

基于大模型的大数据分析方向课程融合与教学探索

王佩瑶 李欣 汤金 余欢 邓新玉

成都锦城学院, 中国·四川 成都 610097

摘要: 在数字化时代高速发展与国家教育创新战略驱动下, 大数据分析课程面临知识更新快、实践脱节、跨学科融合难等挑战。论文提出基于大模型的系统性教学改革框架, 从课程体系、教学方法、实践平台等方面进行创新。经实践验证, 该框架有效提升学生能力, 为高校培养复合型数据人才提供可行路径。

关键词: 大语言模型; 大数据分析; 课程融合; 教学探索

Exploration of Course Integration and Teaching in the Field of Big Data Analysis based on Big Models

Peiyao Wang Xin Li Jin Tang Huan Yu Xinyu Deng

Chengdu Jincheng College, Chengdu, Sichuan, 610097, China

Abstract: Driven by the rapid development of the digital age and the national education innovation strategy, big data analysis courses are facing challenges such as fast knowledge updates, disconnected practices, and difficult interdisciplinary integration. The paper proposes a systematic teaching reform framework based on large models, which innovates from aspects such as curriculum system, teaching methods, and practical platforms. Through practical verification, this framework effectively enhances students' abilities and provides a feasible path for universities to cultivate composite data talents.

Keywords: big language model; big data analysis; curriculum integration; teaching exploration

0 前言

在大数据时代, 数据分析能力成为各行业人才必备技能, 高校大数据分析课程对于培养“技术与业务融合”的复合型人才至关重要。然而, 传统教学模式存在课程内容碎片化、实践与实际业务脱节、个性化教学不足等问题。大模型技术的兴起, 为解决这些问题提供了新契机。成都锦城学院开展的“基于大模型的大数据分析方向课程融合与教学探索”项目, 致力于探索大模型与大数据分析教学的深度融合, 提升教学质量, 培养适应时代需求的数据科学人才。

1 大模型在大数据分析中的能力剖析

ChatGPT 是由人工智能实验室 OpenAI 研发的通用聊天机器人, 其所具有的自然语言生成、深度学习、情感分析、学习迁移等诸多优势功能, 一方面正在对现有部分人才的能力形成巨大的“高压”态势; 另一方面, 其风靡于不同层级的学生之中, 对传统的教学产生巨大影响^[1]。自 ChatGPT 发布引发大模型研究热潮以来, 不同类型大模型在大数据分析领域展现出独特优势, 形成互补态势, 重塑数据分析流程。

数据预处理环节, 通用大模型如 GPT-4 凭借强大的自然语言处理和图像识别能力, 高效处理非结构化数据。以企业合同和社交媒体文本为例, 能精准抽取关键信息, 大幅提升数据处理效率与准确性。代码大模型, 像 StarCoder2, 熟悉分布式计算框架底层机制, 可生成优化的分布式数据清洗脚本, 满足海量结构化数据清洗需求。领域大模型在特定领

域更具专业性, 如医疗领域的大模型, 能严格遵循行业规范对医疗文本去标识化, 保障数据合规性。

特征工程中, 通用大模型依据行业知识和数据特点, 生成高价值特征组合建议。在电商用户分析场景下, 可提出创新的特征组合。代码大模型专注技术实现, 自动生成特征重要性分析代码框架, 助力筛选关键特征。领域大模型在金融数据处理时, 增强时序特征, 精准识别周期性模式, 为金融分析提供有力支持。

模型开发阶段, 通用大模型的少样本学习能力突出, 能快速生成基线模型架构, 缩短开发周期。代码大模型优化分布式训练代码, 集成相关工具, 提升模型运行效率。领域大模型则针对行业需求, 构建专属评估指标, 使模型更贴合业务实际。

成果输出方面, 通用大模型以自然语言生成多维度分析报告, 便于非技术人员理解。代码大模型生成可部署代码, 推动分析成果落地应用。领域大模型生成符合行业规范的可视化脚本, 为行业决策提供直观依据。

2 大模型引入带来的教学变革与挑战

大模型对程序设计类课程带来了根本性挑战: 一方面, 其强大的代码生成能力降低了传统编程技能的必要性, 引发对课程目标的重新审视; 另一方面, 它作为智能“第三主体”介入教学过程, 不仅削弱了教师的知识权威和指导作用, 还重构了师生互动模式, 迫使教学重点从基础语法训练转向更高阶的工程思维和 Prompt 设计能力培养^[2]。

表 1 大模型在大数据分析中的能力剖析

分析环节	通用大模型（如 GPT-4）	代码大模型（如 StarCoder2）	领域大模型（如医疗领域专用模型）
数据预处理	抽取非结构化数据关键信息	生成分布式数据清洗脚本	医疗文本去标识化
特征工程	生成特征组合建议	提供特征分析代码框架	增强特定领域数据特征
模型开发	少样本生成模型架构	优化分布式训练代码	构建行业专属评估指标
成果输出	自然语言生成报告	生成可部署代码	生成行业可视化脚本

