

## AI+环境下学生科学探究能力测评研究

黄 华<sup>1</sup> 张 睿<sup>2</sup> 潘 鑫<sup>2</sup>

(<sup>1</sup> 同济大学职业教育技术学院, 上海 200092; <sup>2</sup> 同济大学物理科学与工程学院, 上海 200092)

**摘 要** 降低教师重复劳动, 提高学习评测的自动化和智能化水平是人工智能在教育评测领域的研究热点之一。在此背景下, 本研究设计并开发了基于斜抛运动的探究学习场景, 通过搭建人工神经网络, 提取参考答案的关键短语辅助分词, 训练人工标注的数据完成智能批改算法的设计, 并将操作数量与探究结论的比值作为学生探究能力的指标。结果表明, 基于辅助分词的人工智能批改算法对问答短句评判有较高的准确性, 进而能实现对学生探究能力的评判; 高中生中探究能力强的学生比例低于大学生, 但相同能力聚类的学生之间能力没有显著差异。

**关键词** 人工神经网络; 探究学习; 教育评价; 智能批改

## RESEARCH ON THE ASSESSMENT OF STUDENTS' SCIENTIFIC INQUIRY ABILITY VIA AI

HUANG Hua<sup>1</sup> ZHANG Rui<sup>2</sup> PAN Xin<sup>2</sup>

(<sup>1</sup> College of Vocational Education and Technique, Tongji University, Shanghai 200092;

<sup>2</sup> School of Physical Science and Engineering, Tongji University, Shanghai 200092)

**Abstract** One of the hotspots of artificial intelligence in the field of education evaluation is to reduce the repetitive work of teachers and improve the automation and intelligence level of learning evaluation. In this background, this paper designs and develops an inquiry learning scenario based on oblique throwing movement. By building an artificial neural network, the key phrases of reference answers are extracted to assist word segmentation, and the design of an intelligent correction algorithm is completed by training manually labeled data. The ratio of the operating times to the inquiry conclusion is taken as an indicator of students' inquiry ability. The results show that the artificial intelligence correction algorithm based on auxiliary word segmentation has high accuracy in the evaluation of short answer sentences, and then can realize the evaluation of students' scientific inquiry ability. the proportion of high school students with strong exploration ability is lower than that of college students, but there is no significant differences in ability among the students in the same cluster.

**Key words** artificial neural network; inquiry learning; educational evaluation; intelligent correction

2021 年人社部、教育部等部门发布《十四五职业技能培训规划》<sup>[1]</sup>, 指出提高职业培训的质量, 增强职业培训的针对性和有效性, 探索 AI+职业培训的方法是今后职业技能培训的重要任务。高素质的劳动者不仅应该有丰富的知识储备

和过硬的动手能力, 也需要通过生产实践活动建构知识, 发现规律。以 VR/AR/MR 技术为代表的虚拟仿真技术在汽车维修培训<sup>[2]</sup>、外科手术教学<sup>[3]</sup>、电工实践<sup>[4]</sup>和工业机器人编程教育<sup>[5]</sup>等领域得到广泛应用, 使用虚拟技术有利于细致展现

收稿日期: 2022-04-07

基金项目: 2020 年度上海市教育科研市级课题“大学物理互动式电子书设计与教学实践”(编号: C20122)。

通讯作者: 张睿, 副教授, 主要从事教育技术与物理教育研究工作, zr2002\_7@163.com。

引文格式: 黄华, 张睿, 潘鑫. AI+环境下学生科学探究能力测评研究[J]. 物理与工程, 2022, 32(6): 49-56, 65.

Cite this article: HUANG H, ZHANG R, PAN X. Research on the assessment of students' scientific inquiry ability via AI[J]. Physics and Engineering, 2022, 32(6): 49-56, 65. (in Chinese)

实际训练场景,降低训练成本,丰富训练内容。当前大多数职教应用案例的评价仅限于学生答题或操作的结果,较少涉及能力评估等。如何通过虚拟实践场景在训练学生的同时实现对学生探究能力的测评,是职业教育需要解决的问题。

随着计算机和人工智能技术的发展,自动化及智能化的评测方法逐步成为教育评测领域的重点研究内容,自动化和智能化评测不仅降低了人工评测的重复劳动,还降低了由个体认知、思维模式和情感态度等不同造成的评价差异,提高了评价的一致性。此外,智能评测系统支持不间断工作,适用于频次较高的过程性评价监测和结果汇总,具有广阔的应用空间。

本研究以探究实验的形式提供探究场景,通过探究过程和探究结论实现对学生探究能力的智能化测评,希望能对职业教育学生技能的智能化测评提供一定的参考价值。

## 1 研究背景

### 1.1 科学探究能力评价体系

科学探究能力指的是从事科学探究活动应具备的能力,包括发现问题,收集信息,设计实验,实施实验,归纳总结等多个方面技能。国际上对科学探究能力评测的主要方式包含观察法、纸笔测验、工作单和计算机模拟<sup>[6]</sup>。其中计算机模拟方法不仅可以真实模拟探究中的物体材质、环境和运动特征,也能以较低成本记录用户行为数据、基于算法规则开展评测,在科学探究能力评价领域得到广泛应用。

