Advances in Education 教育进展, 2025, 15(5), 1482-1486
Published Online May 2025 in Hans. https://www.hanspub.org/journal/ae
https://doi.org/10.12677/ae.2025.155928

Hans汉斯

基于机器学习的初中数学课后作业诊断

梁佳媛[1]，龚　笛[2]，姜　荣[3*]

[1]西藏大学经济与管理学院，西藏　拉萨
[2]上海市尚文中学，上海
[3]上海对外经贸大学统计与数据科学学院，上海

收稿日期：2025年4月25日；录用日期：2025年5月23日；发布日期：2025年5月30日

摘　要

随着教育数字化转型步伐的加速推进，教育过程中产生的数据规模呈现爆发式增长态势。在此背景下，如何高效挖掘并运用教育大数据的深层价值，以推动教学质量的持续优化与革新，已成为教育界亟待破解的核心命题。本文聚焦初中数学课后作业数据，创新性地引入基于众数回归统计诊断框架的机器学习技术，深度剖析作业数据中的异常模式与潜在规律，旨在为一线教育工作者提供精细化、智能化的教学反馈机制，助力课堂教学实现“减量不减质、增效不增负”的双重目标。

关键词

初中数学，课后作业诊断，机器学习，众数回归

Diagnosis of Junior High School Math Homework Based on Machine Learning

Jiayuan Liang[1], Di Gong[2], Rong Jiang[3*]

[1]School of Economics and Management, Tibet University, Lhasa Tibet
[2]Shanghai Shangwen Middle School, Shanghai
[3]School of Statistics and Data Science, Shanghai University of International Business and Economics, Shanghai

Received: Apr. 25th, 2025; accepted: May 23rd, 2025; published: May 30th, 2025

Abstract

With the acceleration of the digital transformation of education, the scale of data generated in the process of education shows an explosive growth trend. In this context, how to efficiently tap and apply the deep value of educational big data to promote the continuous optimization and innovation

*通讯作者。

文章引用: 梁佳媛, 龚笛, 姜荣. 基于机器学习的初中数学课后作业诊断[J]. 教育进展, 2025, 15(5): 1482-1486.
DOI: 10.12677/ae.2025.155928

of teaching quality has become a core proposition that needs to be solved in the education sector. This paper focuses on junior high school math homework data resources, innovatively introduces machine learning technology based on mode regression statistical diagnosis framework, and deeply analyzes abnormal patterns and potential rules in the homework data, aiming to provide a refined and intelligent teaching feedback mechanism for front-line educators, and help classroom teaching achieve the dual goals of "reducing quantity without reducing quality, increasing efficiency without increasing negative".

Keywords

Junior Math, Homework Diagnostics, Machine Learning, Mode Regression

Copyright © 2025 by author(s) and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

1. 引言

党的二十大对教育工作作出重要战略部署，明确提出推进教育数字化进程，着重强调了构建全民终身学习的学习型社会与学习型大国的战略意义，并明确要求进一步强化教育信息化建设。教育数字化作为信息时代以数字技术为核心驱动力的教育形态，其内涵界定在学术界引发了广泛探讨与辩证思考。学者们通过多维度解析，最终达成共识：教育数字化是教育信息化发展的关键阶段(黄荣怀，2022 [1])。

随着技术迭代升级，教育信息化正深刻重塑教育教学范式。在传统“小数据”时代，教师教学主要依赖经验传承，表现为教师主导课程内容设计，以教科书知识为核心，通过讲授法借助纸笔、黑板等媒介实施知识传递，并以考试测评检验学习成效。然而，教育数字化浪潮下，传统教学模式已难以适应现代教育需求，数据驱动成为教学变革的核心动力(魏亚丽和张亮，2022 [2])。张丽和章志强(2024) [3]基于数据驱动视角揭示学生学习状态与学业质量的关联机制；龚笛(2024) [4]运用 TPACK 理论框架优化初中数学讲评课教学，通过数据分析实现精准施教；张群等(2025) [5]系统梳理中外教育数字资源研究现状，对比分析研究领域、热点、演进趋势、学术共同体等维度，揭示数字资源建设存在的问题并提出发展路径。

