大数据科学与应用丛书

健康医疗大数据
理论与实践

主编 卢朝霞　副主编 姚勇 尹新

近距离行业洞察+35年行业积累+20例深度揭秘
倾囊相授真实实践探索

中国工信出版集团
電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

大数据科学与应用丛书

健康医疗大数据理论与实践

主编　卢朝霞　　副主编　姚　勇　尹　新

电子工业出版社
Publishing House of Electronics Industry
北京 · BEIJING

内 容 简 介

本书围绕健康医疗大数据的理论与实践展开论述。全书共分为 7 章：第 1 章主要描述大数据的基础知识、通用技术以及技术发展趋势；第 2 章主要对健康医疗大数据的概念、特征、分类、主要应用技术、国内外发展现状以及应用需求进行系统阐述；第 3 章～第 6 章分别对临床大数据、精细化运营大数据、健康管理大数据以及基因检测大数据的应用实践案例进行详细论述；第 7 章对健康医疗大数据的未来发展趋势进行展望。

本书是很多应用实例和经验的总结，案例丰富翔实，将理论与实际紧密结合，对互联网技术人员、健康医疗行业的从业人士，以及高等院校相关专业的学生均有很大帮助。

图书在版编目（CIP）数据

健康医疗大数据 ：理论与实践 / 卢朝霞主编. —北京 ：电子工业出版社, 2017.7
（大数据科学与应用丛书）
ISBN 978-7-121-31486-5

Ⅰ. ①健… Ⅱ. ①卢… Ⅲ. ①医学－数据处理 Ⅳ. ①R319

中国版本图书馆 CIP 数据核字（2017）第 096930 号

责任编辑：王敬栋
印　　刷：三河市双峰印刷装订有限公司
装　　订：三河市双峰印刷装订有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　邮编　100036
开　　本：720×1 000　1/16　印张：16.25　字数：266 千字
版　　次：2017 年 7 月第 1 版
印　　次：2017 年 7 月第 2 次印刷
定　　价：49.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888，88258888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：（010）88254459；qianwy@phei.com.cn。

推荐序

当前，我国社会整体信息化程度不断加深，以云计算、物联网、移动互联网和人工智能为代表的新一代信息技术对健康医疗事业的革命性影响日趋明显，海量的数据资源正在以超乎人们想象的速度进行积累和汇聚。可以说，我们已经真正进入了“大数据时代”。以美国、英国为代表的发达国家已经将健康医疗大数据作为国家公共事业的重要组成部分，投入了大量的人力和物力发展健康医疗大数据，把对数据的利用看作衡量国家经济发展的新指标。我国政府同样高度重视健康医疗大数据的发展。2016 年 6 月，国务院正式印发了《关于促进和规范健康医疗大数据应用发展的指导意见》，首次将健康医疗大数据定位为“国家重要的基础性战略资源”。可以预见，健康医疗大数据的应用与发展势必带动我国医疗服务模式的深刻变革和健康服务新业态的发展，极大地提升医疗健康服务的质量和效率，不断满足人民群众多层次、多样化的健康需求，为实现“健康中国 2030”的宏伟目标提供有力支撑。

我国经过了二十余年的医疗卫生信息化建设，已经积累了非常丰富的数据资源。但是与发达国家相比，我国对健康医疗大数据的应用还处于初步的探索阶段，应用水平整体不高。医疗机构掌握着大量的数据资源，但往往不知道该如何让数据发挥真正的价值。目前我国出版的健康医疗大数据的书籍，大多还是从宏观层面和理论层面进行论述，涉及数据应用与实践的比较少，相关内容还不够具体和丰富。而本书最大的不同之处，在于能够通过专业的分

析、丰富的案例和深入浅出的技术语言，生动地展现大数据在临床科研、机构运营、健康管理（体检）、基因检测等诸多细分领域的应用背景、设计思想、应用过程、创新点以及最终的应用效果。例如在健康管理（体检）领域，通过阅读本书，读者能够清晰地了解到应该如何对健康体检、慢性病和睡眠监测的相关数据进行采集、分析和挖掘，最终为用户创造价值，这对于我国健康服务业的从业者来说具有很好的参考意义。大数据的具体应用，不仅能够有效解决我国健康服务业同质化竞争严重的问题，而且可以推动医疗模式从“被动治疗”向“主动预防”转变，真正实现对全人群的健康信息覆盖和全面、全程、全生命周期健康管理的目的。

