面向大数据人才培养的融合式教学模式

胡艳丽，白　亮，谭　真，葛　斌，唐九阳
（国防科技大学　系统工程学院，湖南　长沙　410073）

摘　要：数据科学的快速发展以及海量数据分析需求的日益增长，对大数据人才培养提出了严峻挑战，也带来了重大机遇。本文探索了在教学中如何有机融合知识构建和实践创新环节，培养大数据人才创新能力的教学模式。在数据科学课程教学中，采用反映领域前沿的大数据分析竞赛应用问题，实施“实践驱动的知识构建”与“问题导向的实践创新”教学设计，通过知识构建和实践创新深度融合提升教学质量，促进大数据人才培养。

关键词：大数据人才培养；实践驱动；问题导向；大数据分析竞赛
中图分类号：G642　　**文献标识码**：A　　**文章编号**：1672－8874（2020）01－0101－03

Integrated Teaching Mode for Training Big Data Talents

HU Yan-li, BAI Liang, TAN Zhen, GE Bin, TANG Jiu-yang
(*College of Systems Engineering*, *National University of Defense Technology*, *Changsha* 410073, *China*)

Abstract: The rapid development of data science and the growing demand for massive data analysis have brought a severe challenge and an important opportunity for the cultivation of big data talents. This paper explores how to integrate knowledge building and innovation practice in the teaching in cultivating the innovation ability of those talents. In the data science course, up-to-date questions in big data analysis competition were introduced. The teaching design combined practice-driven knowledge building and problem-oriented practice innovation. Teaching quality has improved due to the integration of knowledge building and practice innovation, which promoted the cultivation of big data talents.

Key words: cultivation of big data talents; practice-driven; problem-oriented; big data analysis competition

一、引言

信息技术（Information Technology，IT）的研发应用日新月异，人类所拥有的数据规模正以史无前例的速度迅猛增长。数据密集型科学成为继实验、理论、计算模拟之后的新的科学研究范式[1-2]，在现实应用的各个领域，对思维方式、研究方法、应用模式等方面带来了巨大的冲击和挑战。培养精通大数据分析和实践的高素质复合型人才成为社会对人才培养的迫切需求，也对高校传统的教学方法带来了新的挑战。

针对大数据人才培养需求，本文探索了知识构建和实践创新相融合的教学模式，采用反映领域发展前沿的大数据分析竞赛应用问题，实施“实践驱动的知识构建”与“问题导向的实践创新”教学设计，应用于数据科学课程教学，培养大数据人才实践创新能力。

收稿日期：2019－11－01
基金项目：全国教育科学“十三五”规划课题（JYKYD2018007）；系统工程学院教改课题资助
作者简介：胡艳丽（1979－），女，河南夏邑人。国防科技大学系统工程学院副教授，博士，主要从事大数据分析和文本挖掘研究和教学工作。

二、构建面向大数据人才培养的融合式教学模式

培养大数据专业人才需要建立问题意识和数据思维，系统掌握大数据分析技术和方法，具备构建领域知识、解决复杂工程应用问题的能力和素质。针对上述人才培养目标，我们提出了“学生中心、问题导向、实践驱动、能力提升”的教学理念，设计了知识构建与实践创新相融合的教学模式（如图 1 所示）：

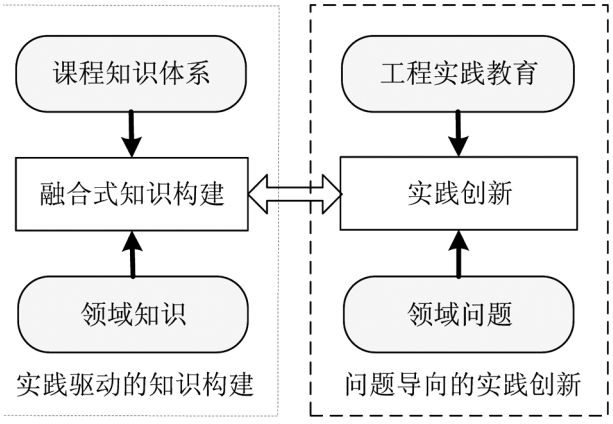

图 1　面向大数据人才培养的融合式教学模式

1. *实践驱动的知识构建*。面向实践创新需求开展融合式知识构建，在课程知识体系构建的同时，引入实践所需的领域知识。课程知识构建培养专业技能，领域知识牵引知识构建围绕应用问题开展，建立实践所需的理论基础。

