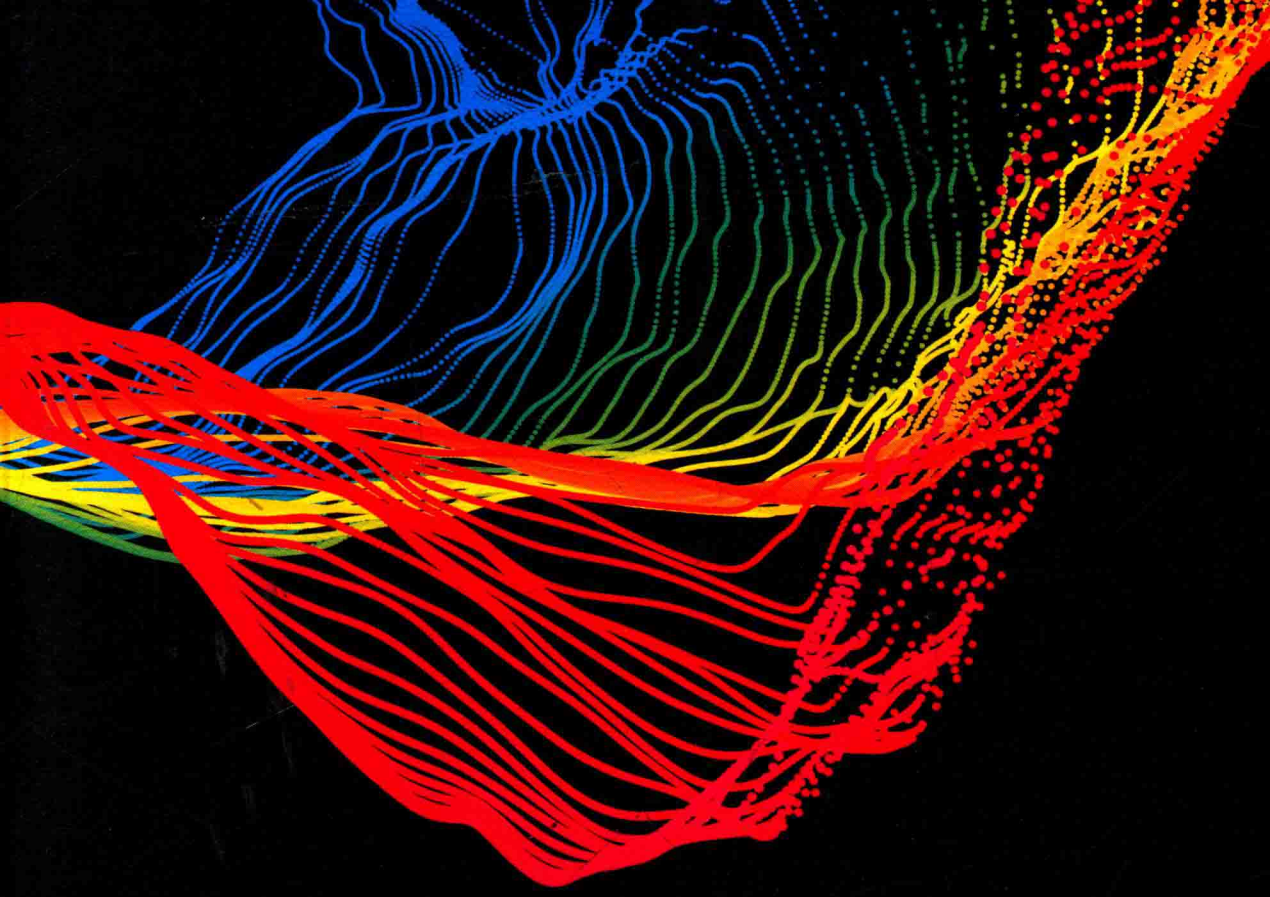

大数据、数据挖掘与智慧运营

梁栋 张兆静 彭木根◎编著

BIG DATA MINING AND INTELLIGENT OPERATION

清華大學出版社

BIG DATA MINING
AND INTELLIGENT OPERATION

大数据、数据挖掘与智慧运营

梁栋 张兆静 彭木根◎编著

清華大學出版社
北京

内 容 简 介

本书系统地介绍了大数据挖掘的基本概念、经典挖掘算法、挖掘工具和企业智慧运营应用案例。

全书分为9章，内容包括：大数据挖掘与智慧运营的概念，数据预处理，数据挖掘中的四种主流算法：聚类分析、分类分析、回归分析、关联分析，增强型数据挖掘算法，数据挖掘在运营商智慧运营中的应用案例，未来大数据挖掘的发展趋势等。