本书之所以能够具备很强的应用性和实践性，与作者团队扎实的技术功底和丰富的实战经验是分不开的。尤其是本书的主编卢朝霞女士，作为东北大学的资深教授，我国大健康和信息化领域著名的专家学者，以及东软集团的高级副总裁，她既拥有非常深厚的专业背景和理论功底，同时也拥有数十年所积累的非常丰富的企业实践经验。应该说，以卢朝霞教授为首的专业团队在健康医疗大数据方面做出了大量有意义的探索和实践，最终将经验汇总凝练形成本书。在大数据时代已经到来的今天，本书的出版将为健康医疗领域的从业者和专业人士提供有价值的参考和借鉴。作为我国健康管理学界的一位老兵和健康大数据的追随者，我愿以此为序，向广大读者推荐本书。

中华医学会健康管理学分会前主任委员

中关村新智源健康管理研究院院长

武留信

2017 年 4 月

前 言

我自 1978 年起在东北大学系统地学习计算机技术相关知识，随后留校任教，专注于计算机应用技术的研究，和中国最早一批计算机领域的专家、学者一起在浩瀚而神秘的知识海洋中探索数据之广袤、编程语言之神奇、算法之奇美，并有幸成为其中一员，出版了几本数据库应用方面的书。

1995 年我作为东北大学最年轻的女教授，加入了东软集团，期望着将之前在高校的理论学习与研究成果在实践中得以应用，期望着可以更直接地以信息化为祖国构建发展腾飞之翼。如今我已加盟东软二十多年，和东软人一起，始终坚持围绕国计民生等社会发展大趋势进行规划布局，积极响应国家“信息惠民”、“信息消费”、“发展健康服务业”、“健康中国”等政策，怀着“以信息化助力实现中国梦”的理想，在大健康领域精耕细作，先后为 4 亿社保人群、30 多个省市的卫生厅局、2 000 多家大型医疗机构、30 000 多家基层医疗卫生机构提供有力支撑，并不断探索新业务形态，创新商业模式，布局产业生态，力争做时代发展的引领者和创造者。

目前，面对全球新经济、新技术、新消费的发展趋势，特别是全球人口结构的变化，健康、医疗、养老等产业与云计算、大数据、互联网、人工智能等新一代信息技术的结合，与金融保险行业的结合，与智能制造的结合，与共享经济模式的结合，将会创造一个巨大的产业发展和就业机会。我们将这样的产业融合称为“大健康产业生态”。而

在这个“大健康产业生态”之中，有巨量的信息像血液一样在各个组织之中或之间不停地流动，并不断地产生营养、创造价值，我们将这些“巨量的信息”合称为“健康医疗大数据”。

如今，在国外，健康医疗大数据的发展如火如荼，应用遍地开花，其生态系统相对成熟；而我国健康医疗大数据处于起步阶段和发展初期，国家已发出“促进和规范健康医疗大数据应用发展”的政策号角。东软，作为一家投身健康医疗信息化建设二十余年的民族软件企业，有责任，更有义务依托自身的优势和积累，借鉴国外成熟的经验，积极探索发展适合我国国情的健康医疗大数据应用，为祖国健康医疗大数据的发展与振兴再献绵薄之力。同时，我作为一名大学教授，作为在健康医疗信息化领域奋战数十载的实践者，更有一种使命驱动着我，那就是要把目前我国在健康医疗大数据领域的典型应用案例及实战经验，进行梳理与有效总结，与众人分享，进行知识的传播与传递。因此，我组织了国内外优秀的科研人员、高校教师、知名企业人员、东软团队骨干等共同编撰了本书，希望通过本书的出版，能够理清大数据、健康医疗大数据的基本概念，并通过缜密的分析以及翔实的实践案例，重点阐述健康医疗大数据在相关领域的应用实践以及未来的发展趋势。