教学过程中，可以针对学生现有知识水平进行调查分析，分析学生所掌握的相关知识、技术方法以及学习需求，进而动态确定教学需求和教学内容，遵循从宏观整体概念体系逐层细化、具体化和实例化的渐进式过程，开展课程知识体系和领域知识相融合的知识构建，帮助学生构建实践所需的融合式知识体系，在知识构建过程中建立问题意识；

2. *问题导向的实践创新*。采用领域问题设置知识应用情境，围绕问题指导学生综合运用所学知识开展综合实践。同时，引入解决复杂问题的工程实践框架和流程，通过实践深化对知识的理解和运用，针对实践中出现的问题强化讲解和研讨，引导学生开展实践创新；最终学生作为主体展示实践问题的分析求解过程及结果，由学习、运用知识转化为能力提升和输出。

三、领域前沿问题牵引的融合式教学设计——以“数据科学综合实践”课程为例

“数据科学综合实践”是面向大数据等专业人才培养的核心课程，需要学生构建领域知识，综合运用数据管理、统计分析、数据挖掘以及机器学习等知识进行数据科学实践问题求解，具有极大的挑战性和创新度。经过多年教学实践，我们设计实施了融合式教学设计（如图 2 所示）。

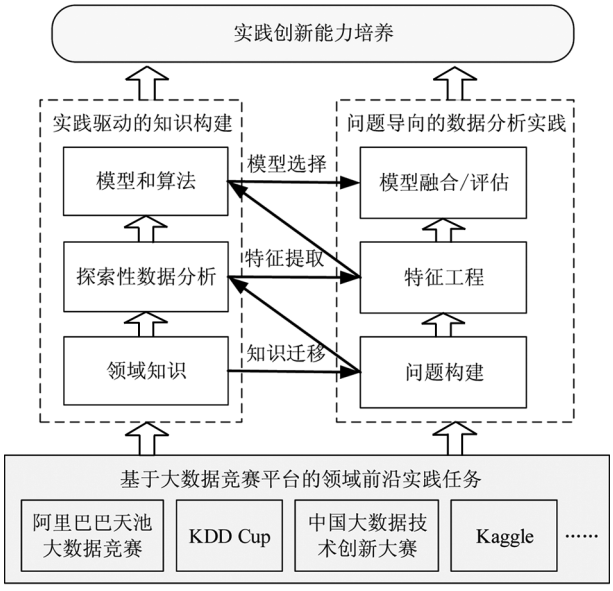

图 2　领域前沿问题牵引的融合式教学设计

教学过程中，我们采用“实践驱动的知识构建”与“问题导向的数据分析实践”融合式教学设计，基于国内外顶级大数据竞赛平台设计符合教学目标的领域前沿实践任务，以领域应用问题为牵引，构建融合课程知识、领域知识和工程实践知识的知识体系，通过问题构建和实践牵引帮助学生整合所学的理论技术和方法，实现理论知识和工程实践的有机融合，跨越理论学习和实践应用的鸿沟，建立大数据专业人才所需的技术方法、领域知识、实践应用综合创新能力。

领域前沿问题牵引的融合式教学设计能够克服传统教学中知识体系的条块分割，以及实践教学中存在的诸多难点问题。课程基于大数据竞赛平台设计实践任务，保证了实践任务的前沿性、科学性和实践效果可评测。这类大数据分析任务源于现实应用，往往蕴含极具价值的学术问题，而且具有通过其他方式难于直接获得的真实数据集，吸引了全球范围内的高水平参与者，在问题

求解过程中支持参与者间的实时研讨，并对解决方案进行公开评测和发布，为学生实践大数据分析技术和方法提供了极好的机会，为实践教学提供了很好的平台。例如，2014 年度阿里巴巴大数据竞赛基于 5.7 亿条数据预测用户的购买行为，吸引了 14 个国家和地区 7276 支队伍参赛。

