“数据科学概论”课程设计

覃雄派[1]，陈跃国[1]，杜小勇[1]，王伟娟[2]
1. 中国人民大学信息学院，北京 100872; 2. 中国人民大学出版社，北京 100872

摘要
大数据时代已经到来，为了挖掘大数据的价值，社会急需大量合格的数据科学家，数据科学家的培养是一个紧迫的问题。提出了三大课程群的课程体系建设思路，其中“ 数据科学概论 ”是数据科学课程群的导论和入门性质的一门课程。本课程通过案例对关键技术的原理进行介绍，提供了中等规模实际问题的全流程实践案例，有利于学生掌握。数据科学是一门交叉学科，课程应该体现学科交叉的特点。对于时间序列数据，从统计学视角和数据挖掘/机器学习视角，对其分析和建模技术进行了介绍和比较。

关键词
数据科学；课程群；数据科学概论；课程设计

中图分类号：TP311.13 文献标识码：A doi: 10.11959/j.issn.2096-0271.2017065

Course design of the "Introduction to Data Science"

QIN Xiongpai[1], CHEN Yueguo[1], DU Xiaoyong[1], WANG Weijuan[2]
1. School of Information, Renmin University of China, Beijing 100872, China
2. China Renmin University Press, Beijing 100872, China

Abstract
Big data era has arrived. In order to extract the value from big data, the community needs a large number of qualified data scientists. The training of data scientists is a pressing problem. School of Information Renmin University of China (Computer Science Department) proposed the construction thinking of building a curriculum system of three course groups, among them “Introduction to Data Science” is an introductory course of the data science course group. Firstly, the key technologies were introduced by cases for students to easily grasp the basic idea were introduced. Besides that, in order to enhance students' ability to analyze real problems (complex engineering problems) and to solve them, a whole-process practice case for a medium-sized practical problem was provided. Data science is an interdisciplinary subject, the course should reflect the interdisciplinary characteristics. For example, for time series data, the methods from statistics perspective and data mining / machine learning perspective to model and analyze the data, some comparison of the methods was given.

Key words
data science, course group, Introduction to Data Science, course design

1 引言

信息技术的进步大大降低了人们获取数据、存储数据和传输数据的成本，使得越来越多的企业/机构有能力从自身的业务系统或通过互联网等其他途径获取规模日益庞大的数据。数据的价值对于企业而言越发重要，人们更加重视对历史数据的积累。

不断堆积的数据在规模和复杂度上逐渐超越了企业/机构采用已有技术方案在执行数据管理和数据分析任务时所能达到的处理能力，形成了大数据。

大数据[1,2]具有3个主要的特点，其中最重要的特点是数据量大（big volume），其规模超出了已有工具的处理能力，需要研发新的工具进行处理。大数据的第二个特点是数据类型多样，人们希望把不同来源、不同类型的数据关联起来，进而分析其中隐藏的规律。大数据的第三个特点是数据生成速度快，比如在传感器网络中，传感设备生成的数据数量大、速度快，需要及时处理。

数据中蕴含着规律性，即数据中包含价值。很多企业/机构对于收集数据乐此不疲，究其原因，是数据带来的价值或者潜在的价值超出了它们收集数据和管理数据的成本。数据的价值体现通过两个实例可见一斑。2012年，早在飓风Frances来临的一周之前，沃尔玛（Wal-Mart）公司的首席信息官（chief information officer，CIO）Linda M Dillman督促她的团队根据几周之前飓风Charley来袭期间沃尔玛的销售数据，对新飓风来袭的销售进行预测。这些销售数据保存在数据仓库中，达到TB级别。基于这些数据，可以预测将要产生的销售情况，其目的是提高公司的销售额。分析人员对数据进行挖掘分析，以发现对某些产品的不同寻常的需求。他们发现，人们确实更多地购买了某些特定的产品，而不是普通的手电筒等。比如，他们以前并没有了解到，飓风到来前，草莓馅饼的销售量出现了增长，是平时销量的7倍左右，而最畅销产品则是啤酒。据此，他们提前备货，并且及时销售出去，极大地提高了公司的销售额。2016年，谷歌公司的AlphaGo围棋程序击败了人类棋手李世石九段，给人们留下了深刻的印象。Deep Mind公司开发的AlphaGo程序利用深度学习、增强学习、蒙特卡洛树搜索等技术建立了学习模型，然后用成千上万的实际对弈棋局对其进行训练，使其棋艺不断得到增强，最后达到甚至超过人类九段的水平。

