计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2023,32(9):239−247 [doi: 10.15888/j.cnki.csa.009261]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于编程认知诊断模型的学生表现预测①

张雨婷, 李　征, 刘　勇, 吴永豪

(北京化工大学 信息科学与技术学院, 北京 100029)
通信作者: 刘　勇, E-mail: lyong@mail.buct.edu.cn

摘　要: 近年来, 学生认知诊断是教育数据挖掘领域的重要研究课题, 对现代教育的精准反馈有重要的意义. 然而, 传统的认知诊断模型存在预测准确性低和处理大规模数据时效率低等问题, 且现有研究主要围绕传统线下教学展开, 缺少针对程序设计教育领域的研究. 为了解决上述问题, 本文从程序设计教育的特点分析出发, 提出了一种基于编程表现的模糊认知诊断模型 P-FuzzyCDF (programming-performance-based fuzzy cognitive diagnosis framework). 具体来说, 为了处理编程题部分正确的情况, 该模型首先模糊了学生对知识点的掌握情况. 随后, P-FuzzyCDF 将模糊集合理论与教育假设相结合, 对学生对问题的掌握情况进行了建模. 除此之外, 本文还考虑抄袭因素, 并最终生成学生在每个问题上的得分. 值得注意的是, 该模型利用编程教育数据可视化和精确性的特点, 对模型中每个部分的参数进行了量化. 本文基于真实数据集进行实验, 实验结果表明 P-FuzzyCDF 可以实现较高的精度, 其中 *MAE*、*MSE* 和 *RMSE* 评估指标的值分别为 0.07、0.09 和 0.01. 此外, 将 P-FuzzyCDF 与现有经典方法 (如 DINA, IRT 和 FuzzyCDF) 进行比较时, P-FuzzyCDF 的结果在 *MAE*、*MSE* 和 *RMSE* 等指标上取得了明显优势.

关键词: 教育数据挖掘; 认知诊断; 学生表现; 在线教育; 学生行为特征

引用格式: 张雨婷,李征,刘勇,吴永豪.基于编程认知诊断模型的学生表现预测.计算机系统应用,2023,32(9):239–247. http://www.c-s-a.org.cn/1003-3254/9261.html

Student Performance Prediction Based on Cognitive Diagnosis Model

ZHANG Yu-Ting, LI Zheng, LIU Yong, WU Yong-Hao

(College of Information Science and Technology, Beijing University of Chemical Technology, Beijing 100029, China)

Abstract: In recent years, student cognitive diagnosis has been an important research topic in educational data mining, which is of great significance for accurate feedback in modern education. However, traditional cognitive diagnosis models have problems such as low prediction accuracy and low efficiency when dealing with large-scale data. Moreover, the existing research is mainly focused on traditional offline teaching and learning, and more research is needed in programming education. To solve the above problems, a programming-performance-based fuzzy cognitive diagnosis framework (P-FuzzyCDF) is proposed from the analysis of the characteristics of programming education. First, to deal with the case of partially correct programming questions, the model fuzzes the students' mastery of the knowledge points. Second, fuzzy set theory is combined with educational assumptions to model student mastery of the questions. Finally, students' scores on each problem are generated by considering plagiarism factors. Notably, the model takes advantage of the visualization and accuracy of programming education data to quantify the parameters for each model component. Experiments are conducted based on real data sets, and the results show that P-FuzzyCDF can achieve high accuracy, where the values of *MAE*, *MSE*, and *RMSE* assessment indexes are 0.07, 0.09, and 0.01, respectively. In addition, when comparing P-FuzzyCDF with existing classical methods such as DINA, IRT, and FuzzyCDF, the results of P-FuzzyCDF are significantly better than these methods in terms of *MAE*, *MSE*, and *RMSE*.

① 基金项目: 北京化工大学校级教改项目 (2021BHDJGYB16, G-JG-PTKC202107)
收稿时间: 2023-02-10; 修改时间: 2023-04-07; 采用时间: 2023-05-11; csa 在线出版时间: 2023-07-14
CNKI 网络首发时间: 2023-07-17

Key words: educational data mining; cognitive diagnosis; student performance; online education; student behavior characteristics

1 引言

在大数据时代背景下, 教育数据挖掘领域相关研究迅速发展, 其目的是从大规模的教育数据中提取有价值的信息. 教育数据挖掘的关键任务之一就是利用学生的考试数据进行建模, 从而获知学生的潜在认知状态[1]. 其中, 认知状态是指学生在学习过程中对所学知识的学习情况.

传统的考试评估通常只报告一个笼统的总分或能力分数, 而忽略了个体之间存在的认知状态差异. 但是即使考试成绩相同的学生, 也有可能具有不同的认知状态[2]. 为了解决传统考试评估的缺陷, 最新的研究提出了认知诊断模型, 该模型能把认知过程与测量手段结合起来, 不仅能对学生的整体水平做出评估, 同时还可以将学生的认知结构模式化. 其利用合适的测量模型对不同的认知结构模式进行诊断, 从而定量地考察学生的认知状态与个体差异[3].

具体来说, 认知诊断模型的有效性是通过预测学生表现 (predict student performance, PSP) 来实现的. 此外, PSP 可以进一步应用于许多方向, 如个性化的补救建议和教学计划的改进[4]. 由于认知诊断结果具有很强的可解释性, 研究人员投入了大量精力来设计合适的认知诊断模型, 以提高个性化学习的质量[5].

