产学合作协同育人项目　高等院校数据科学与大数据技术系列规划教材

教育部-普开数据教学内容与课程体系改革项目数据科学与大数据技术专业示范课程

# 大数据挖掘

赵志升　主编

梁俊花　李静　刘洋　副主编

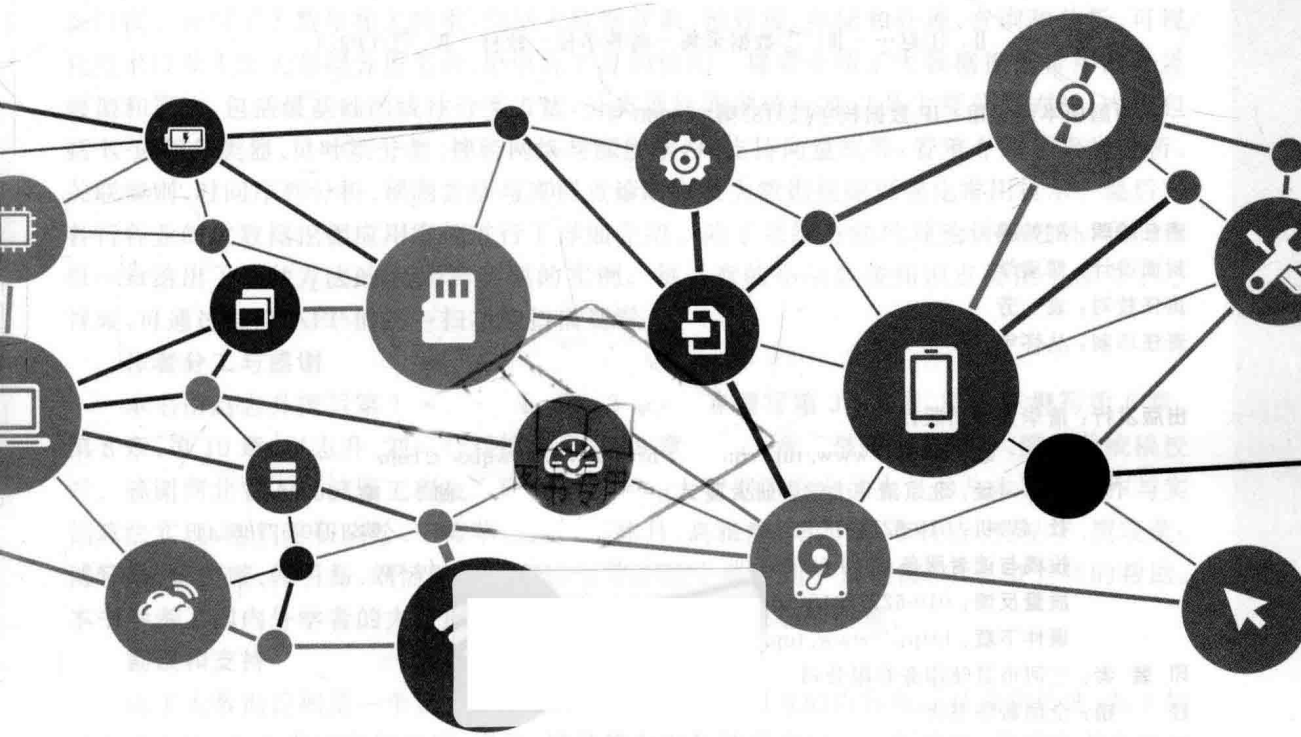
清华大学出版社

高等院校数据科学与大数据技术系列规划教材

# 大数据挖掘

赵志升　主编

梁俊花　李静　刘洋　副主编



清华大学出版社

北京

## 内容简介

本书详细介绍了大数据挖掘技术，全书分为3篇，共12章。第1篇为大数据分析基础，包括第1～4章，分别为大数据概述、大数据相关技术、数据预处理、R语言工具的使用。第2篇为大数据挖掘技术，包括第5～11章，分别为线性分类方法、分类方法、聚类分析、关联规则、预测方法与离群点诊断、时间序列分析、大数据挖掘可视化。第3篇为大数据挖掘案例，包括第12章，介绍了大数据挖掘应用案例。