当前国外科学探究能力的评价体系主要包含美国的 PADI (Principled Assessment Designs for Inquiry) 与英国的 APU (Assessment of Performance Unit)、GCE (General Certificate Examination)、三大学业评价项目 [TIMSS (Third International Mathematics and Science Study)、NAEP (National Assessment of Educational Progress) 和 PISA (Program for International Student Assessment)]<sup>[7]</sup>。APU 强调能力探究的六个方面:符号表征与使用、器材及工具使用、观察能力、解释和应用、探究设计和整体表现,这六个方面都具备清晰的评价标准并循序渐进<sup>[8]</sup>。PISA2015 科学素养评价框架主要包含情境、能力、知识和态度四个方面,它们之间的关系密不可分<sup>[9]</sup>,强调从理解学校、社会情境,掌握解释探究和寻找证据的能

力,掌握学科、程序性和认知性知识,具备科学的兴趣、环境意识等<sup>[10]</sup>。TIMSS2019 科学评估测试主要包含科学内容、认知领域和科学实践三大方面,强调从理解问题、提出问题、处理数据和证据论证等系列流程<sup>[11]</sup>。

此外,英国哲学家斯蒂芬·图尔敏提出一个由主张、资料、根据、支援、限定词和反驳等 6 个功能要素构成的过程性模式,称为图尔敏论证模型 (Toulmin's Argument Pattern, TAP)<sup>[12]</sup>。国内学者马双永在分解要素和征求高校教师的基础上制定了 PTA (Primary Trait Analysis) 量表<sup>[13]</sup>,李春梅等人给“探究日志”制定了详细的评分标准<sup>[14]</sup>,林崇德等基于青少年思维结构特征,从方法、内容和品质三个维度构建了青少年科学思维模型<sup>[15]</sup>,胡卫平等基于科学论证内容、科学论证方法及科学论证品质三个维度,提出了以科学概念为中介,以科学推理和批判性思维为核心,要求必须以科学思维品质参与的科学论证能力结构模型<sup>[16]</sup>。国内科学探究能力评价多借鉴外国评价标准,并根据学科特色进行模块设计。

### 1.2 教育评测中的人工智能技术类型

20 世纪 60 年代以来,基于依据一定的目标和标准,对写作过程和结果进行价值判断的自动作文评分 (Automated Essay Scoring, AES) 系统逐步发展,并逐步参与实际作文评价<sup>[17]</sup>。基于自然语言处理技术,自动作文评分系统还衍生出教学质量评价、语音表达评价<sup>[18]</sup>、诗词文章创作评价等。此外,借助摄像头,教育评测还加入了表情、动作、专注度<sup>[19,20]</sup>等识别,对教学质量的评价提供了更多维度。

总体上,教育评测中的人工智能技术主要有计算机视觉和自然语言处理两大领域。计算机视觉领域通过运用计算机视觉技术捕捉教师和学生表情、板书和动作行为等,用于观察学生听课、教师讲授情况、捕捉答题关键词等,用视觉处理技术提取关键特征并进行评价。自然语言处理技术则侧重于理解表达的句式、逻辑,发掘表达内容之间的关联,从而通过机器习得的判定标准进行评价。

智能批改主要包含以自动作文评分为代表的主观评价和以问答评分为主的客观评价。主观评价主要包含作文自动评分、翻译自动评分和口语自动评价。主观评价大多建立在词汇水平、切题程度、句型难度和表达准确度等主观性判断基础上,自然语言处理技术的应用目标在于降低机器评价与平均评价水平的偏差。主观评价中的语料库相对广泛,专业词汇较少,允许较大的误差幅度。

国外开发了多个自动评价系统,广泛应用于 TOEFL(Test of English as a Foreign Language)、GRE (Graduate Record Examination)、GMAT(Graduate Management Admission Test)等大规模考试<sup>[21,22]</sup>,国内出现了基于汉语和英文的自动作文评分系统<sup>[23,24]</sup>,并逐步得到推广应用。

基于问答评分的客观评价主要包含简答题、填空题、应用题等题型的自动评价,这类问题专业性通常较强,表达方式简单明快,人为评价差异较少,对评价结果的精度要求高,这类问题也被称为自动短句评分(Automatic short answer grading)<sup>[25]</sup>。

### 1.3 智能批改中的算法研究

Burrows, Gurevych 和 Stein 提出使用计算机评估主观答案减少了下一代教育系统中评估的时间和精力<sup>[26]</sup>, Leacock 和 Chodorow 开发了处理文本中语义信息的简答系统——C-rater<sup>[27]</sup>。Bin, Jun, Jian-Min 和 Qiao-Ming 使用 KNN 分类器对文章自动分类<sup>[28]</sup>。Kakkonen, Myller, Sutinen 和 Timonen 使用比较了 LSA、PLSA 和 LDA 模型与 KNN 在自动评价系统中的表现<sup>[29]</sup>。Noorbehbahani 和 Kardan 提出了基于 BLEU 改进算法的 M-BLEU 用于评价学生的自主回答<sup>[30]</sup>。Islam 和 Hoque 开发了 GLSA 算法用于评价自动化作文评价<sup>[31]</sup>。

