“十二五”
国家重点图书出版规划项目

sas WILEY

新信息时代商业经济与管理译丛

# 大数据分析

## 数据科学应用场景与实践精髓

# ANALYTICS IN A BIG DATA WORLD

## THE ESSENTIAL GUIDE TO DATA SCIENCE AND ITS APPLICATIONS

【英】Bart Baesens◎著
柯晓燕 张纪元◎译　邓煜熙◎审校

中国工信出版集团
人民邮电出版社
POSTS & TELECOM PRESS

“十二五”国家重点图书出版规划项目
新信息时代商业经济与管理译丛

# 大数据分析：
## 数据科学应用场景与实践精髓

[英] Bart Baesens 著
柯晓燕　张纪元 译
邓煜熙 审校

人 民 邮 电 出 版 社
北 京

**图书在版编目（CIP）数据**

大数据分析 ：数据科学应用场景与实践精髓 / （英）贝森斯（Baesens, B.）著 ；柯晓燕，张纪元译. -- 北京：人民邮电出版社，2016.1
（新信息时代商业经济与管理译丛）
ISBN 978-7-115-40745-0

Ⅰ. ①大… Ⅱ. ①贝… ②柯… ③张… Ⅲ. ①商业信息－数据处理 Ⅳ. ①F715.51

中国版本图书馆CIP数据核字(2015)第246023号

## 版权声明



◆ 著　　[英] Bart Baesens
译　　柯晓燕　张纪元
审　　校　邓煜熙
责任编辑　刘　洋
责任印制　彭志环
◆ 人民邮电出版社出版发行　　北京市丰台区成寿寺路 11 号
邮编　100164　　电子邮件　315@ptpress.com.cn
网址　http://www.ptpress.com.cn
北京隆昌伟业印刷有限公司印刷
◆ 开本：700×1000　1/16
印张：15.75　　2016 年 1 月第 1 版
字数：231 千字　　2016 年 1 月北京第 1 次印刷
著作权合同登记号　图字：01-2014-5083 号

定价：59.00 元

**读者服务热线：(010)81055488　印装质量热线：(010)81055316**
**反盗版热线：(010)81055315**

# 内容提要

本书是一本讨论大数据理论及应用实践的专著，从讨论理论界的前沿观点开始，之后转向讨论这些理论在日常商业活动中的实践应用。

本书首先介绍了大数据分析的业务应用场景、分析建模过程和主要任务，以及模型商用的关键点；接着讲述了数据收集、抽样和预处理的实施要点；之后系统性地讨论了各种模型技术及其应用，包括预测分析、描述分析、生存分析、社交网络分析等。在完成了这些理论知识和模型技术方法铺垫之后，就进入到实践应用部分，包括把分析活动转化为生产力的关键事项，以及各种应用实例。

本书帮助读者系统地梳理了各类模型方法的技术要点和应用要点，包括线性回归、Logistic 回归、决策树、聚类、关联规则、序列规则、神经网络、支持向量机、套袋算法、Boosting 算法、随机森林算法、生存分析等；本书还介绍了大量的应用实例，如信用风险建模、欺诈检测、营销响应提升模型、客户流失预测、自动推荐、网页分析、社交媒体分析，以及业务流程分析等。因此，对于从事大数据分析相关工作的人士来说，本书是一本难得的实务指南；对于高等院校相关专业的师生来说，本书是一本非常好的课外阅读材料，特别是书中关于如何把分析变成生产力的章节部分，相信一定能给他们很多的启发和思考。

# 献词

献给我的爱妻凯特（Katrien）和孩子们，他们是安—索菲（Ann-Sophie）、维克多（Victor）和汉娜萝拉（Hannelore），还有我的父亲、母亲、岳父和岳母。

