普通高等院校数据科学与大数据技术专业“十三五”规划教材

数据科学与大数据技术导论

方志军 ■ 主编

SHUJU KEXUE YU DASHUJU JISHU DAOLUN

華中科技大學出版社
http://www.hustp.com

普通高等院校数据科学与大数据技术专业“十三五”规划教材

数据科学与大数据技术导论

主　编　方志军
副主编　董新华　俞　雷
　　于　为　黄　勃

华中科技大学出版社
中国·武汉

内容简介

本书以 Python 为主线，按照学习者的知识逻辑展开，呈现数据科学与大数据技术的基本知识、基本概念、基本方法。本书内容主要包括：什么是大数据、Python 基础知识、数据分析与可视化、数据挖掘、机器学习、大数据处理。本书可作为普通高等院校计算机、数据科学与大数据技术、人工智能等专业的教材，也可作为数据科学、大数据技术、数据管理及应用等方面的自学教材或参考书。

图书在版编目(CIP)数据

数据科学与大数据技术导论/方志军主编. —武汉：华中科技大学出版社，2019.8
普通高等院校数据科学与大数据技术专业“十三五”规划教材
ISBN 978-7-5680-5220-7

Ⅰ. ①数… Ⅱ. ①方… Ⅲ. ①数据处理-高等学校-教材 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字(2019)第 179248 号

数据科学与大数据技术导论
Shuju Kexue yu Dashuju Jishu Daolun

方志军 主编

策划编辑：李 露 廖佳妮
责任编辑：李 露
封面设计：原色设计
责任校对：阮 敏
责任监印：徐 露
出版发行：华中科技大学出版社(中国·武汉) 电话：(027)81321913
武汉市东湖新技术开发区华工科技园 邮编：430223
录 排：华中科技大学惠友文印中心
印 刷：武汉华工鑫宏印务有限公司
开 本：787mm×1092mm 1/16
印 张：16.25
字 数：406 千字
版 次：2019 年 8 月第 1 版第 1 次印刷
定 价：41.00 元

前言

PREFACE

信息化时代的到来使得人类的生活进入了历史上前所未有的新领域，数据资源的重要性越发突显出来。大数据和人工智能的结合，更是将数据的发掘与利用推向了新的高潮。全球范围内数据人才奇缺，教育部和各大高校也意识到这一问题，因而将数据科学与大数据技术的人才培养与学科建设提上日程。

从2013年起，在国内教育领域掀起了利用大数据来促进教育改革和创新发展的研究热潮，大数据的教育与应用研究迅速发展起来，直接表现为相关论文的数量和质量倍增。2014年3月，教育部办公厅印发的《2014年教育信息化工作要点》指出：加强对动态监测、决策应用、教育预测等相关数据资源的整合与集成，为教育决策提供及时和准确的数据支持，推动教育基础数据在全国的共享。近年来，教育部积极采取措施，加强大数据人才的培养，支持大数据技术产业的发展。自2014年起，为贯彻落实教育规划纲要，创新产学合作协同育人机制，教育部组织有关企业和高校实施产学合作协同育人项目。在相关专业设置方面，2015年本科特设新专业——数据科学与大数据技术；同年10月，教育部公布了新修订的《普通高等学校职业教育（专科）专业目录（2015年）》，主动适应大数据时代发展的需要，新设了云计算技术与应用、电子商务技术专业。2016年，北京大学、对外经济贸易大学、中南大学首次成功申请到“数据科学与大数据技术”本科新专业；2017年，另外32所高校获批；2018年，248所学校获批。

近年来，国内出版的数据处理、数据分析等相关书籍层出不穷，观其内容，不同的专家和学者从不同的角度提出了对于大数据的理解和认识。其中，有的专家侧重在“数据分析”上，重点讨论了统计学、机器学习等相关内容；有的专家侧重在“数据处理”上，重点讨论了数据挖掘、数据库管理等相关内容；有的专家侧重在“数据平台”上，重点讨论了各种计算平台和硬件设备等相关内容。

多样化的观点支撑了多样化的教材，编者团队在面向一线教学的工作中发现，一本涉猎广泛、由浅入深、适合入门者研习和一线教师使用的教材亟待出现。

2018年，华中科技大学出版社不弃浅陋，邀请几位编者参与到本书的编撰工作当中。大家一致认为这是一个非常宝贵的机会，希望能够跟同行们分享多年的心得和体会，也希望能帮助相关专业的学生，使之对大数据领域产生兴趣。

本书以当下大数据发展的最新科研教学成果为基础，从培养学生大数据思维入手组织内容。本书采用“理论＋提升＋实践”的模式，以理解大数据理论为基础，以知识扩展为提升，以数据处理、数据挖掘案例为实践途径，努力做到既促进数据思维的培养，又避免流于形式；既适应总体知识需求，又满足个体深层要求。每章章前设计了学习目标与内容介绍，章