数据科学家是伴随大数据技术的崛起和数据科学的兴起而出现的新的就业岗位。近年来，对数据科学家的需求持续增长。数据科学家被誉为21世纪最性感的职业①。他们使用各种技术对不同来源的数据进行分析，帮助企业做出更加明智的决策。

2 数据科学的创立

数据科学②-④是2010年以来逐渐兴起的科学分支，人们普遍认为该门科学正在逐步形成，其知识体系仍在创立之中。

根据维基百科的释义，数据科学是一个交叉的领域，它研究具体的方法、过程和系统，以便从不同形式的数据（包括结构化数据和非结构化数据）中抽取知识，获得对事物的理解和洞察。从这个意义上讲，数据科学和传统的数据挖掘是类似的。但数据科学的内涵有所扩大，它是一个全新的概念，试图把统计方法和数据

① https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

② https://en.wikipedia.org/wiki/Data_science

③ http://courses.csail.mit.edu/18.337/2015/docs/50YearsDataScience.pdf

④ http://dblab.xmu.edu.cn/wp-content/uploads/2014/10/Data-science-and-its-relationship-to-big-data.pdf

分析方法统一起来，目的是分析和理解客观现象产生的数据。它吸收了来自若干传统领域的技术和方法，这些领域包括数学、统计学、计算机科学，特别是计算机科学领域的数据库、大数据、机器学习、数据挖掘、数据可视化等子领域的相关技术和方法。

就笔者的理解，数据科学的本质是从数据中挖掘和抽取价值。数据科学是对数据分析、抽取信息和知识的过程提供指导和支持的基本原则和方法的科学。数据科学研究各种类型数据的不同状态、属性及其变化规律，研究各种方法和技术手段以对数据进行简单以及复杂的分析，从而揭示自然界和人类行为等不同现象背后的规律。

数据科学的核心任务是从数据中抽取信息、发现知识。它的研究对象是各种各样的数据及其特性。数据科学包含一组概念、原则、过程、技术/方法以及工具，为其核心任务服务。其中，概念和基本原则给予人们观察问题、解决问题的一套完整的思想框架，而大量的数据分析技术/方法和工具则帮助人们切实实现数据科学的目标。

简而言之，数据科学是以各类数据作为研究对象，建立在应对数据分析挑战的众多关键技术基础上的一般意义上的科学。为了建立数据科学，人们需要从深层次梳理关键的数据分析处理技术，解析它们的定位和相互关联关系，在理论层面把这些技术联系起来，也就是对基本概念、理论和技术加以系统化的整理。

数据科学不是凭空发展起来的，它是一门新兴的交叉学科。它从数学/统计学、计算机科学等传统学科领域，特别是从数据库、数据挖掘、大数据分析、人工智能/机器学习、可视化等领域借鉴了大量的理论和技术，吸收了有效的成分，逐步建立起自己的学科体系。由于相关的理论和技术来自不同的研究方向，相互之间存在较大的差异，比如研究的基本假设等，数据科学试图在此基础上，构建和谐自洽的理论体系。

3 数据科学专业与课程

数据是新的石油，正成为一种生产资料、稀有资产，是重要的战略资源，全面融入社会、生产、生活各个方面，深刻改变着世界的经济格局、利益格局、安全格局。数据包含信息，可以为人们的决策服务。发挥数据的潜在价值需要大量的数据科学家，他们的工作是结合相关领域的背景知识，对数据进行建模、分析、展现等。