## 2 实验内容

### 2.1 研究对象

本研究以上海市某中学 41 名中学生、华东地区 T 大学 90 名大学生在斜抛实验中的回答情况作为研究对象,实验主要记录其回答短句和发射

的参数以及结果的点位,所有实验均在实验室使用电脑完成,旁边有一位老师负责监督学生独立答题并解答使用过程中的疑问,实验界面见图 1。

### 2.2 实验介绍

斜抛实验是探究抛掷高度、角度、速度、质量与抛掷距离关系的一种探究实验,实验中实验者修改高度、角度、速度、质量等参数并发射小球,通过观察小球轨迹变化总结抛掷高度、角度、速度、质量与抛掷距离关系。实验者通过访问特定网页参与模拟实验,在允许的 12 次实验完成后,实验者需要使用文字分别回答抛掷小球高度、角度、速度、质量与抛掷距离之间的关系。同时,实验者发射的实验数据也将被录入系统中。

### 2.3 实验原理及评价方式

斜抛运动的公式为

$$L = \frac{\cos^2 \theta}{g} (v^2 \tan \theta + v \sqrt{2gh \sec^2 \theta + v^2 \tan^2 \theta}) \quad (1)$$

由此可以看出初速度与抛掷距离成正相关,初始高度与抛掷距离成正相关,抛掷距离随着抛掷角度先增大后减小,经过反复实验可以得出在 45 度时达到最大,抛掷质量与抛掷距离无关。根据以上原理,可以编制评价说明如下:

(1) “抛掷高度与距离成正相关”或“正相关”得 1 分,“成正比例关系”不得分。

(2) “抛掷速度与距离成正相关”或“正相关”得 1 分,“成正比例关系”及“平方成正比”也不得分。

(3) “抛掷距离随着抛掷角度先增大后减小”或“先增大后减小”得 1 分,“达到 45 度最大”得 1 分,答对两者之一均可得 1 分,全部答对得 2 分。



图1 实验界面

(4)“抛掷距离与无力质量无关”或“无关”、“没有关系”得1分,其余回答不得分。

### 3 自动批改算法研究

#### 3.1 算法流程

本算法由分词、人工神经网络训练和语义推测三部分组成。用户回答的文字首先通过分词算法切分,构成词袋的基本元素。人工神经网络通过录入训练集,记录训练集中词袋单词和句式的相关关系,并生成模型文件。语义推测时通过加载模型文件,判定其语意所属的类型,完成自动打分,算法流程见图2。

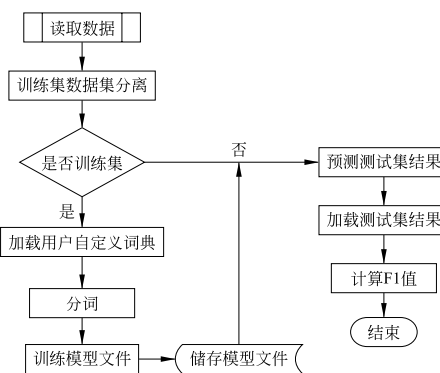


图2 算法流程图

#### 3.2 自定义词典与分词

本次分词采用 Jieba 分词工具。由于物理短语涉及较多专业名词及符号(如用  $x$ 、 $X$  代表位移,  $V$ 、 $v$  代表速度,  $H$ 、 $h$  代表投掷高度等),专业名词(特别是物理符号)分词识别率相对较低,见表1。

表1 常规分词工具的分词识别率

|       | 总词数  | 错误数 | 非英文错误 | 分词识别率 | 非英文识别率 |
|-------|------|-----|-------|-------|--------|
| jieba | 3151 | 280 | 21    | 91.1% | 99.3%  |
| hanlp | 2792 | 21  | 17    | 99.3% | 99.5%  |

借助自定义词典,可有效提高分词的正确率,避免算法对特定名词进行错误划分。在自定义词典中,主要定义词的类型、词性及重要程度,使用逗号分隔,自定义词典见表2。

表2 自定义词典

| 高度  | 速度  | 角度    | 质量  | 关系      |
|-----|-----|-------|-----|---------|
| $H$ | $V$ | 夹角    | 重量  | 平方/开方   |
| $h$ | $v$ | Theta | $m$ | 正比/反比   |
| 高   | 初速度 | theta | $M$ | 正相关/负相关 |

#### 3.3 词袋模型

人工神经网络算法基于词袋划分,词袋即分词后训练集的所有词汇集合,相同词汇不再重复录入,语意预测时,通过查找分词后单词在词袋中的位置并加载相关模型函数判断语意。设置目标语句词袋为  $W$ ,  $W = (w_1 \ w_2 \ \cdots \ w_n)$ ,  $w_n$  表示词袋中包含切分的单词,  $T_w$  为模型中所有的词汇,  $L$  表示特定语法中分词对应的词袋模型集合,当词袋模型包含单词时,表示为1,不包含单词时表示为0,  $L = (l_1 \ l_2 \ \cdots \ l_n)$

$$l_n = \begin{cases} 1, & w_n \in T_w \\ 0, & \text{others} \end{cases} \quad (2)$$

#### 3.4 人工神经网络算法

人工神经网络算法主要由输入层、网络层和输出层三部分组成。输入层包含训练集和测试集的切分词袋,网络层主要包含两层神经元,使用 sigmoid 作为神经元激活函数,通过设置训练集比例、神经元个数、梯度下降参数、训练次数和拾取比例改变最终训练结果。算法运行流程及激活函数示意图如图3。

