计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2019.20180366
56(5):1007-1019,2019

数据驱动的数学试题难度预测

佟 威 汪 飞 刘 淇 陈恩红
(中国科学技术大学计算机学院 合肥 230027)
(tongw@mail.neea.edu.cn)

Data Driven Prediction for the Difficulty of Mathematical Items

Tong Wei, Wang Fei, Liu Qi, and Chen Enhong
(*School of Computer Science, University of Science and Technology of China, Hefei* 230027)

Abstract The construction of item banking system is an important guarantee for the reform and development of educational examination, and meanwhile, is also an essential means to promote the modernization of examination. In such a system, item difficulty is one of the most important parameters, which has a direct influence on item designing, test paper organization, result report and even the fairness guarantee. Unfortunately, due to the unique education background and test characteristics in China, it is difficult to evaluate item difficulty through pre-test organization like some foreign countries. Thus, traditional efforts usually refer to the manual evaluation by expertise (e.g., experienced teachers). However, this way tends to be laborious, time-consuming and subjective in some way. Therefore, it is of great value to automatically judge the difficulty of items by information technology. Along this line, in this paper, we aim to propose a data-driven solution to predict the item difficulty in mathematics leveraged by the historical test logs and the corresponding item materials. Specifically, we propose a C-MIDP model and a R-MIDP model, which are based on CNN and RNN respectively, and further a hybrid H-MIDP model combined with both C-MIDP and R-MIDP. In the models, we directly learn item sematic representation from its text and train its difficulty with the statistic score rates among tests, where the whole modeling do not need any expertise, such as knowledge labeling. Then, we adopt a context-dependent training strategy considering the incomparability between different groups. Finally, with the trained models, we can predict each item difficulty only with its text input. Extensive experiments on a real-world dataset demonstrate that the proposed models perform very well.

Key words education; item banking system; difficulty prediction; deep learning; text mining

摘 要 现代化国家题库系统建设是教育考试改革发展的重要保障,也是促进我国教育考试现代化的重要手段.试题难度是入库试题的核心参数,对于命题、组卷、分数报告甚至是考试公平性保障都有着直接

收稿日期:2018-05-22;修回日期:2018-09-06
基金项目:全国教育科学规划基金项目(FCB160610);国家自然科学基金项目(61672483,U1605251);中国科协青年人才托举工程 &CCF 青年人才发展计划项目(CCF-QNRCFZ(17-19)03);中国科学院青年创新促进会会员专项基金项目(2014299)
This work was supported by the Fund for National Plan of Education Science (FCB160610), the National Natural Science Foundation of China (61672483, U1605251), the Young Talent Promotion Program of China Association for Science and Technology & the Young Talent Development Program of CCF (CCF-QNRCFZ(17-19)03), and the Special Fund for the Member of Youth Innovation Promotion Association of Chinese Academy of Sciences (2014299).
通信作者:刘淇(qiliuql@ustc.edu.cn)

影响. 由于我国国家考试的特点,很难通过类似国外考试机构的考前试测等方式提前获取试题难度参数,传统的试题难度评估任务通常由人工完成,即由命题专家对试题难度进行评估. 这样的做法耗时耗力,且难以保证客观性,因此借助先进信息技术手段探索试题难度的自动化判断具有较大的研究意义,更是体现着中国特色教育考试背景下的中国智慧和中国解决方案. 以利用试题文本和答题记录数据实现数据驱动的数学试题难度自动化预测模型为目标,提出了分别基于卷积神经网络(convolutional neural network, CNN)和循环神经网络(recurrent neural network, RNN)的数学试题难度预测模型 C-MIDP(CNN for mathematical item difficulty prediction)和 R-MIDP(RNN for mathematical item difficulty prediction),以及二者的混合模型 H-MIDP(hybrid model for mathematical item difficulty prediction). 具体地,利用所提出的模型直接学习试题文本表征,将考试试题得分率作为标签训练模型,整个过程不需要提供知识标注等教育先验信息. 然后,考虑到不同考试中学生群体的不可比性,在训练时提出一种基于 context 的训练方式;最后,可通过输入试题特征到训练好的模型中进行难度预测. 模型在真实的试题数据上取得了较好的实验结果.

关键词 教育;题库;难度预测;深度学习;文本挖掘

中图法分类号 TP399

教育是人才培养的重要途径,而考试自古以来就是评价教育成果、进行人才选拔的重要方式,在国家经济社会发展中发挥着重要的作用. 党和国家高度重视教育工作,提出了加快建设教育现代化、建设教育强国以及办好人民满意的教育的总体要求. 新时代的教育考试改革要紧密结合当前和今后一个时期国家和社会层面对人才价值的需求和判断,紧密结合先进的信息技术手段,为新一轮高考改革和政策制定提供更多的体现着中国智慧的中国解决方案.

长久以来,试题难度,特别是高考试题难度,都是教育考试国家题库建设,甚至全社会重点关注的指标参数,对保障考试安全平稳顺利实施、服务高校人才选拔、合理引导中学教学都有关键影响. 如今教育越来越受重视,对教育质量的要求逐渐增加,如何高效、准确地评估试题难度自然也成为了一个重要的研究问题.

传统方法中,试题难度评估大多是由人工进行[1]. 通常考试的命题人员和审校人员由具有充足专业知识和丰富教学经验的老师或专家担任,在设计试题时除了考虑涵盖的必备知识和关键能力等内容相关的属性和维度,也需要控制试题难度在合理范围,命题和审校人员以自身知识和经验评估试题难度. 另外也有以试测的形式请部分样本学生试做样题,根据学生实际答题情况评估试题难度,之后对样题稍作更改和重组投入使用,例如 TOEFL 考试和 SAT(scholastic assessment test)考试题等[1].

在教育数据挖掘领域,试题评估是一个重要的研究方向,现有方法已经对试题多种参数(如难度、区分度、猜测度等)进行了评估分析[2-3]. 其中应用最为广泛的是来自教育心理学的认知诊断理论. 认知诊断通过利用学生答题记录对学生试题得分进行建模,从而评估试题参数和学生能力. 常见的认知诊断模型包括基于项目反映理论(item response theory, IRT)[4]的潜在特质模型和以 DINA(deterministic inputs, noisy "and" gate)模型[5]为代表的潜在分类模型等. 其中 IRT 通过类逻辑斯蒂回归模型,结合学生的潜在能力,可以评估试题在难度、区分度和猜测度属性上的数值;而 DINA 进一步结合 Q 矩阵(或称"试题关联知识点矩阵"),且将学生能力描述成多维知识点掌握向量,建模学生得分,可以得到试题失误率、猜测率等参数. 其中 Q 矩阵是人工标注的用以表示试题包含知识点的矩阵. 表 1 是一个简单的 Q 矩阵示例,其中每一行代表一个试题,每一列代表一个知识点. 如表 1 第 1 行表示试题 q_1 包含知识点 s_1 和 s_4,但不包含知识点 s_2 和 s_3. Q 矩阵的完备性将影响到建模结果的准确性,然而 Q 矩阵通常由人工提供,因此其完备性也常常难以保证. 另外,也有学者通过特征工程的方式,提取试题诸如考察点、迷惑性、复杂性等特征后利用机器学习方法(如线性回归、神经网络等)实现难度预测[1,6].

