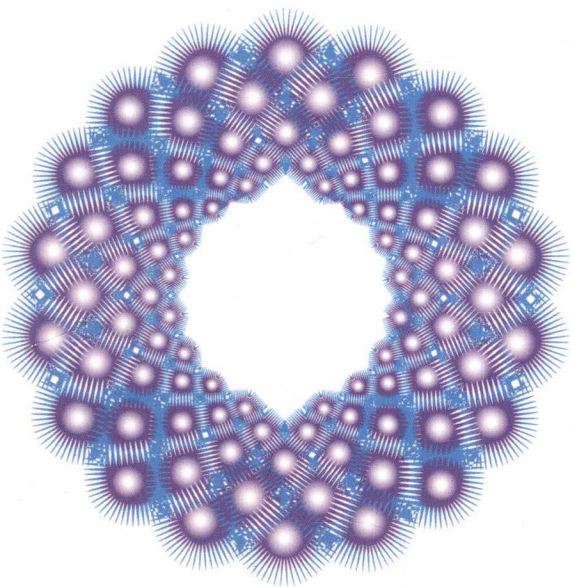
数据中国“百校工程”项目系列教材
数据科学与大数据技术专业系列规划教材

瑞翼教育

数据挖掘与机器学习

吴建生 许桂秋 ◉ 主编
黄楠 霍雷刚 王彦超 张军 王照文 ◉ 副主编



BIG DATA
Technology

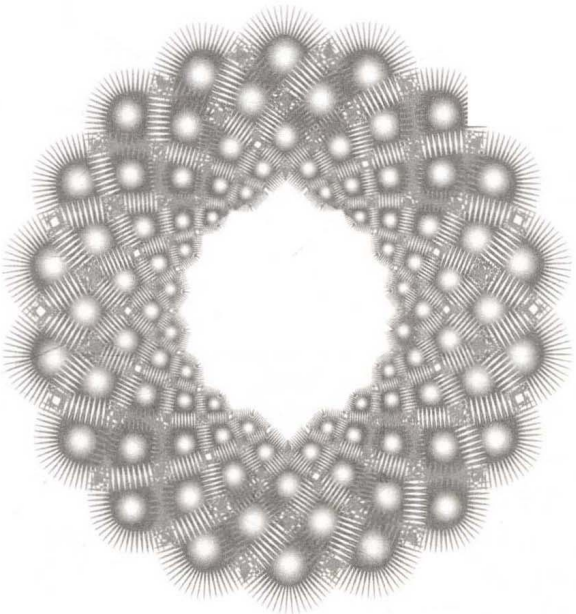
中国工信出版集团
人民邮电出版社
POSTS & TELECOM PRESS

数据中国“百校工程”项目系列教材
数据科学与大数据技术专业系列规划教材

瑞翼教育

数据挖掘与机器学习

吴建生 许桂秋 ◉ 主编
黄楠 霍雷刚 王彦超 张军 王照文 ◉ 副主编



BIG DATA
Technology

人 民 邮 电 出 版 社
北 京

图书在版编目（CIP）数据

数据挖掘与机器学习 / 吴建生，许桂秋主编. -- 北京 : 人民邮电出版社，2019.4
数据科学与大数据技术专业系列规划教材
ISBN 978-7-115-50352-7

Ⅰ. ①数… Ⅱ. ①吴… ②许… Ⅲ. ①数据采集－教材②机器学习－教材 Ⅳ. ①TP274②TP181

中国版本图书馆CIP数据核字(2019)第029350号

内 容 提 要

本书从实用角度出发，采用理论与实践相结合的方式，介绍数据挖掘与机器学习的基础知识，力求培养读者数据思维的能力。全书内容包括数据挖掘概述、Pandas 数据分析、机器学习、分类算法与应用、回归算法与应用、无监督学习、关联规则和协同过滤、图像数据分析、自然语言处理与NLTK。

本书既可作为各类高校数据挖掘与机器学习的课程教材，又可供对数据挖掘和机器学习感兴趣的读者学习参考。

◆ 主　　编　吴建生　许桂秋
副 主 编　黄　楠　霍雷刚　王彦超　张　军　王照文
责任编辑　张　斌
责任印制　陈　犇

◆ 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
固安县铭成印刷有限公司印刷

◆ 开本：787×1092　1/16
印张：11.25　2019 年 4 月第 1 版
字数：289 千字　2019 年 4 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

前　言

信息技术的高速发展，引发了近几年的大数据和人工智能浪潮。IT行业的工作人员作为时代的弄潮儿，在为这些波澜壮阔的景象感到兴奋的同时，又深切地感受到技术的飞速变化所带来的巨大压力。