本书既可作为高等学校计算机科学与技术、数据科学与大数据技术、统计学、数据分析等专业的高等教育教材，也可作为科研人员、从事大数据相关工作的技术人员的参考书。



**图书在版编目(CIP)数据**

大数据挖掘/赵志升主编. —北京：清华大学出版社，2019
(高等院校数据科学与大数据技术系列规划教材)
ISBN 978-7-302-51179-3

Ⅰ. ①大…　Ⅱ. ①赵…　Ⅲ. ①数据采集－高等学校－教材　Ⅳ. ①TP274

中国版本图书馆CIP数据核字(2018)第210080号

**责任编辑**：刘翰鹏
**封面设计**：傅瑞学
**责任校对**：袁　芳
**责任印制**：丛怀宇

**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
**课件下载**：http://www.tup.com.cn，010-62770175-4278
**印 装 者**：三河市君旺印务有限公司
**经　　销**：全国新华书店
**开　　本**：185mm×260mm　　**印　　张**：23.5　　**字　　数**：568千字
**版　　次**：2019年3月第1版　　**印　　次**：2019年3月第1次印刷
**定　　价**：59.00元

产品编号：078814-01

**为什么要写这本书**

大数据时代的到来，使我们的生活在政治、经济、社会、文化等各个领域都发生了很大的变化。如何从大数据中挖掘出隐含的丰富知识与价值，更好地得出结论并作出智能决策已成为相关工作者面临的机遇与挑战。

本书基于教育部“2016 年产学合作协同育人项目”——普开数据教学内容和课程体系改革项目，作为项目成果公开出版。

**读者对象**

本书适合作为高等教育“大数据处理”与“大数据分析”课程的教材，也可作为其他领域有数据分析需求的人员培训教材以及大数据从业人员的参考书。

**如何阅读本书**

本书首先介绍大数据，包括大数据的业务应用场景、云计算与大数据挖掘以及大数据挖掘过程。介绍了大数据相关技术，包括大数据获取、预处理、存储和处理、查询和分析、可视化技术以及主流大数据分析平台、R 语言工具的使用。接着介绍了大数据挖掘常用的分类模型和算法，包括最基础的线性分类方法，分类器性能评价标准以及主要分类方法，内容包括 K-近邻分类器、贝叶斯分类、神经网络与深度学习、支持向量机等，着重介绍了聚类分析、关联规则、时间序列分析、预测方法与离群点诊断以及大数据挖掘可视化常用技术。最后对各行各业的大数据挖掘应用案例进行了详细介绍。除了系统方法的理论讲解之外，我们在每一章给出了每种方法的 R 语言实现的实例。每一章的小结处按知识点提供了参考学习视频，可通过微信 APP 的扫一扫功能扫描观看。

**作者分工与感谢**

本书由赵志升撰写第 1 章、第 2 章、第 12 章，李静撰写第 3～5 章，梁俊花撰写第 6 章、第 8 章、第 10 章，赵志升、刘洋合写第 7 章、第 9 章、第 11 章。最终由赵志升、梁俊花统稿校对。感谢河北省人口健康工程技术研究中心医疗大数据研究室的人员参与本书的写作与实例算法实现，他们是靳晓松、王秀苹、吴仪、韩冰月、高雅静、李凯璇、李佳垚、樊亚宁、贾晓莹、傅轩昂、张艺璇、郭明磊、刘恬恬等。在编写的过程中也得到了刘艳霞、宋玉玺老师的帮助。本书参考了国内外学者的大量成果文献，在此一并表示诚挚的谢意。

**勘误和支持**

由于大数据挖掘是一个正在蓬勃发展的学科领域，涉及的内容宽泛且变化迅速，鉴于作者水平有限，在本书内容的安排、表述、推导等方面各种不当之处在所难免，敬请作者在阅读本书的过程中不吝赐教，以改进此书，读者的意见和建议请发至邮箱 zzsbigdata@sina.com。

编　者

2018 年 11 月

## 第 1 篇　大数据分析基础

## 第3篇 大数据挖掘案例

# 第1篇

# 大数据分析基础

# 第1篇

# 大数据分析基础

第1章

# 大数据概述

【内容摘要】 本章对大数据的产生及特征、现状及趋势及其面临的技术问题业务应用场景进行了简述，并对云计算与大数据挖掘进行了比较，对大数据挖掘过程进行了概述。

【学习目标】 理解大数据、云计算与大数据挖掘的基本概念与特征，掌握大数据挖掘过程与方法，了解大数据现状及趋势以及面临的技术问题。

## 1.1 大数据的业务应用场景

### 1.1.1 大数据的产生及特征

1. 什么是大数据

大数据(Big Data)或称巨量资料，是指需要用新处理模式才能具有更强的决策力、洞察力和流程优化能力的海量、高增长率和多样化的信息资产。大数据是指无法用现有的软件工具提取、存储、搜索、共享、分析和处理的海量的复杂的数据集合。