当前，教育信息化技术已广泛渗透至学校作业管理领域，推动数据驱动教学研究快速发展。但初中数学课后作业数据分析仍存在明显短板：研究数量有限、分析手段单一，机器学习等高级技术尚未普及，且缺乏针对教育场景的深度优化。更关键的是，研究成果与教学实践存在显著脱节，教师数据分析应用能力不足，缺乏系统性指导。在此背景下，基于机器学习的统计诊断方法凭借其异常值识别优势，为破解上述难题提供了新思路。数据删除模型作为经典工具，通过对比删除特定数据点后模型与原始模型的统计量差异，评估异常程度，其中 Cook 距离作为核心诊断统计量得到广泛应用(韦博成等，2009 [6])。Zhao 等(2019) [7]在高维数据背景下研究多影响点检测方法。陈实与姜荣(2024) [8]则利用相关系数探索分位数回归的群组诊断问题。现有分析方法多基于均值或分位数回归，但在数据分布失衡时，均值与中位数难以准确反映数据特征。众数回归(Lee, 1989 [9])适用于偏态数据，且在相同区间长度下，条件众数相较于条件均值或分位数能提供更短的预测区间。

本研究聚焦初中生课后作业数据，创新地构建基于众数回归的统计诊断方法，系统诊断学生在初中数学学习中的具体问题及根源。通过多维度数据挖掘，揭示学生在数学知识点上的薄弱环节，提出针对

性教学改进建议。研究成果旨在为教师优化教学设计、实施精准辅导提供科学依据，最终提升学生数学学习效果。

2. 方法设计

假设有 K 次作业数据集记为$\{D_1,\cdots,D_K\}$，每个子数据集 $D_j=\{\boldsymbol{x}_{ij}^{\top},y_{ij}\},i=1,\cdots,n_j$，且数据来自以下线性回归模型：

$$y_{ij}=\boldsymbol{x}_{ij}^{\top}\beta_0+\varepsilon_{ij},\ i=1,\cdots,n_j \tag{1}$$

其中 y_{ij} 是响应变量，$\boldsymbol{x}_{ij}$ 是 p 维协变量，β_0 是未知参数，ε_{ij} 为随机误差向量。利用 Yao 和 Li (2014) [10] 基于核密度方法估计模型(1)中的未知参数 β_0 如下：

$$\hat{\beta}=\arg\max_{\beta}\sum_{j=1}^{K}\sum_{i=1}^{n_j}\phi_h\left(y_{ij}-x_{ij}^{\mathrm{T}}\beta\right), \tag{2}$$

其中 $\phi_h(\cdot)=\phi(\cdot/h)/h$，$\phi(\cdot)$ 是关于 0 对称的核函数，h 是窗宽。根据 Yao 和 Li (2014) [10]，本文选取核函数为高斯核 $\varphi(u)=\mathrm{e}^{-\frac{u^2}{2}}/\sqrt{2\pi}$，带宽为 $h=n^{-1/7}$。由此可定义广义 Cook 距离(见韦博成等，2009 [6])：

$$GD_j=\left(\hat{\beta}-\hat{\beta}_{-j}\right)^{\top}\left\{\sum_{j=1}^{K}\sum_{i=1}^{n_j}\frac{x_{ij}x_{ij}^{\mathrm{T}}}{\sqrt{2\pi}h^5}\left\{\left(y_{ij}-x_{ij}^{\mathrm{T}}\beta\right)^2-h^2\right\}\mathrm{e}^{-\frac{\left(y_{ij}-x_{ij}^{\mathrm{T}}\beta\right)^2}{2h^2}}\right\}^{-1}\left(\hat{\beta}-\hat{\beta}_{-j}\right), \tag{3}$$

其中 $\hat{\beta}_{-j}$ 是去掉 D_j 数据集后，利用剩余 $K-1$ 个数据集和(2)式所得的估计量。

3. 实例分析

本研究采用的实际数据集涵盖 2023~2024 学年八年级某教学班在第 18 章至第 23 章学习阶段的课后作业情况，具体包括 27 名学生累计 77 次作业的完整记录，总数据量达 2079 条。作业得分作为衡量学生课堂知识掌握程度的核心指标，能够直观反映其学习成效；而平均每题用时则从时间维度侧面揭示学生对知识点的熟练程度。

为验证得分与用时之间的相关性，研究采用皮尔逊相关性检验。原假设设定为“得分与平均每题用时不相关”，备择假设为“两者存在相关性”。检验结果显示 P-value = 0.049，因此可以拒绝原假设，即得分与用时之间存在显著相关性。基于上述结论，研究构建线性回归模型以量化两者关系：$\boldsymbol{Y}=\alpha+\boldsymbol{X}^{\mathrm{T}}\beta+\varepsilon$，其中 $\boldsymbol{Y}$ 是课后作业的总得分(百分制)，$\boldsymbol{X}$ 为平均每题做题时间(单位：秒)，α 和 β 是回归系数，ε 是随机误差。