当前句意相似度函数为  $F(C_1, C_2, \cdots, C_i)$ ,  $C_i$  表示第  $i$  种意图,句义相似度函数主要判别句法含义以及是否出现错误表达。算法第  $m$  层的神经元模型为  $S_m$ ,  $S_m = (\lambda_{1m} \ \lambda_{2m} \ \cdots \ \lambda_{nm})^T$ ,  $\lambda_{nm}$  为第  $m$  层第  $n$  个神经元参数,  $f_m(x)$  为第  $m$  层的激活函数,句意相似度函数表达式为

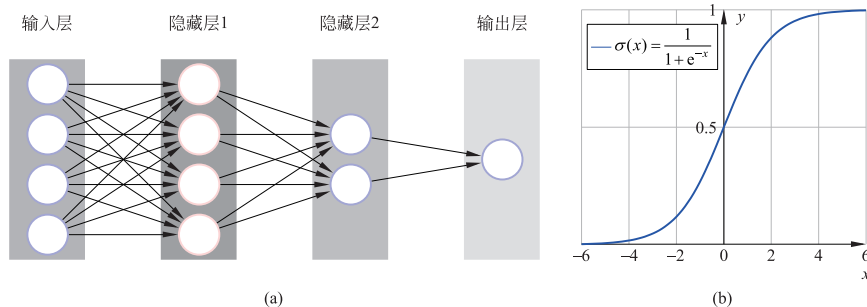


图3 人工神经网络与激活函数

(a) 人工神经网络; (b) 激活函数

$$\begin{aligned}
 F(C_1, C_2, \dots, C_n) &= \prod_{i=1}^m L S_m f_m(x) \\
 &= \prod_{i=1}^m (1 \ 0 \ \dots \ 1) \begin{pmatrix} \lambda_{1m} \\ \lambda_{2m} \\ \dots \\ \lambda_{nm} \end{pmatrix} f_m(x)
 \end{aligned}
 \tag{3}$$

本算法采用全局随机迭代的手段,生成若干代随机神经元,每一代选用采用梯度下降策略,随机生成神经元参数并比较错误率,通过不断优选,取得局部最优。若干代的随机迭代并从中优选避免参数陷入局部最优困境,提高整体算法可信度,算法伪代码如下:

---

**Algorithm 1** 基于人工神经网络的句意评价算法
 

---

**Input:** 训练集  $\mathbf{T}(\mathbf{S}, \mathbf{R})$ , 预测语句  $\mathbf{P}$ , 用户词典  $\mathbf{U}$

**Output:** 输出预判结果  $\mathbf{P}_R$

```

1: 定义训练次数  $e$ , 隐藏神经元数  $h_n$ , 神经元参数  $S_n$ , 测试结果集  $T_R$ , 学习率  $\alpha$ , 梯度下降斜率  $\Delta$ , 神经网络层  $L_n$ , 损失函数  $f_e$ , 激活函数  $f_{sig}$ 
2:  $W \leftarrow Jieba \cup U$  //词典为用户词典和结巴分词词库
3: if  $W.cut(T.S) \notin W_B$  then
4:    $W_B \leftarrow W.cut(T.S)$  //将句中没有的词添加进入词典  $W_B$ 
5: end if
6: for  $W.cut(P) \in W_B$  do
7:    $W_L \leftarrow list(zeros.len(W_B))$  //生成长度和词典相当, 值为零的数列
8:    $W_L[W.cut(P)] \leftarrow 1$  //如果包含词, 则词袋中对应的值为 1
9: end for
10: for  $k \in [0, 10]$  do
11:    $S_1 \leftarrow random.matrix(len(W), len(h_n))$  //随机生成隐藏层 1 的参数
12:    $S_2 \leftarrow random.matrix(len(h_n), len(P_R))$  //随机生成隐藏层 2 的参数
13:   //通过梯度下降求隐藏层 1 和 2 的参数
14:   while  $j \in [1, e/10]$  do
15:      $L_0 \leftarrow W_L(T.S)$  //输入层为词袋模型
16:      $f_e(L_2) \leftarrow T.R - L_2$ 
17:     for  $i \in [1, 2]$  do
18:        $f_e(L_i) \leftarrow \Delta_i \times S_i^T$ 
19:        $L_i \leftarrow f_{sig}(L_{i-1} \times S_i)$ 
20:        $\Delta_i \leftarrow (L_i) * (1 - L_i) * f_e(L_i)$ 
21:        $\Delta_{S_i} \leftarrow L_{i-1} \times \Delta_i$ 
22:        $S_i \leftarrow S_i + \alpha * \Delta_{S_i}$ 
23:     end for
24:     if  $f_{ej}(L_2) < f_e(L_2)$  then
25:        $S_1 \leftarrow S_{1j}, S_2 \leftarrow S_{2j}$  //取损失函数更小的参数模型
26:     end if
27:   end while
28:   if  $f_{ek}(L_2) < f_e(L_2)$  then
29:      $S_1 \leftarrow S_{1k}, S_2 \leftarrow S_{2k}$  //取损失函数更小的参数模型
30:   end if
31: end for
32:  $P_R[n] \leftarrow f_{sig}(f_{sig}(P \times S_1) \times S_2)$  //计算输出矩阵
33: if  $P_R[n].max[0] > 0.6$  then
34:   Return  $P_R[n].max[0]$  //最大概率大于 60%, 判定有效并输出
35: else
36:   Return None //其余情况判定无效
37: end if