全书以运用大数据挖掘方法提升企业运营业绩与效率为主线，从运营商实际工作中选取了大量运营和销售案例，详细讲述了数据采集、挖掘建模、模型落地与精准营销的全部过程。书中大部分案例的代码、软件操作流程和微课视频可以通过扫描本书封底的二维码下载。

本书主要面向运营商及其他高科技企业员工、高等院校相关专业本科生和研究生，以及其他对数据挖掘与精准营销感兴趣的读者。

图书在版编目(CIP)数据

大数据、数据挖掘与智慧运营 / 梁栋，张兆静，彭木根编著. — 北京：清华大学出版社，2017
ISBN 978-7-302-48337-3

Ⅰ. ①大… Ⅱ. ①梁… ②张… ③彭… Ⅲ. ①数据处理 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字（2017）第 218392 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： 王荣静
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 刷 者： 北京富博印刷有限公司
装 订 者： 北京市密云县京文制本装订厂
经 销： 全国新华书店
开 本： 187mm×235mm **印 张：** 26.25 **字 数：** 494 千字
版 次： 2017 年 11 月第 1 版 **印 次：** 2017 年 11 月第 1 次印刷
印 数： 1～4000
定 价： 99.00 元

产品编号：075552-01

前 言

数据挖掘（Data Mining），是指从数据中发现知识的过程（Knowledge Discovery in Databases，KDD)。狭义的数据挖掘一般指从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中，提取隐含其中的、人们事先不知道的、但又是潜在有用知识的过程。自从计算机发明之后，科学家们先后提出了许多优秀的数据挖掘算法。2006 年 12 月，在数据挖掘领域的权威学术会议 *the IEEE International Conference on Data Mining* (*ICDM*) 上，科学家们评选出了该领域的十大经典算法：C4.5、K-Means、SVM、Apriori、EM、PageRank、AdaBoost、kNN、Naive Bayes 和 CART。这是数据挖掘学科的一个重要里程碑，从此数据挖掘在理论研究和实际应用两方面均进入飞速发展时期，并得到广泛关注。

在实际生产活动中，许多问题都可以用数据挖掘方法来建模，从而提升运营效率。例如，某企业在其移动终端应用（App）上售卖各种商品，它希望向不同的客户群体精准推送差异化的产品和服务，从而提升销售业绩。在这个案例中，如何将千万量级的客户划分为不同的客户群体，可以由数据挖掘中的聚类分析算法来完成；针对某个客户群体，如何判断某个产品是否是他们感兴趣的，可以由数据挖掘中的分类分析算法来完成；如何发现某个客户群体感兴趣的各种产品之间的关联性，应该把哪些产品打包为套餐，可以由数据挖掘中的关联分析算法来完成；如何发现某个客户群体的兴趣爱好的长期趋势，可以由数据挖掘中的回归算法来完成；如何综合考虑公司的 KPI 指标、营销政策和 App 页面限制等条件，制订最终的落地营销方案，可以基于数据挖掘中的 ROC 曲线建立数学模型求得最优解来解决。

当前，许多企业正面临前所未有的竞争压力。以运营商企业为例，从政策层面看，国家提出了“提速降费”的战略指示：一方面要提高网络连接速度、提供更好的服务，这意味着公司成本的提高；另一方面要降低资费标准，这意味着单个产品收入的下降，运营商该如何化解这对矛盾？从运营商内部数据统计看，传统的语音和短信、彩信业

务收入占比正不断下降，传统的利润点已经风光不再；流量收入目前已占据主要位置并保持上涨趋势，但单纯的流量经营又将面临“管道化”压力；未来的利润增长点要让位于被称为“第三条曲线”的数字化服务。运营商该如何经营这一新鲜事物？从外部环境看，互联网和电子商务企业借助其在各方面的优势，已经对运营商形成了巨大的压力，特别是在数字化服务营销领域，传统运营商企业已经不再具备优势，又该如何应对互联网企业的全面竞争？

随着移动互联网和物联网时代的来临，人和万事万物被广泛地联系在一起。人们在联系的过程产生了大量的数据，例如用户基础信息、网页浏览记录、历史消费记录、视频监控影像，等等。据此，以 Google 为首的互联网公司提出了“大数据”（Big Data）的概念，并声称人类已经脱离了信息时代（Information Time，IT），进入了大数据时代（Data Time，DT）。显然，海量数据包含了非常丰富的浅层次信息和深层次知识。对于同一竞争领域的企业，谁能获取最大量的数据，展开最精准的数据挖掘与建模分析，并加以精细化的落地实施，谁便能在行业竞争中取得优势。对于运营商企业而言，其具备的一个显著优势便是手握海量数据资源。如果能运用先进的数据挖掘技术找出客户的行为规律，从传统的经验式、粗放式、“一刀切”式的运营决策向数据化、精细化、个性化的运营决策转型，运营商将迎来新的腾飞。上述运营模式转型的目标，便是所谓的“智慧运营”。