为了能够透彻阐述，让读者充分了解每个案例实践，我们力争从以下几个方面入手，进行较为深入的介绍：第一，介绍实践案例的应用背景，例如恶性肿瘤大数据分析的应用背景，要介绍清楚恶性肿瘤的危害、严重程度以及国内外发展情况等；第二，介绍实践案例的设计思想与总体框架，让读者清楚地了解该应用为什么要设计，设计的时候出于哪些角度考虑，如何进行数据的抽取，底层采用了哪些模型以及应用的总体框架结构等；第三，介绍实践案例的数据建模与算法优化，从技术角度介绍清楚采集获取数据之后，如何进行数据的清洗转换，如何进行数据模型的建立，采用了哪些数据算法，针对算法进行了哪些优化等；最后，介绍实践案例所取得的效果，通过具体的数据有效地论述案例所达到的实际效果。因此，本书是很多应用实例和经验的总结，案例丰富翔实，将理论与实际紧密结合，希望能够为健康医疗大数据领域相关人员提供有价值的参考，以此达到传道授业解惑的目的。

本书共分为 7 章。第 1 章主要描述大数据的基础知识、通用技术以及技术发展趋势；第 2 章主要对健康医疗大数据的概念、特征、分类、主要应用技术、国内外发展现状以及应用需求进行系统阐述；第 3 章从恶性肿瘤大数据分析、药物应用大数据分析、疾病辅助诊断分析三个方面，对临床大数据应用的实践案例进行详细论述；第 4 章详细论述精细化运营大数据的应用背景、设计思想、应用案例以及应用效果；第 5 章从健康体检大数据分析、慢病管理大数据分析、睡眠大数据分析三个方面，对健康管理大数据应用的实践案例进行详细论述；第 6 章从精准医疗、“电子病历与基因组学”两个领域，对基因检测大数据应用的实践案例进行详细论述；第 7 章对健康医疗大数据的未来发展趋势进行展望。

本书由卢朝霞主编，姚勇、尹新为副主编，主要编委还包括毕丹、陈禹、窦元珠、何璇、赫阳、刘芬、孙传海、王敏、吴一多、徐华、杨风雷、于洪勇、张一鸣、赵力维（按姓氏拼音排序）。

最后，要特别感谢 IBM、美国得克萨斯州立大学休斯敦健康科学中心以及我所在的团队，感谢所有编委半年多来呕心沥血的付出，保证了本书出版工作的顺利完成。同时感谢本书的读者，感谢你们积极投身健康医疗大数据的应用与实践之中。让我们携起手来，共同推动我国健康医疗大数据的发展，提升健康医疗服务效率和质量，不断满足人民群众多层次、多样化的健康需求，培育新的业态和经济增长点，为实现中华民族伟大复兴的中国梦贡献一份力量。

由于时间有限，书中内容难免存在疏漏，不足之处请多指正。

卢朝霞

2017 年 5 月

目 录

第 1 章　大数据概述

随着新一代信息技术的迅猛发展，无处不在的移动终端、智能设备、无线传感器等，每分每秒都在产生大量的数据。并且，在互联网上数以亿计的用户时时刻刻在产生大量的交互。2016 年淘宝“双十一”当天的销售额高达 1 207 亿元人民币。百度每天大约要处理几十 PB 的数据，Twitter 每天会产生 7 TB 的数据，而 Facebook 每天生成 300 TB 以上的日志数据。这些数据产生的速度快，需要处理的数据量巨大，并且数据的价值也在不断显现。大数据时代的到来，为金融服务、健康、教育、农业、医疗等多个重要领域带来了前所未有的机遇。与此同时，大数据时代的到来也为传统的数据处理技术带来了更大的挑战。大数据处理需要更高的实时性、有效性和安全性，需要融合多个学科的关键技术来满足大数据的发展。