Table 1 Example of Item Associated Q-matrix

表 1 试题关联知识点 Q 矩阵示例

Item	Knowledge Points			
	s_1	s_2	s_3	s_4
q_1	1	0	0	1
q_2	0	1	1	1

然而,不论是传统的人工评估,还是现有的认知

诊断或机器学习建模,在国家教育日益深化改革的背景下,应对试题难度预测这个问题上,都有各自的局限性,具体体现在3个方面:

1) 人力、时间消耗大. 人工的试题难度评估较为耗时耗力,而入库试题资源量庞大,且某些学科试题更迭频繁,这些都使得纯人工的试题难度预测变得不切实际. 且认知诊断中的Q矩阵也由人工标注,同样需要消耗较多的人力与时间.

2) 对先验知识的依赖. 人工的试题难度评估结果除试题本身外,很大程度上依赖于评估者自身的水平和对试题的认知程度;同样,认知诊断模型通常也需要预先提供试题的Q矩阵. 这些都使得评估或预测结果客观性或准确性不足.

3) 特征工程中人工定义的特征较为缺少试题语义,是试题的浅层表示. 且部分特征(如试题复杂性、灵活性、干扰性等)的判定仍然需要人工进行,非客观性和界限模糊等问题同样存在.

我国国家考试具有高利害性、社会关注度极高等特点,很难通过考前试测等方式提前获取试题难度参数,目前仍然按照传统的试题难度评估方式,由人工进行[1]. 随着大数据、人工智能时代的到来,众多先进的机器学习、深度学习算法为国家题库现代化建设和入库试题的难度参数估计赋予了更多方法和途径. 基于人工智能的试题难度预测以往年产生的大量数据作为训练样本,能够有效解决试题安全保密要求和试测曝光两者之间的矛盾,有效调整传统人工估计难度中存在的偏差和波动. 要实现高效、准确的试题难度评估,需要解决3个挑战:

1) 如何从包含复杂语义的试题文本出发,挖掘其中可用于难度预测的重要信息. 高效的试题难度预测自动化方法应尽量避免知识点标注等人工劳动,因此要求模型具有较强的文本信息挖掘能力.

2) 如何减少人工干预,使得评价结果更加客观. 诸如试题知识点标注或经验性的特征设计等都难以避免地引入个人倾向,使得结果客观性难以保证.

3) 如何克服不同考生群体在不同试卷版本中作答数据的比较. 这些数据得到的试题得分率往往具有样本依赖性,实际难度差异很大的试题从数据呈现的结果来看可能非常接近,反之亦然. 如果不能克服这个问题,预估结果会出现很大误差.

各项考试,特别是国家考试,都在一定程度上存在此类问题. 本文从数学试题难度预测着手,提出了针对数学试题的模型C-MIDP(CNN for mathematical item difficulty prediction),R-MIDP(RNN for mathematical item difficulty prediction)和H-MIDP(hybrid model for mathematical item difficulty prediction),利用试题文本和学生答题记录进行难度预测. 3种模型均为神经网络结构,其中C-MIDP以CNN(convolutional neural network)为基础,R-MIDP以RNN(recurrent neural network)为基础,H-MIDP则为二者的混合模型. 难度的预测分为3步:1)使用word2vec词向量对训练集的试题文本进行表征,作为模型输入. 以word2vec词向量构建的试题表征,可以较好地保留试题语义,使得神经网络能够基于试题文本自身挖掘出重要信息,同时保证客观性. 2)从答题记录中获取各场考试中试题的得分率,考虑得分率的适用范围,设计context相关的方式进行模型训练,将"以偏概全"变为"以小见大". 3)将需要预测难度的试题文本进行表征,输入到训练好的模型中,获得难度预测值. 本文的主要贡献点有3个方面:

1) 提出针对数学试题的难度预测模型,实现高效的数学试题难度预测,并在真实数据集上取得了较好的实验结果;

2) 模型是数据驱动的,训练和预测都不需要人工提供关于试题的先验知识,提高了预测结果的客观性,且因减少了人工参与因而提高了预测效率;

3) 考虑到不同考试中学生群体能力的差异性,训练时采用的是context相关的训练方式,提高了预测的准确率.

1　相关工作

本节将从难度预测和文本建模2个方面介绍相关工作.

1.1　难度预测

传统教育中,难度评估大多是人工进行的. 教育者利用自己的知识储备和教学经验评估试题难度,以设计或选择合适的试题,评估的结果通常随评估者知识、经验的差异出现不同.

在教育学领域中,有学者研究影响试题难度的具体因素,如Beck等人[7]认为试题特征和学生能力都是试题难度的影响因素. 在试题方面,Kubinger等人[8]指出试题类型、试题结构以及知识深度等因素都与试题难度有关;而在学生能力方面,也有许多理论和模型被提出,其中认知诊断是重要的研究方向,其目标是利用试题和学生的答题记录,对学生的学习过程进行建模,挖掘学生对知识或技能的掌握程度.

在教育数据挖掘领域,认知诊断是一类重要的研究方向,其目标是利用试题和学生的答题记录,对学生的学习过程进行建模,挖掘学生对知识或技能的掌握程度,从而通过能力分析、试题推荐、学生分组等方式优化学生的学习过程[9-10]. 认知诊断模型根据不同的分类方式可分为离散模型和连续模型,或分为一维技能模型和多维技能模型. 常见认知诊断模型包括基于项目反应理论(item response theory, IRT)的模型、DINA 模型和它们的改进模型[1,4,11-13]等,模型中通常会考虑试题的难度、区分度、失误可能性、猜对可能性等因素[11,14],有些研究中还会融合教育学理论,如学习曲线和遗忘曲线[15]等. 尽管这些模型考虑了试题难度等因素,但通常作为参数,或是通过已知的 Q 矩阵计算,因而需要人为提供较多的先验知识.

有学者将传统机器学习结合特征工程的方法运用到试题难度预测中. 文献[1]中作者定义了试题考察的能力、知识点重要程度、试题迷惑性、复杂性、灵活性等特征,将这些特征值作为神经网络的输入,预测试题难度. 尽管这些人工定义的特征能够反映试题的一些重要信息,但是基于经验人工筛选出的试题表征,对试题语义没有加以利用. 且部分此类特征值的确定并非是可统计的,而是由经验判断的,其客观性和准确性难以保证.

以上工作具有相同的局限性:即都需要较多的人为干预,如提供先验知识或教学经验和劳动力. 而本文所提出的模型是数据驱动的,所需要的只是试题文本和答题记录,从而避免上述问题.

目前已有学者进行了针对英语试题的难度预测工作[16],受其启发,本文提出了针对数学试题的难度预测模型.

1.2 文本建模

本文提出的模型针对试题的纯文本输入,且不需要提供试题的诸如知识点等先验信息,因此对模型的文本建模与信息提取能力要求较高.

随着大数据时代的到来,文本数据挖掘现已广泛运用于互联网[17]、教育[18]、医疗[19]、媒体[20]等领域,涉及的技术包括文本聚类、文本分类[21]、情感分析[22]、文本推荐[23]等. 与之相关的自然语言处理(natural language process, NLP)也在文本处理、自然语言理解、人机交互等领域具有重要意义. Mikolov 等人[24-25]提出 word2vec 和 doc2vec,尽管作为语言模型训练的副产物,但由于其维度低和保持部分语义特征等优点,被大量运用到文本建模的数据表征中,使得许多模型的效果得以提升.

在模型方面,过去文本数据挖掘方法通常需要分析文本的词法、语法、语义特征,人为地构造一些具体的结构. 近年来,深度学习的兴起使得文本数据挖掘有了新的探索路径,CNN[26]和 RNN[27]对文本类数据具有较好的拟合能力,避免了对词法、语法等先验知识的要求. 相关工作如情感识别[28]、文本蕴含[29]、机器理解[30]等.