```

---

3.5 评价指标

当前算法评价指标主要采用预测结果和实际结果的比较,采用准确率、精确率模型,通过计算准确率( $P$ )和召回率( $R$ )的调和平均数得到  $F_1$  值,通过  $F_1$  值的高低确定算法的有效程度,其计算方法见表 3 及公式(4)(5)(6)。

表 3 答案预测表

| 实际    | 预测                               |                                  | 合计                            |
|-------|----------------------------------|----------------------------------|-------------------------------|
|       | Positive                         | Negative                         |                               |
| True  | TP                               | TN                               | Actual<br>Postive<br>(TP+TN)  |
| False | FP                               | FN                               | Actual<br>Negative<br>(FP+FN) |
| 合计    | Predicted<br>Positive<br>(TP+FP) | Predicted<br>Negative<br>(TN+FN) | TP+TN+<br>FP+FN               |

$$P = \frac{TP}{TP + FP} \tag{4}$$

$$R = \frac{TP}{TP + FN} \tag{5}$$

$$F_1 = \frac{2 \times P \times R}{P + R} \tag{6}$$

3.6 实验结果及分析

通过上述算法对当前实验的计算,针对实验数据选择使用 66.7%、75%、80%和 83.3%的数据作为测试集,分别计算四个评价标准——“发射高度与射程的关系”“初速度与射程的关系”“发射角与射程的关系”“质量与射程的关系”,并分别进行 8 次独立实验,结果见表 4。由此可得出的结果如下:通过上述实验,可以看出当前算法对整体评  $F_1$  值为 90.59%,其中初速度与射程、质量与射程关系回答语句相对简单,因此  $F_1$  值均超过 90%,发射角与射程及发射高度与射程回答语句相对复杂, $F_1$  值略低于 90%。

通过比较大学生和高中生的探究得分,发现大学生平均分为 2.97 分,而高中生的平均分为 1.51 分(满分为 5 分),主要原因可能有以下两点:(1)某高中学生整体生源素质与 T 大学有较大距离(某高中能够考取 T 大学毕业生人数年均个位数);(2)科学探究能力与知识掌握程度存在正相关关系。

表 4 实验结果表

| 评价条目           | 训练集<br>比例 | 准确率           | 召回率           | $F_1$ 值       |
|----------------|-----------|---------------|---------------|---------------|
| 初速度与<br>射程的关系  |           | <b>95.88%</b> | <b>94.95%</b> | <b>95.32%</b> |
|                | 66.67%    | 97.26%        | 86.59%        | 91.61%        |
|                | 75.00%    | 95.56%        | 100.00%       | 97.73%        |
|                | 80.00%    | 94.87%        | 97.37%        | 96.10%        |
|                | 83.33%    | 95.83%        | 95.83%        | 95.83%        |
| 发射高度与<br>射程的关系 |           | <b>82.82%</b> | <b>96.49%</b> | <b>87.28%</b> |
|                | 66.67%    | 94.33%        | 93.29%        | 93.81%        |
|                | 75.00%    | 97.32%        | 95.09%        | 96.15%        |
|                | 80.00%    | 90.69%        | 97.57%        | 93.43%        |
|                | 83.33%    | 48.94%        | 100.00%       | 65.71%        |
| 发射角与<br>射程的关系  |           | <b>85.30%</b> | <b>91.96%</b> | <b>87.73%</b> |
|                | 66.67%    | 95.00%        | 90.48%        | 92.68%        |
|                | 75.00%    | 94.12%        | 80.00%        | 86.49%        |
|                | 80.00%    | 77.08%        | 97.37%        | 86.05%        |
|                | 83.33%    | 75.00%        | 100.00%       | 85.71%        |
| 质量与射<br>程的关系   |           | <b>86.51%</b> | <b>98.76%</b> | <b>92.04%</b> |
|                | 66.67%    | 96.00%        | 95.05%        | 95.52%        |
|                | 75.00%    | 79.75%        | 100.00%       | 88.73%        |
|                | 80.00%    | 83.05%        | 100.00%       | 90.74%        |
|                | 83.33%    | 87.23%        | 100.00%       | 93.18%        |
| 总计             |           | <b>87.63%</b> | <b>95.54%</b> | <b>90.59%</b> |

4 学生探究能力的测评

通过操作数据分析显示,学生的探究结论得分和学生的操作次数成正相关关系<sup>[7]</sup>,学生结论和操作次数的比值反应了学生通过学习活动构建知识的能力,将大学生的结论和操作次数的比值做聚类分析,图 4 为聚类方差与聚类数量关系的手肘图。

结果显示将 90 个大学生样本分成 3 个聚类比较合适。聚类结果中,第一聚类包括样本 56 人(62%),第二聚类包括样本 8 人(10%),第三聚类包括样本 25 人(27%),一个样本得分为 0 不属于任何聚类。分别对不同聚类做点击次数与探究结论得分的回归方程,结果如表 5 所示(表中  $B$  表示回归方程的系数)。其中第一聚类操作数量与对应得分比值较高的学生, $r^2$  为 0.31,解释性较弱,

