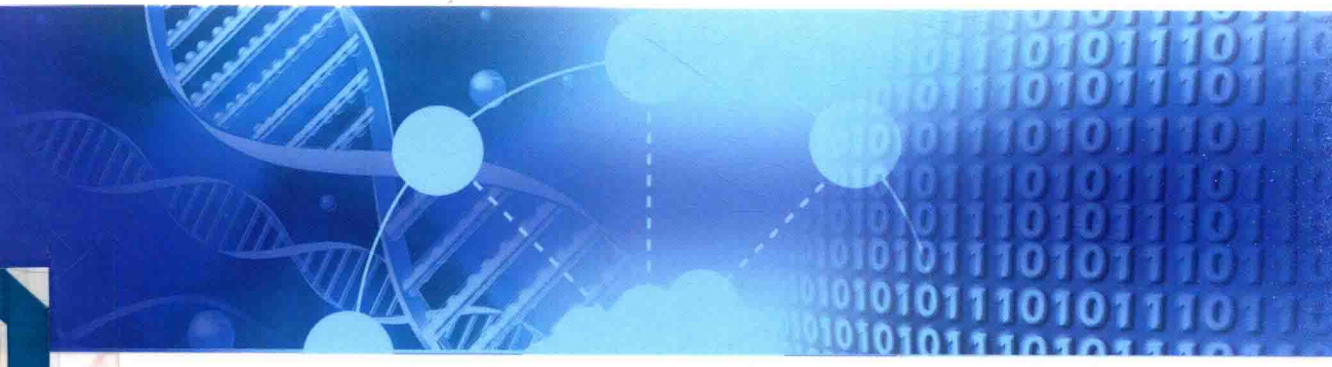
普通高等教育“十三五”规划教材

Introduction to Medical Big Data

医学大数据应用概论



◎ 娄岩 主编

非外借

科学出版社

普通高等教育“十三五”规划教材

医学大数据应用概论

娄　岩　主编

张志常　马　瑾　副主编

科学出版社

北　京

内 容 简 介

本书是继2015年中国医科大学计算机教研室编写的《医学大数据挖掘与应用》之后的又一本面向大数据在医学领域应用的教材。本书遵循定义、特征、技术流程和医学应用典型案例分析的逻辑，抽丝剥茧，由易到难，有助于读者理解和掌握大数据技术。本书应用案例围绕医学大数据及其相关应用这一主线，递进展开，内容具体，过程详尽，并且具有一定的操作性，既方便教师教学，又能引起读者自主学习的兴趣，加深对知识的理解，以及对学习效果的检验。

本书可作为医学院校本科生、研究生的教学用书，也可供医学从业人员，尤其是致力于医学数据处理的人员自学和参考。

图书在版编目（CIP）数据

医学大数据应用概论/娄岩主编. —北京：科学出版社，2017

（普通高等教育“十三五”规划教材）

ISBN 978-7-03-055999-9

Ⅰ. ①医… Ⅱ. ①娄… Ⅲ. ①医学-数据处理-高等学校-教材

Ⅳ. ①R319

中国版本图书馆CIP数据核字（2017）第310786号

责任编辑：宋 丽 陈将浪 / 责任校对：马英菊

责任印制：吕春珉 / 封面设计：东方人华平面设计部

科学出版社出版

北京东黄城根北街16号

邮政编码：100717

http://www.sciencep.com

三河市骏杰印刷有限公司印刷

科学出版社发行 各地新华书店经销

*

2017年11月第 一 版 开本：787×1092 1/16

2017年11月第一次印刷 印张：10 1/2

字数：250 000

定价：30.00元

（如有印装质量问题，我社负责调换〈骏杰〉）

销售部电话 010-62136230 编辑部电话 010-62135927-2014

本书编写人员

主　编　娄　岩

副主编　张志常　马　瑾

参　编　郑琳琳　刘尚辉　李　静　丁　林

　　徐东雨　曹　阳　庞东兴　霍　妍

前　言

我们正处在一个新技术和传统行业相融合的智能时代，大数据、AR、VR 和人工智能等信息技术必将撬动传统行业的各个板块，为社会发展和时代进步注入新的血液。习近平总书记在十九大报告中提出要“推动互联网、大数据、人工智能和实体经济深度融合”，强调“贯彻新发展理念，建设现代化经济体系”。