麦肯锡咨询公司发布了一份分析报告，预计到2018年，大数据或者数据分析人员的岗位需求将激增，其中数据科学家的缺口为140 000~190 000人。懂得如何利用大数据做决策的管理人员的岗位缺口，则将达到1 500 000人左右。对数据处理需求最旺盛的行业包括制药业、计算机软件、互联网、科研、IT技术服务、生物技术、金融业等。为了满足企事业单位对数据科学人才的需求，国内外各知名高校设立了专门的数据科学类专业，或在相关专业开设了数据科学课程。数据科学专业或课程在高校中越来越受到学生的欢迎和重视[3]⑤。

3.1 数据科学专业的创立

一些知名大学创立了新的数据科学相关专业，设计了一整套课程体系，招收和培养硕士生和博士生。比如，哥伦比亚大学专门成立了数据科学研究所（Data Science Institute），体现了对数据科学的

⑤ http://dblab.xmu.edu.cn/post/3007/

重视。他们于2014年秋季开始招生，培养数据科学硕士。该专业开设的课程主要包括传统的数学和统计学课程以及相关的计算机课程，具体包括概率论、统计分析与建模、算法、计算机系统、机器学习、探索式数据分析与数据可视化、数据科学伦理、数据科学大作业等方面的课程。

根据调研材料，笔者注意到，国外数据科学专业的创立，有些是由工程学院或者计算机学院发起的，以美国大学为例，如哥伦比亚大学、斯坦福大学、美国西北大学、加州大学伯克利分校、弗吉尼亚大学等，有些则是由管理学院或者商学院发起的，如卡内基梅隆大学、纽约大学、普渡大学、亚利桑那州立大学、康涅狄格大学等。这一方面体现了数据科学专业跨学科的特点，同时也体现了各个专业渴望对数据科学相关技术方法进行了解、掌握和运用。其他开设大数据和数据科学相关专业的高校，还有芝加哥大学、约翰霍普金斯大学、罗彻斯特大学、伊利诺伊大学厄巴纳-香槟分校、乔治·华盛顿大学、德克萨斯大学奥斯汀分校、明尼苏达大学双城分校等。

在国内，2017年3月教育部批准新增32所高校第二批开设“数据科学与大数据技术”本科新专业。至此，总共有35所高校开设“数据科学与大数据技术”专业。中国人民大学成为第二批获得教育部“数据科学与大数据技术”本科新专业批准的院校之一，并且于2017年9月开始招生。这个新专业是由中国人民大学统计学院和信息学院联合申请的。

3.2 数据科学课程的开设

数据科学专业学生的培养需要一系列的课程。其中，数据科学（概论）课程起到一个统领的作用。以美国哈佛大学“数据科学”课程为例，其内容全面广泛，强调学生动手实践能力的培养。

- 广博是该课程内容的突出特点。具体涉及统计推断、代数理论、算法编程、机器学习、人工智能、数据可视化等多个学科，在数据可视化部分，甚至还涉及一些美学和社会学知识。这些内容充分体现了数据科学本身是一门综合性的新兴学科，“数据科学”课程需要给予学生一个全景式的介绍。
- 特别重视学生动手实践能力的培养，课程项目是该课程教学中的重要组成部分。该课程专门配备了一支由25名助教组成的指导团队，对项目小组实现“一对一”的指导。由于数据科学常常面对的是开放性的问题，这些问题没有唯一、确定的答案，通过对实际生活中遇到的数据问题进行分析和解决，学生能够切身体会到数据科学家的工作内容和思想方法。在实践过程中，学生对知识的理解和掌握程度将大大加深，解决问题的能力会得到极大提高。
- 过程是评判成绩的重要依据。该课程强调对过程进行细致的考核与评判，及时发现学生存在的知识漏洞，从而有针对性地进行辅导。