# 作者简介

巴特·贝森斯（Bart Baesens）是比利时鲁汶大学的副教授，英国南安普敦大学的讲师。他承担过很多客户关系管理、网络分析、欺诈检测以及信用风险管理等领域的分析项目，积累了丰富的研究分析及应用实践经验（相关介绍可查阅 www.dataminingapps.com）。他的研究发现，已经见诸于多本世界知名期刊（如《Machine Learning》《Management Science》《IEEE Transactions on Neural Networks》《IEEE Transactions on Knowledge and Data Engineering》《IEEE Transactions on Evolutionary Computation》《Journal of Machine Learning Research》等），并多次出席顶级的国际研讨会，发表主题演讲。他还是《信用风险管理精要》（牛津大学出版社，2008 年出版）一书的作者。他经常为国际知名企业提供咨询服务，就商业分析、信用风险管理等业务管理问题，提供策略建议以及项目实施辅导等专业服务。

# 致谢

在这里，我要深深地感谢为本书做出贡献的所有同事，他们是赛普·万登·布鲁克（Seppe vanden Broucke）、亚历克斯·塞雷特（Alex Seret）、托马斯·佛贝瑞肯（Thomas Verbraken）、艾米·拜克尔（Aimée Backiel）、佛欧尼克·范·佛拉瑟乐儿（Véronique Van Vlasselaer）、海伦·莫盖思（Helen Moges）以及芭芭拉·德根特（Barbara Dergent）。

# 中文版序

在移动互联网时代，社交网络成为推动移动互联网迅猛发展的生力军。互联网花了 30 年时间达到 7.5 亿用户，成立于 2004 年的 Facebook 只花了 8 年时间便达到与之不相上下的用户数。

社交网络的核心价值在于人和人的社交关系，马克·扎克伯格说："人们分享得越多，他们就能够通过自己信赖的人，获得更多有关产品和服务的信息。他们能够更加轻松地找到最佳产品，并提高生活品质和效率。在这一过程中，企业获得的益处是，他们能够制造更好的产品，即以人为本的个性化产品。与传统商品相比，那些基于社交关系、社交图谱、社交圈推广的产品更富有吸引力。"可见，社交网络为人们开拓了新的信息分享和交流空间，也为企业创造了利用社交关系更开阔、更深入、更高效地开展客户销售、服务和营销的机会。对于企业来说，谁更早抓住机会研究了解自身的客户社交网络关系，谁就更具核心市场竞争力。

博雅公关 Burson-Marsteller 和互联网监测分析公司 Visible 联合发布的 2012 年度财富 100 强公司社会化媒体使用报告显示，2010 ~ 2012 年，100 家公司平均拥有 Twitter 账号分别为 4.2 个、5.8 个和 10.1 个，Facebook 账号分别为 2.1 个、4.2 个和 10.4 个，YouTube 账号分别为 1.6 个、2.7 个和 8.1 个。而根据 LinkedIn 与市场研究公司 TNS 于 2014 年 2 月发布的合作研究成果，在美国中小型企业中，81%的被调查者使用社交媒体促进业务增长，94% 将社交媒体作为营销工具，而 49% 为了教育目的使用社

交媒体，并获取业务洞察力。可见，确实如制定企业社会化媒体实践“黄金标准”、著有《营销和公共关系的新规则》一书的营销专家大卫•米尔曼•斯科特（David Meerman Scott）所言：“我们正在经历一场人类沟通方式的变革。我认为这是自印刷机发明以来人类沟通方式最显著的革命……社会化媒体已经在革命性地改变商业沟通。”

我们知道，这是移动互联网时代，这是社交网络时代，而同时，人们的数字化生存让有关人们生活甚至工作的行为信息都数字化，而这些以单个个体为对象的形形色色、包罗万象、细致入微、支撑个体兴趣需求和喜好的数字化信息构成了大数据，所以，这个时代更是一个大数据时代：到今天，世界上所有印刷材料的数据量是 200PB，全人类说过所有对话的数据量大约是 5EB；每天我们产生的数据大约是 2.5PB，这就意味着当今世界全部数据的 90%都在近两年产生。