在如今这个处处以数据驱动的世界中，数据挖掘正变得越来越大众化。它被广泛地应用于不同领域，如搜索引擎、机器人、无人驾驶汽车等。本书不仅可以帮助读者了解现实生活中数据挖掘的应用场景，还可以帮助读者掌握处理具体问题的算法。

全书共9章，阐述了数据挖掘与机器学习的核心概念与技术，以及运用这些技术分析实际问题的思路。书中还介绍了9个综合性的案例，其难易程度和侧重点与书中各章的知识点相匹配。

第1章介绍数据挖掘的基本概念、功能与应用领域，使读者掌握数据挖掘的基本理念、流程和方法。

第2章介绍Pandas模块的语法结构，并通过对自行车行驶数据与服务热线数据的分析，使读者掌握通过Pandas模块对数据进行统计分析的方法。

第3章介绍机器学习的框架，以及Sklearn模块的语法结构，使读者掌握搭建机器学习流水线的方法。

第4章介绍分类算法，使读者掌握通过Sklearn模块搭建一个分类器并实现分类功能的方法。

第5章介绍回归算法，使读者掌握通过Sklearn模块搭建一个回归模型并实现回归功能的方法。

第6章介绍无监督学习，使读者掌握通过Sklearn模块搭建一个聚类模型并实现聚类功能的方法。

第7章介绍关联规则和协同过滤，使读者掌握通过这些算法实现电影推荐的方法。

第8章介绍图像数据分析的相关技术，使读者掌握进行图像特征提取和人脸识别的方法。

第9章介绍自然语言处理的相关技术，使读者掌握进行文本特征提取和文本分类的方法。

本书建议安排64课时。根据学生的接受能力及高校培养方案的设置，教师可以把某些内容作为学生的自学材料。

由于编者水平有限，编写时间仓促，书中难免存在疏漏和不足之处，恳请广大读者不吝赐教。

编　者

2019年1月

试读结束：需要全本请在线购买：　www.ertongbook.com

目　录

第 1 章 数据挖掘概述

本章主要讲解数据挖掘技术的概念、功能、应用领域等基础知识。

本章重点内容如下。

（1）数据分析技术的发展。

（2）数据挖掘的定义。

（3）知识发现的步骤。

（4）数据挖掘的功能与应用领域。

（5）数据挖掘的模型。

（6）数据挖掘的数据类型。

1.1 数据分析技术发展简介

我们生活在一个数据爆炸的时代。如何在大量的数据中获取我们想要的知识，是当今时代的一个重要需求。

1.1.1 数据时代

随着各行各业技术的发展，这个时代的数据量已经发生跨越式的增长。

（1）天文数据：2000 年，斯隆数字巡天（Sloan Digital Sky Survey，SDSS）项目启动的时候，位于美国新墨西哥州的望远镜在几周内收集到的数据，比天文学历史上总共收集的数据还要多。到了 2010 年，这个项目信息档案数据量已经高达 1.4×2^{42} 字节。再看另一组数据，哈勃太空望远镜（Hubble Space Telescope，HST）每天产生 3～5GB 数据，郭守敬望远镜（LAMOST）每年生产 10TB 数据，天眼（FAST）仅 4 小时就可产生 10TB 数据。天文数据已成为天文学研究的重要部分，预计到 2025 年，全球天文数据采集量将达到每年 2.5×10^{10}TB。

试读结束：需要全本请在线购买： www.ertongbook.com

（2）互联网数据：谷歌公司在2014年曾经指出："截至2000年，人类仅存储大约12EB的数据，但如今，我们每天产生2EB的数据。过去两年的时间里产生了世界上90%以上的数据。"这只是2014年的数据，时至今日，数据量更是大得惊人。Excelcom公司在2016年发布了一份"互联网一分钟产生数据"的图表，图表展示了一分钟内互联网产生的数据：40万人登录微信，2万人使用视频或语音聊天，百度有416万个搜索请求，1.5亿封电子邮件进行了发送，YouTube上有278万个视频被观看，WhatsApp上发送了2000万条信息。以上仅仅列举了部分知名的互联网巨头公司的数据，如果统计整个互联网数据，数据量将会更大。2017年，淘宝每天产生的数据都高达7TB。