华盛顿大学开设的“数据科学导论”课程⑥同样表现出内容的丰富性。该课程是数据科学课程群的第一门课程，数据科学课程群包括“数据科学导论（Introduction to Data Science）”“数据分析方法（Methods for Data Analysis）”“从大规模数据中获取知识（Deriving Knowledge from Data at Scale）”3门课程。其中，数据科学导论课程讲授数据存储、管理和操作的相关技术和工具，并且把这些技术和工具应用到实际场景中，包括关系数据库技术和各类新型的NoSQL技术，目的是使学生可以根据

⑥ https://www.pce.uw.edu/courses/introduction-to-data-science

问题选择合适的工具。课程的具体内容包括数据的基本概念、数据的类型、关系数据库系统、NoSQL数据库、Hadoop大数据平台、探索式数据分析等。该课程没有纠结于什么是数据科学、数据科学的内涵是什么，而是通过介绍工具和实际应用场景使学生迅速获得利用现有工具解决实际问题的经验。

麻省理工学院开设了“计算思维和数据科学导论（Introduction to Computational Thinking and Data Science）”课程。该课程强调涉猎的范围，而不是一味增加深度。它为学生提供许多主题的浅显介绍，这样学生就可以知道在他们的职业生涯中可以用计算机完成什么样的任务。课程的内容包括绘图、随机程序、概率和统计、随机漫步、蒙特卡洛模拟、数据模型化、优化问题和分类归并等。该课程要求学生具备一定的Python编程经验，掌握计算复杂度的基础知识。

教材为课程提供了内容支撑，国外出版的数据科学方面的教材可以分为如下几类。

● 一些教材专注于数据科学基本原理、技术和方法的讨论，比如《数据分析的要素（The Elements of Data Analytic Style）》《数据科学的艺术：数据工作者指南（The Art of Data Science: a Guide for Anyone Who Works with Data）》《数据智能：利用数据科学把信息转换为洞察力（Data Smart: Using Data Science to Transform Information into Insight）》等。

● 一些作者专门为数据科学调整和重新编写统计分析、数据挖掘和机器学习方面的教材，比如《统计学和贝叶斯数据分析方法（Statistics and Bayesian Data Analysis）》《数据科学中的统计推断（Statistical Inference for Data Science）》《应用预测式建模技术（Applied Predictive Modeling）》《统计思维（Think Stats）》等。

● 大量的教材通过具体的编程语言、工具和案例介绍数据科学，使用的语言主要有R、Python、MATLAB等。这些教材包括《数据科学的R语言实践（Practical Data Science with R）》《精通数据科学的Python语言编程（Mastering Python for Data Science）》《使用R语言建立机器学习系统（Building Machine Learning Systems with Python）》等。这类教材很多，本文不在此一一列出。

● 有些教材特别介绍数据科学技术在具体领域的应用，比如《商业中的数据科学：你必须了解的数据挖掘技术与数据分析思维（Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking）》等。

● 数据可视化是数据科学的一个重要方面，一部分教材专门介绍这方面的技术和原理，比如《量化信息的可视化（The Visual Display of Quantitative Information）》《可视化：关于设计、统计分析、可视化中的数据流动的指南（Visualize This: The Flowing Data Guide to Design, Visualization, and Statistics）》《如你所见：量化分析的简单可视化技术（Now You See It: Simple Visualization Techniques for Quantitative Analysis）》等。

● 少量教材从文化、社会、法律、伦理等方面对数据科学进行讨论，如《数学成为毁灭性的武器：大数据是如何加剧不平等和对民主造成威胁的（Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy）》等。

可以看出，国外出版的大量教材注

重实践性和可操作性，并未纠结于数据科学理论体系的创立。就笔者的认识，数据科学理论体系的创立正在进行，远未完成。

国内一些高校，包括清华大学、北京大学、中国科学院大学等开设了大数据和数据科学相关课程。一些专家开始编写相关讲义和教材，其中，中国人民大学信息资源管理学院（即档案学院）朝乐门老师编写的《数据科学》，是国内较早的关于数据科学的教材。该教材共包括8个部分（基础知识、数据预处理、数据统计、机器学习、数据可视化、数据计算、数据管理以及R编程），既涵盖了数据科学的基本内容，又避免了与相关课程的低级重复。每章设有综合例题，做到理论学习与动手操作相结合。

