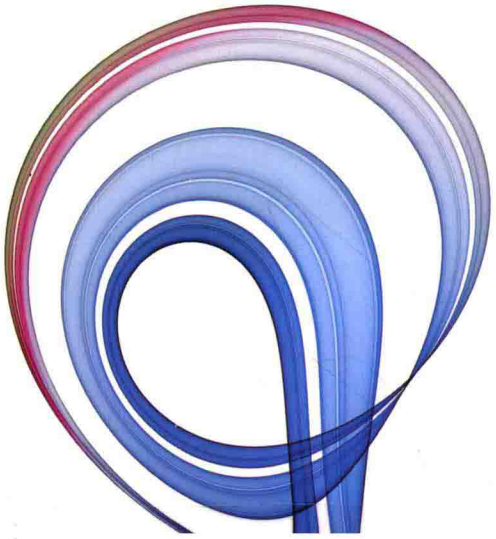
HZ BOOKS
华章IT

MATLAB官方（MathWorks）资深大数据挖掘专家撰写，MathWorks官方及多位专家鼎力推荐

从技术、方法、案例、最佳实践4个维度循序渐进地讲解了大数据挖掘技术

大数据技术丛书



Using Big Data to Build Your Business

大数据挖掘

系统方法与实例分析

周英 卓金武 卞月青◎著

机械工业出版社
China Machine Press

技术丛书

Using Big Data to Build Your Business

大数据挖掘

系统方法与实例分析

周英 卓金武 卞月青◎著

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

大数据挖掘：系统方法与实例分析 / 周英，卓金武，卞月青著，—北京：机械工业出版社，2016.4
（大数据技术丛书）

ISBN 978-7-111-53267-5

I. 大… II. ①周… ②卓… ③卞… III. 数据处理 IV. TP274

中国版本图书馆 CIP 数据核字（2016）第 057140 号

大数据挖掘：系统方法与实例分析

出版发行：机械工业出版社（北京市西城区百万庄大街 22 号 邮政编码：100037）
责任编辑：李 艺
责任校对：殷 虹
印 刷：北京文昌阁彩色印刷有限责任公司
版 次：2016 年 5 月第 1 版第 1 次印刷
开 本：186mm × 240mm 1/16
印 张：25.25
书 号：ISBN 978-7-111-53267-5
定 价：79.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
客服热线：（010）88379426 88361066
投稿热线：（010）88379604
购书热线：（010）68326294 88379649 68995259
读者信箱：hzit@hzbook.com

序　言 Preface

欣闻三位好友的新书即将出版，很荣幸能为本书撰写序言。这是一本不再仅仅是概念介绍而是实实在在介绍如何利用大数据的书籍，希望通过本书让更多的读者能够更具体地了解大数据，了解大数据的价值，并利用大数据挖掘技术来让大数据更好地服务我们的生产和生活，从而提升整个社会价值体系。

大数据是最近几年兴起的概念，虽然有被过分炒作之嫌，但我觉得是有客观原因的。因为随着信息技术的发展，各行业都已经有足够的数据积累，而且有的行业已经体验到了数据的巨大能量。国内最直接体验到大数据价值的当属 BAT（百度、阿里、腾讯），在传统行业，大数据也已经开始应用。比如，银行利用大数据进行风险管理；电力公司利用大数据进行负载预测，从而分时定价，并可以根据预测结果优化电能的储蓄和调配；矿业公司利用大数据进行精细加工，提高产品竞争力。总之，大数据已对各行业产生了十分明显的影响，无论是银行、证券、通讯、铁路、航空，还是军事、政治、工业、商业，基于大数据的决策已经成为现代社会各行业运行的基础。纵然这样，各行业对大数据的利用还处于初期阶段，如何更有效地利用这些数据已成为各行业的一个大课题！

浏览一下本书的目录，顿时令人振奋起来！概念、技术、项目、经验四位一体，层层递进，非常符合我们的阅读习惯。基础篇让大家知道大数据的基本概念、分类和挖掘流程。技术篇系统地介绍了整个大数据挖掘理论体系里的具体技术，包括数据预处理和六大类核心算法，即关联、回归、分类、聚类、预测、诊断，每类算法中又详细讲解了常用算法的原理、实现步骤、应用实例，并且每个实例都有一个 MATLAB 实现实例，对于当代的读者来说，这些实例太有价值了，可以直接借鉴、研读、修改、提升。技术学习的同时也可以深化对概念的理解，从而与基础篇的内容相得益彰。项目篇相当于大数据挖掘技术在各行业的具体应用，技术与应用融会贯通，既可启发读者在各行业如何应用大数据又可让读者知道如何去使用这些技术，并且这些项目本身都是各行业的经典，可以直接加以借鉴、拓展和推广。理念篇起

