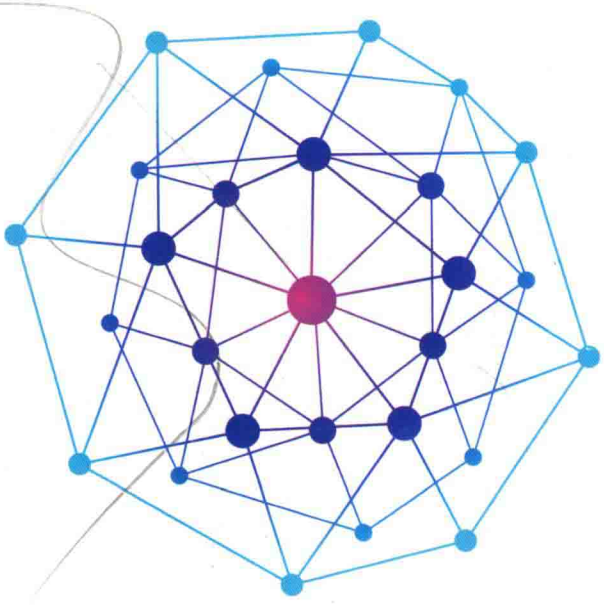
教育部高等学校计算机类专业教学指导委员会-华为ICT产学合作项目
数据科学与大数据技术专业系列规划教材
华为信息与网络技术学院指定教材

大数据分析与挖掘

石胜飞◉编著



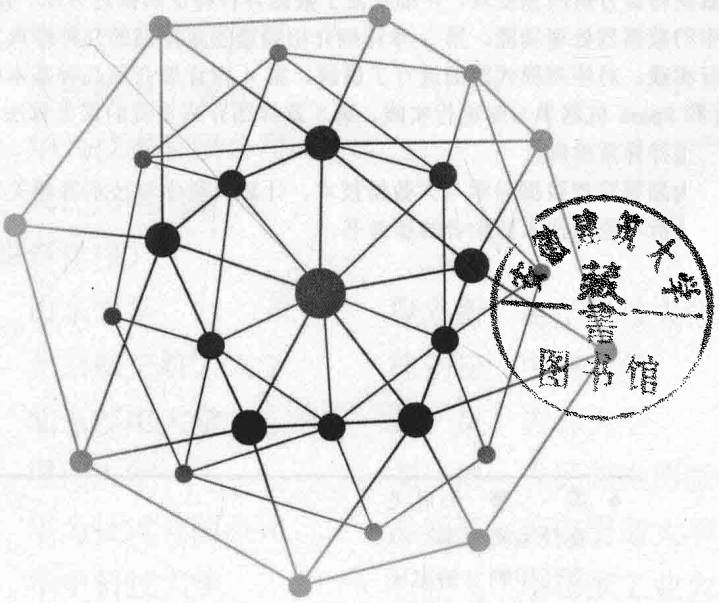
系统、完整的数据科学与大数据技术专业解决方案
名校名师打造大数据领域精品力作
强调基本概念+算法实践
“小数据”上会“算”，“大数据”上“算得快”

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

教育部高等学校计算机类专业教学指导委员会-华为ICT产学合作项目
数据科学与大数据技术专业系列规划教材
华为信息与网络技术学院指定教材

大数据分析与挖掘

石胜飞◉编著



人民邮电出版社
北京

图书在版编目（CIP）数据

大数据分析与挖掘 / 石胜飞编著. -- 北京 : 人民邮电出版社, 2018.8
数据科学与大数据技术专业系列规划教材
ISBN 978-7-115-48305-8

Ⅰ. ①大… Ⅱ. ①石… Ⅲ. ①数据处理－教材 Ⅳ. ①TP274

中国版本图书馆CIP数据核字(2018)第162492号

内 容 提 要

本书是大数据分析与挖掘领域的入门教材，全书共 6 章，内容主要涵盖大数据分析与挖掘过程中用到的基本算法，目的是通过算法原理的介绍，使学生能更高效地将它们运用于数据分析与挖掘的实践中。第 1 章主要介绍大数据分析与挖掘技术发展与应用的特点，以及三种主流的工具。第 2 章主要讲解数据特征分析与预处理，详细介绍了数据各种特征的描述方法、预处理技术，以及 Spark 机器学习库中的数据预处理功能。第 3 章详细介绍频繁模式挖掘的几种经典算法，并结合 Spark 机器学习库进行实践，对序列模式挖掘进行了讲解。第 4 章详细介绍几种基本的分类与回归算法，并结合 Sklearn 和 Spark 机器学习库进行实践。第 5 章详细介绍主流的聚类算法。第 6 章综合运用多种数据挖掘算法进行异常检测。