4 “数据科学概论”课程定位内容与教学设计

数据科学家需要什么样的具体技能呢？这是人们关心的问题，也为数据科学这门课程应该提供什么样的内容提供一个指引。下面介绍笔者在“数据科学概论”课程设计方面的想法和实践。

4.1 数据科学课程群与“数据科学概论”课程定位

中国人民大学信息学院在数据库研究方面具有悠久的历史。在大数据时代，信息学院计算机系与时俱进，对课程体系进行了梳理，提出了建设三大课程群的课程体系建设思路，包括算法课程群、系统结构课程群和数据科学课程群。其中，数据科学课程群将由一系列课程构成，包括数据库、大数据、商务智能、数据挖掘、统计分析、机器学习与深度学习等。课程体系的改革为2017年招生的数据科学专业方向学生的培养打下了基础。

在数据科学课程群中，把“数据科学概论”定位为一门入门和导论性质的课程。通过该课程的学习，学生了解了数据科学的内涵，掌握了数据处理的技术原理，并且通过一些实践案例增强了动手能力，为深入学习后续课程打下了良好的基础。

从2013年起，笔者已经在中国人民大学信息学院本科生开设了4年的“数据科学概论”课程，开始时作为一门选修课。选修的人数逐年上升，由最初的十几名到现在的四十几名。

4.2 “数据科学概论”课程内容

“数据科学概论”课程的内容，分为四大模块，分别如下。

- 数据科学基础（fundamentals）：讲述数据科学的基本概念和原则。
- 数据和数据上的计算（data and computing on data）：讲述不同的数据类型及其分析方法。数据类型包括结构化数据、非结构化数据、半结构化数据，具体包括表格（关系数据库）、文本、社交网络、时间序列数据、轨迹数据等。分析方法包括统计学方法、数据挖掘和机器学习方法等。
- 数据处理基础设施、平台和工具（infrastructure, platforms and tools）：讲述云平台、数据库、大数据平台及工具以及编程语言Python。
- 大数据案例和实践（applications and practice）：讲述大数据应用的成功案例，并且面向金融领域的量化交易应用，从数据采集、模型训练、预测、评价到可视化等环节，带领读者完成数据分析处理全流程的实践。