目前常见的认知诊断模型可以分为两类: 离散型和连续型. 离散型认知诊断模型包括 DINA 模型 (deterministic inputs, noisy and gate), 连续型认知诊断模型包括 IRT 模型 (item response theory). 最近, 也有许多新的模型被提出, 如 FuzzyCDF[6], R-FuzzyCDF[7] 和 NeuralCDM[8].

然而, 在上述认知诊断模型中仍存在一些限制. 首先, 现有的认知诊断模型考虑了对客观题和主观题的预测情况, 但缺少对这两类题目的细化研究. 例如, 编程题属于主观题, 但是和传统意义上的主观题又存在差异, 编程题具有更加严格的约束性和限制性. 因此, 现有的认知诊断模型在分析编程题时存在精度不足的问题. 其次, 尽管传统的认知诊断模型在小规模数据环境中具有良好的性能, 但是由于收敛速度慢, 此类模型在处理大规模数据时的执行效率显著降低.

为了解决这些问题, 本文提出了一种基于编程表现的模糊认知诊断模型 P-FuzzyCDF (programming-performance-based fuzzy cognitive diagnosis framework). 具体来说, 该模型首先基于学生的潜在特质来模糊化学生对特定知识点的掌握程度. 随后, 该模型基于编程题掌握知识点越多, 分数越高的特性, 模糊化了学生对问题的掌握程度. 此外, 该模型会基于学生由于抄袭而答对题目的情况, 生成学生在每道题目上的最终得分. 其中, 我们使用了学生在校期间的学习数据以及编程数据来分析学生的认知状态, 从而得到模型中的重要参数.

为了评估提出的方法, 本文在 4 个真实的数据集上进行了实验. 实验结果表明, 本文提出的方法与现有经典方法相比, 在保证预测准确性的基础上, 减少了时间开销. 综上所述, 本文的贡献如下.

(1) 提出了一种基于编程表现的模糊认知诊断模型 P-FuzzyCDF, 并将该模型应用在了学生表现预测方面, P-FuzzyCDF 弥补了传统认知诊断模型在程序设计教育领域研究的不足.

(2) 引入了学生在校期间的编程数据信息, 将其应用在 P-FuzzyCDF 中, 从而提高模型的准确率.

(3) 多个数据集上的实验结果表明, 本文提出的 P-FuzzyCDF 方法优于基准方法.

2 相关工作

本节介绍了现有的几种被广泛应用的认知诊断模型: DINA, IRT, FuzzyCDF (fuzzy cognitive diagnosis framework).

2.1 DINA 模型

DINA 模型是一种典型的离散型认知诊断模型. 该模型将学生描述成一个多维的知识点掌握向量, 从学生实际作答结果入手进行诊断. DINA 模型简单, 参数的可解释性较好, 且 DINA 模型的复杂性不受属性个数的影响[9].

式 (1) 展示了 DINA 模型定义的学生i在问题j上的作答情况:

$$\eta_{ij}=\prod_{k=1}^{K}\alpha_{ik}^{q_{jk}} \tag{1}$$

其中, η_{ij}是指学生i在问题j上的潜在作答情况, α_{ik}是指学生i对知识点k的掌握情况, q_{jk}是指问题j对知识点k的考察情况. $\eta_{ij}=1$表示学生i已经掌握问题j包含的所有知识点; $\eta_{ij}=0$表示答错, 学生i对问题j中的知识点至少有一个没有掌握.

DINA 模型联合试题知识点关联矩阵Q和学生答题情况X矩阵对学生进行建模, 引入问题参数s_j, g_j. s_j表示学生在掌握了问题j所考察的所有知识点的情况下做错的概率; g_j表示学生在并不完全掌握问题j所考察的所有知识点下猜对的概率.

式 (2) 表示在已知学生i的知识点掌握情况α_i的条件下, 答对问题j的概率:

$$P_j(\alpha_i)=P_j(X_{ij}=1|\alpha_i)=g_j^{1-\eta_{ij}}(1-s_j)^{\eta_{ij}} \quad (2)$$

其中, X_{ij}表示学生i在问题j上的得分情况[10].

由于 DINA 模型在客观题的预测中准确性高, 可解释性强, 所以被广泛应用在认知诊断中. 但是, DINA 模型只能把学生的潜在认知状态分为两类, 即完全未掌握 (0) 或完全掌握 (1). 这并不符合编程题目需要多级评分的特点, 从而使得 DINA 模型在预测学生在编程题的认知状态时的准确性和精度都有所下降[11].

2.2 项目反应理论

项目反应理论 (IRT) 是一种典型的连续型认知诊断模型, 被广泛应用在心理学和教育测量领域. IRT 根据学生回答问题的情况, 通过对题目特征函数的运算, 来推测学生的能力. IRT 的题目参数有: 区分度a、难度b和猜测系数c. 根据参数的不同, 特征函数可分为单参数模型 (难度)、双参数模型 (难度、区分度) 和三参数模型 (难度、区分度、猜测参数) 等[12].

式 (3) 展示了 IRT 的双参数模型:

$$\alpha=1/(1+\exp[-D\times a\times(\theta-b)]) \quad (3)$$

其中, α为学生的学习状态, θ为学生的潜在特征水平, D为经验参数, 一般为 1.7[13].

IRT 模型相比于 DINA 模型可以进行多级评分, 使用潜在变量来描述一个学生. 但是, IRT 模型对测验条件要求较为严格, 样本容量要大, 被试者的能力分布范围要广, 测试题目数量要多, 这些条件如果没被满足则会影响其精确性[14].