本书可作为高等院校数据科学与大数据技术、计算机科学与技术等相关专业的本科生教材，也可作为大数据分析与挖掘技术初学者的参考书。

◆ 编　　著　石胜飞
责任编辑　李　召
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
涿州市京南印刷厂印刷

◆ 开本：787×1092　1/16
印张：17.75　　2018 年 8 月第 1 版
字数：459 千字　　2018 年 8 月河北第 1 次印刷

定价：49.80 元

读者服务热线：(010)81055256　印装质量热线：(010)81055316
反盗版热线：(010)81055315
广告经营许可证：京东工商广登字 20170147 号

教育部高等学校计算机类专业教学指导委员会-华为 ICT 产学合作项目
数据科学与大数据技术专业系列规划教材

编　委　会

丛书序一 PREFACE

毫无疑问，我们正处在一个新时代。新一轮科技革命和产业变革正在加速推进，技术创新日益成为重塑经济发展模式和促进经济增长的重要驱动力量，而“大数据”无疑是第一核心推动力。

当前，发展大数据已经成为国家战略，大数据在引领经济社会发展中的新引擎作用更加突显。大数据重塑了传统产业的结构和形态，催生了众多的新产业、新业态、新模式，推动了共享经济的蓬勃发展，也给我们的衣食住行带来根本改变。同时，大数据是带动国家竞争力整体跃升和跨越式发展的巨大推动力，已成为全球科技和产业竞争的重要制高点。可以大胆预测，未来，大数据将会进一步激起全球科技和产业发展浪潮，进一步渗透到我们国计民生的各个领域，其发展扩张势不可挡。可以说，我们处在一个“大数据”时代。

大数据不仅仅是单一的技术发展领域和战略新兴产业，它还涉及科技、社会、伦理等诸多方面。发展大数据是一个复杂的系统工程，需要科技界、教育界和产业界等社会各界的广泛参与和通力合作，需要我们以更加开放的心态，以进步发展的理念，积极主动适应大数据时代所带来的深刻变革。总体而言，从全面协调可持续健康发展的角度，推动大数据发展需要注重以下五个方面的辩证统一和统筹兼顾。

一是要注重“长与短结合”。所谓“长”就是要目标长远，要注重制定大数据发展的顶层设计和中长期发展规划，明确发展方向和总体目标；所谓“短”就是要着眼当前，注重短期收益，从实处着手，快速起效，并形成效益反哺的良性循环。

二是要注重“快与慢结合”。所谓“快”就是要注重发挥新一代信息技术产业爆炸性增长的特点，发展大数据要时不我待，以实际应用需求为牵引加快推进，力争快速占领大数据技术和产业制高点；所谓“慢”就是防止急功近利，欲速而不达，要注重夯实大数据发展的基础，着重积累发展大数据基础理论与核心共性关键技术，培养行业领域发展中的大数据思维，潜心培育大数据专业人才。

三是要注重“高与低结合”。所谓“高”就是要打造大数据创新发展高地，要结合国家重大战略需求和国民经济主战场核心需求，部署高端大数据公共服务平台，组织开展国家级大数据重大示范工程，提升国民经济重点领域和标志性行业的大数据技术水平和应用能力；所谓“低”就是要坚持“润物细无声”，推进大数据在各行各业和民生领域的广泛应用，推进大数据发展的广度和深度。