大模型的应用促使大数据分析教学从传统知识传授向培养综合能力转变，要求学生具备从问题定义到业务验证的闭环思维。这需要在课程中新增多个知识模块：大模型技术层，涵盖模型架构、API 调用和成本控制；伦理规范层，涉及数据隐私保护、模型可解释性和算法偏见检测；跨课融合层，促进不同课程在大模型应用中的融合。

传统教学方式注重理论讲授，学生实践机会有限。引入大模型后，教学面临新挑战，如如何引导学生正确使用大模型，避免过度依赖；怎样确保大模型辅助教学的技术稳定性和教学适应性；如何在教学中平衡大模型应用与学生自主思考能力的培养等。

3 融合大模型的教学改革实践

3.1 课程体系重构

以能力培养为核心，重新构建课程体系，划分为基础理论、技术工具、应用实践和前沿拓展四个层次。基础理论层巩固数据结构、算法和统计学等基础知识；技术工具层教授 Hadoop/Spark 分布式计算和大模型 API 集成等技术；应用实践层通过实际项目提升学生解决问题的能力；前沿拓展层关注可信 AI 和伦理规范等前沿领域。

建立动态知识更新机制，利用大模型监测行业动态。收集学术会议论文、技术报告和企业招聘要求等信息，经处理后每学期更新部分课程内容，确保课程与时俱进。

3.2 教学方法创新

构建包含基础、综合和真实案例的三级案例库。基础案例聚焦工具操作，如使用大模型生成数据库操作脚本；综合案例培养跨技术栈整合能力；真实案例基于企业实际问题，借助大模型提出解决方案，提升学生业务问题解决能力。通过代码测试、分析报告等方式考核学生。

主动引入大模型使用，在课程案例声波灭火系统测试项目中，提供基本数据集，要求学生通过文心一言或其他大模型去解读并了解数据分析流程，以文字及流程图的形式进行描述；以问答的形式对文心一言进行提问，并使用 python 进行数据查看，见图 1。

对数据集进行合理的数据分析并完成大数据分析报告。以问答的形式对文心一言进行提问，辅助完成大数据分析任务，见图 2。

开发“大模型教学助手”，实现智能答疑、学习规划推荐和协作支持功能。基于课程知识库和大模型检索增强技术，提供符合教学大纲的解答；依据学生错误类型推荐学

习资源；自动组建学习小组并生成讨论提纲，促进学生协作学习。

根据锦域在线提供的数据集，若需要对数据集进行分析，流程如下

1.明确问题与目标：在本数据当中，涉及到了尺寸大小、燃料类型、距离、分贝、气流、频率以及状态。本次探讨：不同燃料的灭火器在不同距离下的灭火效果。

2.数据集收集：本次数据为锦域在线数据

3.数据预处理：1.数据清洗：处理缺失值、异常值、重复值。

2.数据转换：格式化、标准化、离散化数据，使其适合分析。

3.数据归约：选择重要特征、数据抽样、降维以减少数据量。

4.建立模型

1.选择适当的模型（如线性回归模型来用于预测、朴素贝叶斯分类器来分类等）。

2.将数据进行特征工程等，比如数据按照一定的比较进行分割。

3.训练模型并进行参数调优。

5.模型评估与优化：

1.使用验证集或测试集来评估模型的性能。

2.分析模型的误差和偏差，进行参数的调整或收集更多数据。

将 excel 数据在 python 中进行读取

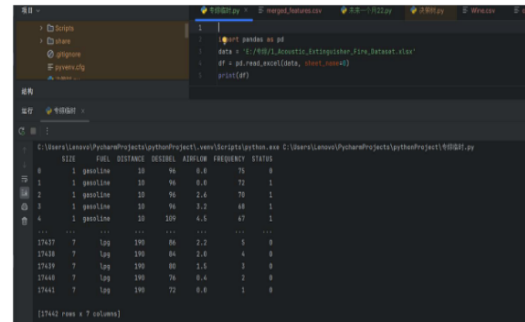
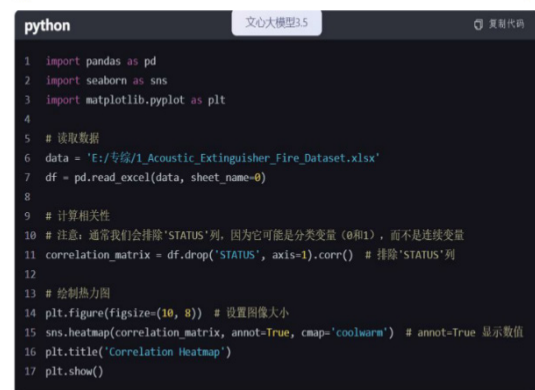