多层 CNN 神经网络可从词、短语、句子等不同层次挖掘文本信息;RNN 则适合挖掘长程的逻辑关系. 因此 2 种模型都可用于试题难度预测的建模当中. 基于此,本文提出了基于 CNN 的难度预测模型 C-MIDP 和基于 RNN 的难度预测模型 R-MIDP,并且考虑到 CNN 和 RNN 各自的优缺点,将 CNN 和 RNN 结合,提出 H-MIDP,进一步提高预测的准确率.

2 数据驱动的试题难度预测模型

本节中将给出问题的形式化定义,介绍模型的整体框架,具体介绍 3 种不同的难度预测模型.

2.1 问题定义

模型训练所需要的数据为真实的数学考试试题及答题记录,考试为正式的统一测评(如期中考试、期末考试、月考等),试题为常规考试题型(如选择、填空或简答题). 表 2 为 1 道数学试题文本数据示例,数据包括试题 ID、题面、答案和解析. 表 3 为答题记录结构示例,1 条记录代表 1 个学生在 1 场考试中某道题的得分,将具有相同试卷 ID、学校 ID 和考试日期的答题记录集合定义为同一场考试 T_i 记录集合.

对于考试、试题、得分率等概念的形式化定义及本文应对的问题定义如下:

定义 $Q=\{Q_1,Q_2,\cdots,Q_n\}$ 为试题集合,$T=\{T_1,T_2,\cdots,T_m\}$ 为数学考试集合. $T_i=\{\widetilde{Q}_i,\widetilde{R}_i\}$,其中 $\widetilde{Q}_i$ 和 $\widetilde{R}_i$ 分别为考试 T_i 中的习题集合和得分率集合. $\widetilde{Q}_i=\{\widetilde{Q}_{i1},\widetilde{Q}_{i2},\cdots,\widetilde{Q}_{im_i}\}$,$\widetilde{Q}_{ij}\in Q$,$m_i$ 为考试 T_i 中的试题数;$\widetilde{R}_i=\{\widetilde{R}_{i1},\widetilde{R}_{i2},\cdots,\widetilde{R}_{im_i}\}$,$\widetilde{R}_{ij}$ 为考试 T_i 中第 j 道题的得分率,以得分率作为考试 T_i 中试题难度的真实值. 得分率的计算为

$$\widetilde{R}_{ij}=\frac{\text{考试 } i \text{ 中试题 } j \text{ 的得分之和}}{\text{考试 } i \text{ 中试题 } j \text{ 答题记录数}\times\text{试题 } j \text{ 总分}}. \quad (1)$$

Table 2　Example of Mathematical Item

表 2　数学试题示例

Attribute	Value
Item ID(Q_i)	00004b28-4a67-4a51
Question	已知 a,b,c 分别为$\triangle ABC$ 内角 A,B,C 的对边，$a^2=bc$，则$\angle A$ 的最大值为____.
Solution	$\frac{\pi}{3}$
Analysis	本题考查余弦定理及基本不等式，根据题意利用余弦定理及基本不等式即可求得结果. 解答：在$\triangle ABC$ 中，$\cos A=\frac{b^2+c^2-a^2}{2bc}\geqslant\frac{2bc-bc}{2bc}=\frac{1}{2}$，因此 A 的最大值为$\frac{\pi}{3}$

Table 3　Example of Answer Log

表 3　答题记录示例

Attribute	Value
Paper ID	ff3110a0-fd51-4dfd-a6c7
School ID	2300000001000002
Test Date	2017-1-17
Student ID	4444000020013967
Item ID	a8da5256-fe26-451b
Full Score	12
Get Score	10

定义 1. 给定数学试题集合 Q 和数学考试记录集合 T，其中 Q 包含每道试题的文本，T 包含每场考试的试题和对应的得分率，目标是对数学试题建模，使得通过输入试题特征到模型中可以得到试题的难度预测值.

表 4 给出了问题涉及到的符号和对应的描述：

Table 4　Related Symbols and Explanations

表 4　试题难度预测问题涉及的符号及解释

Symbol	Explanation
Q	Set of all items
Q_i	Item i
R_i	Average score of item i
P_i	Predicted difficulty of item i
T	Set of all mathematical tests
T_i	Mathematical test i
$\tilde{Q}_i$	Item set in T_i
$\tilde{Q}_{ij}$	Item j in T_i
$\tilde{R}_i$	Set of average scores in T_i
$\tilde{R}_{ij}$	Average score of item j in T_i
$\boldsymbol{X}_q$	Feature of item q

2.2　模型整体框架

本节介绍本文提出的数学试题难度预测模型的整体框架，整体流程如图 1，分成 2 个阶段：训练阶段和预测阶段. 在训练阶段，根据将答题记录中的试题文本进行表征后得到训练特征，作为模型训练的输入，并从答题记录获取每一场考试中各道试题的得分率作为试题难度的标签，考虑不同考试中试题得分率的不可比性，训练时采用 context 相关的成对试题目标函数；在预测阶段，将待预测试题的文本经同样的表征方式得到预测特征，将其输入训练得到的模型，获得难度的预测值. 模型分 3 部分介绍：

1）模型结构. C-MIDP，R-MIDP，H-MIDP 这 3 个模型均为神经网络模型，其中 C-MIDP 以 CNN 网络为基础，R-MIDP 以 RNN 网络为基础，H-MIDP 为前两者的融合.

2）模型训练. 训练时以试题文本的词向量特征作为输入，试题得分率作为标签. 考虑到不同考试中不同学生群体的得分率具有一定的不可比性，本模型采用 context 相关(context-dependent)的方式，将同一场考试中成对试题预测难度的差值与实际差值比较，计算目标函数值.

3）预测. 试题难度预测是 context 无关的，将预处理过的试题特征作为输入，得到试题的绝对难度.

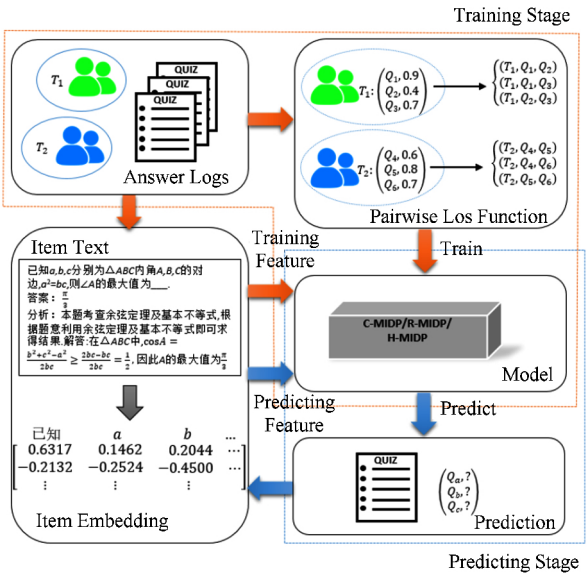


Fig. 1　Model framework

图 1　模型框架图

2.3　模型结构

本文提出的 3 种模型接受试题特征作为输入，输出为试题的预测难度. 试题特征通过对文本字符的词向量拼接获得，具体步骤：

步骤 1. 获取试题中文词语或英文字符的表征向量. 使用 word2vec[24] 方法，以数学试题文本集合

为训练数据，训练得到文本库中每个词语或字符的词向量 $w_i \in \mathrm{R}^{d_0 \times 1}$. 以训练得到的词向量作为对应词语或字符的表征向量，相比 one-hot 的表征方式，维数更低，且能够保持一定的语义特征.