（3）物联网数据：物联网（Internet of Things，IoT）是新一代信息技术的重要组成部分。物联网的含义就是物物相连的互联网。物联网产生大量数据，数据时代的到来使物联网获得极大的发展。在投资方面，物联网的资金投入预计从2015年的2150亿美元增长到2020年的8320亿美元；而物联网上设备的数量，高通公司预计2020年连网设备数量有望达到250亿台以上，阿里云预计2020年物联网连接设备将达到200亿台以上。物联网中的每台设备都会产生大量的数据，物联网的发展是推动电子数据爆炸式增长的主要动力。

如此庞大的数据，蕴含着巨大的价值。随着大量数据存储和采集技术的发展，不同的机构都可以较容易地收集到大量的数据，但对大量数据的信息分析成为一个较为困难的事情。针对大量数据的分析，传统的数据分析技术存在不足，主要体现在对这些大数据无法分析或者处理性能低等。另外，即使有些数据量较小，但也可能因为数据的一些特点，不适用传统的数据分析方法。在这种情况下，大数据技术的出现很好地解决了大量数据的计算问题，针对大量数据的挖掘工作也取得长足的进步。

1.1.2 数据分析技术的发展

随着数据分析技术的发展，尤其是数据库技术的发展，数据挖掘的出现也是一个必然的趋势。

数据库技术始于20世纪60年代中期，距今已有几十年，经历三代演变，出现了巴克曼（C.W. Bachman）、科德（E. F. Codd）和格雷（J. Gray）三位图灵奖得主，发展成了以数据建模和数据库管理系统（DBMS）等核心技术为主，内容丰富的一门学科。20世纪60年代，传统的文件系统已经不能满足人们对数据管理和数据共享的需求（文件系统存在数据冗余、数据联系弱等问题）。在这种需求下，能够统一管理和共享数据的数据库管理系统（DBMS）应运而生。代表数据管理技术进入数据库阶段的标志是20世纪60年代末和70年代初的三件大事：1968年IMS系统（层次模型）的研制成功，1969年DBTG报告（网状系统）的发布，1970年科德（E. F. Codd）文章（关系模型）的发表。

1968年，IBM公司研制出的数据库管理系统（Information Management System，IMS）

是层次数据库系统的典型代表。1969 年，数据系统语言会议（Conference on Data Systems Languages，CODASYL）提出了一份“DBTG 报告”，根据其实现的系统一般称为 DBTG 系统，现有的网状数据库系统大都是采用 DBTG 方案。层次和网状数据库系统是第一代数据库系统。层次数据库开拓了数据库系统，而网状数据库则是数据库概念、方法、技术的奠基者。网状数据库和层次数据库很好地解决了文件系统存在的一些问题（集中和共享），但在数据独立性和抽象级别上仍有较大的欠缺。

1970 年，IBM 公司的科德发表了一篇名为“*A Relational Model of Data for Large Shared Data Banks*”的论文，提出了关系模型的概念，奠定了关系模型的理论基础。之后科德又陆续发表多篇文章，论述了范式理论和衡量关系系统的 12 条标准，用数学理论奠定了关系型数据库的基础。1970 年，关系模型建立之后，IBM 公司在圣何塞实验室确立了著名的 System R 项目，其目标是论证一个全功能关系 DBMS 的可行性。该项目结束于 1979 年，完成了第一个实现 SQL 的 DBMS。关系型数据库系统的代表产品有 Oracle、DB2、SQL Server 以及 Informix 等。

随着信息技术和市场的发展，关系型数据库系统的局限性也逐渐显露出来：它能很好地处理“结构化的数据”，却对越来越多的复杂类型的数据无能为力。20 世纪 90 年代以后，在相当长的一段时间内，技术领域将重点放在研究“面向对象的数据库系统（Object Oriented Database）”上。然而，理论的完善并未带来市场的响应。面向对象数据库产品没有获得普遍认可的主要原因在于，其主要设计思想是取代现有的数据库系统，对很多企业来说，改变一个现有的成熟的系统，同时使用一种全新的产品，是工作量巨大且充满未知的。

20 世纪 60 年代后期，决策支持系统（Decision Support System，DSS）出现了，决策支持系统解决非结构化问题，是服务于高层决策的管理信息系统。一般 DSS 包括数据库、模型库、方法库、知识库和会话部件。DSS 数据库不同于一般 DBMS，它有很高的性能要求，一般由数据仓库（Data Warehouse）来充当 DSS 数据库。1988 年，为解决企业集成问题，IBM 公司的研究员巴里·德夫林（Barry Devlin）和保罗·墨菲（Paul Murphy）提出了数据仓库（DataWarehouse）的概念。1991 年，恩门（W. H. Inmon）出版了《构建数据仓库》一书，意味着数据仓库真正开始应用。数据仓库是决策支持系统和联机分析应用数据源的结构化数据环境，是一个面向主题（Subject Oriented）的、集成（Integrated）的、相对稳定（Non-Volatile）的、反映历史变化（TimeVariant）的数据集合，用于支持管理决策（Decision Making Support）。