后附有小结和习题。学习目标与内容介绍部分紧密结合教学目标和特点，紧扣教学重点，突出计算思维方法；小结部分对每章知识进行归纳总结、突出重点；习题部分的题目大多选自一些国外经典参考资料，力求使读者全面地巩固所学知识。

在本书的编写过程中，编者从系统的视角介绍了数据科学与大数据技术的相关基础理论和应用，同时注意突出语言文字应用的规范性。在选择内容时，既注意到基础性，又注意吸收比较成熟的、有价值的新成果，同时编写适合教学和巩固知识的习题。本书内容力求保证较强的系统性，对基本概念的阐述力求严谨、清晰，叙述力求通俗易懂，以增强可读性和启发性。

“大数据导论”是计算机科学与技术、数据科学与大数据技术等相关专业本科生的专业课程和其他专业的选修课程，是国内外大学计算机学科教育体系中的核心课程之一。它系统、全面地介绍大数据的基础知识和数据挖掘、数据处理的基础知识及简单应用，使学生能够具备基本理论知识和简单编程的能力，同时提高学生的综合素质与创新思维。

本书第 1 章简述了大数据的基本概念，从“什么是大数据”这一问题入手，从其定义、相关科学、应用领域等方面系统地介绍了大数据的基本概念，简要地对后续章节中所述的内容进行了阐述。第 2 章和第 3 章简要地讲述了 Python 在大数据中的应用，第 2 章对 Python 的基础知识进行了讲解，以便学生能尽快对 Python 有一个简单的认识，第 3 章描述了如何利用 Python 对数据进行处理、分析以及可视化等。第 4 章从如何进行数据挖掘入手，描述了数据挖掘的源起、相关工具以及如何对数据进行存储、利用。第 5 章概述了机器学习的几种算法，从其所讲述的算法中可以看到，无论是传统的机器学习算法还是新兴的神经网络算法，数据是不可或缺的一部分。第 6 章介绍了能够对海量数据进行分析的软件框架——Hadoop，Hadoop 平台释放了前所未有的计算能力，同时大大降低了计算成本。

致本书的使用者：

(1) 学生使用者。本书涉及大数据领域的多个方面，编写团队希望帮助大家尽快入门。为了让大家能够更好地理解相关的知识点，每个部分的写法和阐述方式略有不同。同时，在阅读本书前，希望大家具备一定的线性代数和概率论知识，这样学习本书将轻松许多。

(2) 教师使用者。本书适用于数据科学与大数据技术专业、计算机科学与技术等专业的学生，也适用于不同层次的学生，教师可针对不同学生对知识点和讲授深度有所侧重。

(3) 专业技术人员。本书可以作为专业技术人员的参考书。本书内容宽泛，既涉及软件的安装和配置方法，又涉及大量的常用算法。本书的章节编排有序，方便专业人士直接查询相关内容。

在本书的成书过程中，编写团队和华中科技大学出版社保持了愉快的合作，在此感谢出版社各位编辑的帮助和支持。

本书获得了贵州省科技计划项目(黔科合 LH 字[2017]7049)以及贵州省创新群体项目(黔教合 KY 字[2018]034)的支持，在此表示感谢！

本书由上海工程技术大学方志军教授担任主编，安顺学院于为、湖北工业大学董新华、上海工程技术大学黄勃和俞雷共同参与编写。

本书的编写参考了大量的文献资料，一并向文献作者表示感谢！由于编者水平有限，在内容安排、表达等方面难免存在不当之处，敬请广大读者朋友不吝赐教。

编　者

2019 年 6 月

目 录

CONTENTS

第1章 什么是大数据

本章学习目标

- 了解数据、大数据、数据挖掘的含义
- 理解大数据与统计学之间的关系
- 掌握大数据、机器学习和人工智能之间的关系
- 了解大数据的应用领域

当今时代是一个充斥着庞大信息的时代，身处这样一个时代，如若能够站在数据链的顶端，便能够应用数据来解决一些现实问题，如减少决策误差、量化风险、减少损失，并通过大数据分析解决社会问题等。本章节作为全书的开篇，浅谈了大数据的应用领域、大数据与统计学之间的关系等问题，并在本章结束部分简单介绍了大数据、机器学习与人工智能的微妙联系，希望读者通过这一章的学习能够对大数据有一个大致的了解。

1.1 数据、大数据及数据挖掘

1.1.1 数据

21世纪是一个信息化的时代，作为信息的表现形式和载体，数据是当下研究的主要课题。数据和信息之间是相互依赖、密不可分的，数据是事实或是观察的结果，是客观事物的逻辑归纳，是用于表示客观事物的未经加工的原始素材。

近年来，互联网、物联网及云计算的快速发展带动了几乎所有产业和商业领域的数据急剧增长，数据的存储单位也由B、KB、MB、GB、TB扩充到PB、EB、ZB、YB。据不完全统计，过去三年的信息数据总量比在此之前的所有数据的总和还要多。例如，2003年，科学家为完成对30亿对碱基对的排序，花费了十年的时间，这是人类首次尝试破解人类基因组，而现在，世界范围内的基因仪每15分钟就可以完成与最初十年时间相同的工作量。在金融领域，美国股市每天交易70亿股，其中2/3是由基于数学模型和算法的计算机程序自动完成的，这些程序利用大量数据预算收益和规避风险。互联网公司的数据增速之快，让人叹为观