步骤 2. 构建试题的初始表征向量. 将试题分词并去除停用词，剩余的词语或字符按原文顺序以对应的词向量替换，得到试题初始表征向量 $\boldsymbol{X}'_q \in \mathrm{R}^{d_0 \times n}$，其中 n 为词向量数量.

步骤 3. 修改 $\boldsymbol{X}'_q$ 长度得到试题最终表示向量 $\boldsymbol{X}_q$. 由于模型输入的特征需为固定长度，因此选择合适的长度 N（N 为词向量数量，实验中设置 $N=600$，具体见 3.1 节），若 $n<N$，则用 $N-n$ 个零向量填充，反之若 $n>N$，则删去最后 $n-N$ 个词向量. 最终得到的 $\boldsymbol{X}_q$ 作为试题的表征向量.

将试题文本转换成向量特征后，输入模型进行语义理解. 图 2 是 3 种模型的结构图，其中图2(a)～(c)分别是 C-MIDP 模型、R-MIDP 模型和 H-MIDP 模型.

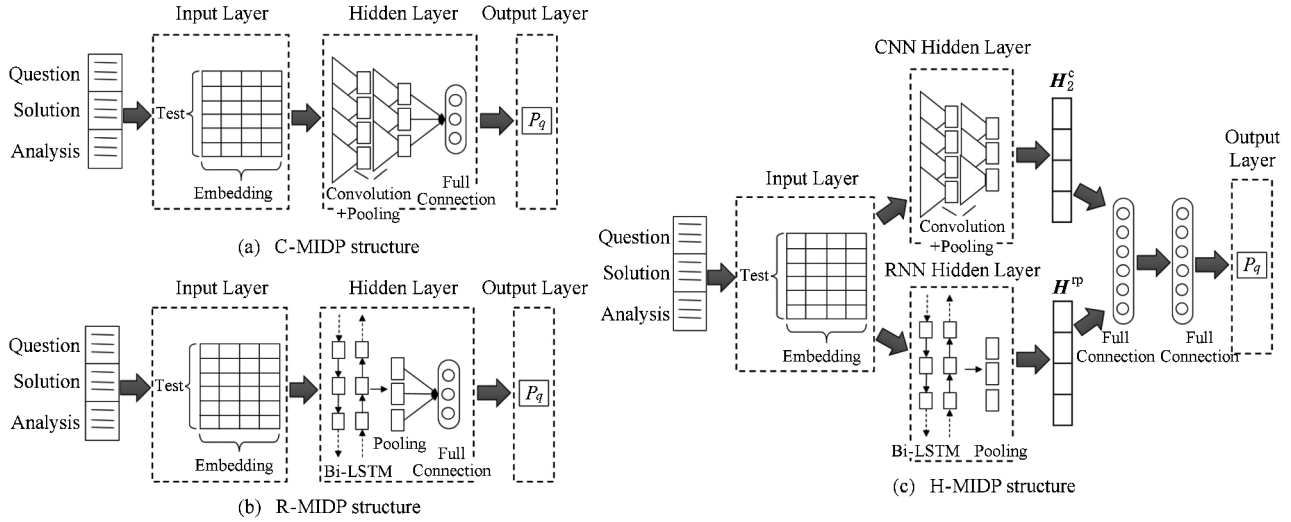


Fig. 2 Model structures

图 2 模型结构

2.3.1 C-MIDP 模型

试题文本包含较丰富的语义，要使模型能够不依赖 Q 矩阵等先验知识，就必须能够从文本中挖掘足够的信息. 相关研究表示，局部重要的词句对于文本理解具有重要的意义[30]. 例如在理解试题时，我们只需理解其中最重要的知识概念描述（如公式、定义等）即可理解整个试题的语义. 因此，本文利用 CNN 中的卷积-池化从局部到整体的方式挖掘试题文本中的主要信息[30]. 具体地，本文提出 C-MIDP 模型，它以 CNN 为基础，使用的多层卷积与池化层可以从不同层次学习试题信息. 例如 C-MIDP 可以以试题中的数字或运算符为基础扩大范围，提取由这些数字或运算符等组成公式信息；再进一步联系公式的上下文获取更大范围的信息，逐步获取整个试题的主要信息，这个过程也符合人真实的阅读习惯.

C-MIDP 模型结构如图 2(a)所示，模型包括输入层、2 层卷积层和 Max Pooling 层、全连接层和输出层. 输入层接受试题特征 $\boldsymbol{X}_q \in \mathrm{R}^{d_0 \times N}$. $\boldsymbol{X}_q$ 在第 1 个卷积层通过 d_1 个 $k \times 1$ 的卷积核执行卷积操作，输出 d_1 个 channel 的隐层.

具体地，给定输入 $\boldsymbol{X}_q=(\boldsymbol{w}_1, \boldsymbol{w}_2, \cdots, \boldsymbol{w}_N)$，$\boldsymbol{w}_i \in \mathrm{R}^{d_0 \times 1}$，通过卷积获得隐层 $\boldsymbol{H}^c=(\boldsymbol{h}_1^c, \boldsymbol{h}_2^c, \cdots, \boldsymbol{h}_{N+k-1}^c) \in \mathrm{R}^{d_1 \times (N+k-1)}$，其中：

$$\boldsymbol{h}_i^c = ReLU\Big(\sum_{j=1}^{d_0} \boldsymbol{G} \times (w_{i-k+1}^j, w_{i-k+2}^j, \cdots, w_{i-1}^j, w_i^j)^{\mathrm{T}} + \boldsymbol{b}\Big), \tag{2}$$

$\boldsymbol{G} \in \mathrm{R}^{d_1 \times k}$，即 d_1 个长度为 k 的卷积核；w_i^j 表示 $\boldsymbol{w}_i$ 的第 j 维元素值；$\boldsymbol{b} \in \mathrm{R}^{d_1 \times 1}$，为偏置项；$ReLU=\max(0, x)$ 为非线性激活函数.

第 1 层卷积层的输出 $\boldsymbol{H}^c$ 经过 p-max 池化层，以选出其中最重要的信息，得到新的隐层 $\boldsymbol{H}^{cp}=(\boldsymbol{h}_1^{cp}, \boldsymbol{h}_2^{cp}, \cdots, \boldsymbol{h}_{\lfloor (N+k-1)/p \rfloor}^{cp})$，其中：

$$\boldsymbol{h}_i^{cp} = \left[\max\begin{bmatrix} h_{i\times p-p+1,1}^c \\ \cdots \\ h_{i\times p,1}^c \end{bmatrix}, \cdots, \max\begin{bmatrix} h_{i\times p-p+1,d_1}^c \\ \cdots \\ h_{i\times p,d_1}^c \end{bmatrix}\right]^{\mathrm{T}}. \tag{3}$$

$\boldsymbol{H}^{\mathrm{cp}}$再次经过一层卷积层和 p-max 池化层,具体操作与第 1 层类似.设第 2 层卷积层的卷积核数量为 d_2,取第 2 层 Max Pooling 层的窗口大小为 $\lfloor (N+k-1)/p \rfloor$,此时输出的隐层 $\boldsymbol{H}_2^{\mathrm{c}} \in \mathrm{R}^{d_2 \times 1}$,转置后经过一层全连接层,输出试题难度的预测值 P_q.

