数据治理测试数据生成策略与实践研究

张律
九江职业大学，江西 九江 332499
DOI: 10.61369/TACS.2025060009

摘　　要： 随着教育数字化转型的深入推进，数据治理平台已成为学校信息化建设的核心基础设施。本文聚焦数据治理平台测试阶段的关键环节——海量学生数据生成，系统探讨了符合教育行业规范的数据生成方法论，通过分析标准，结合分布式计算与 AI 生成技术，提出了“规范约束－算法驱动－场景模拟”三位一体的生成框架。
关 键 词： 教育数据治理；测试数据生成；预设结果模拟

Research on Strategies and Practices of Test Data Generation for Data Governance

Zhang Lv
Jiujiang Vocational University, Jiujiang, Jiangxi 332499

Abstract: With the deep development of digital transformation in education, the data governance platform has become the core infrastructure of the school's information construction. This paper focuses on the key link of the data governance platform testing stage – a large amount of student data generation, systematically discusses the data generation methodology in line with the education industry specifications,and proposes a "norm-constraint-algorithm-driven-scene-simulation" trinity generation framework through the analysis of standards, combined with distributed computing and AI generation technology.
Keywords: educational data governance; test data generation; preset result simulation

引言

教育信息化2.0时代背景下，《中国教育现代化2025》明确提出要“推进教育治理方式变革，加快形成现代化的教育管理与监测体系”。数据治理平台作为实现这一目标的关键载体，其功能完整性与性能稳定性直接决定了教育数据应用的质量[1]。

当前学校在数据治理平台测试中普遍面临三大挑战：一是缺乏符合《教育数据管理办法》要求的标准化测试数据集；二是难以生成足以验证系统极限性能的海量数据；三是无法模拟真实教育场景中的数据关联关系与异常模式。

本文通过构建科学的数据生成方法论，为教育数据治理平台提供高质量的测试素材，从而确保平台在实际运行中能够准确处理学生信息、学业成绩、教学资源等多维度数据，为教育决策提供可靠支撑。

一、教育测试数据的标准框架与生成原则

通过对《教育系统人员基础数据》等标准的分析，学生测试数据应至少包含六个核心维度即身份标识维度、学籍信息维度、个人特征维度、学业表现维度、行为记录维度、隐私保护维度。

教育测试数据的特殊性要求生成过程必须遵循四项基本原则。

合规性原则是测试数据生成的法律基础[2]。教育部明确要求“强化数据安全与隐私保护，符合国家网络安全法及教育行业相关标准”。在数据生成中具体体现为：一是敏感信息脱敏，如对联系电话和电子邮箱采用加密存储；二是数据权限控制，通过身份认证与最小权限原则防止未授权访问；三是跨境传输限制，即使测试环境也必须严格执行数据保护措施。

真实性原则要求测试数据能够反映教育场景的实际规律[3]。如通过分析学生行为数据发现，学习时长与成绩呈正相关（相关系数0.63），但超过每日4小时后相关性显著下降，这些统计特征应在数据生成中通过算法模拟实现。

关联性原则强调不同数据实体间的逻辑一致性[4]。学生表与成绩表通过学号建立外键关联；课程表与成绩表通过课程号关联；这些关系在生成时必须严格保持。在海量数据生成中，可采用图数据库技术构建实体关系网络，确保跨表数据的一致性。

可控性原则是实现预设结果生成的关键[5]。测试数据不仅要模拟正常场景，还需包含各类异常模式，如成绩突降、作业提交时间异常波动、数据缺失等。通过在生成引擎中植入规则引擎，可

精确控制异常数据的比例和类型，从而全面测试平台的数据校验与容错能力。

基于上述标准和原则，构建包含基础信息层、学业表现层、行为特征层和关系网络层的四维学生数据模型，该模型一是引入时间维度，通过记录各指标的变化轨迹支持趋势分析；二是增加场景化标签，如“考试周”、“寒暑假”等时间特征，使生成数据能够模拟不同教学阶段的特点；三是嵌入质量校验规则，如“学分不得大于课程上限值”，在数据生成过程中实时拦截无效数据。

二、海量教育测试数据生成的技术架构与算法

面对海量学生数据的生成需求，传统单机工具已无法满足性能要求。基于 Hadoop 生态构建分布式数据生成平台，采用“主从架构 + 任务分片”的设计模式 [6]。主控节点负责任务分配与结果聚合，从节点执行具体数据生成，通过 ZooKeeper 实现分布式协调。平台采用数据并行策略，将生成任务按学生 ID 范围分片，每个从节点处理学生全量数据，这种架构使系统理论上可支持无限水平扩展。