具体而言，实践驱动的知识构建从领域知识、探索性数据分析以及模型算法维度，构建实践所需的核心知识和工程实践框架流程等动态知识，并与问题导向的实践创新有机结合，始终引导学生围绕领域前沿问题开展探究式学习，在领域知识的指导下归约实践问题，进行探索性分析，提取影响问题目标的若干属性及其组合构造特征，然后选择适合问题的模型和算法进行求解，在求解过程中评估模型的效果，融合多种模型以获得更好的分析结果。

在教学中采用领域前沿问题牵引的融合式教学设计，围绕学生实践创新能力培养，先后组织学生参加了国际知识发现和数据挖掘大数据分析竞赛（KDD Cup）、Kaggle 国际大数据竞赛平台、阿里巴巴天池大数据竞赛平台[3-5]等发布的开放式实践任务，培养学生创造性应用大数据分析技术解决实践问题的能力。

四、融合式教学设计实施案例

本节以新浪微博互动预测大赛（以下简称预测大赛）为例，说明领域前沿问题牵引的融合式教学的组织和实施过程。

1. 基于领域知识的问题构建

在微博数据分析实践过程中，首先引导学生从应用维度分析影响微博互动数的因素：

（1）大赛要求预测每一条博文在发表一天后的互动数（即转发、评论、点赞总数）属于的档位，不同的互动数对应不同的档位。直观看来该分析问题属于回归预测，即建立模型预测互动数，然后根据互动数划分博文所属档位；但比赛的精确度评估指标基于档位预测正确的博文数计算，而非博文的互动数，因此深入分析表明该问题更适合采用分类方法进行建模分析，即不同的档位对应博文所属的类别；

（2）从微博用户角度分析，用户往往具有不同的年龄层次、领域背景、兴趣爱好等，通过博文发布信息、获得粉丝关注，其博文质量、粉丝数等因素的差异导致不同用户的影响力不同，因此博文的互动数也不同。例如，两条内容相似但被不同用户发布的微博，知名博主获得的互动数一般要远远大于普通用户；同时，微博用户中存在机器人和专门发布广告的非常规用户，博文数量大于常规用户，但互动数非常少；

（3）从博文角度分析，微博可以包含 140 汉字（280 字符）以内的信息，实际博文内容往往只有几十个字，甚至几个字，话题覆盖面广泛，其中包含大量停用词、数字、电话号码、日期、邮箱、URL 等噪声；

（4）从粉丝角度分析，用户的粉丝数对博文的互动具有很大影响，拥有大量粉丝的博主，其博文互动数往往较大；但不能简单统计粉丝数，因为存在大量僵尸粉，活跃粉丝是影响互动数的主要因素。

在问题构建过程中，根据评估指标，学生认识到该问题的目标是将不同互动数的用户划分为不同的档次，因此更适合用分类模型进行求解，进而对微博用户及其博文内容、粉丝情况如何影响微博互动数有了直观的认识，对于如何消除微博中的噪声，并对机器人、广告用户、僵尸粉等离群点进行分类处理有了清晰的认识，为从数据中提取特征打下基础。

2. 探索性数据分析驱动的特征工程

在此基础上，引导学生进行探索性数据分析，分析数据的基本特征以及数据之间的关联如何影响微博互动数，例如：

（1）对数据进行预处理和清洗，去除离群点，消除博文内容中的噪声；

（2）探索用户微博数与微博互动数的特征，分析用户类型，并检测机器人、广告用户和僵尸粉等离群点；

（3）探索用户粉丝数、粉丝活跃度与微博互动数的关系，通过曲线拟合粉丝数和粉丝活跃度与互动数的关系；

（4）探索微博互动数随时间的变化，统计微博发布后互动数随小时、天、星期的变化；

（5）探索博文内容与互动数的关系，采用文本分析、话题发现等技术抽取特征，分析关键词、频率及其与微博互动数的关系。

通过探索性数据分析，学生对数据进行预处理和清洗，分别从用户属性、博文属性、粉丝属

（下转第 120 页）