四是要注重“内与外结合”。所谓“内”就是要向内深度挖掘和深入研究大数据作为一门学科领域的深刻技术内涵，构建和完善大数据发展的完整理论体系和技术支撑体系；所谓“外”就是要加强开放创新，由于大数据涉及众多学科领域和产业行业门类，也涉及国家、社会、个人等诸多问题，因此，需要推动国际国内科技界、产业界的深入合作和各级政府广泛参与，共同研究制定标准规范，推动大数据与人工智能、云计算、物联网、网络安全等信息技术领域的协同发展，促进数据科学与计算机科学、基础科学和各种应用科学的深度融合。

五是要注重“开与闭结合”。所谓“开”就是要坚持开放共享，要鼓励打破现有体制机制障碍，推动政府建立完善开放共享的大数据平台，加强科研机构、企业间技术交流和合作，推动大数据资源高效利用，打破数据壁垒，普惠数据服务，缩小数据鸿沟，破除数据孤岛；所谓“闭”就是要形成价值链生态闭环，充分发挥大数据发展中技术驱动与需求牵引的双引擎作用，积极运用市场机制，形成技术创新链、产业发展链和资金服务链协同发展的态势，构建大数据产业良性发展的闭环生态圈。

总之，推动大数据的创新发展，已经成为了新时代的新诉求。刚刚闭幕的党的十九大更是明确提出要推动大数据、人工智能等信息技术产业与实体经济深度融合，培育新增长点，为建设网络强国、数字中国、智慧社会形成新动能。这一指导思想为我们未来发展大数据技术和产业指明了前进方向，提供了根本遵循。

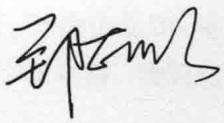
习近平总书记多次强调“人才是创新的根基”“创新驱动实质上是人才驱动”。绘制大数据发展的宏伟蓝图迫切需要创新人才培养体制机制的支撑。因此，需要把高端人才队伍建设作为大数据技术和产业发展的重中之重，需要进一步完善大数据教育体系，加强人才储备和梯队建设，将以大数据为代表的新兴产业发展对人才的创新性、实践性需求渗透融入人才培养各个环节，加快形成我国大数据人才高地。

国家有关部门“与时俱进，因时施策”。近期，国务院办公厅正式印发《关于深化产教融合的若干意见》，推进人才和人力资源供给侧结构性改革，以适应创新驱动发展战略的新形势、新任务、新要求。教育部高等学校计算机类专业教学指导委员会、华为公司和人民邮电出版社组织编写的《教育部高等学校计算机类专业教学指导委员会-华为 ICT 产学合作项目——数据科学与大数据技术专业系列规划教材》的出版发行，就是落实国务院文件精神，深化教育供给

侧结构性改革的积极探索和实践。它是国内第一套成专业课程体系规划的数据科学与大数据技术专业系列教材，作者均来自国内一流高校，且具有丰富的大数据教学、科研、实践经验。它的出版发行，对完善大数据人才培养体系，加强人才储备和梯队建设，推进贯通大数据理论、方法、技术、产品与应用等的复合型人才培养，完善大数据领域学科布局，推动大数据领域学科建设具有重要意义。同时，本次产教融合的成功经验，对其他学科领域的人才培养也具有重要的参考价值。

我们有理由相信，在国家战略指引下，在社会各界的广泛参与和推动下，我国的大数据技术和产业发展一定会有光明的未来。

是为序。



中国科学院院士　郑志明

2018 年 4 月 16 日

丛书序二 PREFACE

在500年前的大航海时代，哥伦布发现了新大陆，麦哲伦实现了环球航行，全球各大洲从此连接了起来，人类文明的进程得以推进。今天，在云计算、大数据、物联网、人工智能等新技术推动下，人类开启了智能时代。

面对这个以“万物感知、万物互联、万物智能”为特征的智能时代，“数字化转型”已是企业寻求突破和创新的必由之路，数字化带来的海量数据成为企业乃至整个社会最重要的核心资产。大数据已上升为国家战略，成为推动经济社会发展的新引擎，如何获取、存储、分析、应用这些大数据将是这个时代最热门的话题。

国家大数据战略和企业数字化转型成功的关键是培养多层次的大数据人才，然而，根据计世资讯的研究，2018年中国大数据领域的人才缺口将超过150万人，人才短缺已成为制约产业发展的突出问题。