2.3 FuzzyCDF

FuzzyCDF 模型将模糊理论应用到认知诊断中, 可以同时对学生作答客观题和主观题进行诊断, 解决了传统认知诊断模型无法有效诊断主观题的问题[5]. FuzzyCDF 模型假设在客观题作答中, 学生要掌握题目所涉及的全部知识点才能掌握题目. 而在主观题作答中, 学生仅需掌握题目所涉及的部分知识点即表明该学生掌握该题目.

在 FuzzyCDF 模型中, 学生正确回答客观题和主观题的概率公式分别为式 (4) 和式 (5):

$$P(X_{ij}=1|\eta_{ij},s_j,g_j)=(1-s_j)\eta_{ij}+g_j(1-\eta_{ij}) \quad (4)$$

$$P(X_{ij}|\eta_{ij},s_j,g_j)=N(X|[(1-s_j)\eta_{ij}+g_j(1-\eta_{ij})],\sigma^2) \quad (5)$$

其中, σ^2为主观题得分的方差[15].

FuzzyCDF 模型相比于传统的认知诊断模型, 考虑了客观题和主观题的不同情况, 但是缺少对编程题这一特殊类型的题目的考虑. 此外, FuzzyCDF 模型需要应用采样算法进行参数估计, 所以面临着高计算复杂度的问题和需要大量训练数据的问题[15].

3 个性化的模糊认知诊断模型

本节将介绍本文提出的基于编程表现的模糊认知诊断模型 (P-FuzzyCDF). 如图 1 所示, P-FuzzyCDF 由 4 个部分组成, 自上而下分别是学生的潜在特质、学生对知识点的掌握程度、学生对问题的掌握程度以及预测得到的问题得分. 其中, 我们使用了学生在校期间的学习数据以及编程数据来分析学生的认知状态, 从而得到模型中的重要参数. 为了更好地说明, 表 1 展示了建模过程中的一些重要的数学符号, P-FuzzyCDF 的每个步骤将在第 3.1–3.4 节中详细说明.

3.1 模糊化知识点掌握程度

本节将展示如何获知学生对特定知识点的掌握情况. 在基于 DINA 的认知诊断模型中, 其把学生对知识点的掌握程度假设为完全掌握和完全未掌握, 该模型适用于只有正确和错误两种选项的客观题目[16]. 然而, 在计算机编程课程中, 对于仅满足了部分题目要求的编程题, DINA 模型并不适用. 因此, 为了解决这个问题, P-FuzzyCDF 将模糊集合理论引入认知诊断模型中, 从而使二元变量 (0 或 1) 模糊为[0,1]之间的连续变量. 根据项目反应理论, 每一个学生都有一个高阶潜在特质和对知识点的潜在认知程度[17]. 同时, 每一个知识点都有一个自身属性 (即知识点难度). 三者共同影响着学生对知识点的掌握程度. 依据在项目反应理论中采用的双参数模型[17], 本文将学生j对知识点k的掌握程

度α_{jk}和$\mu_k(j)$定义为:

$$\alpha_{jk}=\mu_k(j)=1/(1+\exp[-1.7\times a_{jk}\times(\theta_j-b_k)]) \quad (6)$$

式 (6) 表明学生对知识点的掌握程度受到学生的潜在特质 (θ_j)、学生对知识点的辨别力 (a_{jk}) 和知识点的难度系数 (b_k) 的共同作用. 其中, −1.7 为经验参数, 能够最小化正态分布函数与逻辑斯谛分布函数的最大差异. 此外, 为了解决参数估计面临的高计算复杂度和需要大量训练数据的问题, P-FuzzyCDF 利用教育数据以及编程数据精确性的特点, 提出了以下 3 个教育假设对θ_j, a_{jk}和b_k进行参数估计.

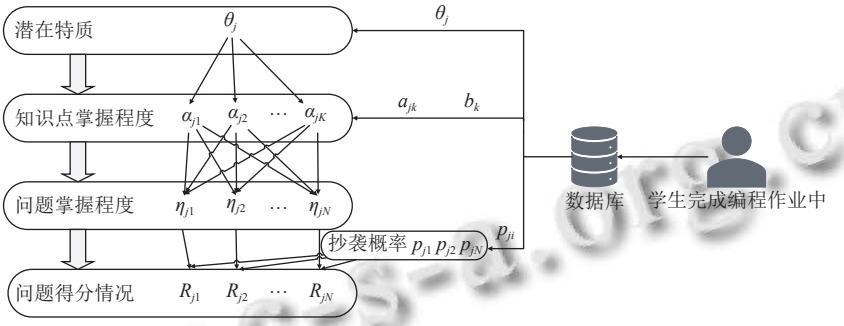

图 1 4 层模糊认知诊断模型的框架

表 1 一些重要的数学符号

数学符号	描述
θ_j	学生j的潜在特质
a_{jk}	学生j对知识点k的辨别力
b_k	知识点k的难度系数
α_{jk}	学生j对知识点k的掌握情况
μ_k	与知识点k有关的模糊集的隶属函数
η_{ji}	学生j对编程题i的掌握情况
p_{ji}	学生j在编程题i上的抄袭情况
R_{ji}	学生j在编程题i上的得分情况
Q_{jk}	每道编程题i包含的知识点k

假设 1: 学生的高阶潜在特质可以由学生当前的学业 GPA 决定.

学生的高阶潜在特质指学生的能力水平, 而学生的能力水平一般表现在学生的考试成绩 (即 GPA). 因此, θ_j可以由学生的 GPA 来量化.

假设 2: 学生对知识点的潜在认知程度可以由含有该知识点的题目的历史通过率决定.