2.3.2　R-MIDP 模型

除此之外,文本的序列语义与逻辑信息对于理解试题也非常重要.例如公式中的一个数字本身可能不包含多少信息,但若与它前面的若干个字符联系,可能就表现出重要的语义.基于此,本文提出 R-MIDP 模型,它以 RNN 为基础,利用 RNN 中的 Cell 模块保存历史信息,学习到试题文本的序列语义或逻辑信息.具体地,R-MIDP 模型是一个双向 LSTM 的网络结构,LSTM 采用经典的 3 门结构[31-32],在理解试题的过程中,可以从正向和反向 2 个方向学习试题语义逻辑,使语义更加完整.

如图 2(b)所示,R-MIDP 模型包括输入层、双向 LSTM 隐层、Max Pooling 层、全连接层和输出层.输入层接受试题特征 $\boldsymbol{X}_q = (\boldsymbol{w}_1, \boldsymbol{w}_2, \cdots, \boldsymbol{w}_N) \in \mathrm{R}^{d_0 \times N}$,LSTM 层输出维度为 d,经过单层 LSTM 得到隐层 $\boldsymbol{H}^{\mathrm{r}} = (\boldsymbol{h}_1^{\mathrm{r}}, \boldsymbol{h}_2^{\mathrm{r}}, \cdots, \boldsymbol{h}_d^{\mathrm{r}})^{\mathrm{T}} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_N) \in \mathrm{R}^{d \times N}$,其中 $\boldsymbol{h}_j^{\mathrm{r}} = (h_{j,1}^{\mathrm{r}}, h_{j,2}^{\mathrm{r}}, \cdots, h_{j,N}^{\mathrm{r}})$,正向输入时的 $\boldsymbol{y}_t$ 由下列 LSTM 计算公式获得:

$$\boldsymbol{i}_t = \sigma(\boldsymbol{W}_{\mathrm{ii}} \boldsymbol{w}_t + \boldsymbol{b}_{\mathrm{ii}} + \boldsymbol{W}_{\mathrm{hi}} \boldsymbol{y}_{t-1} + \boldsymbol{b}_{\mathrm{hi}}), \tag{4}$$

$$\boldsymbol{f}_t = \sigma(\boldsymbol{W}_{\mathrm{if}} \boldsymbol{w}_t + \boldsymbol{b}_{\mathrm{if}} + \boldsymbol{W}_{\mathrm{hf}} \boldsymbol{y}_{t-1} + \boldsymbol{b}_{\mathrm{hf}}), \tag{5}$$

$$\boldsymbol{g}_t = \sigma(\boldsymbol{W}_{\mathrm{ig}} \boldsymbol{w}_t + \boldsymbol{b}_{\mathrm{ig}} + \boldsymbol{W}_{\mathrm{hg}} \boldsymbol{y}_{t-1} + \boldsymbol{b}_{\mathrm{hg}}), \tag{6}$$

$$\boldsymbol{o}_t = \sigma(\boldsymbol{W}_{\mathrm{io}} \boldsymbol{w}_t + \boldsymbol{b}_{\mathrm{io}} + \boldsymbol{W}_{\mathrm{ho}} \boldsymbol{y}_{t-1} + \boldsymbol{b}_{\mathrm{ho}}), \tag{7}$$

$$\boldsymbol{c}_t = \boldsymbol{f}_t * \boldsymbol{c}_{t-1} + \boldsymbol{i}_t * \boldsymbol{g}_t, \tag{8}$$

$$\boldsymbol{y}_t = \boldsymbol{o}_t * \tanh(\boldsymbol{c}_t), \tag{9}$$

其中 $\boldsymbol{i}_t, \boldsymbol{f}_t, \boldsymbol{o}_t$ 分别为输入门、遗忘门、输出门,$\boldsymbol{w}_t$ 为时刻 t 输入,$\boldsymbol{y}_{t-1}$ 为时刻 $t-1$ 时 LSTM Cell 的输出,$\boldsymbol{c}_t$ 为时刻 t 时 Cell 的状态,σ 为 sigmoid 函数,$*$ 为卷积运算.$\boldsymbol{H}^{\mathrm{r}}$ 经过 p-max 池化层并转置后得到新的隐层 $\boldsymbol{H}^{\mathrm{rp}} = (h_1^{\mathrm{rp}}, h_2^{\mathrm{rp}}, \cdots, h_d^{\mathrm{rp}})$,其中:

$$h_j^{\mathrm{rp}} = \max(h_{j,1}^{\mathrm{r}}, h_{j,2}^{\mathrm{r}}, \cdots, h_{j,N}^{\mathrm{r}}).$$

$\boldsymbol{H}^{\mathrm{rp}}$ 再经过一层全连接层,最终输出试题难度的预测值 P_q.

2.3.3　H-MIDP 模型

更进一步,本文结合 C-MIDP 和 R-MIDP 这 2 个模型的优势,提出一种混合模型 H-MIDP,以期同时对试题文本的局部重要语义和序列逻辑信息进行有效建模.H-MIDP 结构如图 2(c)所示.模型前半部分分为并行的 2 部分,其中一部分与 C-MIDP 相同,输入特征经 2 层卷积层和 Max Pooling 层后得到隐层值 $\boldsymbol{H}_2^{\mathrm{c}}$;另一部分与 R-MIDP 相同,输入特征经 LSTM 和 Max Pooling 层后得到隐层值 $\boldsymbol{H}^{\mathrm{rp}}$,将 $\boldsymbol{H}_2^{\mathrm{c}}$ 与 $\boldsymbol{H}^{\mathrm{rp}}$ 拼接得到 $\boldsymbol{H} \in \mathrm{R}^{1 \times (d_2 + d)}$,经过 2 层全连接层得到试题难度的预测值 P_q.

2.4　模型训练

在通常的有监督模型中,常规的训练方法是以训练数据的试题表征向量作为输入,以试题得分率作为标签,模型的损失函数(loss function):

$$L(T) = \sum_q (P_q - R_q)^2, \tag{10}$$

其中,T 为整个数学考试训练集,P_q 和 R_q 分别为试题 q 的预测难度和实际得分率.

这种方式在计算试题得分率时常以试题为单位进行,其训练时其实是不区分不同学生群体或不同场考试的.但实际上,不同考试中由于学生群体的不同,得分率是具有一定不可比性的.例如假设 A 校和 B 校使用同一份试卷进行考试,A 校的试题 a 得分率为 0.8,B 校的试题 b 得分率为 0.7,不能简单地认为试题 b 比试题 a 更难,因为 A 校学生的整体水平可能强于 B 校学生,而实际 A 校的试题 b 得分率为 0.9,B 校的试题 a 得分率 0.6,因而判断试题 a 的难于试题 b 更合理.

由此可知,试题得分率受到学生群体水平差异性的影响.为了能够消除这种影响,本文认为,当考试学生群体处于相同的 context 范围下,通过考试计算的试题得分率才具有可比性.此处,context 可以定义为同一个班级、同一所学校、同一场考试等.例如,在同一场考试中,若试题 a 得分率低于试题 b,即可认为 a 比 b 难.本文将在实验部分中具体对此范围进行实验说明.