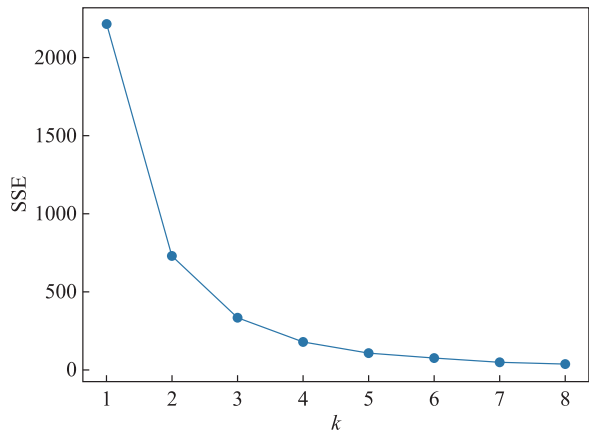


图 4 聚类方差与聚类数关系曲线

表 5 大学生不同聚类得分与操作次数的回归关系<sup>a</sup>

| 模型   | 非标准化系数       |        | 标准系数    | <i>t</i> | Sig.  |
|------|--------------|--------|---------|----------|-------|
|      | <i>B</i>     | 标准误差   | $\beta$ |          |       |
| 聚类 3 | 聚类 3<br>(常量) | 29.444 | 10.462  | 2.814    | 0.031 |
|      | 聚类 3<br>得分   | 12.903 | 1.631   | 7.909    | 0.000 |
| 聚类 2 | 聚类 2<br>(常量) | -1.851 | 10.301  | -0.180   | 0.859 |
|      | 聚类 2<br>得分   | 10.193 | 1.116   | 9.136    | 0.000 |
| 聚类 1 | 聚类 1<br>(常量) | -5.041 | 8.538   | -0.590   | 0.557 |
|      | 聚类 1<br>得分   | 4.562  | 0.918   | 4.969    | 0.000 |

a. 因变量\：操作次数

第二聚类对应操作数量与得分比值居中的学生， $r^2$  为 0.78，解释性较强，第三聚类对应操作数量与得分的比值较高的学生， $r^2$  为 0.91，解释性最强。第一聚类的探究能力明显高于第二聚类和第三聚类。但探究过程中，第一聚类的同学可能使用了其他方法(比如查找资料)来帮助他们寻找探究结论。

采取同样的办法对高中生进行处理，发现十四名高中生的研究结论得分为 0，对于剩余的学生做三聚类分析，再根据聚类情况做回归分析，结果见表 6，其中聚类 1 人数为 15 人，聚类 2 人数 7 人，聚类 3 人数 5 人。聚类 3 因得分均为 1 分，无法回归分析。从数据上看，前两个聚类的探究能

表 6 高中生不同聚类得分与操作次数的回归关系<sup>a</sup>

| 模型   |              | 非标准化系数   |       | 标准系数    | <i>t</i> | Sig.  |
|------|--------------|----------|-------|---------|----------|-------|
|      |              | <i>B</i> | 标准误差  | $\beta$ |          |       |
| 聚类 3 | 聚类 3<br>(常量) | —        | —     | —       | —        | —     |
|      | 聚类 3<br>得分   | —        | —     |         | —        | —     |
| 聚类 2 | 聚类 2<br>(常量) | 0.917    | 3.043 |         | 0.301    | 0.775 |
|      | 聚类 2<br>得分   | 10.583   | 1.446 | 0.956   | 7.320    | 0.001 |
| 聚类 1 | 聚类 1<br>(常量) | −0.029   | 2.291 |         | −0.013   | 0.990 |
|      | 聚类 1<br>得分   | 4.043    | 0.930 | 0.770   | 4.348    | 0.001 |

a. 因变量\：操作次数

力对应于大学生样本的相同聚类。高中生中聚类 1 的学生占总人数的比例为 37.5%，低于大学生聚类 1 比例。高中生中聚类 2 学生比例为 17.5%，高于大学生聚类 2 比例，高中生得分为 0 的学生远高于大学生的比例。说明探究能力仍然和学生的知识背景有关，但并不完全取决于知识背景，知识背景较好的学生中，探究能力强的学生比例更大，探究失败的学生比例更小。但是对于同一类型的学生，大学生与中学生探究能力是相当的。

5 结语

本研究将人工智能技术中的自然语言处理，人工神经网络、聚类分析、回归分析等技术相结合，探索出通过学生的探究活动有效测评学生探究能力的方法。尽管本研究是基于中学生与大学生的虚拟实验探究活动，但研究方法对于职业教育的虚拟操作精准测评仍然具有一定的参考价值。考虑到学生非探究活动对于探究结论的影响，建议在探究能力测评中采用相对封闭的探究环境，比如在单机环境下进行探究。

基于人工神经网络的自动批改算法能自动完成用户回答的评价，整体准确率达到 90%，能够应用于教育游戏的自动批改与评价。但随着回答句式的复杂化，标准词袋库的词袋增多会造成训练时间变长、批改正确率降低的情况，这一点需要在

后续的研究中进一步完善。本研究采取的人工神经网络仍然相对简单,没有使用基于后向传播的深度学习技术,采用深度学习技术能够进一步提高人工神经网络的性能,这一点也将在后续的研究工作中加以改善。