2018年初，华为公司提出新的愿景与使命，即“把数字世界带入每个人、每个家庭、每个组织，构建万物互联的智能世界”，它承载了华为公司的历史使命和社会责任。华为企业BG将长期坚持“平台+生态”战略，协同生态伙伴，共同为行业客户打造云计算、大数据、物联网和传统ICT技术高度融合的数字化转型平台。

人才生态建设是支撑“平台+生态”战略的核心基石，是保持产业链活力和持续增长的根本，华为以ICT产业长期积累的技术、知识、经验和成功实践为基础，持续投入，构建ICT人才生态良性发展的使能平台，打造全球有影响力的ICT人才认证标准。面对未来人才的挑战，华为坚持与全球广大院校、伙伴加强合作，打造引领未来的ICT人才生态，助力行业数字化转型。

一套好的教材是人才培养的基础，也是教学质量的重要保障。本套教材的出版，是华为在大数据人才培养领域的重要举措，是华为集合产业与教育界的高端智力，全力奉献的结晶和成果。在此，让我对本套教材的各位作者表示由衷的感谢！此外，我们还要特别感谢教育部高等学校计算机类专业教学指导委员会副主任、北京大学陈钟教授以及秘书长、北京航空航天大学马殿富教授，没有你们的努力和推动，本套教材无法成型！

同学们、朋友们，翻过这篇序言，开启学习旅程，祝愿在大数据的海洋里，尽情展示你们的才华，实现你们的梦想！

华为公司董事、企业BG总裁　阎力大

2018年5月

前言 FOREWORD

随着互联网+、物联网的广泛应用，以及生命科学、工业 4.0 等领域的快速发展，在越来越多的应用中数据量将达到 Terabyte、Petabyte 甚至更高量级。如何快速、准确、实时、方便地从庞大的、分散的数据中获取所需要的知识，是当前科技领域面临的重要问题，也是科学技术及产业领域研究的前沿课题之一。面对这一挑战，数据分析与挖掘技术显示出强大的生命力。数据挖掘使数据处理技术进入了一个更高级的阶段，能够找出过去数据之间的潜在联系，进行更高层次的分析，以便更好地决策、预测各种问题。麻省理工学院的《科技评论》提出，数据挖掘技术是对人类未来产生重大影响的十大新兴技术之一。数据挖掘也必将成为支撑大数据分析的重要及核心技术。

2018 年 3 月，教育部公布在 283 所高校设立数据科学与大数据技术专业。数据科学与大数据技术专业旨在培养具有大数据思维、运用大数据思维及分析应用技术的高层次大数据人才。本教材编写的目的是培养学生掌握大数据分析与挖掘技术，提升学生解决实际问题的能力。

“大数据分析与挖掘”是面向本科高年级的课程。这门课程覆盖的知识面较广，和其他课程的衔接也比较密切，同时，这门课程又具有其明显的应用特点。

本教材的编写符合“大数据分析与挖掘”课程自身的特点。从“厚基础、强实践、严过程、求创新”的人才培养目标出发，能够促进学生对于相关专业基础课程的掌握和提升，如数据库原理、数据结构、算法原理，以及相关的数学基础课程等，使得学生能够将所学的基础知识用于前沿的研究领域，加深对基础课程的理解和掌握。另外，本教材突出大数据计算框架下的实践特点，深入浅出地讲述数据挖掘的基本算法，要求学生进行算法的实践，增强实践动手能力。同时，引导学生找出算法存在的问题，勇于对其进行改进，从而促进学生创新能力的培养。

本教材针对以往在课程教学过程中发现的问题，确立教材的主要编写目标是大数据分析与挖掘的入门级教材。通过简单易学的例子，让学生快速入门，并在动手实践的过程中培养学生对大数据分析与挖掘技术的兴趣。通过教材的介绍，努力弥合理论与实践之间的缝隙，夯实理论基础，强调基本概念与算法的学习。教材在内容组织上，注重提高学生的实践能力。通过单机环境 Python Sklearn 工具的实践，体验在“小数据”上如何“算”的过程，理解算法的基本