具体地,本文的 3 种模型采用 context 相关的训练方式,模型的损失函数:

$$L(T) = \sum_{(T_t, Q_i, Q_j)} ((P_{ti} - P_{tj}) - (\widetilde{R}_{ti} - \widetilde{R}_{tj}))^2, \tag{11}$$

其中,T_t 表示 context 范围 t,P_{ti} 和 P_{tj} 分别指 context T_t 中试题 Q_i 和 Q_j 的预测难度,$\widetilde{R}_{ti}$ 和 $\widetilde{R}_{tj}$ 分别指 context T_t 中试题 Q_i 和 Q_j 的实际难度(得分率).

使用这样的模型损失函数可以消除不同学生群体的差异性,获取其中的共性,使得训练得到的模型能够预测试题的真实难度(对于所有答题记录涉及到的学生全体而言的难度,而不是对于其中某场考试的学生群体).

2.5　难度预测

模型训练完毕,进行试题难度的预测时,将需要预测的试题表征向量输入训练得到的模型中(C-MIDP 或 R-MIDP 或 H-MIDP),得到的模型输出值

即为试题难度的预测值. 在实际应用情境下,如果收集的群体答题数据量充足且答题分布均匀,则可以认为模型的输出值可以预测试题对于该群体的难度值(或得分率).

3 模型验证实验

3.1 数据集介绍

数据来自科大讯飞股份有限公司采集的国内多个中学 2014—2017 年的考试试题和答题记录,相关统计见表 5.

Table 5 Statistical Analysis of Data Set

表 5 数据集相关统计分析

Attribute	Value
Amount of Schools	1 314
Amount of Tests	5 185
Average Amount of Items per Test	18.33
Amount of Different Items	53 027
Amount of Logs	57 457 353
Amount of Students	1 035 526

对试题文本数据预处理后统计每道题的特征长度(即分词后有效词项数目),得到其分布如图 3 所示,图 3 中横坐标为特征长度,纵坐标为试题数量. 由统计结果知特征长度大于 600 的不到总试题数的 0.2%,因此实验中取特征向量长度 $N=600$,实际少于 600 的试题用零填充,多于 600 的试题截取前 600 个词项作为试题特征.

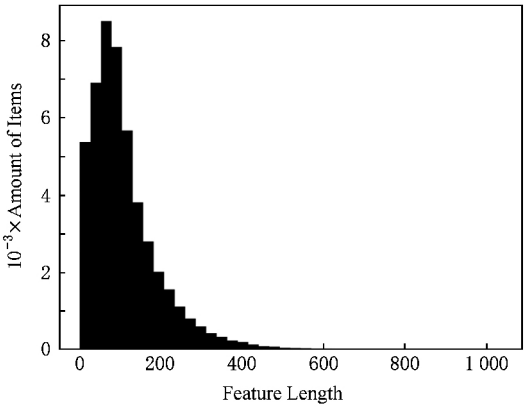


Fig. 3 Distribution of item feature length

图 3 试题特征长度分布

选取使用某一份试卷不同场考试的答题记录,绘制不同学校的试题得分率折线图如图 4 所示,可以看到,不同学校在各个试题上的得分率虽有明显差异,但试题之间的得分率相对差异却相近. 图 4 中 A 校(最上方绿色折线)的试题 Q_{10} 的得分率为 0.3, B 校(最下方橙色折线)的试题 Q_9 的得分率为0.22, 但不能简单以此判断试题 Q_{10} 的难度低于 Q_9,因为 A 校的整体能力强于 B 校. 实际上,A 校的试题 Q_9 的得分率为 0.4,B 校的试题 Q_{10} 的得分率为 0.08, 可以看到不论是 A 校还是 B 校,试题 Q_9 的得分率高于试题 Q_{10} 的得分率,因此判断试题 Q_9 的难度低于 Q_{10} 更合理. 这正验证了 2.4 节中的观点.

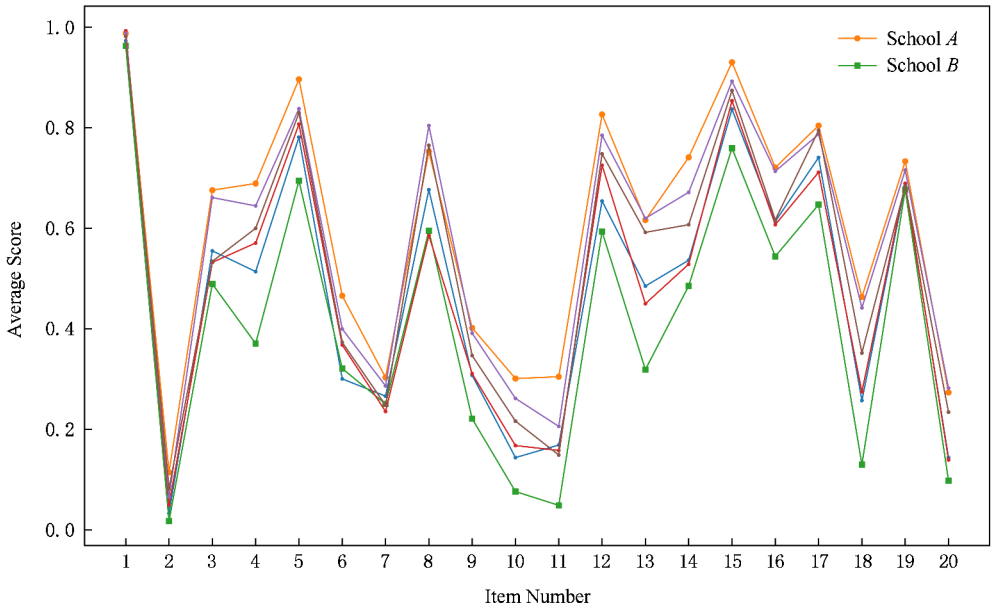


Fig. 4 Scoring rates of 6 schools in a final exam

图 4 6 所学校在同场期末考试中的得分率

3.2　实验评价指标

3.2.1　皮尔森相关系数(Pearson correlation coefficient, PCC)

PCC 是教育学常用的评价指标,可以衡量每一场考试中试题实际难度与模型预测难度之间的相关性[33]. 实验中 PCC 具体定义为

$$L_{\mathrm{PCC}}=\frac{\sum_{j=1}^{m_i}(P_{ij}-\bar{P}_i)(\tilde{R}_{ij}-\bar{R}_i)}{\sqrt{\sum_{j=1}^{m_i}(P_{ij}-\bar{P}_i)^2}\sqrt{\sum_{j=1}^{m_i}(\tilde{R}_{ij}-\bar{R}_i)^2}},\quad(12)$$

其中,m_i 为某场考试 i 中的试题数,P_{ij} 为该场考试中试题 j 的预测难度,$\bar{P}_i$ 为该场考试中试题的平均预测难度,$\tilde{R}_{ij}$ 为该场考试中试题 j 的实际难度,$\bar{R}_i$ 为该场考试中试题的平均实际难度.

PCC 取值在区间$[-1,1]$,越大的绝对值意味着越高的线性相关性,且 $\mathrm{PCC}>0$ 表示正相关,$\mathrm{PCC}<0$ 表示负相关.