相比于传统的线下教育, 在编程教育领域中, 教师可以得到详细的学生在编程过程中的学习数据. 因此, a_{jk}可以通过具体的计算来量化, 即a_{jk}可以通过学生j对含有知识点k的题目的通过率获得.

假设 3: 知识点的难度系数可以由教师对知识点的难易程度进行人工评级来决定.

教师作为知识的传播者对每个知识的都有着很深入的认识, 因此, b_k可以通过老师对知识点进行人工评级获得.

3.2 模糊化问题掌握程度

基于第 3.1 节中模糊化的知识点掌握程度, 我们可以进一步模糊化学生对问题的掌握程度 (即能够解决问题的概率). 在模糊认知诊断模型中, 学生对问题的掌握程度受到学生对该问题所需知识点的掌握程度的影响.

知识点在问题上的相互作用主要分为联结型和补偿型[18]. 联结型是指学生只有掌握了解决问题所需要的全部知识点才能答对问题, 补偿型是指学生只要掌握了解决问题所需要的部分知识点就可以获得该题目的部分分数. 对于编程题而言, 学生掌握的知识点越多, 在这道题目上的得分就越高. 因此, 我们假设知识点对编程题的相互作用是补偿型的. 那么学生在这一道题目上的掌握程度就是学生对这道题目所需知识点的掌握程度的并集. 学生j对编程题i的掌握程度为:

$$\eta_{ji}=\mu_{\cup_{1\leqslant k\leqslant K,q_{ik}=1}k}(j) \quad (7)$$

其中, q_{ik}表示解决问题i是否需要掌握知识点k, 0 表示不需要掌握, 1 表示需要掌握. 采用标准模糊并运算[19], 公式为:

$$\mu_{A\cup B}(x)=\max(\mu_A(x),\mu_B(x)) \quad (8)$$

3.3 预测学生问题得分

由第 3.2 节可以确定学生在问题上的掌握程度. 在本节中, 考虑了一种例外情况 (即学生在作答题目的过

程中存在抄袭行为), 并采用高斯分布预测编程题的分数. 在实际答题过程中, 学生的题目分数不仅与学生对题目的掌握程度有关, 还与学生是否抄袭有关. 同时考虑到编程题具有多级评分的需求, 将题目的得分划分为 [0, 1] 之间的连续变量来归一化编程题的分数. 然后, 假设学生在编程题上的得分服从高斯分布, 这在研究中被广泛应用[6,7,15]. 结合学生对题目的掌握情况以及抄袭因素可以得到学生的真实作答得分:

$$P(R_{ji}|\eta_{ji},p_{ji})=N(R_{ji}|p_{ji}(1-\eta_{ji}),\sigma^2) \tag{9}$$

其中, R_{ji}是指学生j在编程题i上的得分情况, p_{ji}是指学生的抄袭概率. $p_{ji}(1-\eta_{ji})$表示学生通过抄袭得到了正确答案. σ^2表示题目标准化得分的方差. 对于p_{ji}的计算做了以下教育假设.

假设 4: 学生在一道题目上的抄袭概率可以由学生历史答题的抄袭比率决定.

在计算机编程教育领域, 可以通过在线编程平台获得每个学生在每道题目上的抄袭概率, 当抄袭概率大于等于 80% 时, 学生的抄袭可能性较高. 因此, 如果学生在一道题目上的抄袭概率大于等于 80%, 则假定该学生通过抄袭来回答题目. 那么p_{ji}可以由学生抄袭回答题目总数除以学生回答题目总数来得到.

3.4 模型总结

为了更好地说明本文提出的 P-FuzzyCDF 模型, 使用如图 2 所示的模型图来表示. 得分矩阵R_{ji}包括M个学生在N道编程题上的分数. 知识点矩阵Q_{jk}表示每道编程题包含的知识点, 如果答对编程题i需要掌握知识点k, 那么$q_{jk}=1$. 学生对知识点的掌握程度α_{jk} ($k=1,2,\cdots,K$)取决于学生j的潜在特质θ_j, 学生j对知识点k的辨别力a_{jk}和知识点k的难度系数b_k. 学生对编程题的掌握程度η_{ji}由α_{jk} ($q_{jk}=1$)决定. 学生在编程题上的得分R_{ji}由η_{ji}和p_{ji} (学生j在编程题i上的抄袭情况) 决定.

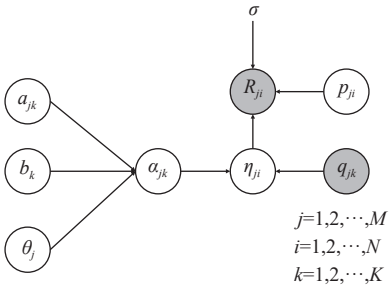

图 2 P-FuzzyCDF 模型

4 实验设计

为了验证 P-FuzzyCDF 方法的有效性, 在真实数据集上设置了对比实验. 本节首先在 PSP 任务上, 将 P-FuzzyCDF 与基准方法进行对比. 其次, 使用假设检验验证了 P-FuzzyCDF 方法的有效性. 最后, 通过案例分析评估预测结果的可解释性.

4.1 数据集

实验数据来自北京化工大学在教学过程中学生与 OJ 系统交互所产生的学习行为数据以及教务管理系统中的学生基础数据. 如表 2 所示, 共包含 4 个数据集, 分别来自 4 门编程课程, 共有 206 道编程题, 总计 531 名学生参与. 每个数据集中包括了得分矩阵R_{ji}和知识点矩阵Q_{jk}. 如表 3 所示, 得分矩阵R_{ji}每一行代表一个学生, 每一列代表一道编程题的得分, 得分区间在[0,1]之间. 如表 4 所示, 知识点矩阵Q_{jk}每一行代表一道编程题, 每一列代表编程题考察的知识点. 1 表示该题目考察了这个知识点, 0 表示该题目没有考察这个知识点.