平台核心组件包括元数据管理模块、任务调度模块、数据生成引擎、质量监控模块和结果存储模块。

针对不同类型的教育数据，设计三类核心生成算法如下：

身份特征数据生成采用规则引擎 + 概率模型的混合方法。对于学号、课程号等标识符，通过“前缀 + 序列码”的规则生成，如“CS20230001”表示计算机专业2023级第1号学生；对于姓名、地址等自然属性，基于统计语料库随机组合，如中文姓名采用“姓氏库（300个常见姓）+ 名字库（2000个常用字）”的组合策略；对于性别、民族等分类数据，按照实际分布比例生成，如高校学生性别比例约为1.12:1（男：女）。实践表明，组合使用规则和概率模型可使身份数据的真实性提升78%。

学业成绩数据生成引入教育测量理论，通过项目反应理论（IRT）模拟真实答题过程。首先建立各学科的知识点网络，然后为每个知识点设置难度参数（-3~3）和区分度参数（0.2~1.2）；最后根据学生能力参数（-2~2）计算答题正确率，生成最终成绩。

行为序列数据生成采用隐马尔可夫模型（HMM）构建学生行为模式。将“预习 - 学习 - 复习 - 作业 - 测试”定义为状态集合，通过转移概率矩阵控制状态流转；观测值生成则结合时间特征，如作业提交时间呈现双峰分布（18:00-20:00和22:00-23:00）；课堂互动数据引入泊松过程，模拟随机事件发生频率。

预设结果生成是教育测试的特殊需求，要求数据能够按照预期目标呈现特定规律，用于验证平台的分析和预测功能 [7]。本文提出基于目标函数的数据反向生成算法。在目标定义阶段，明确预设结果的具体指标，这些目标可通过 JSON 格式定义，支持多指标组合。在约束构建阶段，建立目标与生成参数的映射关系。在参数优化阶段，采用遗传算法求解最优生成参数。将预设目标的偏差作为适应度函数，通过选择、交叉、变异操作寻找最优参数组合。在数据生成阶段，使用优化后的参数驱动基础生成算法，同时引入扰动因子模拟真实数据的随机性。

三、数据质量评估体系

为确保生成数据的有效性，构建包含技术指标和业务指标的双层评估体系 [8]。

技术指标量化数据的物理特性，一是完整性，字段缺失率 = 缺失记录数 / 总记录数 ×100%，核心字段（如学号、姓名）需达到100% 完整。二是准确性，逻辑矛盾检测合格率 = 通过校验记录数 / 总记录数 ×100%，如“学分不得大于课程上限值”。三是一致性，跨表数据一致率 = 关联字段匹配数 / 总关联数 ×100%，如学生表与成绩表的学号匹配。四是时效性，数据更新延迟时长 = 采集时间戳 - 业务发生时间戳，实时数据要求延迟 <5分钟。

业务指标评估数据的场景适用性，一是分布相似度，生成数据与真实数据的分布差异，采用 KS 检验，P 值 >0.05为通过。二是异常覆盖率，预设异常模式的生成比例，如离群值、缺失值、逻辑错误等类型的覆盖率≥ 95%。三是决策有效性，基于生成数据的平台决策与预期结果的吻合度，如预警准确率、资源推荐精度等。四是隐私保护度，敏感信息泄露风险评估，采用数据安全影响评估（DSIA）方法。

四、数据治理平台测试数据生成实践

某校为推进智慧校园建设，于2023年启动数据治理平台项目，旨在整合教务、学工、科研等12个部门的业务系统数据，实现“一数一源、一源多用”[9]。平台测试阶段面临三大需求：一是生成2万在校生的全维度数据，验证系统存储能力；二是模拟各类异常数据场景，测试平台容错机制；三是按照预设教学目标生成数据，评估分析功能有效性。

针对测试需求，设计了“分层并行”的数据生成方案 [10]。在硬件架构上，采用3台物理服务器构建分布式集群，每台配置24核 CPU、128GB 内存和4TB SSD 存储，通过 Kubernetes 实现容器编排。软件层面部署了本文设计的四维数据模型引擎，集成 MySQL（关系型数据）、MongoDB（非结构化行为日志）和 Neo4j（实体关系网络）三种数据库，形成多模态数据存储架构。