原理以及各个参数设置对算法的影响；通过 Spark 机器学习库的实践，体验如何在“大数据”计算平台上对大数据集合也能“算得快”。

本教材的教学计划为 50 学时。通过课程的学习，学生能够掌握大数据分析与挖掘的基础理论，能够运用 Sklearn 数据挖掘软件包从事基本的数据分析与挖掘任务，能够利用 Spark 机器学习库在大数据集合上进行分析与挖掘工作，并为学生从事大数据分析与挖掘领域的更深层次的工作打下坚实的基础。

最后，感谢李克果、李东升、李天禹和范佳欢同学在本书文献整理和示例代码撰写方面所提供的大量帮助；感谢教育部高等学校计算机类专业教指委-华为 ICT 产学合作项目对本教材出版提供的帮助；感谢读者选择本教材，并欢迎读者对本教材内容提出批评和改进建议。

编者

2018 年 5 月

目 录 CONTENTS

01 第1章 绪论

学习大数据分析与挖掘技术，首先要对大数据有一个基本的认识，了解大数据产生的背景以及它所带来的挑战。在充分掌握数据分析与挖掘的原理基础上，结合大数据处理技术，才能有效地完成大数据分析与挖掘的任务。本章介绍了完成上述学习过程的三种主流技术。在学习数据分析与挖掘的算法过程中，可以通过 Sklearn 这个数据挖掘工具包，在小规模数据上完成各种基本的分析与挖掘任务，并且通过实践，加深对各种基本算法的理解。在此基础上，运用 Spark 平台所提供的机器学习组件（ML）来完成大数据集合的高效分析与挖掘任务，并加深对大数据计算平台的理解。最后，本章也详细介绍了华为云所提供的机器学习服务（MLS），可以为企业用户提供更加高效的大数据分析与挖掘功能。

1.1 大数据分析与挖掘简介

大数据研究机构高德纳（Gartner）将大数据（Big Data）定义为需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产。

大数据不仅意味着数据的大容量，还具有一些区别于海量数据（Mass Data）和非常大的数据（Very Large Data）的特点。

国际数据中心（IDC）也定义了大数据："大数据技术描述了一个技术和体系的新时代，被设计用于从大规模、多样化的数据中通过高速捕获、发现和分析技术提取数据的价值"。这个定义刻画了大数据的 4 个显著特点，即容量（Volume）、多样性（Variety）、速度（Velocity）和价值（Value），这个由"4V"描述的大数据定义使用最为广泛。

既然大数据是一种资产，人们自然希望从中挖掘出更多有价值的信息，因此大数据分析与挖掘越来越引起人们广泛的关注。

数据分析是用适当的统计分析方法，对收集来的大量数据进行分析，提取有用信息和形成结论并对数据加以详细研究和概括总结的过程。在这个过程中，用户会有一个明确的目标，通过"数据清理、转换、建模、统计"等

一系列复杂的操作，获得对数据的洞察，从而协助用户进行决策。数据分析可以分为三个层次，即描述分析、预测分析和规范分析。大数据分析是指对规模巨大的数据进行分析是从大数据到信息、再到知识的关键步骤。

数据挖掘（Data Mining）是指从数据集合中提取人们感兴趣的知识，这些知识是隐含的、事先未知的、潜在有用的信息。提取出来的知识一般可表示为概念（Concepts）、规则（Rules）、规律（Regularities）、模式（Patterns）等形式。

在大数据的背景下，知识的获取与传统的学习方式有了很大不同。在很多情况下，只要数据足够多，不再需要通过具体问题的专业知识建模，就可以直接从数据中发现事先未知的知识。以对流感疫情的预测为例，在大数据时代之前，我们要根据数理统计的要求，通过对人群和医院的抽样调查获得数据，然后根据其抽样分布和经验模型来进行预测。谷歌公司则另辟蹊径，运用大数据分析的方法来展开预测。谷歌搜索引擎每天会执行超过数十亿次的搜索，公司从搜索记录中筛选出 5000 万条频繁词，然后与美国疾控中心公布的流感数据进行相关性分析，挖掘出高度相关的 45 种搜索词组合，构建流感预测的挖掘算法。在 2007 年～2008 年，公司根据网民的搜索记录进行了准确的预测。由此可见，与数理统计相比，大数据分析不需要具备概率分布的先验知识，限制条件更少，更为灵活高效。