在本章中，我们将重点介绍大数据的概念、特征、分类，所涉及的通用技术，以及大数据技术未来的发展趋势。

1.1 大数据基础知识

1.1.1 大数据概念和特征

随着互联网、移动互联网、物联网、云计算的快速兴起，以及移动智能终端的快速发展，数据的增长速度远比人类社会以往任何时候都要迅速；数据的规模变得越来越大，内容越来越丰富，关系越来越复杂，更新速度越来越快。这些新的特征促使一个新的概念诞生，那就是大数据。

2008 年，《Nature》推出了大数据（Big Data）专刊。计算机社区联盟阐述了在数据驱动背景下解决大数据问题所需的技术及其将面临的一系列挑战。

2011 年，《Science》推出“Dealing with Data”专刊，围绕着科学研究中的大数据问题展开讨论，并说明大数据对于科学研究的重要性。美国数据管理领域的知名专家联合发布了一份白皮书《Challenges and Opportunities with Big Data》，详细分析了大数据产生的原因、处理流程以及大数据所面临的挑战。

2011 年，麦肯锡全球研究院发布的《Big data: The next frontier for innovation, competition, and productivity》正式对大数据进行了定义，即大数据是指在一定时间内无法用传统数据库软件工具采集、存储、管理和分析其内容的数据集合。大数据技术则特指新一代的创新型的技术，能够突破常规软件的限制，是对大数据进行采集、存储和处理的技术的统称。

研究机构 Gartner 认为：大数据是指需要借助新的处理模式才能拥有更强的决策力、洞察发现力和流程优化能力的，具有海量、多样化和高增长率等特点的信息资产。而维基百科认为：大数据指的是需要处理的资料量规模巨

大，无法在合理时间内通过当前主流的软件工具采集、管理、处理并整理的资料，它成为帮助企业经营决策的资讯。

从大数据的概念看，对大数据的概念界定各有各的看法，目前尚未出现一个公认的定义；但大都是从大数据的特征出发，通过对这些特征的阐述和归纳，试图给出其定义。在这些特征中，比较有代表性的是3V定义，即认为大数据需满足 3 个特征——规模性（Volume）、多样性（Variety）和高速性（Velocity），但是这没有体现出大数据的巨大价值。以国际数据公司 IDC 为代表的业界在 3V 的基础上增加价值性（Value）特征，表示大数据虽然价值总量高，但其价值密度低。

因此，目前公认的大数据具有 4V 特征，即数据规模大（Volume）、数据种类多（Variety）、处理速度快（Velocity）及数据价值高密度低（Value），具体如图 1-1 所示。

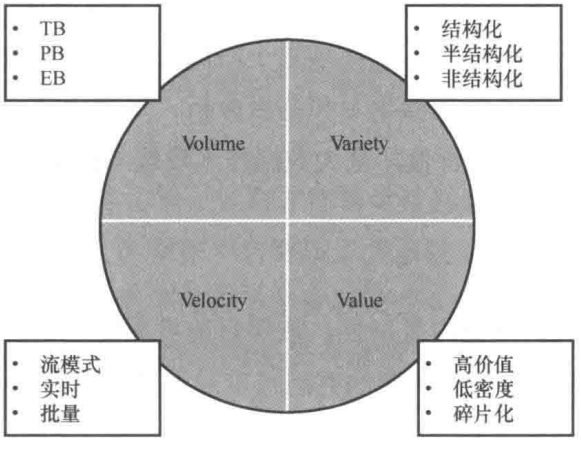

图 1-1　大数据的 4V 特性

1. 数据规模大（Volume）

数据量大是大数据的基本属性，它是指数据的采集、存储和计算的量都非常大；大数据通常指 10 TB 以上规模的数据量。根据 IDC 公司监测，全球数据量大约每两年就翻一番，预计到 2020 年，全球将拥有 40 ZB 的数据，并且 85%以上的数据以非结构化或半结构化的形式存在。

2. 数据种类多（Variety）

随着传感器种类的增多以及智能设备、社交网络等的流行，数据种类也变

得更加复杂。相对于以往便于存储的文本形式或者结构化数据，如今非结构化数据越来越多，包括网络日志、音频、视频、图片、机器数据、地理位置等各种复杂结构的数据，这些多类型的数据对数据的处理能力提出了更高要求。