表 2 数据集信息汇总

项目	Python国际化课程	程序设计基础	ACM/ICPC程序设计竞赛方法与实践	数据结构	总计
题目数量	54	56	45	51	206
参与人数	28	63	213	227	531
知识点个数	48	32	47	58	185

表 3 得分矩阵示例

学生	题目1	题目2	题目3
学生1	0.6	0.9	0.45
学生2	0.78	0.5	0.7
学生3	0.9	0.73	0.56

表 4 知识点矩阵示例

题目	知识点1	知识点2	知识点3	知识点4
题目1	1	0	1	1
题目2	0	0	1	1
题目3	1	0	1	0

4.2 标签

由于在 OJ 系统中学生每道编程题的得分情况只有两种 (完全正确和完全错误), 只是用 OJ 系统中的得分不符合编程题作为主观题的特性. 因此, 我们提出了一种计算得分矩阵R_{ji}的方法.

首先, 根据提交次数与题目是否正确之间的关系来定义编程题的初始分数C_j. 如表 5 所示, 分为两种情

况, 第 1 种为学生最终完全答对编程题, 第 2 种为学生最终答错编程题.

其次, 我们根据学生在 OJ 系统中的排名对学生的编程题得分进行二次定义. 按照排名顺序将分数定义为等差数列, 排名第一的学生分数定义为$S_{1i}=1$, 排名最后的学生分数定义为$S_{Mi}=0.1$, 公差d为: $d=(S_{Mi}-S_{1i})/M$, 其余学生的得分为$S_{ji}=S_{1i}+(M-1)\times d$.

最后, 通过将上述两个得分加和取平均即可得到学生的最终分数$R_{ji}=(C_{ji}+S_{ji})/2$.

表 5 定义初始分数

回答正确/错误	提交次数	初始分数
正确	1	1
	2	0.95
	3	0.85
	4	0.8
	5	0.7
	6–9	0.6
	10–12	0.55
	13–15	0.5
	>15	0.45
错误	0	0
	1	0.05
	2	0.1
	3	0.15
	4	0.2
	5–6	0.25
	7–8	0.3
	9–10	0.35
	>10	0.4

4.3 评价指标

我们使用 3 种不同的指标 (即 *MAE*、*MSE* 和 *RMSE*) 来评估性能. 这 3 个性能指标在现有关于认知诊断的研究中被广泛使用[15].

如式 (10) 所示, *MAE* 是预测得分和实际得分之间的绝对差值的平均值, 它衡量的是预测误差的大小. *MAE* 值越小, 表示预测误差越小. 如式 (11) 所示, *MSE* 是预测得分与实际得分之差平方的期望值, 它可以评价数据的变化程度, *MSE* 值越小, 说明预测模型具有更好的精确度. 如式 (12) 所示, *RMSE* 是 *MSE* 的算数平方根, 用于指示模型在预测中会产生的误差规模, 对于较大的误差, 权重较高, *RMSE* 越小越好.

$$MAE=\sum_{j=1}^{m}|y_j-\bar{y}_j|/M \tag{10}$$

$$MSE=\sum_{j=1}^{m}(y_j-\bar{y}_j)^2/M \tag{11}$$

$$RMSE=\sqrt{\sum_{j=1}^{m}(y_j-\bar{y}_j)^2/M} \tag{12}$$

4.4 基准方法

在实验中考虑了 3 个方法进行对比实验, 分别是 DINA, IRT 和 FuzzyCDF, 它们参数都是通过参数估计算法得到的[15]. 具体描述如下:

(1) DINA[7]: 一种经典的离散型认知诊断模型. 该模型在给定知识点矩阵的情况下, 对学生的认知状态进行建模, 结合回答问题时的例外情况 (失误因素s_j, 猜测因素g_j) 预测学生表现. 预测得到的学生题目分数仅分为两种情况 (1 满分, 0 不得分). 采用最大期望算法对模型中的参数 (s_j, g_j) 进行估计. 此外, 在参数估计时, 每一个问题都会有一个失误因素s_j和一个猜测因素g_j.

(2) IRT[13]: 一种经典的连续型认知诊断模型. 该模型通过评估学生的潜在特征θ与题目参数 (区分度a、难度b) 来预测学生表现. 预测得到的题目分数处于 [0, 1] 之间. 采用最大期望算法对模型中的参数 (θ, a和b) 进行估计. 其中, 每一个学生都会有一个潜在特征θ, 每一个问题都会有一个区分度参数a和难度参数b.

(3) FuzzyCDF[15]: 该模型将模糊理论应用到认知诊断中, 基于学生的潜在特征θ, 题目参数 (区分度a、难度b) 和回答问题时的例外情况 (失误因素s_j, 猜测因素g_j) 预测学生表现. 预测得到的题目分数处于 [0, 1] 之间. 该模型使用蒙特卡罗和马尔科夫链来估计上述参数. 在进行参数估计时, 训练数据为 80%, 测试数据为 20%. 同样, 每一个学生都会有一个潜在特征θ, 每一个问题都会有一个区分度参数a, 难度参数b, 失误因素s_j和猜测因素g_j.