### 参 考 文 献

- [1] 人力资源社会保障部,教育部,发展改革委等. 关于印发“十四五”职业技能培训规划的通知[EB/OL]. [http://www.gov.cn/zhengce/zhengceku/2021-12/17/content\\_5661662.htm](http://www.gov.cn/zhengce/zhengceku/2021-12/17/content_5661662.htm).
- [2] BORSCI S, LAWSON G, BROOME S. Empirical evidence, evaluation criteria and challenges for the effectiveness of virtual and mixed reality tools for training operators of car service maintenance [J]. *Computers in Industry*, 2015, 67.
- [3] SHARAF S A, HILARY E, CHRIS P. Next-generation Virtual and Augmented Reality in Surgical Education: A Narrative Review [J]. *Surgical technology international*, 2019, 35: 27-35.
- [4] 章彬. 基于 Leap Motion 手势识别的 VR 电工实验系统设计 [D]. 大连:大连理工大学, 2021.
- [5] MONETTI F M, GIORGIO A, YU H, et al. An experimental study of the impact of virtual reality training on manufacturing operators on industrial robotic tasks[J]. *Procedia CIRP*, 2022, 106: 33-38.
- [6] 罗国忠. 科学探究能力的评价方法[J]. *教育科学*, 2007(3): 7-10.  
LUO G Z. The Assessment Methods of Scientific Investigation Abilities[J]. *Education Science*, 2007(3): 7-10. (in Chinese)
- [7] 潘鑫, 张睿, 张志华. 基于教育游戏数据挖掘的科学探究能力测评研究[J]. *物理与工程*, 2021, 31(6): 25-32.  
PAN X, ZHANG R, ZHANG Z H. Research on evaluation of scientific inquiry ability based on educational game data mining [J]. *Physics and Engineering*, 2021, 31(6): 25-32. (in Chinese)
- [8] GOTTR, MURPHY P. Sciencereport for teachers: 9-Assessing investigations at ages 13 and 15[R]. Letchworth, Hertfordshire, Garden City Press, 1987: 24. 5.
- [9] OECD(2016b). PISA 2015 Results (Volume I): Excellence and Equity in Education, PISA, OECD Publishing, Paris [EB/OL]. <http://www.oecd.org/education/pisa-2015-results-volume-i-9789264266490-en.htm>.
- [10] 陈红君, 田曼钰, 于海波. PISA 科学素养测试题的分析与启示[J]. *物理教师*, 2021, 42(12): 72-74.  
CHEN H J, TIAN M Y, YU H B. Analysis and Enlightenment of PISA Scientific Literacy Test Questions [J]. *Physics Teachers*, 2021, 42(12): 72-74. (in Chinese)
- [11] IEA. TIMSS 2019 Assessment Frameworks[EB/OL]. <http://timss2019.org/wpcontent/uploads/frameworks/T19-Assessment-Frameworks.pdf>.
- [12] TOULMIN S E. Rationality and scientific discovery[J]. *Boston Studies in the Philosophy of Science*, 1974(20): 387-406.
- [13] 黄璩, 唐劲军. 量表法评价中学生化学科学探究能力[J]. *化学教育*, 2007(4): 30-32.  
HUANG C, TANG J J. Evaluation of Scale Method to Middle School Students' Chemical Science Inquiry Ability [J]. *Chemical Education*, 2007(4): 30-32. (in Chinese)
- [14] 李雪梅, 范春玲. 运用“探究日志”评价中学生的科学探究能力[J]. *物理教学探讨*, 2008(7): 5-8.  
LI X M, FAN C L. Evaluation of Middle School Students' Scientific Inquiry Ability via "Inquiry Journal"[J]. *Discussion on Physics Teaching*, 2008(7): 5-8. (in Chinese)
- [15] 胡卫平, 林崇德. 青少年的科学思维能力研究[J]. *教育研究*, 2003, (12): 19-23.  
HU W P, LIN C D. A Study of Teenagers' Capacity of Scientific Thinking [J]. *Educational Research*, 2003, (12): 19-23. (in Chinese)
- [16] 韩葵葵, 胡卫平. 科学论证能力结构模型建构研究[J]. *教育理论与实践*, 2019, 39(22): 3-7.  
HAN K K, HU W P. Exploring the Construction of Scientific Argumentation Ability Model[J]. *Theory and Practice of Education*, 2019, 39(22): 3-7. (in Chinese)
- [17] BESEISO M, ALZUBI O A, RASHAIDEH H. A novel automated essay scoring approach for reliable higher educational assessments[J]. *Journal of Computing in Higher Education*, 2021, 33(3): 1-20.
- [18] 孙海洋. 国内外英语口语自动评分研究综述[J]. *外语教育研究前沿*, 2021, 4(2): 28-36+89-90.  
SUN H Y. Automated scoring of speech: An overview of research at home and abroad[J]. *Foreign Language Education in China* 2021, 4(2): 28-36+89-90. (in Chinese)
- [19] 周国雄, 李家永, 朱俊杰, 等. 智慧教室学生状态检测建设研究[J]. *现代电子技术*, 2021, 44(22): 71-76.  
ZHOU G X, LI J Y, ZHU J J, et al. Research on student status detection in smart classroom[J]. *Modern Electronics Technique*, 2021, 44(22): 71-76. (in Chinese)
- [20] KUMAR G K, NARAYANAM B. FPGA implementation of eye movement detection algorithm, *Microprocessors and Microsystems*[M]. 2021.
- [21] SHERMIS M D, BURSTEIN J. Automated Essay Scoring: A Cross-disciplinary Perspective[C]. Mahwah, NJ: Lawrence Erlbaum Associates, 2003: 113-121.
- [22] DIKLI S. An overview of automated scoring of essays[J]. *Journal of Technology, Learning, and Assessment*, 2006(1): 3-35.
- [23] 梁茂成. 中国学生英语作文自动评分模型的构建[M]. 北京: 外语教学与研究出版社, 2011.

