大数据技术与应用专业规划教材
教育部-阿里云产学合作专业综合改革项目规划教材

大数据挖掘与应用

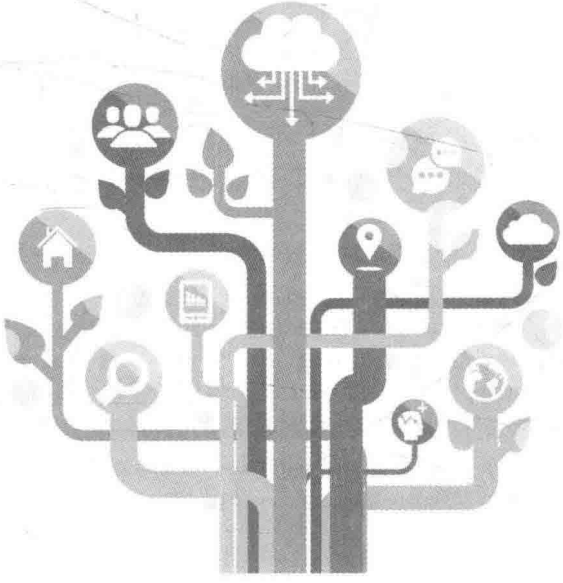
◎ 王振武　编著

清华大学出版社

大数据技术与应用专业规划教材
教育部-阿里云产学合作专业综合改革项目规划教材

大数据挖掘与应用

◎ 王振武　编著



清華大學出版社
北京

内 容 简 介

本书对大数据挖掘与应用的基本算法进行了系统的介绍，每种算法不仅包括对算法基本原理的介绍，而且配有大量的例题以及基于阿里云数加平台的演示，这种理论与实践相结合的方式极大地方便了读者对抽象的数据挖掘算法的理解和掌握。

本书共17章，内容覆盖了数据预处理、关联规则挖掘算法、分类算法和聚类算法及常见的数据挖掘应用，具体章节包括大数据简介、数据预处理技术、关联规则挖掘、逻辑回归方法、KNN算法、朴素贝叶斯分类算法、随机森林分类算法、支持向量机、人工神经网络算法、决策树分类算法、K-means聚类算法、K-中心点聚类算法、自组织神经网络聚类算法、DBSCAN聚类算法以及社交网络分析方法及应用、文本分析方法及应用和推荐系统方法及应用等内容。

本书可作为高等院校数据挖掘课程的教材，也可作为从事数据挖掘工作及其他相关工程技术工作的人员的参考书。

图书在版编目(CIP)数据

大数据挖掘与应用/王振武编著. —北京：清华大学出版社，2017
(大数据技术与应用专业规划教材)
ISBN 978-7-302-46043-5

Ⅰ. ①大… Ⅱ. ①王… Ⅲ. ①数据采集－高等教育－教材 Ⅳ. ①TP274

中国版本图书馆CIP数据核字(2016)第315830号

责任编辑：刘　星　王冰飞
封面设计：刘　键
责任校对：梁　毅
责任印制：王静怡

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，010-62795954
印 装 者：北京鑫海金澳胶印有限公司
经　　销：全国新华书店
开　　本：185mm×260mm　**印　　张**：24.25　**插　　页**：1　**字　　数**：588千字
版　　次：2017年6月第1版　　**印　　次**：2017年6月第1次印刷
印　　数：1～2000
定　　价：49.50元

产品编号：072070-01

序言 1

DT 时代的数据思维与智能思维

本套云计算大数据丛书出版正值信息科技领域进入新一轮巨变，中国经济面临转型机遇的特殊时期。全球信息科技行业伴随着云计算、大数据、物联网、人工智能的发展即将进入一个泛智能的时代，云计算成为数字经济的基础设施；数据驱动、泛在智能成为各行各业转型升级的基础，不仅传统的 IT 从业人员面临能力升级，大多数在校大学生也面临新一轮知识体系的更新，各个垂直行业面临新一轮的人才升级。新一代人才教育与培训，需要一套产学一体的培训课程体系，这是阿里云愿意投身云计算大数据网络安全人才培养体系的时代背景。云计算、大数据、网络安全不仅关乎网络强国的大使命，也逐步成为各行各业专业人才的“元学科”，会逐步成为高等与职业教育的通识课程，一些发达国家已经在中小学立法普及编程课，已经开始指向这个趋势。“懂云计算，有数据思维，理解智能化”，未来可能是每一个工程技术人员与专业人士的必要素质。

2016 年开始，全球信息科技进入一个新的加速爆发周期，可能发生的大概率事件是：二十年之内，有一半的人类知识工作者会被人工智能替代，有服务能力的机器人会诞生，全世界的产业工人会少于机器人；虚拟现实和增强现实会替代今天的智能手机，变成一个新的入口；各行各业都会需要基于物联网的智能化，“中国制造”会成为广泛意义的“中国智造”。

新一轮科技带来了生活方式的变革，生产方式的变革，还有学习方式的变革，这几个趋势的背后，是云计算作为一种普惠科技的基础设施，大数据成为新能源，智能化成为一种新常识。

2016 年，全世界的短视频总量增长了 6 倍，直播业务在中国增长了 10 倍，远在偏远小镇的青年可以通过直播做电子商务，转化率可以提升十倍以上。当一个技术的使用成本趋近于零的时候，会带来广泛的社会效应。十年以前的直播只有电视台能做，需要专门的摄像机等设备，而今天的直播只需要一个手机，而且是多对多带互动的。无论是短视频，还是直播，背后都有云计算作为普惠科技的支撑作用，由此带来的，所有与知识传播有关的教育，包括整个内容行业，都会被它改变，随着大数据和人工智能的加入，人类学习的方式交互性会更强，“学习系统”会根据不同人的理解程度做个性化的推荐与辅导。