止，简直可以说是“数据风暴”了，Google 公司每天要处理的数据量超过 24PB，处理的数据量是美国国家图书馆的纸质出版物的数千倍。Facebook 这个不过于 2004 年上线的社交网路服务网站每天更新的照片量已超过 1000 万张，每天人们在网站上留下评论或点击“喜欢”、“不喜欢”按钮大约三十亿次，这些评论就成为了 Facebook 公司挖掘用户喜好的数据线索。除此之外，世界最大的视频网站 YouTube 每周的访问量高达两亿人次，平均每秒都有一段时长超过 60 分钟的视频被上传至网站，而 Twitter 每天都有多达 4 亿条动态信息要发布，并且每年的信息量都会翻一倍。

大数据时代萌生了一些专属于它的名词，如图 1-1 所示，这个时代的新生词汇有“人工智能”、“商业智能”、“神经网络”等，它们彰显了这个时代的特征。

图 1-1　大数据时代

数据的急剧增长形成了庞大而又复杂的数据王国，想要从这些冗杂的数据中提取有用信息，首先要进行的就是分类存储，形成庞大的数据集。这些容量足够大的数据集就是“大数据”。

1.1.2　大数据

大数据是由数目庞大、结构复杂、类型繁杂的数据组成的数据集合，是基于云计算的数据处理与应用模式，通过对数据的整合共享、交叉复用，形成智力资源和知识服务能力。对于大数据特点的描述有很多，其中最著名的是 Gartner 公司的分析师道格·兰尼（Doug Laney）提出的 3V 特征：Volume（数据规模）、Velocity（数据转输速度）和 Variety（数据形式）。Volume 意在描述数据集的规模；Velocity 是形容数据产生速度的参数；Variety 是指数据的种类和资源的多样性。虽然这三个特性能够用来描述所有数据集的特征，但是有时 3V 特征并不能够很好地诠释一些数据集的特点，因此人们根据特殊需求又加了第四项特征，即 Value（数据价值），如图 1-2 所示。通俗来讲，大数据就是一个庞大的、种类繁多的数据集合体，是一个无论用传统数据

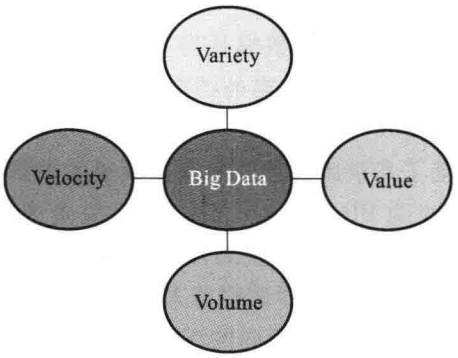

图 1-2　4V 特征图

处理平台或技术，还是新式数据处理方法都难以圆满处理和加工的数据库。随着对大数据的逐步了解，在 2012 年，Gartner 公司进一步提出了更为详尽的定义："大数据是大容量、高速度和多样化的信息资源，它需要一种新的加工形式来提高决策力、洞察力和过程优化技术对其进行获取、管理、分析、可视化，这样的一个数据集就能够称之为大数据。"

从 2012 年开始，大数据就成为了 IT 界的热点词汇。目前，大数据已经成为学术界、商界乃至政治界的新宠。在 2013 年，Gartner 公司列举了"2013 年十大战略技术趋势"和"未来五年十大关键技术趋势"，大数据均在这两项列表中名列第二。由此，我们可以推断大数据能够在生活中的很多方面掀起"革命"的潮流，如在商业、科学研究、公共管理领域等。

1.1.3　数据挖掘

这是一个数据"疯狂增长"的时代，数据量不仅巨大，而且种类繁多，比如传感器网络、科学实验、高通量仪器等，无时无刻不在进行着数据的更新，这些数据是呈指数形式增长的，而有价值的信息就隐藏在这冗杂而又庞大的数据中，因此需要利用数据挖掘技术来探查大型数据库，以获取有用的信息。

挖掘技术的思想主要起源于统计学中的抽样、估计及假设检验，但同时它又以机器学习、人工智能和模式识别的学习理论、建模技术和搜索为依据。挖掘技术以算法为依据，此外还融合了来自信号处理、最优化、可视化、进化计算、信息论和信息检索等领域的思想。目前数据收集和数据存储技术已经能够满足几乎所有组织机构积累海量数据的需求，而如何从数据库中提取有效信息才是我们研究的主题。数据挖掘是数据库中的知识发现(Knowledge Discovery in Database，KDD)不可或缺的一部分。