国务院在 2015 年印发的《促进大数据发展行动纲要》中明确指出，大数据成为推动经济转型发展的新动力、重塑国家竞争优势的新机遇、提升政府治理能力的新途径。坚持创新驱动发展，加快大数据部署，深化大数据应用，已成为稳增长、促改革、调结构、惠民生和推动政府治理能力现代化的内在需要和必然选择。

在智能医学与健康服务的大潮中，构建电子健康档案、电子病历数据库，建设覆盖公共卫生、医疗服务、医疗保障、药品供应、计划生育和综合管理业务的医疗健康管理和服务大数据应用体系势在必行；探索预约挂号、分级诊疗、远程医疗、检查检验结果共享、防治结合、医养结合、健康咨询等服务，优化形成包括规范、共享、互信的诊疗流程在内的医学大数据应用也摆到了我们面前。作为医学院校的教育工作者，应该为创新医学院校人才培养模式，建立健全多层次、多形态的应用人才培养体系，培养具有统计分析、计算机技术、医学知识等多学科知识的跨界复合型人才做出贡献。我们希望能够在医学学生中开展大数据知识普及和教育培训，培育具备大数据技术的应用创新型人才，提高医学学生对大数据的整体认知和应用水平。

为此，本书围绕医学大数据应用，从理论、相关技术和实际应用 3 个层面进行了简明扼要的阐述，目的是让广大师生对大数据在医学领域的应用方法和相关知识有所了解，更好地把握科学发展的方向。

我校已连续 4 年将大数据技术及相关课程纳入大学计算机基础教育中，为国家培养了一批又一批掌握最新科学发展动态和技能的数字化医学人才，同时积累了一定的教学经验。本书针对医学学生的特点和大数据在医学领域的应用策略编写，理论联系实际，书中全部案例和解决问题方法均采用与数字医学密切相关的内容。

本书的一大亮点是每章中将大数据在医学领域中的应用落地，注重方法运用、案例解析及可操作性。另外，本书注重启发式的学习策略，便于读者理解和掌握。全书在每章均附有实际应用案例与关键词注释，方便读者查阅和自学。

本书由娄岩担任主编，张志常、马瑾担任副主编。全书包括 11 章，具体编写分工如下：第 1 章大数据概论由娄岩编写，第 2 章医学大数据采集由郑琳琳编写，第 3 章大数据分析由刘尚辉编写，第 4 章大数据可视化由李静编写，第 5 章 Hadoop 由马瑾编写，第 6 章 HDFS 和 Common 由丁林编写，第 7 章 MapReduce 由徐东雨编写，第 8 章 NoSQL 由曹阳编写，第 9 章 Spark 由庞东兴编写，第 10 章云计算与大数据由张志常编写，第 11 章大数据在医疗领域的应用由霍妍编写。

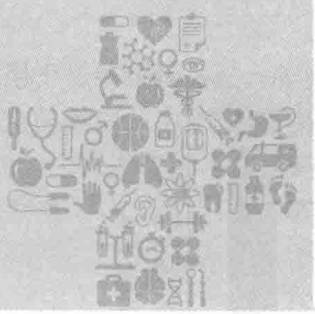
科学出版社对本书的出版做了精心策划和充分论证，在此向所有参加编写的同事们、帮助和指导过我们工作的朋友们及参考文献中的作者们表示衷心的感谢！