利用广义 Cook 距离(3)计算可得图 1，其中 K = 3，16，72 和 74 为异常数据。观察图 2 发现回归系数 β 的估计值与其余 73 次课后作业的估计值有明显的区别。查看数据发现第 K = 3 是 18.3 (1)反比例函数；第 K = 16 是 19.3 逆命题和逆定理；第 K = 72 是 23.2 事件发生的可能性；第 K = 74 是 23.3 事件的概率(2)。对 4 个异常作业的具体分析如下：得分与平均每题用时两者呈现负相关关系，即解题耗时增加往往伴随得分下降。这一现象表明，课后作业主要考察基础知识点，多数学生能够快速完成题目；而耗时显著增加的案例通常反映知识理解不足，导致得分偏低。这一结论与课后作业的基础题型的定位高度契合。进一步分析发现，异常课后作业(标记为 K = 3、16、72)呈现显著正相关关系(见图 2)，即解题耗时与得分同步上升。结合原始数据观察，涉及的知识点难度较高，学生普遍需要更多时间思考推导。针对此类现象，任课教师应在后续复习中重点强化这三个知识点的教学深度。值得注意的是，K = 74 模块呈现较强负相关特征(见图 2)，班级平均得分高达 89 分表明该知识点掌握情况良好。通过原始数据溯源发现，

该模块题型设计侧重基础概念，学生已形成稳定的知识图谱。因此，教师可在复习阶段适当精简该模块内容，将教学资源向薄弱环节倾斜。

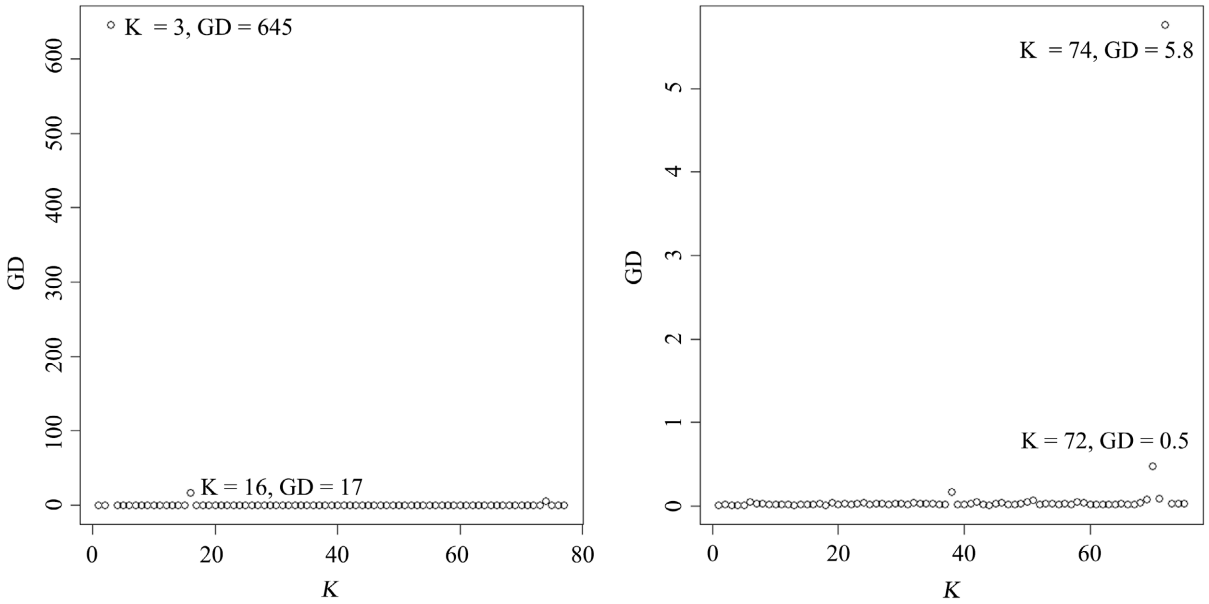

Figure 1. The left diagram shows the generalized Cook distance (GD) of 77 take-home assignments (*K*), and the right diagram shows GD of the remaining 75 after-class assignments after deleting the 3rd and 16th ones
图 1. 左图为 77 次课后作业(第 *K* 次)的广义 Cook 距离(GD)，右图为删除第 3 和第 16 次剩余 75 次课后作业的 GD