数据挖掘是在数据库技术长期积累、数据量快速增长、数据挖掘算法逐步成熟，这三者条件都具备的情况下的直接产物。通过结合数据仓库技术，数据挖掘在商业领域产生了越来越大的作用。

数据库技术的演变如图 1-1 所示。从图 1-1 中可以清楚地了解数据库技术的发展，从网状、层次数据库到关系型数据库，再到后期的专用数据库、NoSQL 数据库的出现等。

在数据库发展的每个阶段，用户的需求也是不一样的，表 1-1 给出了这些年来用户需求以及数据库技术发展的对比。数据挖掘技术与数据仓库和 OLAP 技术的结合，在商业、电信、银行、科研等领域均有应用，并且能处理的数据类型也不仅仅是结构化的二维表，流、时间、空间、多媒体等数据均可进行知识的挖掘。

图 1-1　数据库技术的演变

表 1-1　　数据库技术发展与用户需求对比

进化阶段	商业问题	支持技术	产品厂家（公司）	产品特点
数据搜集（20 世纪 60 年代）	过去五年中我的总收入是多少	计算机、磁带和磁盘	IBM、CDC	提供历史性的、静态的数据信息
数据访问（20 世纪 80 年代）	在北京的分部去年 3 月的销售额是多少	关系型数据库（RDBMS）、结构化查询语言（SQL）、ODBC	Oracle、Sybase、Informix、IBM、Microsoft	在记录级提供历史性的、动态的数据信息
数据仓库；决策支持（20 世纪 90 年代）	在北京的分部去年的销售额是多少？具体分析每个月的销售额	联机分析处理（OLAP）、多维数据库、数据仓库	Pilot、IBM、Arbor、Oracle	在各种层次上提供回溯的、动态的数据信息
数据挖掘（正在流行）	下个月广州的销售会怎么样？为什么	高级算法、多处理器计算机、海量数据库	Pilot、Lockheed、IBM、SGI	挖掘数据中反映的内在规律； 提供预测性的信息

1.2　数据挖掘的概念

数据挖掘（Data Mining）是 20 世纪 80 年代出现的一种技术。作为一个跨学科的领域，数据挖掘有着多种定义。数据挖掘可翻译为资料探勘、数据采矿。它是数据库知识发现（Knowledge Discovery in Database，KDD）中的一个步骤。数据挖掘一般是指从大量数据中通过算法搜索出隐藏于其中的信息的过程。数据挖掘通常与计算机科学有关，并通过统计、联机分析处理、情报检索、机器学习、专家系统（依靠过去的经验法则）和模式识别等诸多方法来实现上述目标。这种关于数据挖掘的定义是狭义的定义，将数据挖掘当成知识发现的一个步骤。还有一种广义的定义，认为数据挖掘就是一个完整的知识发现，包括数据清理、建模、评估等全过程。

1.2.1　数据挖掘的定义与 OLAP

数据挖掘是从大量的、不完全的、有噪声的、模糊的、随机的应用数据中，提取出潜在且有用的信息的过程，并且这个过程是自动的，这些信息的表现形式可以为规则、概念、模型、模式等。数据挖掘是一种综合技术，在对业务数据进行处理的过程中，需要用到很多领域的知识，如数据库、统计学、应用数学、机器学习、模式识别、数据可视化、信息科学、程序开发等领域的理论和技术，如图 1-2 所示。

数据挖掘的核心是利用算法模型对预处理后的数据进行训练，训练后获得数据模型。

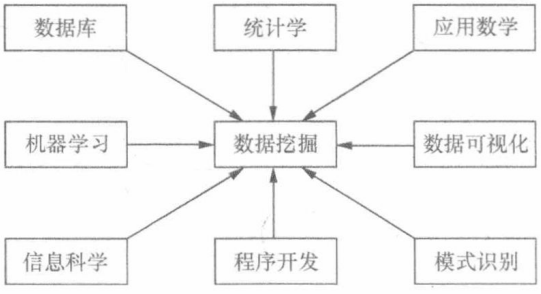


图 1-2　数据挖掘与其他学科的关系

企业数据的数据量巨大，但真正具有价值的信息却比较少，想要获得有用的信息，需要对大量的数据进行深层分析。商业信息的处理技术可以分为两个层次。在浅层次上，我们可利用数据库管理系统的查询、检索功能，与多维分析、统计分析方法相结合，进行联机分析处理（On-Line Analytical Processing，OLAP），得出可供决策参考的统计分析数据。在深层次上，我们可从数据中发现前所未有的、隐含的知识。OLAP 的出现早于数据挖掘，它们都是从数据中抽取有用信息的方法，就决策支持的需要而言，两者是相辅相成的。OLAP 可以看作一种广义的数据挖掘方法，旨在简化和支持联机分析，而数据挖掘的目的是使这一过程尽可能自动化。