这意味着知识生产与知识传播方式的根本性转变，这个恰恰是云计算、人工智能等科技与各行各业产生化学反应的交叉点，数据是这个转变的新能源。

在 2016 年 10 月，阿里云和法院系统合作，发布了一个面向法律服务的智能应用“法小

淘”，通过把数千万份法律判例文本化，“法小淘”智能应用可以为普通老百姓以及初级律师提供“打官司”的咨询服务，根据用户输入的案件信息给出建议，包括推荐合适的律师。貌似与科技远离的法律服务也用上了人工智能，这是垂直行业泛智能化的一个小例子。

中国制造进入智能时代

在工业界，阿里云跟中石化合作，协助他们做了企业的电商平台；与徐工合作，推动工厂基于工业云的智能化；与上汽合作，推出具有智能服务的互联网汽车，都收到积极的市场反馈。中国制造，面临智能化的产业机遇，借助互联网人口和产业布局两大优势成为未来的第一个智能产品制造国。

在接下来的几年，互联网＋智能制造的叠加会在很多个垂直领域出现，数据智能与制造业结合，产生“跨界重混”的效果，甚至制造业就不是以制造为主，而是以服务化为主。这个巨大的重构背后依赖云和大数据。也因为这个需求，我们可预见到工业企业对云计算大数据人才的需求会越来越强烈。

“创业化生存”与共享经济的兴起

创业化，会成为一种常态，越来越多的年轻人开始告别公司，兴起中的数字经济体，都是基于云平台的网络化协作组织；云计算成为共享经济的超级容器，催生新一代创业者和“斜杠青年”。十年以后，或许一半以上的从业者都是“斜杠青年”，今天美国就有数千万人是跨工作、跨公司的“斜杠青年”。

过去十年，云计算使得创业公司的创业门槛降低了 10 倍，没有云计算，Airbnb、NetFlix、推特、Uber 等公司不可能这么快成长壮大，新一代创业者的一个核心能力就是要懂技术，理解数据和算法的价值，缺少技术理解力的创业者将面临更大的同质化压力。一句话，无论是草根创业，还是做一个“斜杠青年”，必要的数据思维是生存本能。

创业化和共享经济的崛起，有赖于云计算作为基础设施，大数据作为新能源的全新范式，新一代创业公司需要大量的科技人才。

在未来的经济环境里，普惠云科技的基础设施化、制造的智能化、软件的泛化以及数据无处不在，是一个大趋势，并且不断向各行各业渗透。本套丛书就是希望在这个普惠科技与各行各业深度融合的时代为下一代科技人才的培养提供更多产业界的经验与实践。

感谢清华大学出版社出版本套云计算与大数据方面的系列教材。感谢各位高校老师的辛苦努力和用心付出，使得本系列教材能够付梓出版。

——阿里云业务总经理　刘松

序 言 2

近年来，随着移动互联网、物联网以及云计算的迅猛发展，大数据成为了世界各地学术界、产业界以及政府部门的关注热点。早在2009年，联合国就启动了“全球脉动计划”，拟通过大数据推动落后地区的发展，2012年，美国政府提出“大数据研究和发展倡议”，发起了全球开放政府数据的运动，2015年，中国政府也通过了“促进大数据发展行动纲要”；在学术界，美国麻省理工大学计算机科学与人工智能实验室建立了大数据科学技术中心(ISTC)，英国牛津大学成立了首个综合运用大数据的医药卫生科研中心；在产业界，IBM、Microsoft、阿里巴巴等都提出了各自的大数据解决方案或应用。

在对大数据的研究和应用过程中，数据挖掘始终占据着核心位置。与传统的数据相比，大数据的特征可以总结为5个V，即体量大(Volume)、速度快(Velocity)、模态多(Variety)、难辨识(Veracity)和价值大(Value)，这些特征给数据挖掘工作带来了新的挑战。

本教材以数据挖掘的研究任务为主线，系统介绍了常见的各类算法及应用。在介绍过程中，本书不仅通过深入浅出的例题讲解、完整规范的源程序实现展示了众多算法的原理，而且基于阿里云数加平台，给出了算法平台级的实现及应用，以便读者更深入地理解大数据挖掘的原理及应用。

相信此书可以帮助读者学习和掌握大数据挖掘的常用算法，更好地理解大数据挖掘的应用场景，进而推动大数据在更加广泛范围内的应用。

——中国矿业大学(北京)机电与信息工程学院院长　钱旭

前　言

大数据泛指大规模、超大规模数据集，因可从中挖掘出有价值的信息而备受关注。数据挖掘是一个涉及数据库技术、人工智能、统计学、机器学习等多个学科的领域，并且已经在各行各业有着非常广泛的应用。为适应我国数据挖掘的教学工作，笔者在数据挖掘教学实践的基础上，参阅了多种国内外最新版本的教材，编写了本书。本书可以作为高等院校研究生的教材，也可以为相关行业的工程技术人员提供有益的参考。

本书是教育部阿里云产学合作项目，在内容安排上循序渐进，对大数据挖掘的基本算法进行详细的讲解。本书的最大特点是理论与实践相结合，算法理论与产业一线实践相结合，全书几乎所有的算法都配有实例和基于阿里云数加平台的演示。这种理论与实践相结合的方法克服了重理论、轻实践的内容组织方式，极大地方便了读者的理解。具体而言，本书17章内容之间的关系如下图所示。

本书提供教学课件，读者可从 www.tup.com.cn 网站自行下载。由于编者水平有限，本书必定存在不妥和不足之处，恳请专家和读者批评指正。

编　者

2017 年 3 月

目　录

第一篇　基　础　篇

第二篇 技 术 篇

第三篇 综合应用篇