图 1 python 数据流程分析

问：（讲数据读取结果放入文心一言后）请帮我对数据进行处理，处理之后进行相关性分析并绘制热力图



但由于没有涉及到数据预处理部分，因此我对数据预处理部分再一次进行了提问。问：你没有进行预处理，请帮我对数据进行处理之后再相关性分析和热力图的绘制

图 2 大数据分析任务

3.3 实践平台搭建

搭建虚拟仿真环境,使用合成数据和流程沙箱。合成数据模拟真实行业数据,包含内在规则;流程沙箱允许学生自由组合大模型工具和传统组件进行数据处理。同时,实时监控数据流向,检测隐私风险,保障数据安全。

与企业合作开展真实项目实训,采用“双导师制”。企业提供真实脱敏数据集,发布项目需求。学生在企业和高校导师指导下完成项目,企业导师侧重业务价值,高校导师关注技术规范。优秀项目成果可转化应用,学生还可申请专利或参与创业孵化。

4 教学实施与评价体系

4.1 教学实施三阶模型

教学实施三阶模型见图 3。

认知阶段(第 1~4 周):借助大模型生成知识图谱,展示大数据分析技术全景及与大模型的协同关系。将经典案例故事化,学生通过提问探索技术细节,建立大模型技术认知框架。

实践阶段(第 5~12 周):开展双轨实验。基础实验要求学生完成大模型生成的标准化任务,提升基本技能;创新实验鼓励学生设计“大模型+传统算法”混合方案,对比效果并撰写技术白皮书,培养创新能力。

创新阶段(第 13~16 周):举办“大模型数据分析马拉松”。学生调用大模型生成的爬虫代码获取数据,使用大模型构建模型训练流程,最后由大模型生成汇报材料,进行 48 小时工程化实战。

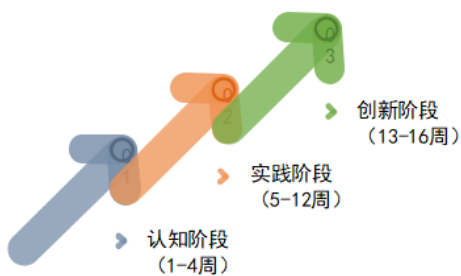


图 3 教学实施三阶模型

4.2 评价体系：三维度动态评估模型

价体系：三维度动态评估模型见图 4。

过程性评价(30%)：关注学生与大模型的交互数据,如提问深度、代码优化采纳率和学习规划执行度,以及在小组协作中的贡献。

成果性评价(50%)：从代码质量、分析报告和模型性能等方面进行评估。包括代码规范符合度、分布式效率优化、报告逻辑和业务可行性、模型准确率和可解释性等指标,并对照行业标准。

反思性评价(20%)：要求学生撰写“大模型使用日志”,记录工具优势、局限性和改进建议,培养反思和总结能力。

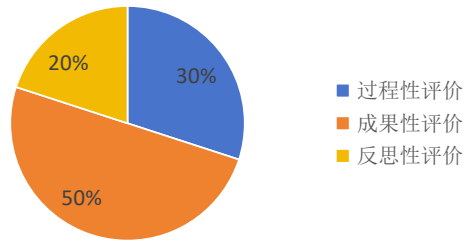


图 4 价体系：三维度动态评估模型

5 教学效果验证

截至目前,论文提出的教改方法已在 2025 级本科生中进行了教学实践,其效果从学生主观评价问卷调查结果进行分析。

共发放调查问卷 120 份,收回有效结果 107 份,结果见下表。调查显示,绝大部分学生认同大模型在大数据分析上有很大影响;支持在大数据分析课堂上进行相关的探索 and 改革;同时,认为论文提出的相关方法是有益的(见表 1)。

表 1 问卷调查

调查问卷问题	结果展示
你认为大模型在大数据分析上的影响程度如何	非常巨大(45.9%) 在部分领域影响巨大(53.27%) 在大部分领域影响较小(0.93%)
你是否支持教学内容和方法中引入大模型	支持(97.2%) 不支持(2.8%)
遇到问题你更喜欢询问谁	老师(6.54%) 同学(4.67%) 网络、大模型等数字资源(88.79%)

6 结语

本研究构建的融合大模型的教学体系,有效解决了大数据分析课程教学中的诸多问题,提升了学生的实践和创新能力。未来研究可从以下方向展开:一是深化跨学科融合,探索大模型在多学科交叉领域的课程设计;二是开发轻量化教学大模型,降低算力需求,提高教学适用性;三是构建教育生态,促进高校、企业和开源社区合作,共享教学资源,形成人才培养闭环。通过持续探索,为大数据分析教育注入新活力,培养更多适应时代发展的复合型数据人才。

参考文献：

- [1] 胡影.ChatGPT对教育的影响:冲击、反思与展望[J].继续教育研究,2023(10):42-48.
- [2] 王聪,万聪.大模型时代计算机程序设计类课程教学模式探索[J].计算机教育,2025(4):137-141.

作者简介:王佩瑶(1993-),女,中国黑龙江黑河人,硕士,讲师,从事通信工程、人工智能与大数据研究。

课题项目:教学研究与教学改革项目(校级),来源:成都锦城学院教学研究与教学改革项目(校级),名称:基于大模型的大数据分析方向课程融合与教学探索(项目编号:2024JCJY0010)。