数据挖掘是一套从数据中提取有效信息的技术，其中包括聚类分析、分类、回归和关联规则学习。它所涉及的统计和机器学习的方法是其根基。与传统的数据挖掘算法相比，大数据挖掘具有更精准的预测性，也因此更具挑战性。以聚类为例，对大数据进行聚类的一种自然方式是扩展现有的方法(如分层聚类、k-Means 和模糊 C 均值等)，使它们能够应付巨大的工作量。这种聚类算法包括 CLARA(大型应用聚类)算法、CLARANS(基于随机搜索的大型应用)算法和 BIRCH(使用动态建模的多阶段层次聚类)算法等，其他的工作将在后面逐步展开阐述。数据处理的步骤如图 1-3 所示，KDD 是将还未进行加工的数据加以处理，进而转换为有用信息的过程，即需历经数据预处理、数据挖掘、后处理这三个步骤。

图 1-3　数据处理的步骤图

输入数据的存储形式多种多样，并且它们既能够保存在数据存储库中，也能够散布在多个站点上。数据预处理所涉及的步骤包括融合来自多个数据源的不同类型的数据、清洗数据以消除噪声和重复的观测值、保留与当前数据挖掘任务相关的记录和特征以便对数据进行挖掘。图 1-4 所示的为数据挖掘过程的具体步骤。

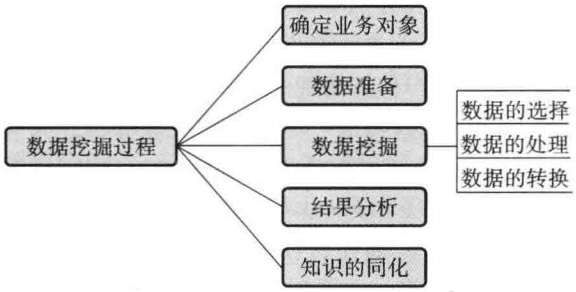

图 1-4 数据挖掘过程的具体步骤

1.1.4 数据挖掘的技术基础

数据挖掘可以说是近年来数据库应用技术中十分热门的话题。但其所用的诸如数据分割(Data Partition)、预测模型(Prediction Model)、偏差侦测(Deviation Detection)、链接分析(Link Analysis)等并不是新兴技术,早在第二次世界大战之前,美国就已将其运用在军事及人口普查等方面。随着信息科技的飞速发展,众多新的计算机分析工具出现,如模糊计算(Fuzzy Computing)、关系数据库(Relational Database)、神经网络(Neural Network)及遗传算法(Genetic Algorithm)等,这些工具的出现使得从数据中发掘"宝藏"成为可能。

一般来说,数据挖掘的理论技术可以分为传统技术和改进技术。传统技术以统计分析为主,统计概率理论包含的序列统计、回归分析和分类数据分析等均属于传统的数据挖掘技术。数据挖掘的对象为变量繁多的大样本数据,因此高等统计学包含的多变量分析中用来精简变量的因素分析(Factor Analysis)、用来分类的判别分析(Discriminant Analysis),以及用来区隔群体的分群分析(Cluster Analysis)等,在数据挖掘过程中经常被用到。

在技术改良方面,应用较为广泛的有决策树(Decision Tree)、神经网络及归纳法(Rule Induction)等。决策树是一种用树型结构图表示数据受各变量的影响情况的预测模型。它是根据对目标对象的不同影响构建的分类规则,一般用于对客户数据的分析。例如,对邮寄对象是否有回函或没有回函进行划分,找出影响其分类结果的变量组合。

常见的分类方法有分类回归树(Classification and Regression Tree,CART)和卡方自动交互检测法(Chi-Square Automatic Interaction Detector,CHAID)等。

神经网络是模拟人脑思维结构的数据分析模型。先输入变量,然后将结果用于自我学习,根据学习经验获得知识,最后通过常数参数优化构造相应的数据模型。神经网络是一种非线性设计。与传统的回归分析相比,它不需要限制固定的分析模式,特别是当数据变量之间存在交互作用时,可以自动进行检测;而其缺点则在于它的分析过程为一个黑盒子,因此常无法以可读的模式呈现,而且每阶段的加权与转换也是不明确的,所以神经网络常用于高度非线性,且带有相当程度的变量交感效应的数据处理。神经网络应用于人工智能领域,成为了一类极具代表性的方法。

归纳法是知识发掘领域中最常见的方式,它是由一连串的"if/then"(如果…/则…)组成的,应用逻辑规则是对数据进行划分的技术,在实际应用中确定规则的有效性是最大的难题,我们通常需要先将数据中发生次数太少的项目剔除,以防止出现无意义的逻辑规则。