数据分层生成策略具体实施如下：基础信息层采用“种子数据 + 批量衍生”模式，先人工构建真实学生样本作为种子，再通过规则引擎批量生成衍生数据，确保核心字段100% 符合标准；学业表现层采用 IRT 模型生成成绩记录，设置高等数学、大学物理等16门公共课的知识点网络，如某知识点难度参数设为1.8，区分度参数0.95等；行为特征层通过 HMM 模型生成365天 ×24小时的时序数据，包括图书馆入馆记录（日均1.2次 / 人）、校园卡消费（日均3.5笔 / 人）等8类行为日志；关系网络层构建“学生 - 课程 - 教师”三元关系，共生成120万条关联记录，其中每位学生平均关联6.8门课程。

为满足预设结果测试需求，定义了三类目标场景：一是教学质量分析场景，“预设计算机专业大三学生操作系统课程平均分

68±3分”，通过反向生成算法将该专业学生能力参数 μ 设为0.3，σ=0.8；二是异常行为监测场景，植入“连续7天未刷卡就餐”（预警失联风险）、“单周图书馆入馆时长>40小时”（预警极端学习行为）等12类异常模式；三是隐私泄露测试场景，在3%的记录中嵌入加密手机号，验证平台的数据脱敏功能。

在技术实现过程中，使用三项关键技术：一是分布式ID生成算法，采用“雪花算法+校区标识”方案，确保学生ID在跨服务器生成时无冲突，时间戳精度达到毫秒级；二是内存优化技术，通过数据分片（每片处理5000学生数据）和对象池化，将内存占用控制在64GB以内，避免错误；三是实时校验引擎，集成200+条业务规则（如“成绩必须在0-100分区间”），在数据写入时进行实时校验并记录异常日志，校验通过率从初始的82%提升至99.7%。

针对生成过程中发现的问题，实施三项优化措施：针对“课程成绩分布异常集中”问题，引入贝叶斯优化算法动态调整IRT模型参数；针对“行为日志时间戳重复”问题，增加随机扰动因子（±5秒）；针对“跨表关联失败”问题，开发分布式事务管理器，采用TCC模式保证数据一致性。这些措施使数据质量综合评分从86分（百分制）提升至95分。

从三个维度评估数据生成效果：规模指标显示共生成840万条记录（2万学生×42项指标），总数据量17GB，平均单条记录20.7KB；质量指标通过自动化测试验证，字段完整率99.98%，逻辑一致性99.7%，异常场景覆盖率95.8%；性能指标显示单批次生成耗时48分钟，平均吞吐量14.6万条/秒，达到设计目标的121.7%。

预设结果生成的准确性验证采用对比分析法：教学质量分析场景中，计算机专业操作系统课程实际生成平均分为67.8分，与预设目标（68±3分）偏差0.2分；异常行为场景中，12类预设模式全部被平台准确识别，平均响应时间0.4秒；隐私保护测试中，平台成功识别并脱敏98.3%的加密手机号。这些结果表明生成数据能够有效验证平台功能。

五、展望

未来研究可向三个方向拓展：一是智能化生成，引入生成式AI技术（如GPT模型）生成更具创造性的异常场景；二是跨模态融合，整合文本、图像、视频等多类型测试数据；三是联邦学习框架，实现多校测试数据的联合生成与共享，同时保护数据隐私。随着教育数字化的深入推进，测试数据生成将从“满足功能验证”向“支撑教育创新”演进，为智慧教育的发展提供更有力的支撑。

参考文献

[1] 何曼 . 教育部信息化专家组成员、华东师范大学教授顾小清：教育信息化进入数字化转型重要时期 [J]. 在线学习 ,2022,(4):14-17、80.

[2] 甄晓非 . 教育大数据平台信息共享影响因素分析 [J]. 现代情报 ,2019,(7):122-127.

[3] 董晓辉，郑小斌，彭义平 . 高校教育大数据治理的框架设计与实施 [J]. 中国电化教育 ,2019,(8):63-71.

[4] 彭雪涛 . 美国高校数据治理及其借鉴 [J]. 电化教育研究 ,2017,(6):78-83.

[5] 吴刚 . 高校大数据治理的价值结构 [J]. 中国成人教育 ,2018,(5):41-44.

[6] 董晓辉，马威 . 高校数据治理的价值与特征 [J]. 网络安全与数据治理 ,2023,(2):43-47.

[7] 刘金松 . 数据治理：高等教育治理工具转型研究 [J]. 中国电化教育 ,2018,(12):39-45.

[8] 余鹏，李艳 . 大数据视域下高校数据治理方案研究 [J]. 现代教育技术 ,2018,(6):60-66.

[9] 张世明，彭雪峰，黄河笑 . 开放大学数据治理框架研究 [J]. 中国电化教育 ,2018,(8):116-126.

[10] 周炜 . 大数据视域下高校数据治理优化路径研究 [J]. 教育发展研究 ,2021,(9):78-84.