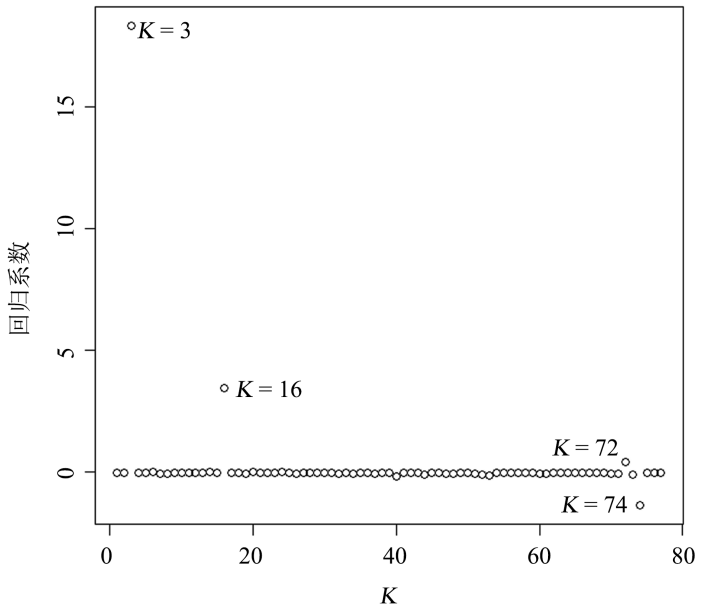

Figure 2. Estimates of regression coefficient β for 77 homework assignments (*K*)
图 2. 77 次课后作业(第 *K* 次)的回归系数 β 的估计值

基于上述诊断结果，建议教师构建动态调整机制：对 $K = 3$、16、72 模块实施“精准攻坚”策略，通过变式训练突破难点；对 $K = 74$ 模块采用“知识巩固+思维拓展”模式，防止重复性训练。这种基于数据驱动的差异化教学，既能保障复习效率，又能切实减轻学生负担，真正实现“减负增效”的教育目标。

4. 总结与展望

本研究运用广义 Cook 距离统计诊断技术，对初中数学课后作业数据开展系统性分析，精准定位了学生在数学认知体系中的知识薄弱环节。研究结果表明，学生在反比例函数、逆命题与逆定理推导、事件可能性分析等知识模块是显著薄弱点，而在事件概率计算领域则展现出相对优势。基于上述发现，特提出以下靶向教学优化策略：

1) 实施精准化分层复习教学。针对反比例函数图象性质、逆命题逻辑推演、事件可能性分析，建议教师采用“三阶递进”复习模式：① 知识重构阶段：通过概念图谱可视化教学，系统梳理知识脉络；② 变式训练阶段：设计“一题多解–多题归一”训练体系，强化思维迁移能力；③ 诊断测评阶段：运用动态测评系统实时监测学习轨迹，实现个性化辅导。

2) 构建动态化错题管理系统。建议搭建“三级错题资源库”：① 班级错题云平台：实时收录学生错题，运用自然语言处理技术自动标注知识点标签；② 错题解析微课库：针对高频错题开发 5 分钟微课视频，提供标准化解析模板；③ 个性化错题本：引导学生建立专属错题档案，运用康奈尔笔记法记录错误归因与改进策略。建议教师实施“错题三步走”教学策略：① 每周开展错题溯源分析会，运用鱼骨图工具定位认知盲区；② 每月组织错题变式挑战赛，通过游戏化机制激发纠错动力；③ 每学期编制《错题白皮书》，形成具有校本特色的教学资源包。

致　谢

感谢教育部人文社会科学研究青年基金(22YJC910005)的支持。

参考文献

[1] 黄荣怀. 教育数字化转型的国际理解与核心关切[J]. 上海教育, 2022(36): 16-17.

[2] 魏亚丽, 张亮. 从“基于经验”到“数据驱动”: 大数据时代的教学新样态[J]. 当代教育科学, 2022(2): 50-56.

[3] 张丽, 章志强. 基于学习状态的初中数学学业质量评价研究[J]. 上海教育科研, 2024(11): 50-55.

[4] 龚笛. 以学生为中心的初中数学 TPACK 课堂——以“七年级下学期数学期末复习”讲评课为例[J]. 现代教育, 2024(13): 20-21.

[5] 张群, 刘康, 蓝方翊, 张慧. 数智化时代教育数字资源: 现状、问题及发展路径[J]. 现代教育, 2025, 39(2): 182-194.

[6] 韦博成, 林金官, 解锋昌. 统计诊断[M]. 北京: 高等教育出版社, 2009.

[7] Zhao, J., Liu, C., Niu, L. and Leng, C. (2019) Multiple Influential Point Detection in High Dimensional Regression Spaces. *Journal of the Royal Statistical Society Series B*: *Statistical Methodology*, **81**, 385-408. https://doi.org/10.1111/rssb.12311

[8] 陈实, 姜荣. 分布式系统下基于分位数回归的统计诊断[J]. 上海第二工业大学学报, 2024, 41(3): 307-314.

[9] Lee, M. (1989) Mode Regression. *Journal of Econometrics*, **42**, 337-349. https://doi.org/10.1016/0304-4076(89)90057-2

[10] Yao, W. and Li, L. (2013) A New Regression Model: Modal Linear Regression. *Scandinavian Journal of Statistics*, **41**, 656-671. https://doi.org/10.1111/sjos.12054