如图 1-5 所示,基于数据挖掘技术能够进行风险预测、业务创新、销售预测、数据挖掘、

需求挖掘、用户行为分析、智能决策等。

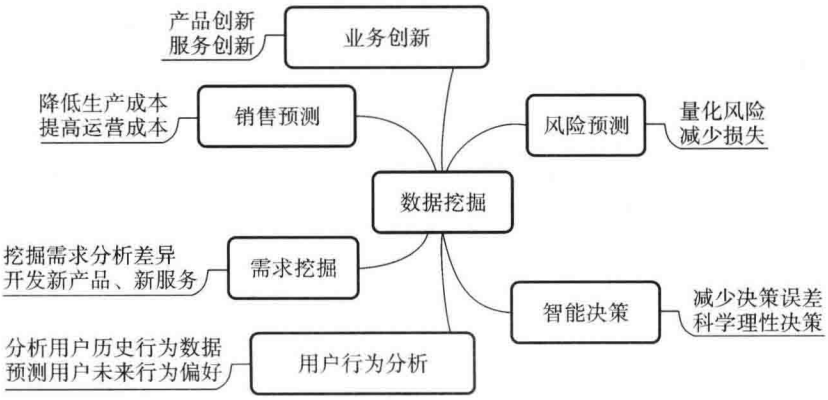

图 1-5　大数据分析决策价值

1.2　大数据与统计学

数据分析与数据统计是密不可分的，在当今这个数据量急速膨胀的时代，统计学的地位显得尤为重要。从本质上来讲，统计是一种系统探究的方法，用来分析离散的、不确定的数据的性质，即从数据库中取出一部分样本，分析其性质，以此来推测数据库的性质。

统计系统一般可分为两类：描述性统计和推论性统计。将数据收集起来，作图、作表、求平均值，用类似这种方式处理数据进而得出结论的方法叫作描述统计学。从数据库中提取一部分样本，通过分析样本的特点去推论总体的特点，这种运用推论的手段去分析数据的方法叫作推论统计学。

随着大数据的兴起，统计数据越来越具有吸引力，这是因为在计划经营策略、市场策略、新产品开发、新业务等方面运用统计分析取得了显著成效。当今时代，经营、决策已不能只单单靠经验了，而必须根据以数据为基础的科学分析方法来进行决策。

根据以往的认知，统计学属于数学的一个分支，但从本质上来看，统计学与数学是处在对立位置上的。这是因为数学是利用已知的公式、定理，通过计算的方法得到确定答案的学科，是一种演绎推理。而统计学是从收集到的零散数据中推导出一般性质的学科，是一种归纳推理。1662 年，英国社会学家约翰·格兰特(John Grant)发表了一篇论文，简要分析了过去 60 年伦敦人口变化与死亡原因之间的关系。他列举出不同死因的人口比例等，进而对死亡率与人口寿命做了分析，这些都是通过观察数据、收集数据、分析数据得到的。同时，通过对数据的观察和分析，他提出新生儿性别比例具有稳定性的论断。这篇文章一经发表就得到了广泛的关注，其涉及的统计学方法也引起了科学家们的关注，由此，统计学开始走进大众的视野。近年来，由于信息技术的迅速发展，一些企业开始应用大数据分析来帮助其进行运营、决策。大数据最鲜明的特点是具有大样本和高维度，针对样本大的问题，统计学采用特别的方式抽样，能做到既减少工作量，又保证必要的精度。

统计技术用于挖掘不同目标之间的关联和因果关系。然而,传统的统计技术实际上并不适合用来管理大数据。许多研究人员提出对古典的技术加以扩展或找寻一个全新的方法来解决现存的问题。例如,Oleg 提出了大规模多元线性回归的有效近似算法。该算法是对输入变量进行重新估计单调函数的一种方法。

1.3 机器学习与人工智能

1.3.1 机器学习与人工智能简介

“人工智能”一词于 1956 年在达特茅斯会议(Dartmouth Conferences)上被提出,自此,关于人工智能的天马行空的想象便不曾停止过。与此同时,研究人员也从不曾停下追逐人工智能的脚步,此后几十年间,人工智能先是被当作人类未来文明的钥匙被追捧,而后又被认为是不切实际的异想天开被摒弃。

但在近几年,人工智能呈现了“爆炸”式的发展,尤其是在 2015 年以后,人工智能的发展掀起了一阵热潮。这主要是由于图形处理器(GPU)的出现使得图形处理更迅速,图形处理器的性价比更高,与之前的技术相比功能更强大。

在达特茅斯会议上,人工智能的先驱们提出了人工智能的研究方向,他们希望能够通过当时新兴的计算机制造出与人类相似的机器。这是一部神奇的机器,它拥有感官、推理能力及人类的思维方式。在电影中已经出现过这样的机器人,例如友好的 C-3PO,及人类的敌人——终结者。虽然我们对机器人这一词汇并不陌生,但是人工智能机器人至今仍只存在于电影和科幻小说里,理由很简单:依靠目前的科技还实现不了“强人工智能”。

若说此前大众对于人工智能的了解仅仅在知道“机器人”这一词汇的层面,那么 2015 年 11 月 9 日,Google 公司发布了人工智能学习系统 TensorFlow。一夜之间,“人工智能”和“机器学习”这两个生僻的词汇传遍大街小巷。机器学习是一种人工智能算法,它允许软件通过分析大量数据来解释或预测未来的情景。如今,科技巨头正在大举投资机器学习的相关研究。