目前，人类对大数据尚没有统一的、公认的定义，但几乎所有学者和企业都认同大数据具备四大特征（四大挑战）：体量巨大（Volume）、类型繁多（Variety）、价值密度低（Value）、需要实时处理（Velocity）。这其中最重要的一点是类型繁多，即过去人类的数据储备以结构化数据为主，而未来将以非结构化数据为主。回到之前提到的 App 营销案例，企业基于用户的基础信息、历史消费信息、简单的网络行为信息等结构化数据展开挖掘建模，被认为是传统的“基于数据挖掘的智慧运营”。随着时代的发展，企业还掌握了用户观看在线视频的内容数据、在营业网点接受营业员推荐的表情信息和语言交流数据、用户在客服热线中的语音咨询数据等。这些数据被统称为非结构化数据，随着语音识别、人脸识别、语义识别等新技术的发展成熟，对非结构化数据的分析挖掘已成为可能，并将获得广阔的商业应用空间。基于非结构化数据的挖掘建模又被称为“基于人工智能的智慧运营”。考虑当前大部分企业的实际运营现状，本书将主要围绕“基于数据挖掘的智慧运营”展开讨论，“基于人工智能的智慧运营”将在后续书籍中展开讨论。

本书共分为九章：第 1 章大数据、数据挖掘与智慧运营综述，讲述数据挖掘的基本概念和发展史、大数据的时代特征、当前结构化数据挖掘进展、非结构化数据挖掘

与人工技能进展、数据挖掘的主流软件等；第 2 章数据统计与数据预处理，讲述在数据挖掘之前的数据集成、数据清洗、数据衍生、数据统计等；第 3 章聚类分析，重点讲述 K-means、BIRCH、DBSCAN、CLIQUE 等几种主流经典聚类算法；第 4 章分类分析，重点讲述决策树、KNN、贝叶斯、神经网络、SVM 等几种主流分类算法；第 5 章回归分析，重点讲述线性回归、非线性回归、逻辑回归等几种主流回归算法；第 6 章关联分析，重点讲述 Apriori、FP-tree 等几种主流关联算法；第 7 章增强型数据挖掘算法，重点讲述随机森林、Bagging、Boosting 等几种主流增强算法；第 8 章数据挖掘在运营商智慧运营中的应用，展开讲述数据挖掘方法在外呼营销、精准推送、套餐适配、客户保有、投诉预警、网络质量监控、室内定位中的应用；第 9 章面向未来大数据的数据挖掘与机器学习发展趋势，简要讲述数据挖掘领域的前沿研究进展。

全书以运用大数据挖掘方法提升企业运营业绩与效率为主线。第 3 章至第 7 章组成本书的理论知识部分，在讲述理论知识的同时，这部分每章都配套列举了大量实际应用案例，及其在 SPSS 等分析软件中的具体操作流程。此外，第 8 章从运营商实际工作中选取了大量运营和销售案例，详细讲述了数据采集、挖掘建模、模型落地与精准营销的全部过程。书中大部分案例的代码、软件操作流程和微课视频可以通过扫描本书封底的二维码下载。

本书基于作者所带领的研究团队多年研究积累和在运营商企业广泛落地应用的基础上提炼而成。全书由曾丽丽博士组织并统稿，梁栋、张兆静和彭木根撰写了主要章节，研究团队中的谢花花、柯联兴、张笑凯、鲁晨、李子凡等在读研究生参与了部分章节的写作，胡林、唐糖等团队外专家参与了部分章节的写作并给出了宝贵的意见。在本书写作过程中，中国移动及许多省市分公司（特别是广西分公司）给予了大力支持。在本书出版前，许多素材被中国移动广西分公司选为教材并展开了广泛落地应用，获得了 2016 年中国移动集团公司颁发的“培训案例最佳实践奖”。在本书出版过程中，得到了深圳市傲举企业管理顾问有限公司的大力支持。在此对有关人员一并表示诚挚的感谢！

由于作者能力所限，疏漏之处在所难免，希望各位读者海涵，并批评指正。

作 者

2017 年 9 月于北京邮电大学

目 录

第1章

大数据、数据挖掘与智慧运营综述

Big Data, Data Mining
And Intelligent Operation

近年来，大数据、数据挖掘、机器学习、云计算和人工智能等词语日渐为人们所熟悉。本章将围绕上述基本概念和话题展开讨论。本章 1.1 节介绍数据挖掘的概念和发展史，1.2 节介绍数据挖掘的主要流程和金字塔模型，1.3 节介绍数据挖掘对企业智慧运营的重要意义，1.4 节介绍大数据的基本概念、特征和挑战，1.5 节介绍非结构化数据挖掘的概念和研究进展，1.6 节介绍结构化数据挖掘与机器学习、深度学习和人工智能之间的关联关系，1.7 节介绍常见的数据挖掘分析软件与系统。