这4个部分的内容相辅相成，构成该课程的内容体系，如图1所示。没有第一部分

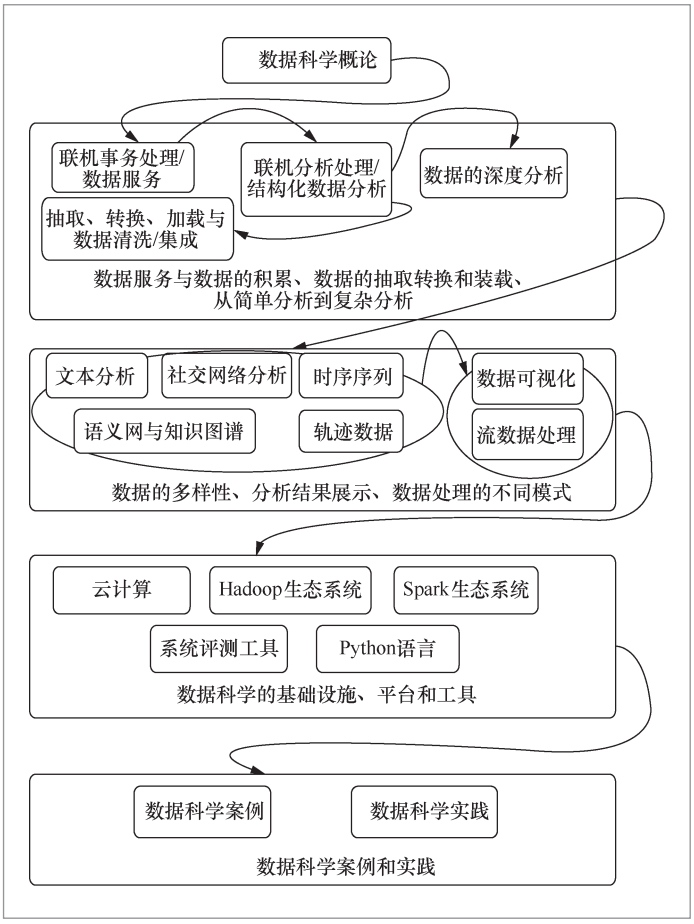

图1 “数据科学概论”课程的内容

的内容，第二部分的内容将是松散的。而没有第三部分的内容，数据分析将无法落地。第四部分的内容则引导读者灵活运用所学知识，解决具体的实际问题，特别是复杂的工程问题。

第一部分对数据科学的定义、数据科学和传统学科的关系、数据科学的原则以及数据科学的4个关键视角进行了介绍。数据科学的定义及其与传统学科的关系第3节已经论述。这里简单介绍数据科学的若干基本原则以及4个关键视角。

笔者认为数据科学包含如下7个基本的原则。

(1) 数据分析阶段

数据分析可以划分成一系列明确的阶段，包括理解业务数据、收集数据、对数据进行集成、对数据进行分析挖掘、对结果进行可视化以及把结果表达给目标受众等。把数据分析任务看作一个工作流，划分成一系列明确的阶段，是结构化地分析问题、解决问题的思想方法。

(2) 描述性分析(descriptive analysis)和预测性分析(predictive analysis)

对数据进行分析有两个方面的目的，即了解过去和预见未来。由此，数据分析分为两类，分别是描述性分析和预测性分析。面向过去，发现隐藏在数据表面之下的历史规律或模式，这类分析称为描述性分析。这些隐藏的模式可以帮助人们更好地进行决策。面向未来，对现有的数据进行深度分析，构建分类/回归模型，对未来趋势进行预测，称为预测性分析。

(3) 实体相似度

在数据科学中经常要计算实体间的相似度。比如在推荐系统中，要计算用户之间的相似度，或者计算商品之间的相似度。在实际工作中，虽然为特定实体建立了高维的刻画模型，但还是有可能遗漏某些信息，没有完整地刻画客观对象。即便这样，人们还是有信心使用已有的属性信息计算实体之间的相似度。因为在一些属性上相似的实体在其他属性上一般也是相似的，这些属性可能是未知的属性，没有进行采集和数字化。

(4) 模型的泛化

一般要避免模型对历史数据的过度匹配，这种现象称为过拟合(over fit)。过拟合导致模型的泛化能力差，也就是模型在新数据上的分类或者预测的效果不好。

(5) 分析结果与场景的结合

对数据进行深入分析以后获得的结果是否具有实际应用价值，是否能够帮助人们做出更好的决策，需要结合具体的应用场景进行评估。

(6)相关性与因果性

从大量的基础数据中,可能分析出变量之间的相关性。相关性很有用,在一定程度上可以帮助人们进行预测。但是相关性和因果关系有重大区别,相关性不意味着因果性。

(7)并行处理

并行处理可以提高数据处理的速度。并行处理分为任务并行(task parallelism)和数据并行(data parallelism)两种类型。所谓任务并行,就是通过多个进程(正在运行的应用程序)对数据进行处理,通过操作系统的多任务处理能力,提高数据处理的效率。数据并行指的是把整个数据集(大规模)划分成一系列小的数据集,然后利用多个进程对这些小的数据集进行并行操作,以达到提高数据处理速度的目的。

4个关键视角包括以下几个方面。

(1)纵向视角(时间维度)