4.5 实验结果与分析

4.5.1 P-FuzzyCDF 模型的有效性

为了评估 P-FuzzyCDF 的有效性, 使用第 4.1 节描述的数据集, 将其与基准方法 (如第 4.4 节所述) 进行了对比. 使用 3 个评价指标来评估 P-FuzzyCDF 的有效性: *MAE*, *MSE* 和 *RMSE*. 这 3 个评价指标的值越接近 0, P-FuzzyCDF 在 PSP 任务上就越准确.

表 6 为在不同的数据集上, P-FuzzyCDF 与基准方法的对比实验结果. 从表中数据可知, P-FuzzyCDF 的表现优于所有基准方法. 具体来说, 在程序设计基础数据集中, 相比于 DINA, P-FuzzyCDF 在 *MAE*, *MSE* 和

RMSE 上效果分别提升了 58.8%, 55% 和 75%. 相比于 IRT, P-FuzzyCDF 在 *MAE*, *MSE* 和 *RMSE* 上效果分别提升了 58.8%, 55% 和 75%. 相比于 FuzzyCDF, P-FuzzyCDF 在 *MAE*, *MSE* 和 *RMSE* 上效果分别提升了 63.2%, 62.5% 和 83.3%.

表 6 Python 国际化课程数据集实验结果

数据集	模型	*MAE*	*MSE*	*RMSE*
Python国际化课程	DINA	0.24	0.29	0.09
	IRT	0.15	0.2	0.04
	FuzzyCDF	0.28	0.34	0.11
	P-FuzzyCDF	**0.13**	**0.17**	**0.03**
程序设计基础	DINA	0.17	0.2	0.04
	IRT	0.17	0.2	0.04
	FuzzyCDF	0.19	0.24	0.06
	P-FuzzyCDF	**0.07**	**0.09**	**0.01**
ACM/ICPC程序设计竞赛方法与实践	DINA	0.2	0.23	0.05
	IRT	0.19	0.22	0.05
	FuzzyCDF	0.22	0.27	0.08
	P-FuzzyCDF	**0.11**	**0.14**	**0.02**
数据结构	DINA	0.18	0.22	0.05
	IRT	0.17	0.21	0.04
	FuzzyCDF	0.23	0.29	0.08
	P-FuzzyCDF	**0.13**	**0.16**	**0.03**

图 3–图 6 为 P-FuzzyCDF 与基准方法在每一道编程题上的详细对比. 如图 4 所示为程序设计基础数据集下的实验结果, 该数据集共包含 56 道题目, 与基准方法相比, P-FuzzyCDF 在每一道题目的预测结果上均有显著优势. 从图 4–图 6 中可以看出, 在程序设计基础数据集, ACM/ICPC 程序设计竞赛方法与实践数据集和数据结构数据集上, P-FuzzyCDF 明显优于基准方法.

具体来说, P-FuzzyCDF 利用在线编程教育数据精确性的特点, 提出了 4 个教育假设进行参数估计, 该参数估计方法更符合编程题的特征, 因此, 相比于基准方法, P-FuzzyCDF 在预测编程题的成绩时表现更好. 此外, P-FuzzyCDF 在对学生的学习认知状态的诊断结果为连续变量 (即学生对知识点的掌握程度为 [0, 1] 区间内的连续值), 而 DINA 模型的诊断结果为离散值 (1 表示完全掌握, 0 表示完全没有掌握). IRT 模型没有考虑学生对知识点的掌握情况, 仅使用一个潜在的连续型数值变量对学生成绩进行表示. FuzzyCDF 模型虽然可以将学生的认知状态表示为连续变量, 但缺少对编程题的具体分析, 忽略了编程题的自身特点. 而 P-FuzzyCDF 在预测学生成绩时, 考虑了学生在答题时的抄袭因素, 使预测结果更接近实际作答结果. 因此, 相较于传统的认知诊断方法, P-FuzzyCDF 模型不仅提高了预测的准确性还保证了结果的可解释性.