由于编者水平有限，加之时间仓促，书中难免存在疏漏之处，恳请广大读者批评斧正！

娄　岩

2017 年 11 月

目　录



第 1 章
大数据概论

导学

本章主要介绍大数据的技术架构、大数据分析的 4 种典型工具及大数据未来的发展趋势。通过对本章的学习，读者可以更好地了解大数据技术。

了解：大数据未来的发展趋势、大数据隐私和安全问题。

掌握：大数据的核心，大数据的数据格式、基本特征，大数据的技术架构，大数据分析的 4 种典型工具，大数据在医学领域的应用。

随着现代科技的发展，人们对数据的认识及处理能力不断提高，开始挖掘、利用、存储、开发、分析大数据（Big Data），从而造福于社会。大数据即目前人们认识的数据全部，其来源广泛，数据格式多元化，用传统的数据挖掘和处理技术已无法对其进行处理，如非结构化，时间敏感或信息量巨大，无法通过关系数据库引擎进行处理的数据。这些类型的数据，需要采用不同的方法和实时且具有分布式处理能力的并行硬件设备来处理。

大数据究竟是什么？有哪些相关技术？医学大数据如何应用？大数据未来的发展趋势如何？本章将一一介绍这些问题。

1.1 大数据技术概述

从技术层面上看，大数据无法用单台计算机进行处理，必须采用分布式计算架构，其特色在于对海量数据的挖掘、分析和处理。同时，大数据必须依托一些现有的数据处理方法，如云式处理、分布式数据库、硬件设备的并行处理等。

大数据在改变人类生活与思考方式的同时，也在推动人类信息管理准则的重新定位。大数据正以不可阻拦的磅礴气势，与当代同样具有革命意义的最新科技（如虚拟现实技术、增强现实技术、人工智能和移动平台应用等）一起，揭开人类新世纪的序幕。

大数据时代已悄然来到我们身边，并渗透到我们每个人的日常生活之中，谁都无法回避。它提供了光怪陆离的全媒体、难以琢磨的云计算、无法抵御的虚拟仿真环境和随处可在的网络服务。随着互联网技术的蓬勃发展，我们一定会迎来大数据的智能时代，即大数据技术和生活紧密相连，它不再只是人们津津乐道的一种时尚，而会成为人们生活中的向导和助手。中国大数据市场的应用如图 1-1 所示。

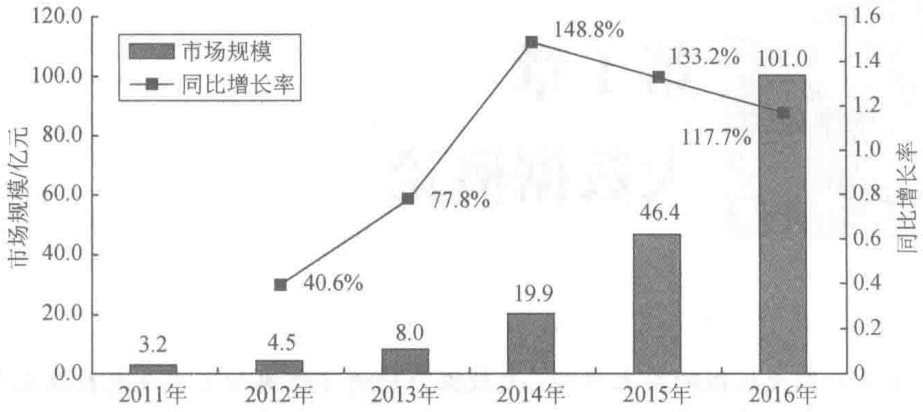


图 1-1 中国大数据市场的应用

1.1.1 大数据的主要来源

大数据的来源非常广泛，如信息管理系统、网络信息系统、物联网系统、科学实验系统等，其数据类型包括结构化数据、半结构化数据和非结构化数据。

1）信息管理系统：企业内部使用的信息系统，包括办公自动化系统、业务管理系统等。信息管理系统主要通过用户输入和系统二次加工的方式产生数据，其产生的大数据大多数为结构化数据，通常存储在数据库中。