数据有其完整的生命周期,数据的生命周期包括数据的产生、数据的表示和保存、数据的销毁等各个阶段。伴随数据整个生命周期的是人们对数据的分析处理流程。在数据存续的整个生命周期内,有可能对数据进行多次分析。分析处理流程划分成数据采集、表示和存储、集成、分析、展现等主要阶段。

(2)计算视角(系统维度)

数据处理系统依赖于计算机系统的存储和计算能力建立。整个系统可以切分成数据库、存储/检索与分析系统、应用系统(数据产品)等主要层次或者子系统。这是对数据处理系统进行观察的一种视角。

(3)横向视角(数据类型维度)

针对不同的应用,采集到的数据类型丰富多样,包括表格数据、HTML网页文件、XML文件、资源描述规范(resource description framework,RDF)数据、文本数据、图(社交网络)数据、多媒体数据(音频/视频/图像)等。这些数据可以划分成结构化数据、非结构化数据和半结构化数据等不同类型。类型多样的数据之间,当它们描述的是现实世界中同样的实体、事件时,便具有内在的联系,必须建立它们之间的关联,以便实现跨媒体的数据分析。

(4)价值提升视角(价值维度)

对于不同的应用来讲,数据价值提升的过程具有共性。首先,原始数据一般数据量较大,但是数据的价值密度低,有可能包含很多的噪声(即错误数据)。这些数据必须经过清洗,以便剔除错误,提高数据的质量。此外,不同来源的数据需要集成起来,删除重复数据。多源异构数据之间还要建立数据之间的关联。掌握的数据越全面,越多样,分析结果越有可能反映客观实际。对数据进行分析的方法,根据分析的复杂度分为简单分析和复杂分析。所谓简单分析,就是对数据进行多维的汇总统计、生成报表等操作。而复杂分析则包括运用统计分析方法、数据挖掘方法、机器学习方法,对数据进行深入分析。通过适当的分析,可以挖掘到数据中隐藏的模式、相关性等。如果数据中反复出现一些模式,可以在此基础上抽象出知识。知识比模式、相关性等更加具有普遍性的规律。数据价值提升的过程伴随着数据(信息)规模的缩小和数据(信息)价值密度的提高。

4.3 “数据科学概论”课程教学设计

在教学设计方面,体现如下特点。

(1)学科交叉

学科交叉在时间序列分析方面表现得尤为明显。传统的统计学研究时间序列的趋势性、季节性、噪声等成分,并且用移动平均线(moving average,MA)模

型、自回归(auto-regressive, AR)模型、自回归移动平均(auto-regressive and moving average, ARMA)模型、自回归积分移动平均(autoregressive integrated moving average, ARIMA)模型、自回归条件异方差(auto-regressive conditional heteroskedasticity, ARCH)模型、广义ARCH(generalized auto-regressive conditional heteroskedasticity, GARCH)模型等对时间序列进行建模,强调模型的严谨性及其自洽性。而数据挖掘和机器学习领域的专家们则通过对时间序列的降维表示、相似度计算以及分类/聚类/关联规则分析等技术手段,解释时间序列数据中蕴含的规律性。近年来,研究人员还把具有优良时间关系建模能力的深度学习模型(如长短期记忆(long short term memory, LSTM)神经网络)应用到了时间序列的建模和预测上,并且取得了良好的效果。在“时间序列”的教学实践中,把来自统计学和机器学习/数据挖掘领域的技术手段和方法进行了对比介绍,并且分析其长处和短处,帮助学生进一步思考。

(2)知识点案例与综合案例

为了帮助学生把握技术原理,并且能够开始运用这些原理,对复杂的工程问题进行求解,笔者从两个方面进行案例式讲解。一方面是针对各个知识点给出实例,通过简单的实例讲解每个技术的原理,使学生迅速把握其本质,而不是陷入艰难的数据推导和绝望中。这并不是说数学推导是不需要的,而是作为一门入门性质的课程,更为重要的是让学生把握技术的原理和思想,而艰难但是必要、深入的数学推导过程可以在后续的课程中进行介绍。