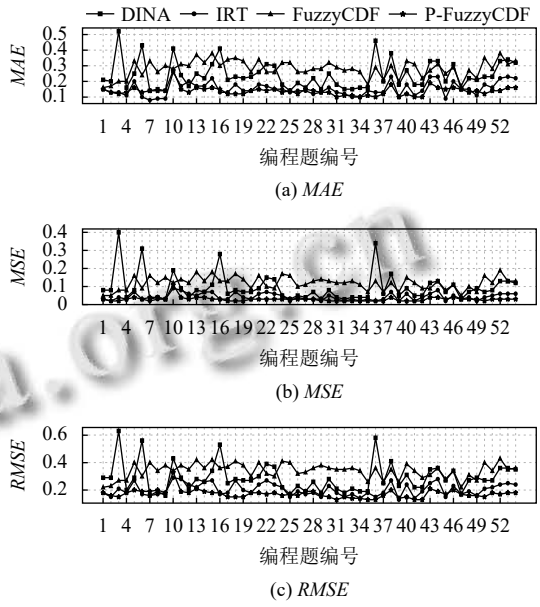

图 3 Python 国际化课程数据集的详细实验结果

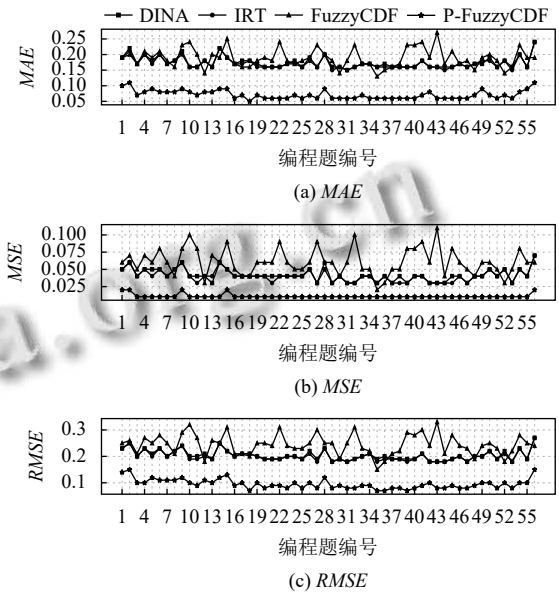

图 4 程序设计基础数据集的详细实验结果

4.5.2 假设检验

从图 3 中可以看出, 在 Python 国际化课程数据集上, P-FuzzyCDF 没有明显的优势. 因此, 在本节进行了 Wilcoxon 符号秩检验[20], 以验证 P-FuzzyCDF 的竞争力. 表 7 显示了上述假设检验在 Python 国际化课程数据集的结果. 注意, 我们实现了 3 种现有方法, 因此我

们只能对这 3 种方法进行假设检验. 我们研究中使用的假设如下: H0: P-FuzzyCDF 和其他方法在 *MAE*, *MSE* 和 *RMSE* 方面没有显著差异. 该检验的显著性水平设置为 0.05. 表 7 显示, 所有 P-value 均低于 0.05, 则统计结果导致拒绝零假设. 这些结果表明, 我们提出的方法与其他方法在 *MAE*, *MSE* 和 *RMSE* 度量方面存在显著差异. 需要注意的是, 图 4–图 6 中展示的结果表明, P-FuzzyCDF 比其他现有方法在评估指标 *MAE*、*MSE*、*RMSE* 上取得了显著的优势, 因此可以安全地得出结论, P-FuzzyCDF 可以比基准方法显著获得更好的性能.

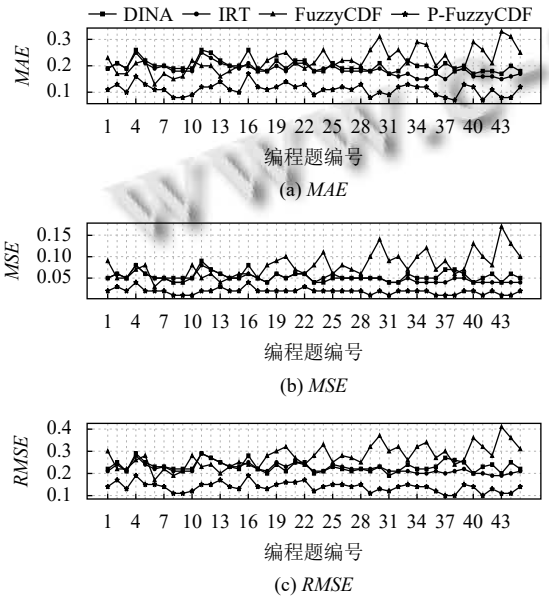

图 5 ACM/ICPC 程序设计竞赛方法与实践数据集的详细实验结果

4.5.3 案例分析

为了验证 PEP 结果的可解释性, 使用了 DINA 和 P-FuzzyCDF 给出了一个学生在数据结构数据集上每个知识点的可视化诊断结果的示例. 可视化结果如图 7 所示, DINA 和 P-FuzzyCDF 都可以获得有意义的诊断结果. 但是, DINA 只能区分学生是否掌握了一个知识点 (1 完全掌握, 0 完全未掌握). 而 P-FuzzyCDF 可以得出一个学生对知识点的具体掌握程度. 因此, 根据诊断结果, 学生可以准确地了解自己的优点和不足. 老师也可以根据我们的诊断结果给出个性化的教学建议. 相比于 DINA 和 P-FuzzyCDF, IRT 方法使用潜在变量来描述一个学生, 因此, 不能为每个学生提供直观可解释性的结果. 另外, 虽然 FuzzyCDF 也可以给出学生对知识点掌握程度的描述, 但是由第 4.5.1 节可知, FuzzyCDF 在 PEP 的准确性方面低于 DINA.