3. 处理速度快（Velocity）

数据每分每秒都在爆炸性地增长，数据的快速动态变化使得流式数据成为大数据的重要特征。与传统数据挖掘不同，全国用户每天产生和更新的微博、微信和股票信息等数据，随时都在传输，这就要求大数据的处理必须具有较强的实时性，能够实时地查询、分析、推荐等。

4. 数据价值高密度低（Value）

在海量的数据中，存在着巨大的待挖掘的商业价值，然而在数据呈指数增长的同时，隐藏在海量数据中的有用信息却没有按相应比例增长。恰恰相反，挖掘大数据的价值类似于沙里淘金，从海量数据中挖掘稀疏珍贵的信息。例如，商场的监控视频，在连续数小时的监控过程中有用的数据可能仅有几秒。如何通过强大的机器学习和高级分析，迅速地完成大数据价值的提取，挖掘出大数据的应用价值，是大数据技术发展与应用的重点。

1.1.2 大数据分类

为了简化大数据类型的复杂性，按照目前业界比较认可的分类方式，可以按照数据结构和处理数据所需的时间跨度对大数据进行分类。

1. 按照数据结构分类

在信息社会，大数据（信息）可以按其数据结构划分为两大类：一类能够用数据或统一的结构加以表示，我们称之为结构化数据，如数字、符号；而另一类无法用数字或统一的结构表示，如文本、图像、声音、网页等，我们称之为非结构化数据。结构化数据属于非结构化数据，是非结构化数据的特例。

结构化数据（即行数据）存储在数据库里，是可以用二维表结构来表达的数据。而不方便用数据库二维逻辑表来表达的数据，称为非结构化数据，包括所有格式的办公文档、文本、图片、XML、HTML、各类报表、图像和音频/视频信息等。非结构化数据又包含半结构化数据和无结构化数据。

1）结构化数据

结构化数据的特点是任何一列数据不可以再细分，并且任何一列数据都具有相同的数据类型。例如，SQL Server、Oracle、MySQL 等关系型数据库中的数据，均为结构化数据。关系型数据库存储的结构化数据示例如表 1-1 所示。

表 1-1　结构化数据示例

学　号	姓　名	科　目	成　绩
1110371	李颖	数学	83
1110412	王庆	语文	92

结构化数据类型是一种用户定义的数据类型，它包含了一系列的属性，每一个属性都有一个数据类型。属性是专门用来帮助描述类型实例的特性。

2）非结构化数据

非结构化数据库是指其字段长度可变，并且每个字段的记录又可以由可重复或不可重复的子字段构成的数据库，用它不仅可以处理结构化数据（如数字、符号等信息），而且更适合处理非结构化数据（全文文本、图像、声音、影视、超媒体等信息）。

非结构化 Web 数据库主要是针对非结构化数据而产生的，它和以往流行的关系数据库相比，最大区别在于它突破了关系数据库结构定义不易改变和数据定长的限制，支持重复字段、子字段以及变长字段，并实现了对变长数据和重复字段进行处理和数据项的变长存储管理。这使它在处理连续信息（包括全文信息）和非结构化信息（包括各种多媒体信息）中有着传统关系型数据库无法比拟的优势。

半结构化数据是介于完全结构化数据（如关系型数据库、面向对象数据库中的数据）和完全无结构的数据（如声音、图像文件等）之间的数据，HTML 文档就属于半结构化数据。它一般是自描述的，数据的结构和内容混在一起，没有明显的区分。半结构化数据虽然也是结构化的数据，但是其结构变化很大。因为我们要了解数据的细节，所以不能将数据简单地组织成一个文件按照非结构化数据处理；又由于其结构变化很大，也不能够简单地建立一个表和它对应。比如存储员工的简历，每个员工的简历大不相同：有的员工的简历很简单，可能只包括教育情况；有的员工的简历却很复杂，可能包括工作情况、婚