2）网络信息系统：基于网络运行的信息系统。网络信息系统是产生大数据的重要方式，如电子商务系统、社交网络、社会媒体、搜索引擎等是常见的网络信息系统。网络信息系统产生的大数据多为半结构化数据或非结构化数据。

3）物联网系统：物联网是新一代信息技术，其核心和基础仍然是互联网，是在互联网基础上延伸和扩展的网络，其用户端延伸和扩展到了物品与物品之间进行信息交换和通信，通过传感技术获取外界的物理、化学、生物等数据信息。

4）科学实验系统：主要用于科学技术研究的补充。可以由真实的实验产生数据，也可以通过模拟方式获取仿真数据。

1.1.2 大数据的核心

大数据的核心就是预测，它使得我们分析信息时需要从不同于以往的角度看待问题。

1. 全新的数据处理理念

1）不只是随机样本，而是全体数据。过去由于受制于技术只能收集与分析随机样本，但是在大数据时代，收集与分析全体数据已成为可能。在大数据时代，我们可以分析更多的数据，甚至可以处理和某个特别现象相关的所有数据，而不再依赖于随机采样，即样本就是总体。

2）不再追求精确性，而是混杂性。大数据时代追求大量数据，而非精确数据，但由于传统处理的信息量较少，因此传统处理方式对数据精确性要求很严格。随着数据量的增加，数据错误率也可能增加，格式也不再单一，只有 5%的数据是结构化数据且适

用于传统统计方法，95%的数据是非结构化数据。因此，只有接受不精确性才能利用大量的数据。由此我们可以断言，大数据时代利用数据快速找出事物的规律更重要。

3）重视的不再是因果关系，而是相关关系，即大数据时代不再热衷于寻找因果关系。

2. 预测

大数据的核心是建立在相关关系分析基础上的预测。相关关系是 A 与 B 经常一起发生，即只要注意到 B 发生，就能预测 A 的发生。

3. 数据价值的获取方式

数据的价值来源于万物数据化和数据交叉复用，大数据的重要价值在于数据深挖掘。

1）数据化。一切事物都可量化，变为数据。数据化不是数字化，数字化即模拟数据转换成用“0”和“1”表示的二进制码，如书页的扫描，无法检索内容；而数据化就是把一种现象转换为可制表分析的量化形式的过程，如书变成数据化文本，可检索。数据化的重点是由 T（Technology，技术）转变为 I（Information，信息）。

2）更有价值。数据价值不会随使用次数的增多而减少，可以重复挖掘。其潜在价值可通过下述 6 种方式释放：数据再利用、重组数据、可扩展数据、数据的折旧值、数据废气、开放数据。

3）角色定位。大数据早期价值来自思维和技术，大数据中后期价值必须从数据本身中挖掘。大数据价值链中主要存在 3 种公司：基于数据本身的公司、基于技能的公司和基于思维的公司。

4. 大数据的安全问题

大数据时代，危险不再是隐私的泄露，而是被预知的可能性，因此需要新的规章制度应对大数据时代的各种隐忧。应用得当，大数据是合理决策的有力武器；应用不当，大数据会变成损害民众利益的工具。大数据时代，告知与许可、模糊化和匿名化三大隐私保护策略都失效。挣脱大数据的困境，是大数据时代人类共同的战争。

1.1.3　大数据的处理流程

大数据的处理流程可以定义为在适合工具的辅助下，对不同结构的数据源进行汲取和集成，并将结果按照一定的标准统一存储，再利用合适的数据分析技术对其进行分析，最后从中提取有益的知识并利用恰当的方式将结果展示给终端前的用户。大数据的处理流程如图 1-2 所示。