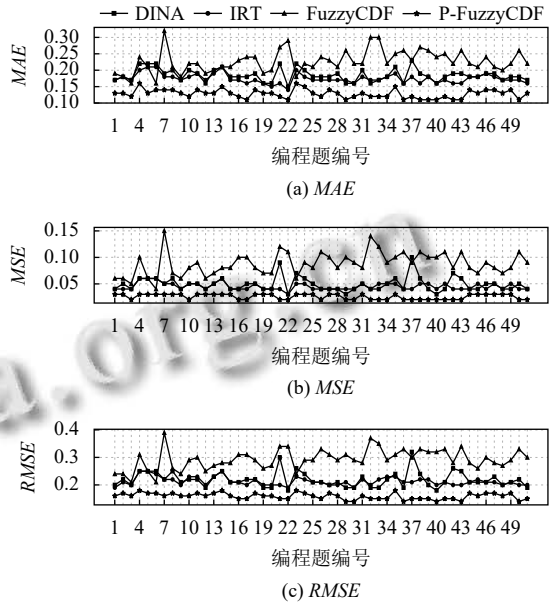

图 6 数据结构数据集的详细实验结果

表 7 第 4.5.2 节中假设的 P-value

模型	*MAE*	*MSE*	*RMSE*
DINA	3.82E–10	6.99E–10	4.62E–10
IRT	1.27E–03	4.46E–06	2.15E–06
FuzzyCDF	8.28E–11	8.19E–11	8.29E–11

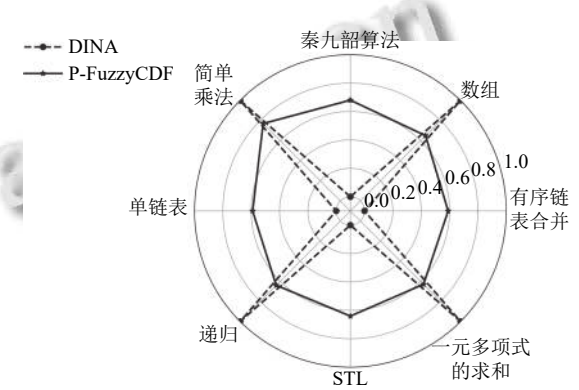

图 7 学生在每个知识点上的诊断结果

5 结论与展望

在本文中, 设计了一个个性化的模糊认知诊断框架 P-FuzzyCDF, 以探索认知诊断模型在编程题上的表现. 首先基于模糊集假设模糊化学生对知识点的掌握程度, 然后通过模糊集运算模糊化学生对编程题的掌握程度, 接下来通过考虑学生抄袭因素来对认知诊断

进行建模. 此外, 在 4 个数据集上进行了评估, 大量的实验结果表明, P-FuzzyCDF 能够定量的, 可解释的分析每个学生的特征, 从而获得更好的预测性能. 今后, 将根据学生量化的学习状态为学生推荐个性化的学习路径和学习活动.

参考文献

1 Xu CJ, Zhu GB, Ye J, *et al*. Educational data mining: Dropout prediction in XuetangX MOOCs. Neural Processing Letters, 2022, 54(4): 2885–2900. [doi: 10.1007/s11063-022-10745-5]

2 刘淇, 陈恩红, 朱天宇, 等. 面向在线智慧学习的教育数据挖掘技术研究. 模式识别与人工智能, 2018, 3(1): 77–90.

3 Zhou SQ, Traynor A. Measuring students' learning progressions in energy using cognitive diagnostic models. Frontiers in Psychology, 2022, 13: 892884. [doi: 10.3389/fpsyg.2022.892884]

4 江培超, 王川, 胡富珍, 等. 基于阅读认知诊断的学生表现预测. 计算机工程与应用, 2022, 58(11): 160–170.

5 李忱喜, 文益民, 易新河, 等. 一种改进的模糊认知诊断模型. 数据采集与处理, 2017, 32(5): 958–969.

6 Liu Q, Wu RZ, Chen EH, *et al*. Fuzzy cognitive diagnosis for modelling examinee performance. ACM Transactions on Intelligent Systems and Technology, 2018, 9(4): 1–26.

7 Liu J, Tang WS, He XP, *et al*. Research on DINA model in online education. Proceedings of the 6th International Conference on E-learning, E-education, and Online Training. Changsha: Springer, 2020. 279–291.

8 Wang F, Liu Q, Chen EH, *et al*. Neural cognitive diagnosis for intelligent education systems. Proceedings of the 34th AAAI Conference on Artificial Intelligence, the 32nd Innovative Applications of Artificial Intelligence Conference, and the 10th AAAI Symposium on Educational Advances in Artificial Intelligence. New York: AAAI Press, 2020. 6153–6161.

9 范士青, 刘华山. 常见的认知诊断模型及其比较. 教育测量与评价, 2015, (7): 4–9.

10 Tatsuoka K K. Rule Space: An approach for dealing with misconceptions based on item response theory. Journal of Educational Measurement, 1983, 20(4): 345–354. [doi: 10.1111/j.1745-3984.1983.tb00212.x]

11 De La Torre J, Minchen N. Cognitively diagnostic assessments and the cognitive diagnosis model framework. Psicología Educativa, 2014, 20(2): 89–97.

12 张兆远, 陶剑. 项目反应理论 (IRT) 甄选试题方法研究. 伊犁师范学院学报 (自然科学版), 2018, 12(3): 10–14.

13 Janssen R, Tuerlinckx F, Meulders M, *et al*. A hierarchical IRT model for criterion-referenced measurement. Journal of Educational and Behavioral Statistics, 2000, 25(3): 285–306. [doi: 10.3102/10769986025003285]

14 刘彦楼, 辛涛, 田伟. 项目反应理论与认知诊断模型的参数估计: 模型整合视角. 北京师范大学学报 (自然科学版), 2017, 53(6): 742–748. [doi: 10.16360/j.cnki.jbnuns.2017.06.017]

15 Wu RZ, Liu Q, Liu YP, *et al*. Cognitive modelling for predicting examinee performance. Proceedings of the 24th International Conference on Artificial Intelligence. Buenos Aires: AAAI Press, 2015. 1017–1024.

16 蔡艳, 赵洋, 刘舒畅, 等. 一种优化的多级评分认知诊断模型. 心理科学, 2017, 40(6): 1491–1497. [doi: 10.16719/j.cnki.1671-6981.20170632]

17 刘彬彬. 几种基于 IRT (项目反应理论) 模型的参数估计方法研究. 硅谷, 2010, (22): 80.

18 Yamaguchi K, Okada K. Hybrid cognitive diagnostic model. Behaviormetrika, 2020, 47(2): 497–518. [doi: 10.1007/s41237-020-00111-x]

19 熊超, 马华. 结合认知诊断和答题行为分析的试题推荐方法. 计算机时代, 2022, (12): 85–88.

20 Divine G, Norton HJ, Hunt R, *et al*. A review of analysis and sample size calculation considerations for Wilcoxon tests. Anesthesia & Analgesia, 2013, 117(3): 699–710.

(校对责编: 孙君艳)