析方法在实际工作中的应用。2007 年年底编著完成《六西格玛管理统计指南》（中国质量协会推荐的专业书籍），积累了统计技术在半导体、电子、医药、汽车、化工、银行、教育等行业应用的宝贵经验，曾受邀为多家国内外著名企业和高等学府辅导统计应用及六西格玛专业知识，并获得广泛好评。

目　录

第一部分　预备知识

第二部分　数据探索与降维

第三部分　性能评估

第四部分　预测与分类方法

第五部分　挖掘记录之间的关系

第六部分　时间序列预测

第七部分 案 例

第一部分

预备知识