如果我们有相应的 IT 技术、分析手段驾驭大数据，大数据就是金矿；如果没有相应的技术和手段，大数据就将成为淹没我们的海洋。谈论大数据在整个社会确实已成为一种时髦，但是根据麦肯锡在 2012 年 4 月的调查，仅有 1/5 的受访者所在公司已经在一个业务单元或职能部门完全部署大数据和分析，以获得客户洞察；仅有 13% 的受访者表示，公司全面使用数据获得洞见。可见，大数据要从谈论和研究到技术和应用实现，路途还很漫长，所以，如何客观审慎地对待已有的大数据优势，提前思考并规划、架构、完善、部署数据从采集、清洗、存储、分析、应用以及管理监控的全企业层面的 BI（商业智能）平台，并培养贯穿企业运营管理流程的 BA（商业分析）体系，用数据说话，实现全企业层面的精确管理和精确营销、销售、服务，也就是大数据时代我们最终能够成为时代弄潮儿抑或被潮水淹没者的“To be or not to be”的关键问题。

中国电信股份有限公司广州研究院市场运营研究所，长年从事电信企业运营管理及市场研究的实践和方法总结，研究时间最长的已达 17 年，并分别在行业竞争、商业模式创新、精确营销、品牌、舆情、口碑营销、数据分析及挖掘、数据仓库/BI 架构及规范等细分领域长年支撑企业运营管理实践，不仅对企业运营有深刻理解和独到见解，且基于企业运营

管理实践完成了大量方法创新和应用研究，发表了多本论著和数百篇专业论文，为各细分专业领域积累了众多的方法、经验和模型。

近几年，随着移动互联网—社交网络—大数据的迅猛发展，也因为企业转型的需要，市场运营研究所在邓煜熙所长带领下，研究人员围绕两大问题开展相关研究：（1）企业如何建立自己的社交媒体策略并进行社交网络分析；（2）为实现精确管理、精确营销、销售和服务，企业如何架构 BI 平台和 BA 体系。部门集中有关资源，有计划、有步骤、层层推进地深入开展研究，完成相关科研项目和撰写论文若干。

接下来，围绕客户关系管理、客户体验管理大体系，以支撑企业生产运营管理流程各环节运作，我们预计对企业大数据体系架构和分析、应用等方面进行深入研究。

最后，借狄更斯的话，“这是最好的时代，也是最坏的时代；这是智慧的年代，也是愚蠢的年代；这是信仰的时期，也是怀疑的时期；这是光明的季节，也是黑暗的季节；这是希望的春天，也是失望的冬天；大伙儿面前应有尽有，大伙儿面前一无所有”，让大伙儿一起，掌握商业智能、商业分析两大工具，驾驭社交媒体，洞察社交网络，弄潮大数据。

**中国电信股份有限公司广州研究院院长　蔡康**

**2015 年 10 月于广州**

# 前言

对很多公司来说，分析潜能尚未开发，单是采集来自多个渠道运营环境的海量数据，就像是在一场数据洪水、海啸中挣扎，更谈不上理解和管理动态的、复杂的客户行为，并从战略高度加以挖掘、利用了。在本书中，我们将讨论如何利用分析技术创建战略优势并发现新的商业机会。

本书的重点不是数学基础知识或理论，而是聚焦实践应用。公式和方程当然也会在本书中出现，但仅限于分析/建模人员必须掌握的部分。本书的目的，不是详尽地给出所有的已开发实现的分析模型技术，而是从帮助企业提升效益出发，选择性地介绍一些常用分析模型技术。