图 1-2　大数据的处理流程

例如，分布式并行处理运算如图 1-3 所示。

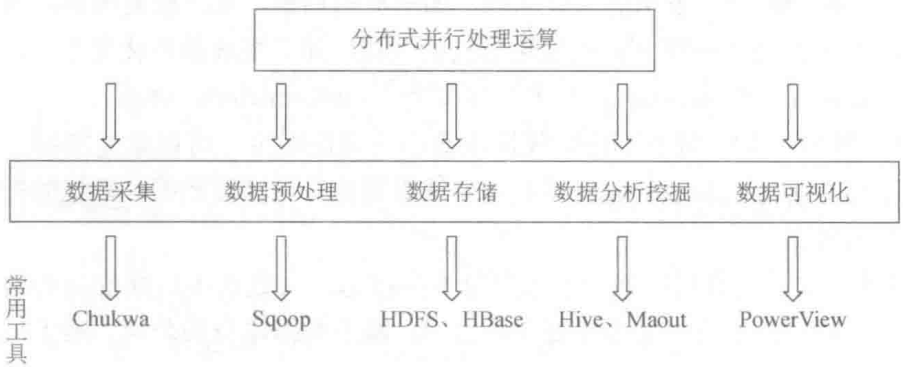


图 1-3　分布式并行处理运算

1. 数据采集

由于大数据处理的数据来源类型广泛，因此其第一步是对数据进行抽取和集成，从中找出关系和实体，经过关联、聚合等操作，再按照统一的格式存储数据。现有的数据抽取和集成引擎有 3 种：基于物化或 ETL（Extraction Transformation Loading，数据抽取、转换和装载）方法的引擎、基于中间件的引擎、基于数据流方法的引擎。

2. 数据预处理

数据预处理的目的是提高数据质量，以便进行数据分析。数据预处理有多种方法，如数据清理、数据集成、数据变换和数据归约等。在数据分析之前使用这些数据处理技术，大大提高了数据分析结果的质量，减少了数据分析所需的时间。

3. 数据存储

大数据存储与管理指用存储器把采集到的数据存储起来，建立相应的数据库，并进

行管理和调用。大数据存储与管理重点解决复杂结构化、半结构化和非结构化大数据管理与处理技术，主要解决大数据的可存储、可表示、可处理、可靠性及有效传输等几个关键问题。开发可靠的分布式文件系统（Distributed File System，DFS）、能效优化的存储、计算融入存储、大数据的去冗余及高效率低成本的大数据存储技术；突破分布式非关系型大数据管理与处理技术、异构数据的数据融合技术、数据组织技术；研究大数据建模技术；突破大数据索引技术；突破大数据移动、备份、复制等技术；开发大数据可视化技术。开发新型数据库技术，主要指的是 NoSQL（Not Only SQL）数据库，分为键值数据库、面向列存储数据库、面向图形存储数据库及面向文档存储数据库等类型。开发大数据安全技术，是指改进数据销毁、透明加解密、分布式访问控制、数据审计等技术，突破隐私保护和推理控制、数据真伪识别和取证、数据持有完整性验证等技术。

4. 数据分析挖掘

数据分析技术包括改进已有数据挖掘和机器学习技术，开发数据网络挖掘、特异群组挖掘、图挖掘等新型数据挖掘技术，突破基于对象的数据连接、相似性连接等大数据融合技术，突破用户兴趣分析、网络行为分析、情感语义分析等面向领域的大数据挖掘技术。数据挖掘就是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中，提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程。数据挖掘涉及的技术方法很多，有多种分类方法。根据挖掘任务，数据挖掘技术可分为分类或预测模型发现、数据总结、聚类、关联规则发现、序列模式发现、依赖关系或依赖模型发现、异常和趋势发现等。