到一个画龙点睛的作用，介绍的都是需要时间和项目磨砺的经验和心得，让读者在共鸣中感知大数据的价值和应用技术的艺术性。

我本人所就职的九次方大数据公司也从事大数据相关工作，我们公司已与不少地方政府联合成立合资公司并建立各地的大数据中心，这些中心负责存储各地政府、企业的重要数据，并对这些数据进行运营，从而实现数据的商业价值，随着《国务院关于印发促进大数据发展行动纲要的通知》的出台，各级政府开始非常重视大数据这项工作，同时也说明我们的大数据资源已经日益丰富。对于如何利用这些数据的课题，本书正好也给我提供了思路，让我知道各行业应该如何挖掘这些大数据，让我坚信大数据未来的发展潜力，也给了我信心继续在大数据这个领域扬帆远航！

此时，突然想起一首古诗，拙改几字，以作为本序的总结：

好书知时节，此时乃出版。随势入眼帘，传知细无声！

张杰，九次方大数据执行副总裁

前　言 Preface

为什么要写这本书

大数据是当前最热的概念之一，在“互联网 +”的背景下，大数据的开放、挖掘和应用已成为趋势。大数据已经成为国家科技竞争的前沿，以及产业竞争力和商业模式创新的源泉。联合国“数据脉动”计划、美国“大数据”战略、英国“数据权”运动、韩国大数据中心战略等先后开启了大数据创新战略的大幕。国务院发布《关于促进大数据发展的行动纲要》，重点强调政府数据的互联互通、共享和开放，并明确提出了具体的时间表。大数据作为目前全球科技创新最主要的战场，有望迎来百花齐放的繁荣盛景。

一花独放不是春，百花齐放春满园，大数据生态系统也生机勃勃。继贵阳大数据交易所成立以来，多个城市相继成立自己的大数据中心，各种数据存储中心和数据评估中心也如雨后春笋。然而，还有相当多的朋友并不了解什么是大数据。市面上介绍大数据概念的书多，但介绍如何应用大数据的书并不多。

大数据的落脚点还是在于应用，如果不能从大数据中挖掘到有利于社会发展的知识，大数据也就没有意义了。数据挖掘技术是从数据中挖掘有用知识的一门系统性的技术，刚好解决了数据利用的问题，所以数据挖掘与大数据便很自然地结合在一起了，故而也就有了本书的构想。

本书特色

纵观全书，可发现本书的特点鲜明，主要表现在以下六个方面：

1）方法务实，学以致用。本书介绍的方法都是数据挖掘中的主流方法，都经过实践的检验，具有较强的实践性。对于每种方法，本书基本都给出了完整、详细的源代码，对于读者来说，具有非常大的参考价值，很多程序可供读者学习并直接套用。

2）知识系统，易于理解。本书的知识体系应该是当前数据挖掘书籍中最全、最完善的，从基本概念与技术，到项目实践，再到理念的整体架构，使得概念、技术、实践、经验四位一体，自然形成一套大数据挖掘的完整体系。而对于具体的技术，也是脉络清晰、循序渐进，不仅包含详细的数据挖掘流程、数据准备方法、数据探索方法，还包含六大类数据挖掘主体方法、时序数据挖掘方法、智能优化方法。正因为有完整的知识体系，读者读起来才有很好的完整感，从而更利于理解数据挖掘的知识体系。

3）结构合理，易于学习。在讲解方法时，由浅入深，循序渐进，让初学者知道入门的切入点，让专业人员又有值得借鉴的干货。本书帮助读者在学习数据挖掘时建立一个循序渐进的过程，使其在短时间内成为一位数据挖掘高手。

4）案例实用，易于借鉴。本书选择的案例都是来自不同行业的经典案例，并且带有数据和程序，所以很容易让读者对案例产生共鸣，同时可以利用案例的数据，进行模仿式的学习，同时，书中的程序也能提高读者的学习效率，可以直接借鉴这些案例，并应用到自己的商业项目中。

5）理论与实践相得益彰。对于本书的每个方法，除了理论的讲解，都配有一个典型的应用案例，读者可以通过案例加深对理论的理解，同时理论也让案例的应用更有信服力。技术的介绍都是以实现实例为目的，同时提供大量技术实现的源程序，方便读者学习，注重实践和应用，秉承笔者务实、切近读者的写作风格。

6）内容独特，趣味横生，文字简洁，易于阅读。很多方法和内容是同类书籍所没有的，这无疑增强了本书的新颖性和趣味性。另外，在本书编写过程中，在保证描述精准的前提下，我们摒弃了那些刻板、索然无味的文字，让文字更有活力，更易于阅读。

如何阅读本书

全书内容分四个部分：

第一部分（基础篇）主要介绍大数据和数据挖掘的基本概念，以及数据挖掘的实现过程、主要内容等基础知识。

第二部分（技术篇）是数据挖掘技术的主体部分，系统介绍了数据挖掘的主流技术，该部分又分三个层次：

1）数据挖掘前期的一些技术，包括数据的准备（收集数据、数据质量分析、数据预处理等）和数据的探索（衍生变量、数据可视化、样本选择、数据降维等）。

2）数据挖掘的六大类核心方法，包括关联规则、回归、分类、聚类、预测和诊断。对于每类方法，则详细介绍了其包含的典型算法，包括基本思想、应用场景、算法步骤、

MATLAB 实现程序、应用实例。

3）数据挖掘中特殊的实用技术，一是关于时序数据挖掘的时间序列技术；二是关于优化的智能优化方法，它们在数据技术体系中不可或缺。时序数据是数据挖掘中的一类特殊数据，所以针对该类特殊的数据类型，又介绍了时间序列方法。另外，数据挖掘离不开优化，所以又介绍了两种比较常用的优化方法——遗传算法和模拟退火算法。