本书为业务专家而写，是分析技术及实践应用的浓缩版，读者应掌握的基本知识包括：描述性统计的基本概念和计算方法，如算术/几何平均值、标准偏差、相关性、置信区间、假设检验等；数据处理工具，如微软的Excel、SQL语言等；数据可视化的应用，如柱状图、饼图、直方图、散点图等图形的适用场景。本书给出了丰富的图表及实例，许多案例来自真实的商业应用，如风险管理、客户关系管理、欺诈侦测、网络分析等。在每一章，作者都采用理论研究成果和咨询服务经验相结合的写作方法，以帮助读者更好地理解和应用，真正做到学以致用。本书的目标读者是高级数据分析师、咨询顾问，企业内部相关的业务人员，以及高等院校的学者和研究生。

第 1 章为大数据与分析应用概述。从一些应用领域的实践案例切入，随后给出了分析建模的过程和工作任务概览，最后讨论并总结了分析模型中的某些关键点。

第 2 章则与数据相关，介绍了数据收集、抽样和预处理的全流程。数据是分析活动的最重要的原材料，因此本章的重要性毋庸置疑。本章首先讨论了抽样技术、数据类型、数据可视化技术和探索性描述统计分析，然后讨论了缺失值、异常值的检测和处理，数据的标准化处理、分类、权重设置等，最后介绍了变量筛选和细分技术。

第 3 章讲述预测分析及其应用。本章从预测目标定义开始，然后讨论了各种预测分析技术，包括线性回归、Logistic 回归、决策树和神经网络，以及支持向量机方法*和集成算法（套袋算法、Boosting 算法、随机森林算法†）。除此之外，本章还介绍了各种多类分类技术（Multiclass Classification Techniques），阐述了 Logistic 回归、决策树、神经网络和支持向量机技术在多类分类中的应用。本章的最后讨论了预测模型性能评估方法。

第 4 章是描述性分析，主要讲述了关联规则、序列规则，以及它们在发现客户行为模式的内在关系中的应用。当然，本章还介绍了细分技术及其应用。

第 5 章介绍生存分析#。首先介绍了主要的生存分析度量指标；然后介绍了卡普兰·梅尔（KM）分析算法、参数生存分析算法，以及比例风险回

---

* 支持向量机（Support Vector Machine，SVM）是 Corinna Cortes 和 Vapnik 等于 1995 年首先提出的，它在解决小样本、非线性及高维模式识别中表现出许多特有的优势，并能够推广应用到函数拟合等其他机器学习问题中。

† 套袋（Bagging）算法、Boosting 算法都是用来提高学习算法准确度的方法，通过构造一个预测函数系列，然后以一定的方式将它们组合成一个预测函数。

随机森林（Random Forests）：在机器学习中，随机森林是一个包含多个决策树的分类器，且其输出的类别是由个别树输出的类别的众数而定。这个术语是根据 1995 年由贝尔实验室的 Tin Kam Ho 所提出的随机决策森林（Random Decision Forests）而来的。

# 生存分析（Survival Analysis）是指根据试验或调查得到的数据，对生物或人的生存时间进行分析和推断，研究生存时间和结局与众多影响因素间的关系及其程度大小的方法，也称生存率分析或存活率分析。

归算法；最后对生存分析模型的延伸应用及模型性能评估方法进行了讨论和总结。

第 6 章的内容是社交网络分析。从社交网络的应用实例、社交网络的定义和度量方法入手，首先介绍了社交网络学习，接着讲述了与社交网络分析相关的算法（如关系近邻分类器及其变异概率等），然后介绍了 Logistic 回归技术在社交网络分析中的应用。本章的最后讨论了图数据处理技术（Egonets）和偶图技术。

第 7 章侧重于实践应用，给出了把分析活动转化为生产力的关键事项。本章从讲述把分析模型投入业务应用的总体要求开始，然后讨论了后验测试、参照标准管理、数据质量、软件工具、隐私保护、模型设计、文档管理以及公司治理等。

第 8 章全面总结了本书给出的各种业务应用实例，如信用风险建模、欺诈检测、营销响应提升模型、客户流失预测、自动推荐、网页分析、社交化媒体分析，以及业务流程分析等。