从挖掘任务和挖掘方法的角度来看，数据挖掘技术包括以下 5 个方面：

1）可视化分析。数据可视化无论对于普通用户还是数据分析专家，都是最基本的功能。数据图像化可以让数据“自己说话”，让用户直观地感受到结果。

2）数据挖掘算法。数据图像化是将机器语言翻译给人看，而数据挖掘就是机器的母语。我们可以通过分割、集群、孤立点分析等各种算法精炼数据，挖掘价值。这些算法一定要能够应付大数据的量，同时应具有很高的处理速度。

3）预测性分析。预测性分析可以让分析师根据图像化分析和数据挖掘的结果作出一些前瞻性判断。

4）语义引擎。语义引擎需要达到较高的人工智能水平，以满足从数据中主动地提取信息的要求。语言处理技术包括机器翻译、情感分析、舆情分析、智能输入、问答系统等。

5）数据质量和数据管理。数据质量和数据管理是管理的优秀实践，透过标准化流程和机器对数据进行处理，可以确保获得一个预设质量的分析结果。

5. 数据可视化

数据可视化主要是指借助图形化手段，清晰有效地传达与沟通信息。数据可视化技术的基本思想是将数据库中的每一个数据项作为单个图元元素表示，大量的数据集合构成数据图像，同时将数据的各个属性值以多维数据的形式表示，可以从不同的维度观察

数据，从而对数据进行更深入的观察和分析。使用可视化技术可以将处理结果通过图形方式直观地呈现给用户，如标签云、历史流、空间信息等。

1.1.4 大数据的结构类型

从信息技术（Information Technology，IT）角度来看，大数据的结构类型大致经历了3个阶段，即结构化信息阶段、半结构化信息阶段和非结构化信息阶段。必须注意的是，旧的阶段仍在不断发展，如关系数据库的使用。因此3种数据结构类型一直存在，只是在不同阶段，其中一种结构类型主导其他结构类型。

1）结构化信息：可以在关系数据库中找到，多年来一直主导着IT应用，是关键任务OLTP（On-Line Transaction Processing，联机事务处理）系统业务所依赖的信息。另外，结构化信息还可对结构数据库信息进行排序和查询。

2）半结构化信息：包括电子邮件、文字处理文件及大量保存和发布在网络上的信息。半结构化信息以内容为基础，可以用于搜索，这也是Google（谷歌）等搜索引擎存在的理由。

3）非结构化信息：在本质形式上可认为主要是位映射数据，数据必须处于一种可感知（如可在音频、视频和多媒体文件中被听到或看到）的形式中。许多大数据是非结构化的，其庞大的规模和复杂性需要高级分析工具来创建或利用的一种更易于人们感知和交互的结构。

1.1.5 大数据的基本特征

从多种类型的数据中快速获得有价值的信息的能力，就是大数据技术。

大数据呈现出“4V1O”的特征，具体如下：

1）数据量大（Volume）：大数据的首要特征，包括采集、存储和计算的数据量非常大。大数据的起始计量单位至少是100TB。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。

2）多样化（Variety）：表示大数据种类和来源多样化，具体表现为网络日志、音频、视频、图片、地理位置信息等多类型的数据。多样化对数据的处理能力提出了更高的要求，其编码方式、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。

3）数据价值密度低（Value）：表示大数据价值密度相对较低，需要很多的过程才能挖掘出来。随着互联网和物联网的广泛应用，信息感知无处不在，信息量大，但价值密度较低。如何结合业务逻辑并通过强大的机器算法挖掘数据价值，是大数据时代最需要解决的问题。

4）速度快，时效高（Velocity）：随着互联网的发展，数据的增长速度非常快，处理速度也较快，时效性要求也更高。例如，搜索引擎要求几分钟前的新闻能够被用户查询到，个性化推荐算法要求实时完成推荐，这些都是大数据区别于传统数据挖掘的显著特征。

5）数据是在线的（On-Line）：表示数据必须随时能调用和计算。这是大数据区别于传统数据的最大特征。大数据不仅大，更重要的是数据是在线的，这是互联网高速发