2016 年,Google 公司旗下的 DeepMind 公司开发出的 AlphaGo 机器人在举世瞩目的围棋比赛中击败了韩国最优秀的职业围棋手李世石。这场比赛引起了很大的轰动,各大媒体争相报道。人们使用“人工智能”、“机器学习”和“深度学习”这几个术语来解释 AlphaGo 机器人获胜的原因,并将这些术语混为一谈。这种说法表面上看起来确实将 AlphaGo 机器人的获胜原因解释通了,但其实这种说法是有所欠缺的,这三者在本质上是有区别的。

机器学习是一类人工智能算法,而神经网络又是机器学习里所包含的一种算法,2016 年战胜韩国围棋选手李世石的 AlphaGo 就是用神经网络算法编写的。目前神经网络已经是一项比较成熟的算法了,而且也已有十分广泛的应用,例如,神经网络在模式识别、图像分析、自适应控制等领域都有着成功的应用,一般神经网络中的隐藏层和节点越多,准确率越高。多层神经网络的应用使得大数据的学习过程耗费了大量的时间,而神经系统的出现经常伴随着大型网络的产生。在这种情况下有两个主要的挑战:一个是传统的训练算法不能

满足大数据处理的需求，另一个是训练时间和记忆的限制越来越难处理。自然地，面对这种情况，我们可以使用以下两种常见的方法：一是通过一些抽样方法来重新确定数据量的大小，如此一来神经网络的结构就有可能保持不变；另一个便是用并行和分布式的方法扩展神经网络，如深度学习模型和并行训练算法的结合提供了处理大数据的潜在方法。

目前要实现完全的、全面的人工智能还存在一定的技术局限性，因此以目前我们所能掌握的技术为基础，只能够实现“弱人工智能”。弱人工智能并不像科幻电影或者科幻小说所描述的那般，能够实现人类所拥有的技能，但是弱人工智能可完成机械的或者单一的任务，在执行特定任务时可以达到与人类相当或在某些方面优于人类的水平。现实生活中有很多弱人工智能的例子，它们给我们的生活带来了极大的便利。

1.3.2　机器学习的定义

一直以来，我们把学习能力视为人类特有的能力，所以是否具备学习能力成了区别人和其他生物的关键。Samuel 于 1959 年设计了历史上第一个国际象棋程序，这个程序可以通过相互对弈来进行学习，进而提高棋艺。经过四年的学习，这个程序击败了 Samuel。三年后，它击败了美国的一个常胜冠军。这一案例向人们展示了机器学习的强大，由此引发了人们对社会问题和伦理问题的讨论，比如常常听到的机器学习能力是否能超越人类的能力，这个问题一直为人们津津乐道，有些人认为机器的学习能力远远在人类之上，但也有一些人持否定态度，他们认为机器是人类创造的，其性能和动作完全是由设计者所规定的，因此无论如何其能力也不会超过设计者本人。对于没有学习能力的机器来说，这一说法并没有错，但是对于具有学习能力的机器来说，这一说法并不是那么准确，因为通过一段时间的学习之后，具有机器学习能力的机器能形成属于自己的知识体系，它不再受控于最初的设定，而且其潜力并不是设计者所能够控制的。

机器学习是一个交叉学科，涉及概率论、统计学、近似理论、凸分析、算法分析与复杂性理论等多个学科。如何专注于计算机模拟或实现人类的学习行为以获取新的知识或技能、重新组织已有的知识结构来改善自己的表现是计算机智能化的根本途径，也是人工智能的核心。它的应用涵盖了人工智能的所有领域，主要使用的统计方法是归纳法、综合法，而不是演绎法。

社会科学家、逻辑学家和心理学家对什么是机器学习存在分歧。例如，Langley 认为“机器学习是人工智能的科学，该领域的主要研究对象是人工智能，尤其是如何提高具体算法在经验学习中的性能”。Tom Mitchell 对信息论中的一些概念进行了详细的解释，其中，机器学习被定义为对计算机算法的研究，可以通过经验自动改进。Alpaydin 提出的机器学习的定义为：“机器学习是利用数据或过去的经验来优化计算机程序的性能标准”。

机器学习的概念来自于早期的人工智能研究者，尽管有这么多关于机器学习的定义，但却没有明确的、统一的定义：为了便于学习，在此我们先对机器学习做一个统一的定义：机器学习是一门研究机器以获得新知识和技能并识别现有知识的学科，是一种实现人工智能的算法，研究的算法包括决策树、逻辑编程、增强学习算法和贝叶斯网络等。机器学习通过算法分析数据，让计算机能够从中获取有效信息，并做出推断或预测。这里的机器指的是计算机、电子计算机、中子计算机、光子计算机以及神经计算机等。