# 目录


1 **第 1 章 大数据及其分析**

1.1 大数据的业务应用场景

1.2 基本的专业术语

1.3 分析过程模型

1.4 分析建模活动中的任务及角色

1.5 分析技术

1.6 分析模型的要求

1.7 本章参考文献

13 **第 2 章 数据采集、抽样和预处理**

2.1 数据源的类型

2.2 数据抽样

2.3 数据类型

2.4 数据可视化及探索性统计分析

2.5 缺失值的处理

2.6 异常值检测及处理

2.7 数据标准化

2.8 粗分类（Categorization）处理

2.9 WOE 值的计算

2.10 变量的选择

2.11 细分

2.12 本章参考文献

35 **第 3 章 预测分析**

3.1 定义目标变量

3.2 线性回归

3.3 Logistic 回归

3.4 决策树

3.5 神经网络

3.6 支持向量机

3.7 集成算法

3.7.1 套袋算法（Bagging）

3.7.2 Boosting 方法

3.7.3 随机森林

3.8 多类分类技术

3.8.1 多类 Logistic 回归

3.8.2 多类决策树

3.8.3 多类神经网络

3.8.4 多类支持向量机

3.9 预测模型的评估

3.9.1 数据集的分割

3.9.2 分类模型的性能评估

3.9.3 回归模型的性能评估

3.10 本章参考文献

89 **第 4 章 描述性分析**

4.1 关联规则

4.1.1 基本概念及假设

4.1.2 支持度和置信度

4.1.3 关联规则的挖掘

4.1.4 提升度的度量

4.1.5 关联规则的后处理

4.1.6 关联规则的扩展

4.1.7 关联规则的应用

4.2 序列规则

4.3 细分技术
4.3.1 分层聚类
4.3.2 K-Means 聚类
4.3.3 自组织映射图（SOM）
4.3.4 聚类解决方案的应用及解释
4.4 本章参考文献
107 **第 5 章 生存分析**
5.1 生存分析的基本概念和函数
5.2 卡普兰·梅尔分析
5.3 参数法生存分析
5.4 比例风险回归模型
5.5 生存分析模型的扩展
5.6 生存分析模型的评估
5.7 本章参考文献
123 **第 6 章 社交网络分析**
6.1 社交网络的定义
6.2 社交网络的度量
6.3 社交网络学习
6.4 关系近邻分类器
6.5 概率关系近邻分类器
6.6 关系逻辑回归
6.7 共同模式推断
6.8 自中心网络（EGO NETS）
6.9 偶图/二分图
6.10 本章参考文献
137 **第 7 章 从分析到生产力**
7.1 模型的后验测试
7.1.1 分类模型的后验测试
7.1.2 回归模型的后验测试
7.1.3 聚类模型的后验测试
7.1.4 设计后验测试方案

7.2 参照管理
7.3 数据质量
7.4 软件工具
7.5 隐私保护
7.6 模型设计相关文档
7.7 公司治理
7.8 本章参考文献
167 **第 8 章 实践与案例**
8.1 信用风险建模
8.2 欺诈检测
8.3 净响应提升建模
8.4 流失预测
8.4.1 流失预测模型
8.4.2 流失预测流程
8.5 推荐系统
8.5.1 协同过滤推荐
8.5.2 基于内容的推荐
8.5.3 基于人口统计信息的推荐
8.5.4 基于知识的推荐
8.5.5 组合推荐
8.5.6 推荐系统的评价
8.5.7 案例介绍
8.6 网页分析
8.6.1 网页数据收集
8.6.2 Web KPI 指标
8.6.3 从 Web KPI 到行动洞察力
8.6.4 导航分析
8.6.5 搜索引擎营销分析
8.6.6 A/B 测试和多变量测试
8.7 社会化媒体分析
8.7.1 社交网站：B2B 广告工具