1.2 大数据应用及挑战

大数据无处不在，已被应用于各个领域，包括宏观经济、金融、电力系统、医疗服务、电子商务以及社交网络等。

在宏观经济领域，淘宝网根据网上成交额比较高的 390 个类目的商品价格得出的 CPI（Consumer Price Index，居民消费价格指数）数据，比国家统计局公布的 CPI 数据更早地预测到经济状况。国家统计局统计的 CPI 数据主要根据的是刚性物品，如食品，百姓都要买，但差别不大；而淘宝网是利用化妆品、电子产品等购买量受经济影响较明显的商品进行预测，因此其 CPI 数据更能反映价格走势。美国印第安纳大学利用谷歌公司提供的心情分析工具，从近千万条的短信和网民留言中归纳出六种心情，来预测道琼斯工业指数，准确率高达 87%。

社交网络近几年突飞猛进的发展，大数据在其中也发挥了巨大的作用。随着互联网用户数量的迅速增加，产生的社交数据也呈现了几何式的增长。对社交网络进行大数据挖掘是当前数据挖掘领域的一个热点，商家通过数据挖掘可以获得与消费者之间更好的互动。越来越多的商家开始将推广渠道转向社交媒体，因为通过社交网络用户之间的转发会产生巨大的社会影响力。

大数据在农业领域也被广泛应用，已经对传统的农业模式产生了巨大的影响。美国推出政府数据开放平台，该平台融合了农业、商业、气候等领域的数据，并且全部免费公开。美国农业部启动土壤数据实时监控项目，建立交互式系统，为农户提供最全面、最新的农业数据，已经成功帮助农民节省了生产成本，同时提高了农产品的质量。英国则发布了《英国农业技术战略》，该战略高度重视利用“大数据”和信息技术来提升农业生产效率，改变农业发展模式。我国的农业也正在向大数据化、精准化农业发展。近年来，国内有关农业大数据的服务类平台相继被开发使用，支持农业大数据的管理、分析、可视化的技术也相继成熟。

大数据在商业模式创新领域发挥了巨大的作用。大数据使商业企业能通过整合网站浏览、购物历史、位置等信息，获得客户的购物偏好，从而为不同的客户定制个性化服务，提供更加精细的产

品和服务，提高购买率，实现更大的商业利润。

大数据在医疗服务领域也正发挥巨大的作用，提高医疗效率和效果。大数据分析技术使临床决策支持系统更加智能。利用图像分析和识别技术，识别医疗影像数据，挖掘医疗文献数据建立医疗专家系统，为医生提出诊疗建议，使医生从重复的咨询工作中解脱出来，提高治疗效率。

日益增长迅速并且繁杂的数据资源，给传统的数据分析、处理技术带来了巨大的挑战。大数据的“4V”特征在数据存储、传输、分析、处理等方面带来本质变化。数据量的快速增长，对存储技术提出了更大的挑战，需要有高速信息传输能力的支持，同时具有对数据的快速分析、处理能力。

为了应对不断涌现的新任务，与大数据相关的大数据技术、大数据工程、大数据科学和大数据应用等迅速成为信息科学领域的热点问题，得到了政府部门、经济领域以及科学领域有关专家的广泛关注。大数据时代的基本特征，决定了其在技术与商业模式上有着巨大的创新空间，将对全球的可持续发展起到关键作用。

大数据还不能全面、准确、真实地反映所有的事物。即使获得了某一事物的所有数据，要挖掘出其中的信息也还存在一定的难度，还取决于数据挖掘的方法和手段。因此，需要将大数据分析与数理统计学相结合，利用数理统计思想优化后的大数据分析，要优于单纯依靠大数据技术分析所得的结果，能有效提高预测的精准度。例如，谷歌公司利用大数据对流感的预测，2008 年的结果与美国疾控中心的数据高度吻合，但在 2009 年、2013 年则出现了很大的偏差，而借助数理统计理论，利用多元线性回归模型改进算法之后则有效消除了这种偏差，从而得到了更加准确的结果。