1.1 数据挖掘的发展史

1.1.1 数据挖掘的定义与起源

什么是数据挖掘，数据挖掘包括哪些范畴？迄今为止不同的学者和公司仍有着不同的理解和定义。例如有的学者认为：数据挖掘即指摆脱传统的经验式、规律式的分析方法，转变为纯粹从数据出发来探索问题的本质。又例如有的公司认为：数据挖掘是一种从数据中榨取价值，提升公司运营效率的重要手段。然而，绝大部分学者和公司都认同数据挖掘的最基本定义：从数据中获取知识。

数据挖掘具体起源于什么年代现在已无从考证。自从有了数据，人类就开始尝试对数据进行分析。随着时代的发展，特别是计算机技术的诞生和发展，人类拥有的数据越来越多，种类越来越复杂，之前传统的浅层次的、以经验式、观察式为主的数据分析方法已不再适用，人类急需一整套深层次的、科学的数据分析方法，这些方法的总和被称为“数据挖掘”。

随着移动互联网时代的来临，我们每天都生活在数据中，时时刻刻都接触着来自生活各个方面的各种数据：早高峰各个十字路口的车流量，各个公司的股市行情、销售票务、产品描述、用户反馈，科学实验记录着的种种信息……数据的产生无时不在，无处不在。爆炸式增长、广泛可用的巨量数据急需功能强大和通用的工具，以便发现它们潜在的巨大价值。交警部门需要通过对车流量数据的观察来决定警力支配；公司需要通过对方方面面商业数据的分析来制订合理的发展计划；科学研究工作者需要对来自实验的种种数据研究来实现实验目的……人们越来越关注如何把海量的数据变为直观、有用的信息。人类的需求是发明之母，人们对数据所蕴含的潜在知识的需求

促使了数据挖掘的诞生。

近年来，数据挖掘引起了信息产业界的极大关注，其主要原因是存在大量数据可以广泛使用，并且迫切需要将这些数据转换成有用的信息和知识。获取的信息和知识可以被广泛用于各种应用，包括商务管理、生产控制、市场分析、工程设计和科学探索等。

数据挖掘利用了来自如下领域的思想：

（1）来自统计学的抽样、估计和假设检验。

（2）人工智能、模式识别和机器学习的搜索算法、建模技术和学习理论。

数据挖掘也迅速地接纳了来自其他领域的思想，这些领域包括最优化、进化计算、信息论、信号处理、可视化和信息检索。一些其他领域也起到重要的支撑作用。特别的，需要数据库系统提供有效的存储、索引和查询处理支持。源于高性能（并行）计算的技术在处理海量数据集方面常常是重要的。分布式技术也能帮助处理海量数据，并且当数据不能集中到一起处理时更是至关重要。

1.1.2　数据挖掘的早期发展

数据挖掘起始于 20 世纪下半叶，是在多个学科发展的基础上逐步发展起来的。随着大数据与数据库技术的发展应用，数据量不断积累与膨胀，这导致基础的查询和统计操作已经无法满足企业的商业需求。如何挖掘出数据隐含的信息是当前亟须解决的难题。与此同时，计算机领域的人工智能（Artificial Intelligence）方向也取得了巨大进展，进入了机器学习的阶段。因此，人们将两者结合起来，用数据库管理系统存储数据，用计算机分析数据，并且尝试挖掘数据背后的信息。这两者的结合促生了一门新的学科，即数据库中的知识发现（Knowledge Discovery in Databases，KDD）。1989 年 8 月召开的第 11 届国际人工智能联合会议的专题讨论会上首次出现了“知识发现”这个术语，到目前为止，知识发现的重点已经从发现方法转向了实践应用。

数据挖掘（Data Mining）则是 KDD 的核心部分，它指的是从数据集合中自动抽取隐藏在数据中那些有用信息的非平凡过程，这些信息的表现形式为：规则、概念、规律及模式等。进入 21 世纪，数据挖掘已经成为一门比较成熟的交叉学科，并且数据挖掘技术也伴随着信息技术的发展日益成熟起来。总体来说，数据挖掘融合了数据库、人工智能、机器学习、统计学、高性能计算、模式识别、神经网络、数据可视化、信息检索和空间数据分析等多个领域的理论和技术，是 21 世纪初期对人类产生重大影响的十大新兴技术之一。