如今，机器学习已被广泛应用于各个领域，如数据挖掘、自然语言处理、计算机视觉、战

略游戏、搜索引擎、语音和手写识别、生物识别、医学诊断、DNA 测序、证券市场分析、信用卡欺诈检测和机器人应用等。机器学习是人工智能的一个重要课题，它的目标是设计算法，使计算机能够根据经验数据进化自己的行为。机器学习的存在对于当前的时代背景来讲是不可或缺的，之前对于数据的处理方式已经不能够满足大数据时代的信息增速，而机器学习最显著的特点正是挖掘并主动汲取知识，进而自主地作出决策。机器学习的蓬勃发展使得数据处理更加方便、快捷，这对于大数据相关的各个行业来讲都是一次革命。

图 1-6 对机器学习进行了一个简单的分类，其分类依据为计算机能否自主进行学习。

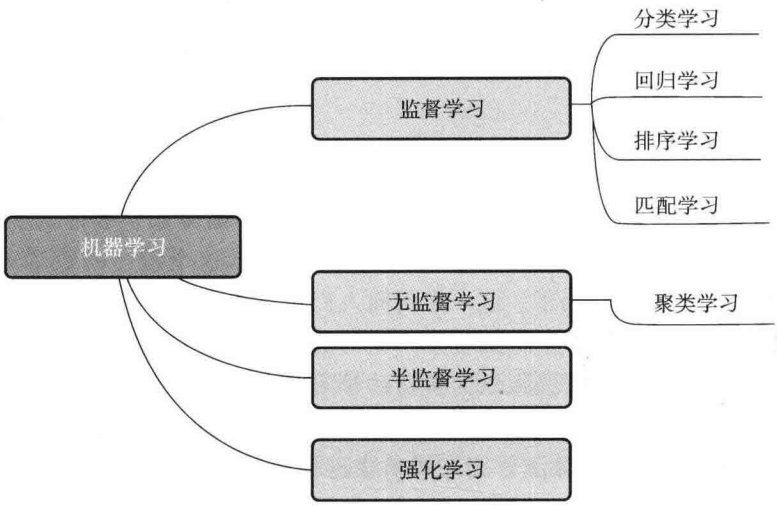

图 1-6　机器学习分类体系

1.4　相关领域应用

1.4.1　数据挖掘的相关案例

Google、Amazon、Facebook、Twitter 公司都是称霸全球互联网的企业。分析它们的运营模式、内部构造及技术支持会发现，正是对数据挖掘这一技术的熟练应用，使得它们能在互联网产业中经久不衰。当然，它们的成功因素还有很多，例如具有新颖的商业模式、优秀的创业者等，但不可否认的是，它们对数据的极高敏锐度及合理利用是促使它们成功，甚至称霸互联网的最主要的原因。它们每天不断地存储和分析大量的数据，了解客户的需求，服务于客户，进而为自身创造更大的利益。

1. 强大的推荐系统

对于能充分应用大数据并由此获得巨额的利益，Google 公司可以称得上是“世界上第一个吃螃蟹的人”。据不完全统计，Google 公司每月要处理 900 亿次 Web 搜索，即每月需要处理的数据量高达 600PB(1PB=1000000GB，这个信息量大约相当于 100 万年《新闻早报》所包含信息的总和)。使用 Google 公司各项服务的用户产生的数据都是其分析的对象。

Google 公司强大的搜索推荐系统是其最普遍且最能体现数据带给我们便利的一项技术。在 Google 搜索框中输入关键字，就会显示一些关于搜索关键字的建议，例如只要输入“基于”二字，系统就会自动提示“基于 MVC 的社团管理系统”、“基于 Web 的学生选课系统源码”、“基于 Java 的社团管理系统”等。像这种搜索关键词的推荐是基于对大量的用户搜索历史数据进行分析所得到的。除了“搜索推荐”，还有“输入修正功能”。“输入修正功能”是指即使输入的关键字是错误的，Google 引擎也会给出正确的搜索推荐。上述这两种推荐方法的原理有异曲同工之处。

在网上购买物品时，通常能看到“购买了此商品的顾客还购买了这些商品”的字眼，这一推荐系统正是由 Amazon 公司创造的商品推荐系统。Amazon 公司分析了大量的历史数据，比如客户的购买记录和浏览历史等，并将这些数据与其他行为模式相似的用户的历史数据进行对比，为用户提供最适合的商品推荐信息。这种以数据挖掘为核心的服务设计理念，推动 Amazon 公司成为全球第二大互联网公司。

2. Facebook 网站和 Twitter 网站的数据对比

Facebook 公司于 2012 年 2 月提出了 IPO 申请，据其公布的数据显示，Facebook 网站每日活跃用户量达到 4.83 亿，每月活跃用户量达到 8.45 亿，可以毫不夸张地说，Facebook 网站是世界上最大的、由用户产生内容的网站。

Facebook 网站的用户每个月在该网站花费的时间总计 7000 亿小时，平均每个用户每个月创建 90 条内容，每个月产生的内容高达 300 亿条。根据已公布的数据显示，Facebook 网站所拥有的数据总量超过了 30PB。Facebook 网站为用户提供的“也许你还认识这些人”的推荐，精准到令人震惊的地步，而这也正是对庞大的数据进行分析的结果。