大数据是一个宽泛的概念，麦肯锡咨询公司是研究大数据的先驱，在其报告 *Big data: The next frontier for innovation, competition, and productivity* 中给出的大数据定义：大数据指的是大小超出常规的数据库工具获取、存储、管理和分析能力的数据集。它同时强调，并不是说一定要超过特定 TB 值的数据集才能算是大数据。国际数据公司(IDC)从大数据的四个特征来定义，即海量的数据规模(Volume)、快速的数据流转和动态的数据体系(Velocity)、多样的数据类型(Variety)和巨大的数据价值(Value)。亚马逊公司的大数据科学家 John Rauser 给出了一个简单的定义：大数据是任何超过了一台计算机处理能力的数据量。在维克托·迈尔-舍恩伯格及肯尼斯·库克耶编写的《大数据时代》中，大数据是指不用随机分析法(抽样调查)这样的捷径，而采用所有数据进行分析处理。

前面几个定义都是从大数据本身出发，我们的定义更关心大数据的功用，即大数据是在多样的或者大量数据中迅速获取信息的能力。在这个定义中，重心是能力。大数据的核心能力，是发现规律和预测未来。

大数据技术是指从各种各样类型的数据中快速获得有价值信息的能力。适用于大数据的技术，包括大规模并行处理(MPP)数据库、数据挖掘技术、分布式文件系统、分布式数据库、云计算平台、互联网和可扩展的存储系统。

2. 大数据的特征

作为《大数据时代》一书的作者，牛津大学网络学院互联网治理与监管专业教授、大数据权威咨询顾问维克托·迈尔-舍恩伯格博士认为，大数据有三个主要特点，分别是全体、混杂

和相关关系。

首先是全体，即收集和分析更多的数据。这个数据都是有关研究问题的数据，其中数据点绝对的数字并不重要，重要的是有多少数据点和研究的现象相关。如果想要研究的现象只有 6000 个数据点，抓住 6000 个数据点就是大数据，因为这就抓住了所有数据。通过这种方式可以看到很多细节，这些细节在之前随机抽样是得不到的。

其次是混杂，即接受混杂。在小数据时代人们总试图收集一些非常干净的数据、高质量的数据，花费很多金钱和精力来确定这些数据是好数据，是高质量的数据。可是在大数据时代，不用去追求那种特别的精确性，当宏观上失去了精确性，微观上却能获得准确性。

最后是相关关系。因为更加混杂，因果关系转向相关关系。人们不要认为可以真正地、容易地找到因果关系，其实那只是发现相关关系。我们应该关注是什么，而不是关注为什么。

业界通常用 4 个 V(Volume、Variety、Value 及 Velocity)来概括大数据的特征。大数据呈现出“4V＋1C”的特点。

(1) 数据体量巨大(Volume)。通过各种设备产生的海量数据，其数据规模极为庞大，远大于目前互联网上的信息流量，PB 级别将是常态。截至目前，人类生产的所有印刷材料的数据量是 200PB(1PB＝210TB)，而历史上全人类说过的所有的话的数据量大约是 5EB(1EB＝210PB)。当前，典型个人计算机硬盘的容量为 TB 量级，而一些大企业的数据量已经接近 EB 量级。

(2) 数据类型繁多(Variety)。大数据种类繁多，在编码方式、数据格式、应用特征等多个方面存在差异性，多信息源并发形成大量的异构数据，这种类型的多样性让数据被分为结构化数据和非结构化数据。相对于以往便于存储的以文本为主的结构化数据，非结构化数据越来越多，包括网络日志、音频、视频、图片、地理位置信息等，这些多类型的数据对数据的处理能力提出了更高要求。

(3) 价值密度低(Value)。价值密度的高低与数据总量的大小成反比，大数据量反而价值密度低。以视频为例，一部时长 1 小时的视频，在连续不间断的监控中，有用数据可能仅有一两秒。如何通过强大的机器算法更迅速地完成数据的价值“提纯”成为目前大数据背景下亟待解决的难题。

(4) 处理速度快(Velocity)。涉及感知、传输、决策、控制开放式循环的大数据，对数据实时处理有着极高的要求，这是大数据区分于传统数据挖掘的最显著特征。根据 IDC 的“数字宇宙”的报告，预计到 2020 年，全球数据使用量将达到 35.2ZB。