第三部分是项目篇，主要讲解数据挖掘技术在各行业的典型应用实例。所介绍的项目分别来自银行、证券、机械、矿业、生命科学和社会科学等行业和学科，基本覆盖数据挖掘技术应用的主流行业，通过这些项目的研学，读者也可以了解各行业数据挖掘技术的应用领域和应用情况，培养对行业的敏感度。

第四部分是理念篇，是数据挖掘应用思想和经验的整合。本篇包含第 20 和 21 两章，第 20 章侧重数据挖掘项目实施过程中各种技术应用的经验和对各方面问题的权衡和拿捏，体现了技术应用中艺术性的一面；第 21 章侧重数据挖掘项目实施过程中的项目管理和团队管理，以及对团队中的个体如何成长的经验分享。

其中，前三篇为本书的重点内容，建议重点研读，第四篇偏经验，适合结合项目实践反复阅读、体会。

读者对象

- 从事大数据挖掘的专业人士。
- 统计、数据挖掘、机器学习等学科的教师和学生。
- 从事数据挖掘、数据分析、数据管理工作的专业人士。
- 需要用到数据挖掘技术的各领域的科研工作者。
- 希望学习 MATLAB 的工程师或科研工作者，因为本书的代码都是用 MATLAB 编写的，所以对于希望学习 MATLAB 的读者来说，也是一本很好的参考书。
- 其他对大数据挖掘感兴趣的人员。

致读者

专业人士

对于从事大数据挖掘的专业人士来说，大家可以关注整个数据挖掘的知识体系流程，因为本书的数据挖掘知识体系应该是当前数据挖掘书籍中最全、最完善的。另外，数据挖掘流程也介绍得很详细，具有很强的操作性。此外，书中的算法实例和项目实例，也是本书的特

色，值得借鉴。

教师

本书系统地介绍了大数据挖掘的理论、技术、项目、工具和理念，可以作为统计、计算机、经管、数学、信息科学等专业本科或研究生的教材。书中的内容虽然系统，但也相对独立，教师可以根据课程的学时安排和专业方向的侧重点，选择合适的内容进行课堂教学，其他内容则可以作为参考章节。授课部分，一般会包含第一篇和第二篇的章节，而如果课时较多，则可以增加其他章节中的一些项目实例的学习。

在备课的过程中，如果您需要书中的一些电子资料作为课件或授课支撑材料，可以直接给笔者发邮件（70263215@qq.com）说明您需要的材料和用途，笔者会根据具体情况，为您提供力所能及的帮助。

学生

作为21世纪的大学生，无论是什么专业，都有必要学习大数据挖掘。在21世纪和未来，很多信息都以数据形式存在，学习并掌握数据挖掘技术，有助于我们从更深层次了解这个社会，也更有助于我们每人从事的工作。所以，无论现在学的什么专业，都建议好好读一下本书或同类书籍。

配套资源

1. 配套程序和数据

为了方便学习，读者可以到MATLAB中文论坛的本书版块下载书中使用的程序和数据，地址为：

http://www.ilovematlab.cn/forum-252-1.html

具体代码下载地址为：

http://www.ilovematlab.cn/thread-452656-1-1.html

如遇到下载问题，也可以直接发邮件与笔者联系：

E-mail：70263215@qq.com

2. 配套教学课件

为了方便教师授课，我们也开发了本书配套的教学课件，如有需要，也可以与笔者联系。

勘误和支持

由于时间仓促，加之笔者水平有限，所以错误和疏漏之处在所难免。在此，诚恳地期待得到广大读者的批评指正。如果您有什么建议也可以直接将您的建议发送至以上邮箱，期待能够得到您的真挚反馈。在技术之路上如能与大家互勉共进，我们将倍感荣幸！

本书勘误地址为：

http://www.ilovematlab.cn/thread-452657-1-1.html

致 谢

感谢 MathWorks 官方文档，在写作期间提供给我最全面、最深入、最准确的参考材料，强大的官方文档支持也是其他资料无法企及的。

感谢机械工业出版社华章公司的副总编杨福川和编辑姜影、高婧雅、李艺在近三年的时间中始终支持我们的写作，你们的鼓励和帮助引导我们顺利完成本书。

特别感谢好友张杰在百忙之中指导本书的编写并为本书写序！在本书的编写过程中，中国科学院金属研究所的王恺博士，MathWorks 的陈建平、董淑成、陈小挺等好友和同事对书稿进行了校对并给出修改建议，在此也向他们表示感谢！

目　录 Contents

第三篇　项目篇