评价量规,形成更加合理的分层教学与过程性评价体系;充分应用信息技术,实现课前学习、课堂研讨和课后拓展的深度融合,促进学生自主学习和创新能力训练。

#### 4 结语

基于华东师范大学高校定位与环境生态工程的新工科人才培养目标,采用“七维一体”进阶式理论-实验混合式课程体系,理念先进、创新;通过分层教学理念、翻转课堂、多维度过程性评价方式,有效地激发了学生的学习兴趣与参与积极性;通过线上、线下学习活动的深度融合,科创项目的参与,全面提升学生自主学习与创新实践能力,形成高阶思维;以学生为中心,动态调整教学模式,教学相长,不断引入研究成果、先进理念、前沿技术,持续革新,保持课程新鲜度与先进性;基于学生专业背景,充分挖掘思政元素,使思政教育全方位沁润专业知识教学过程,实现学生人格、学识、能力的协同培养,使学生获得专业认同感和职业价值感。

#### 参 考 文 献

- [1] 彼格斯,科利斯. 学习质量评价:SOLO 分类理论(可观察的学习成果结构)[M]. 北京:人民教育出版社,2010.
- [2] 布鲁姆 B S. 教育目标分类学:认知领域[M]. 上海:华东师范大学出版社,1986.
- [3] 龚亚群. 分层教学的必要性及其理论依据[J]. 科学大众(科学教育),2009(11):116-117.
- GONG Y Q. The necessity and theoretical basis of layered teaching[J]. Popular Science, 2009(11):116-117. (in Chinese)
- [4] 何光宏,韩忠,杨骏骏,等. 大学物理实验分类分层实验体系建设[J]. 物理与工程:2017(8):168-171.
- HE G H, HAN Z, YANG J J, et al. Construction of stratified system in college physics experiment[J]. Physics and Engineering, 2017, 27(1): 168-171. (in Chinese)
- [5] 周科卫,邓莉,忻静,等. 水质物理参量测量 DIS 演示虚拟仿真实验[J]. 物理实验,2020,40(9):37-43.
- ZHOU K W, DENG L, XIN J, et al. Dis demonstration virtual simulation experiment of water quality physical parameter measurement [J]. Physical Experiment, 2020, 40(9): 37-43. (in Chinese)
- [6] 忻静,孙可,邓莉,等. 基于 STEM 的水质光学综合测量虚拟仿真实验[J]. 物理实验,2022,42(1):50-56.
- XIN J, SUN K, DENG L, et al. Virtual simulation experiment of water quality optical comprehensive measurement based on STEM[J]. Physical Experiment, 2022, 42(1): 50-56. (in Chinese)
- [1] 彼格斯,科利斯. 学习质量评价:SOLO 分类理论(可观察的
- (上接第 56 页)
- [24] 杨丽萍,辛涛. 人工智能辅助能力测量:写作自动化评分研究的核心问题[J]. 现代远程教育研究,2021,33(4): 51-62.
- YANG L P, XIN T. Aided ability measurement by artificial intelligence: The core problems of automatic writing scoring research[J]. Modern Distance Education Research, 2021,33(4): 51-62. (in Chinese)
- [25] SÜZEN N, GORBAN A N, LEVESLEY J, et al. Automatic short answer grading and feedback using text mining methods[J]. Procedia Computer Science, 2020(169): 726-743
- [26] BURROWS S, GUREVYCH I, STEIN B. The eras and trends of automatic short answer grading[J]. International Journal of Artificial Intelligence in Education, 2015; 25(1): 60-117.
- [27] LEACOCK C, CHODOROW M. C-rater: Automated scoring of short-answer questions[J]. Computers and the Humanities, 2003; 37(4), 389-405.
- [28] LI B, LU J, YAO J M, et al. Automated essay scoring using the knn algorithm[C]. In Computer science and software engineering, 2008 international conference, 2008 (volume 1): 735-738.
- [29] KAKKONEN T, SUTINEN M N, TIMONEN J. Comparison of dimension reduction methods for automated essay grading[J]. Journal of Educational Technology & Society, 2008; 11(3), 275.
- [30] NOORBEHBAHANI F, KARDAN A A. The automatic assessment of free text answers using a modified bleu algorithm[J]. Computers & Education, 2011; 56(2), 337-345.
- [31] ISLAM M M, HOQUE A L. Automated essay scoring using generalized latent semantic analysis[C]. In Computer and information technology (ICCIT), 2010 13th international conference. 2010; 358-363.