1.3 大数据分析与挖掘主要技术

大数据分析与挖掘的过程一般分为如下几个步骤。

（1）任务目标的确定

这一步骤主要是进行应用的需求分析，特别是要明确分析的目标，了解与应用有关的先验知识和应用的最终目标。

（2）目标数据集的提取

这一步骤是要根据分析和挖掘的目标，从应用相关的所有数据中抽取数据集，并选择全部数据属性中与目标最相关的属性子集。

（3）数据预处理

这一步骤用来提高数据挖掘过程中所需数据的质量，同时也能够提高挖掘的效率。数据预处理过程包括数据清洗、数据转换、数据集成、数据约减等操作。

（4）建立适当的数据分析与挖掘模型

这一步骤包含了大量的分析与挖掘功能，如统计分析、分类和回归、聚类分析、关联规则挖掘、异常检测等。

（5）模型的解释与评估

这一步骤主要是对挖掘出的模型进行解释，可以用可视化的方式来展示它们以利于人们理解。对模型的评估可以采用自动或半自动方式来进行，目的是找出用户真正感兴趣或有用的模型。

（6）知识的应用

将挖掘出的知识以及确立的模型部署在用户的应用中。但这并不代表数据挖掘过程的结束，还

需要一个不断反馈和迭代的过程，使模型和挖掘出的知识更加完善。

数据挖掘主要包括如下的功能。

（1）对数据的统计分析与特征描述

统计分析与特征描述是对数据的本质进行刻画的方法。统计分析包括对数据分布、集中与发散程度的描述，主成分分析，数据之间的相关性分析等。特征描述的结果可以用多种方式进行展现，例如，散点图、饼状图、直方图、函数曲线、透视图等。

（2）关联规则挖掘和相关性分析

在超市或者网店的商品交易过程中，经常发现有些商品会被同时购买。例如，在购买牛奶时也会购买面包，这些经常一起购买的商品就构成了关联规则。有些商品的购买则是相继出现的。例如，很多消费者先购买一台笔记本电脑，隔了一段时间会接着购买内存卡、蓝牙音箱等。这称为频繁序列模式。

（3）分类和回归

分类是通过对一些已知类别标号的训练数据进行分析，找到一种可以描述和区分数据类别的模型，然后用这个模型来预测未知类别标号的数据所属的类别。分类模型的形式有多种，例如，决策树、贝叶斯分类器、KNN 分类器、组合分类算法等。回归则是对数值型的函数进行建模，常用于数值预测。

（4）聚类分析

分类和回归分析都有处理训练数据的过程，训练数据的类别标号为已知。而聚类分析则是对未知类别标号的数据进行直接处理。聚类的目标是使聚类内数据的相似性最大，聚类间数据的相似性最小。每一个聚类可以看成是一个类别，从中可以导出分类的规则。

（5）异常检测或者离群点分析

一个数据集可能包含这样一些数据，它们与数据模型的总体特性不一致，称为离群点。在很多应用中，例如，信用卡欺诈这类稀有的事件可能更应该引起关注。离群点可以通过统计测试进行检测，即假设数据集服从某一个概率分布，然后看某个对象是否在该分布范围之内。也可以使用距离测量，将那些与任何聚类都相距很远的对象当作离群点。除此之外，基于密度的方法可以检测局部区域内的离群点。

1.4 大数据分析与挖掘工具

目前有很多种大数据分析与挖掘工具，本书重点对三种工具进行介绍。在单机环境下，Sklearn 是高效的数据分析与挖掘工具。在处理大数据的时候，可以采用 Spark 的机器学习模块 Spark ML。对于企业用户，华为云提供的机器学习服务（Machine Learning Service，MLS）则是一个很好的选择。

1.4.1 Sklearn

为了便于初学者实践各种基本算法，本书大部分示例都采用 Sklearn 模块进行演示。

Sklearn 是机器学习中一个常用的 Python 第三方模块，对一些常用的机器学习方法进行了封装，只需要简单地调用 Sklearn 里的模块就可以实现大多数机器学习任务。机器学习任务通常包括分类（Classification）、回归（Regression）、聚类（Clustering）、数据降维（Dimensionality Reduction）、数