1.2.2　数据挖掘与知识发现

数据挖掘与知识发现有很密切的联系，从狭义的角度讲，数据挖掘是知识发现的一个环节；从广义的角度讲，数据挖掘与知识发现的含义是相同的。

知识发现（Knowledge Discovery in Database，KDD）是一个完整的数据分析过程，主要包括以下几个步骤。

1. 确定知识发现的目标

这一步是确定知识发现的目的，要发现哪些知识。对于医疗数据，我们要确定是根据病人的特征预测其患病类型，还要根据关联规则为专家系统提供一些支持；对于电商网站的商品评价，我们的目标可能是对评价进行情感分析，并获得评价关键词。知识发现的第一步是制订目标，目标制订后，就可以根据目标的需求，确定数据采集、预处理、模型选择等后续的步骤。

2. 数据采集

这一步是将可能与知识发现目标相关的数据采集到指定的系统中。这里说的数据采集可以是从网络爬取的数据，也可以是从数据库中直接导出的数据，还可以是常见的 CSV 文件等数据。

采集到的数据维度要满足目标的需求，如果需要的字段特征没有采集到，挖掘出的知识会偏离实际情况，就如数据挖掘领域有一句话："数据质量决定挖掘的上限，而算法仅仅是逼近这个上限。"例如，如果有一个关于房价预测的挖掘过程，在数据的特征中没有房子所处的地理位置信息，那么根据这份数据获得的房价预测的评分一定是很低的。

3. 数据探索

按上述步骤采集到的数据，往往是不可以直接使用的，需要数据分析人员对数据进行探索。探索主要包括数据特征的基本统计描述、数据特征间的相似/相异性等。数据探索阶段可以采用可视化技术，将数据的特征展现出来。离散型数据和连续型数据适用不同的算法模型，数据的分布规律决定其是否符合某些算法模型的要求。通过数据探索后，我们就可以有的放矢地进行下一步——数据预处理的环节了。

4. 数据预处理

数据预处理主要包括数据清理、数据集成、数据归约、数据变换和离散化等几个部分。

（1）数据清理主要包括缺失值与异常值的清理。针对缺失值，可以采用简单的删除，但如果缺失值的比例达到一定阈值，就需要用户去判断在采集过程中是否出现了问题，不可以进行简单的删除操作了。因为一旦删除了数据，数据所代表的信息就无法找回。也可以将缺失值用默认值替换，或是采用拉格朗日插值法对缺失值进行填充等。

（2）数据集成主要是指将多种数据源汇集到一起，放入一个数据仓库的过程。在数据集成的过程中会出现实体识别（Entity Resolution）、冗余属性识别、数据值冲突等问题。在将多种数据源集成时，实体识别是很常见的事情。实体识别可描述成在一个或多个数据源中的不同记录是否被描述为同一个实体，同一实体在数据集成过程中可被用于数据去重和连接键等集成操作中。这里用一个数据库中的实例说明。如果 A 表有一个字段为 stu_id，B 表有一个字段为 stu_num，那么这两个字段是否都为同一个实体的属性呢？如果是同一个属性，那么在集成时，这个字段可以作为多表关联的条件，生成新表时保留两者中的一个值就可以。冗余属性识别是指是否某些属性之间存在相关性，或者一个属性可以由其他的属性推导得出。数据值冲突指的是不同数据源中针对同一个实体的属性值不同，这可能是单位不一致导致的。数据集成就是在多种数据源的集成过程中，解决上述的几个问题，形成一个大的、不冗余的、数值清楚的数据表。

（3）数据归约是指在保证原始数据信息不丢失的前提下，减少分析使用的数据量。数据归约中最常使用的方式是维归约。维归约的含义是将原先高维的数据合理地压缩成低维数据，从而减少数据量，常用的方法为特征的提取，如线性性别分析（Linear

试读结束：需要全本请在线购买：　www.ertongbook.com