Twitter 公司的报告显示，Twitter 网站每日活跃用户量达到 1 亿，每月活跃用户量达到 3.28 亿。Twitter 网站平均每天产生 5 亿条推文，每条推文约有 200 个字节，即 Twitter 网站平均每天会产生约 100GB 的数据流量。

3. Credilogros 公司的客户信用评分系统

Credilogros 公司是阿根廷赫赫有名的信贷公司，其总资产估计值为 9570 万美元。对于 Credilogros 公司来说，识别预付款客户的潜在风险至关重要。如若能够掌握这一风险值，将会把公司的风险降至最低。

该公司数据挖掘的目标是创建一个与公司核心系统和信用报告公司系统交互的决策引擎来处理信贷申请。与此同时，Credilogros 公司也在试图掌握相应的风险评分工具，以对一些低收入客户群体进行评估。除此之外，Credilogros 公司希望这套解决方案能够对其 35 个分支办公地点和 200 多个相关的销售点进行实时操作。

最终，因为 SPSS 公司的数据挖掘软件 PASW Modeler 具有较好的灵活性和可移植性，Credilogros 公司选择了它。通过实现 PASW Modeler，Credilogros 公司将处理信用数据和提供最终信用评分的时间缩短至 8 s 以内，这使它能够在最短时间内做出批准或拒绝信贷请求的决策。

4. DHL 的货箱温度

DHL 是物流行业和国际快递的市场领跑者，它所提供的服务包括快递服务、水陆空三路运输及国际邮件服务等。DHL 通过国际网络将 220 多个国家和地区连接起来，形成一个庞大的物流网。美国食品和药物管理局(Food and Drug Administration，FDA)要求确保药

品装运的温度达标，因此 DHL 的客户要求 DHL 能够提供更可靠且更实惠的选择，这也就意味着 DHL 在运送的各个阶段都要对集装箱的温度进行实时跟踪。

虽然在运输过程中可以确保记录器生成的信息精准无误，但是由于其无法传递实时数据，因此 DHL 和其客户不能够在温度发生偏差时采取有效措施。为了解决这一难题，DHL 的母公司——德国邮政世界网拟定了一个计划：使用射频识别（Radio Frequency Identification，RFID）技术全程跟踪装运药品的温度，并由 IBM 全球企业咨询服务部绘制决定服务的关键功能参数的流程框架。这一改进方案使 DHL 解决了运送过程中药品装运温度达标的问题，切实地增强了运送可靠性。这一举措为 DHL 保持竞争差异奠定了坚实的基础，并成为了 DHL 重要的收入增长来源。

5. Montblanc 的商品促销

高级文具制造商 Montblanc，以及美国大型折扣店 Family Dollar Stores，并不像过去一般只是单一地进行商品促销，它们开始将营销与数据分析结合起来，以期获取更大的利益。这些企业正尝试利用监控数据来分析客户的行为。例如，Montblanc 以前是根据经验和直觉来决定商品布局的。然而通过对监控摄像头数据的分析，他们改变了商品的布局，把最需要销售出去的产品摆放到最能吸引顾客注意的地方。通过这种布局变化，Montblanc 的销售量增加了 20%。

大数据不仅为企业家带来了巨额的利润，也为用户提供了重要信息，例如，Decide. com 就是一家利用大数据为客户提供有效信息的公司。这个成立于 2010 年的创业型公司，能够预测近期某数码产品售价的涨跌趋势，用户可以根据它的分析报告对某款产品的购入时间做出合理的判断。

1.4.2 大数据的应用领域

越来越多的领域涉及大数据问题。从全球经济到社会管理，从科学研究到国家安全，无一不彰显着当下是一个大数据的时代，因此可以很明确地说，在未来几年里，大数据将给通信、金融、零售、制造、交通、物流、医疗、公共服务、农业等领域带来巨大的冲击。最近，麦肯锡的一份报告分析了大数据在美国卫生保健、欧盟公共部门管理、美国零售、全球制造业和个人位置数据这五个领域的变革潜力。麦肯锡的研究认为，大数据能够通过提高企业的生产力和竞争力来推动世界经济的快速发展。

1. 科学研究中的大数据

几千年前，人们对世界的描述仅仅基于人类的经验，所以我们称当时的科学为经验科学，它是科学的开端，被归类为第一范式。第二范式出现在几百年前，这一时期的科学称为理论科学，其代表是牛顿运动定律和开普勒的行星运行三大定律。这一时期，就许多复杂的现象而言，科学家们试图找到科学的解释。但是理论分析是非常复杂的，有时甚至是不可思议的、不可行的，在这种需求下，第三种科学范式作为计算分支应运而生。基于第三种科学范式，科学家们开始进行许多仿真和模拟实验。许多领域的模拟实验产生了大量的数据，同时，越来越多的大数据在各个管道中产生。毫无疑问，科学的世界已经因数据密集的应用而改变。

数据密集型的出现催生了一种新的研究范式。研究人员试图用一种新式工具从大数据中找到或挖掘出所需的信息、知识和情报，这样他们甚至不需要直接访问研究对象就能够得