(5) 数据复杂(Complexity)。通过数据库处理持久存储的数据不再适用于大数据处理，需要有新的方法来满足异构数据统一接入和实时数据处理的需求。

### 1.1.2 大数据现状及趋势

数据价值的凸显和数据获取手段、数据处理技术的改进是大数据应用爆发的根源。随着数据生产要素化，数据科学、数据科技的不断发展和数据价值的深度挖掘及应用，一场大数据革命正在进行，它将带动国家战略及区域经济发展，智慧城市建设，企业转型升级，社会管理及个人工作、生活等各个领域的创新和变革。如何真正应用好大数据，发挥大数据的威力，是当前所有人都在共同研究和探索的问题。

大数据在数据科学理论的指导下，改变创新模式和理念，各个国家都积极推进大数据的战略性产业，利用大数据来提高国家的经济决策和社会服务能力，保障国家安全。互联网、物联网每天都在产生大量的数据，据调查，2015 年有近 200 亿个设备连接到互联网上，这些设备不仅是计算机、智能终端设备，更有汽车、工厂设备、数字标牌等。从产业拓展的角度看，大数据是继云计算、物联网之后的一个新产业领域，其蕴含的机会和挑战将大大多于云计算和物联网。大数据产业（数据产业）具有很强的蜂箱效应，除了产业自身的经济蕴藏量之外，还将大大撬动其他产业的跨越升级。

2009 年，联合国启动“全球脉动计划”，借大数据推动落后地区发展。美国从开放政府数据、开展关键技术研究和推动大数据应用三方面布局大数据产业。美国在开放政府上非常积极，通过 Data.gov 开放 37 万个数据集，并开放网站的 API 和源代码，提供上千个数据应用。除了推动本国政府数据开放，美国倡导发起全球开放政府数据运动，已有 41 个国家响应。美国联邦政府下属的国防部、能源部、卫生总署等 7 部委联合推动，于 2012 年 3 月底发布了大数据研发专项研究计划（Big Data Initiative），投入 2 亿美元用于研究开发科学探索、环境和生物医学、教育和国家安全等重大领域和行业所急需的大数据处理技术和工具，把大数据研究上升为国家发展战略。

在我国，2011 年以来，中国计算机学会、中国通信学会先后成立了大数据委员会，研究大数据中的科学与工程问题。2015 年 9 月国务院出台了《促进大数据发展行动纲要》，通过开放、产业和安全三位一体建设数据强国。2016 年 3 月我国发布的“十三五”规划纲要又对实施网络强国战略、“互联网＋”行动计划、大数据战略等作了部署。实施国家大数据战略，把大数据作为基础性战略资源，全面实施促进大数据发展行动，加快推动数据资源共享开放和开发应用，助力产业转型升级和社会治理创新。全面推进重点领域大数据高效采集、有效整合，深化政府数据和社会数据关联分析、融合利用，提高宏观调控、市场监管、社会治理和公共服务精准性和有效性。

1.“大数据资源”成为重要战略资源，将成为最有价值的资产

互联网时代，“资源”的含义正在发生极大的变化，它不仅仅是指煤、石油、矿产等一些看得见、摸得着的实体，大数据正在演变成不可或缺的战略资源，数据成为新的战略制高点，成为一种新的资产类别，就像货币或黄金一样。大数据已经被视为一种资产、一种财富、一种可以被衡量和计算的价值。一个国家拥有数据的规模和运用数据的能力将成为综合国力的重要组成部分，对数据的占有和控制也将成为国家间和企业间新的争夺焦点。

2.“大数据决策”成为一种新决策方式

依据大数据进行决策，从数据中获取价值，让数据主导决策，是一种前所未有的决策方式，正在推动着人类信息管理准则的重新定位。随着大数据分析和预测分析对管理决策影响力的逐渐加大，依靠直觉做决定的状况将会被彻底改变。

3.“大数据应用”促进信息技术与各行业深度融合

有专家指出，大数据及其分析会在未来 10 年改变几乎每一个行业的业务功能，在制造业、医疗与健康、交通、能源、材料、商业和服务等行业领域甚至在新闻传媒领域，也都在以大数据为发展契机，加速这些行业与信息技术的深度融合。大数据和传统商业智能融合产生大数据商业智能，从而形成一个全面、完整的数据价值发展平台。大数据服务提供商将会以更加定制化的适用于各行业的商业智能解决方案提供大数据服务，在业务运营智能监控、精