3.2.2　一致性(degree of agreement, DOA)

DOA 可以衡量一场考试中试题对之间难度预测值相对大小的准确性[34]. 其计算为

$$L_{\mathrm{DOA}}=\frac{\sum_{1\leqslant a,b\leqslant m_i}\sigma(P_{ia},P_{ib})\wedge\sigma(\tilde{R}_{ia},\tilde{R}_{ib})}{\sum_{1\leqslant a,b\leqslant m_i}\sigma(\tilde{R}_{ia},\tilde{R}_{ib})},\quad(13)$$

其中,m_i 为某场考试中的试题数,P_{ia} 和 P_{ib} 分别为该场考试中试题 a 和试题 b 的预测难度,$\tilde{R}_{ia}$ 和 $\tilde{R}_{ib}$ 分别为该场考试中试题 a 和试题 b 的实际难度. $\sigma(x,y)$的定义为

$$\sigma(x,y)=\begin{cases}1, & x>y.\\ 0, & \text{otherwise}.\end{cases}\quad(14)$$

DOA 取值范围在区间$[0,1]$,DOA 越大表明预测的试题对之间相对难度大小关系越准确.

3.3　对比实验

为验证本文提出的模型效果,将与 4 种 baseline 预测方法进行对比:

1) logistic 回归[35]. 传统的线性回归模型,模型输入特征为试题的词袋特征,采用 context 无关的训练方式.

2) 支持向量机(SVM)[36]. SVM 在线性和非线性回归问题中都比较常见,是机器学习中重要的算法. 对比模型采用非线性高斯核,输入为试题的词袋特征,并采用 context 无关的训练方式.

3) 随机森林(random forest)[37]. 随机森林回归模型是常用的非线性模型,在许多回归任务上具有良好的表现. 模型输入同样采用试题的词袋特征,且采用 context 无关的训练方式.

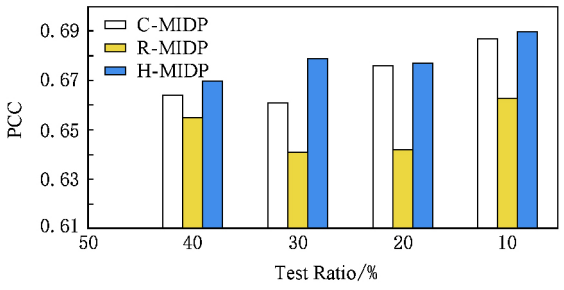
4) 神经网络 context 无关训练方式. 本文的 3 种模型结构不变,但训练方式改为 context 无关,即采用式(10)作为损失函数,以试题的预测难度与实际得分率的差值平方和作为目标函数. 3 种模型分别以 CNN-I, RNN-I, Hybrid-I 指代.

3.4　实验结果及分析

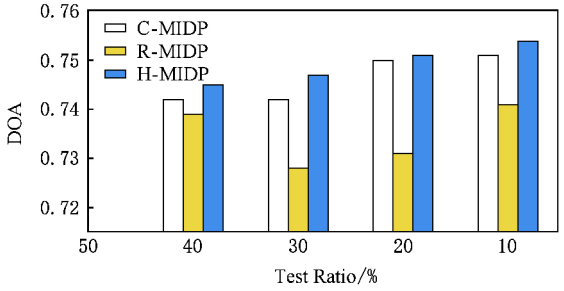
3.4.1　模型对比实验

本节将比较 C-MIDP, R-MIDP, H-MIDP 这 3 种模型的实验结果,以及分析与 baseline 模型实验结果的对比. 此处,C-MIDP, R-MIDP, H-MIDP 这 3 种模型中的 context 定义为同一场考试范围,即式(11)中的 T_t 表示第 t 场考试. 实验分别取数据集中考试数量的 40%, 30%, 20%, 10%作为测试集,同时删除训练集中在测试集出现的试题,这些重复试题若在训练集中得到拟合,将不适合用作模型测试. 注意到,考试可能是一个班级单独的测试,也可能是整个年级统考,或者多所学校联考,这里我们采取的划分方式是:同一所学校同一天使用同一份试卷划分为一场考试,作为计算试题得分率的 context,在此基础上训练 C-MIDP, R-MIDP, H-MIDP 模型. 最终得到各个模型在测试集上的 PCC 与 DOA 指标的值如图 5 所示.

从图 5 中实验结果可知,C-MIDP, R-MIDP, H-



(a)　PCC of MIDP model results



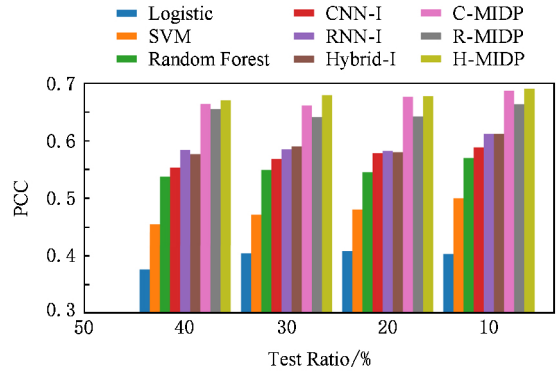
(b)　DOA of MIDP model results

Fig. 5　Experiment results of three models

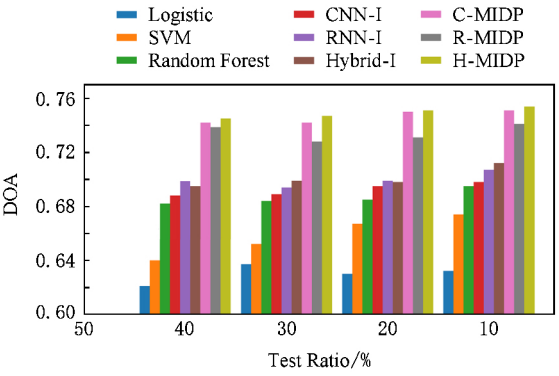
图 5　3 种模型实验结果

MIDP 模型都有良好的表现，并且可以看到，在测试集比例为 40%，30%，20%，10% 情况下，H-MIDP 的测试指标均高于 C-MIDP 和 R-MIDP.

图 6 是本文 3 种模型与对比模型实验结果，从图 6 中可以看出 3 项对比信息：



(a) PCC of baseline and MIDP model results



(b) DOA of baseline and MIDP model results

Fig. 6 Results contrast experiments

图 6 对比实验结果

1) 在使用 context 无关的训练方式前提下，logistic 回归效果最差，显然线性回归不能够胜任试题难度预测任务；SVM 回归效果较 logistic 回归更好；随机森林回归在 3 种非神经网络 baseline 模型中表现最好；CNN-I，RNN-I，Hybrid-I 这 3 种神经网络模型的实验结果明显优于前 3 种非神经网络模型，说明神经网络对此任务的建模能力更强.

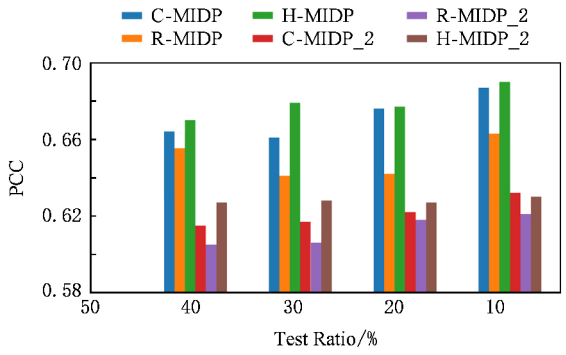
2) 比较 3 种神经网络模型的 context 相关与 context 无关 2 种训练方式的实验结果，可以看到，尽管使用 context 无关训练方式(CNN-I，RNN-I，Hybrid-I)已经获得良好的实验结果，但使用 context 相关训练方式后，模型效果有了进一步的提升，说明在试题难度预测这个任务当中，context 相关的训练方式更适合.