另一方面,从问题出发,展示问题的分析和解决策略及其实现过程,也就是传统意义上的综合实例,而且是面向实际应用的综合实例。在本课程中,用一部分时间讲述数据科学的常用编程语言Python以及几个重要的函数库,包括数据处理函数库pandas、机器学习与数据挖掘函数库Scikit-Learn、数据可视化函数库Matplotlib、社交网络分析函数库NetworkX、文本分析函数库NLTK以及深度学习函数库Theano和Keras等。在此基础上,面向金融领域的量化交易应用,从数据采集、模型训练、预测、评价到可视化等环节,带领学生完成数据处理和分析的实践,打通整个流程。锤炼学生的编程实战能力,使其深刻体会运用数据科学方法解决实际问题的乐趣。

(3)教学内容的深度展开和宽度展开

在教学内容的展开路线上,从简单的数据管理和分析、多维分析和结构化数据分析,到复杂的数据挖掘和机器学习,由浅入深形成了对内容的深度展开。然后,对文本、社交网络、时间序列、轨迹等数据单独进行介绍,完成了内容的宽度展开。

5 结束语

大数据时代已经到来,数据科学正在兴起,时代对数据科学家提出了紧迫的需求。本文把中国人民大学信息学院建设数据科学课程群的思想以及笔者对“数据科学概论”课程设计的想法和经验展现出来,与同行们交流。

参考文献:

[1] 覃雄派, 王会举, 杜小勇, 等. 大数据分析-RDBMS与MapReduce的竞争与共生[J]. 软件学报, 2012, 23(1): 32-45.
QIN X P, WANG H J, DU X Y, et al. Big data analysis-competition and symbiosis of RDBMS and MapReduce[J]. Journal of Software, 2012, 23(1): 32-45.

[2] 王珊，王会举，覃雄派，等. 架构大数据：挑战、现状与展望[J]. 计算机学报，2011，34(10)：1741-1752.
WANG S，WANG H J，QIN X P，et al. Architecting big data：challenges，studies and forecasts[J]. Chinese Journal of Computers，2011，34(10)：1741-1752.
[3] 许嘉，吕品. 哈佛大学数据科学课程教学初探[J]. 教育界：高等教育研究，2015(5)：109-110.
XU J，LV P. Introduction to teaching of data science course in Harvard university[J]. Education Circle，2015(5)：109-110.

作者简介

覃雄派（1971-），男，博士，中国人民大学信息学院讲师，目前主要从事高性能数据库、大数据分析、信息检索等方面的研究工作，主持1项国家自然科学基金面上项目，参与多项国家“973”计划、“863”计划、国家自然科学基金项目，在国内外期刊和会议上发表论文20余篇。

陈跃国（1978-），男，博士，中国人民大学信息学院副教授、博士生导师，中国计算机学会高级会员，数据库专家委员会委员，大数据专家委员会通信委员，Frontiers of Computer Science青年编委，主要研究方向为大数据分析系统和语义搜索。主持国家自然科学基金项目2项，广东省科技应用重大专项1项，参与多项国家核高基（核心电子器件、高端通用芯片及基础软件产品）、“973”计划、“863”计划项目，近年来在SIGMOD、SIGIR、ICDE、AAAI、IEEE TKDE、WWW等国际重要期刊和会议上发表论文30余篇。

杜小勇（1963-），男，博士，中国人民大学信息学院教授、博士生导师，教育部数据工程与知识工程重点实验室主任，中国计算机学会会士，《大数据》期刊编委会副主任。主要研究方向为智能信息检索、高性能数据库、知识工程。主持和参与多项国家核高基（核心电子器件、高端通用芯片及基础软件产品）、“973”计划、“863”计划、国家自然科学基金项目，近年来在SIGMOD、VLDB、AAAI、IEEE TKDE等国际重要期刊和会议上发表论文百余篇。

王伟娟（1979-），女，中国人民大学出版社编辑，主要研究方向为大数据、云计算、统计分析、数据科学。

收稿日期： 2017-05-24