3) 随着测试集比例的降低(即训练数据的增加)，3 种模型的效果均提升. 测试集的比例降到 10%时，3 种神经网络模型的 PCC 达到 0.66 以上，DOA 达到 0.74 以上. 在实际教育环境中，数据量足够的情况下，能够达到良好的预测效果.

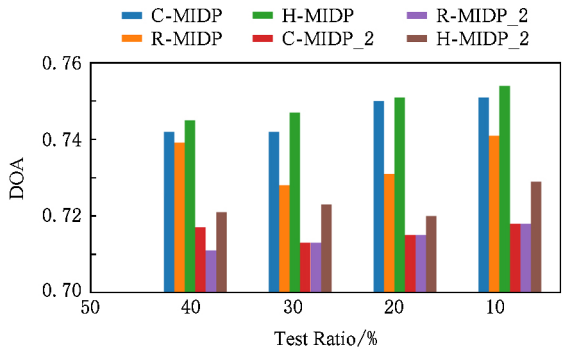
3.4.2 context 划分方式对预测结果的影响

本节将讨论不同的 context 划分对于试题难度预测结果的影响. 这里的 context 划分等价于考试的划分，例如在一场多校联考中，可以将一个班级的记录划分为一场考试，也可以将一所学校的记录划分为一场考试，或者将各个学校的所有记录共同作为一场考试. 本节针对数据采用 2 种不同的划分方式：1)将同一所学校同一天使用相同试卷划分为一个 context；2)将使用相同试卷的所有记录划分为一个 context. 依此进行实验，研究 context 划分方式对试题难度预测结果的影响.

图 7 是 2 种划分方式的在测试集上的 PCC 和 DOA 指标的直方图. 可以看到 2 种划分方式的实验结果有明显差距，第 1 种划分方式的实验结果优于第 2 种划分方式，说明 context 的划分方式对预测结果是有影响的. 在本实验数据集上，若将考试的范围细化到学校层面，可以更好地区分来自不同学校学生群体的差异性，从而获得更稳定的试题难度. 在实际应用中，模型的实际训练与使用中需根据测试结果选择合适的 context 划分方式.



(a) PCC results of models with 2 different contexts



(b) DOA results of models with 2 different contexts

Fig. 7 Experiment results of two different context division

图 7 2 种 context 划分方式实验结果

3.5　案例分析

本节选取测试集比例为 40% 时测试集中的 1 场考试试题,使用 C-MIDP,R-MIDP,H-MIDP 模型进行难度预测,比较预测结果,以说明本文的 3 种模型的有效性. 图 8 是各模型预测结果折线图,其中实际得分率是将数据集中所有使用该份试卷试题的答题记录得分率取平均得到,以更准确反映试题实际难度.

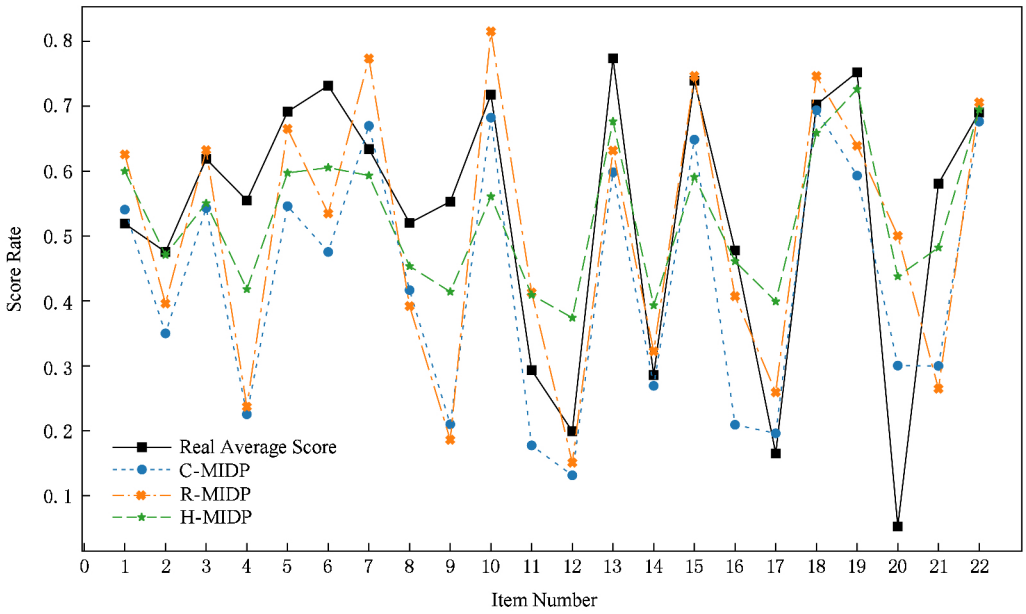


Fig. 8　Comparison between score rates predicted by 3 models and ground truth on a test paper

图 8　某试卷 3 种模型预测得分率与真实值比较

表 6 是评价指标 PCC,DOA,RMSE 值. 可以看到 H-MIDP 的 3 种指标的值均优于 C-MIDP 和 R-MIDP,但 C-MIDP 和 R-MIDP 的评价值也在可接受范围. 观察图 8,可以看到 3 种模型在大多数试题上的预测值能够接近实际得分率,或者在试题相对难度关系上接近,其中 H-MIDP 的预测曲线与真实值最为接近,说明模型能够通过 context 相关的训练方式来预测试题绝对难度.

Table 6　Metrics Values of Models in Case Study

表 6　案例分析各模型评价指标值

Model	PCC	DOA	RMSE
C-MIDP	0.766	0.788	0.171
R-MIDP	0.627	0.723	0.179
H-MIDP	0.797	0.823	0.136

4　结　　论

为解决准确、高效地预测数学试题难度所面临的难题,辅助中国特色教育考试国家题库建设,本文提出了数据驱动的基于神经网络的难度预测模型. 具体地,首先设计了基于卷积神经网络的 C-MIDP 模型和基于循环神经网络的 R-MIDP 模型学习试题文本的序列逻辑信息;进一步,结合 2 种模型的优势,提出混合 H-MIDP 模型. 3 种模型均直接对试题文本进行理解和语义表征,可保留试题描述的局部语义和语序信息;然后,为应对不同考试中学生群体具有不可比性的问题,在模型训练时考虑答题记录的上下文,采用 context 相关的训练方式;最后,所提出的模型只需根据试题文本即可预测新试题难度属性,无需人工标注先验知识信息. 本文在真实数据集上进行了大量实验,实验结果表明了本文所提出的模型具有良好的性能.

本文的模型具有进一步改良的空间和向其他学科扩展的可能性. 在未来研究中,可以考虑新的模型结构对试题文本理解的影响,如 Attention 网络、Memory 网络等. 其次,探索更为准确和稳定的 context 的划分方式,以减少对试题难度预估结果的影响. 我们还将考虑针对不同试题类型设计更为精准的预测模型.

参　考　文　献

[1] Mao Jingfei. Exploration of difficulty prediction methods for questions in college entrance examination [J]. Education Science, 2008, 24(6): 22-26 (in Chinese)

(毛竞飞. 高考命题中试题难度预测方法探索[J]. 教育科学, 2008, 24(6): 22-26)

[2] Liu Qi, Chen Enhong, Zhu Tianyu, et al. Research on educational data mining for online intelligent learning [J]. Pattern Recognition and Artificial Intelligence, 2018, 31(1): 77-90 (in Chinese)
(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[J]. 模式识别与人工智能, 2018, 31(1): 77-90)

[3] Liu Qi, Chen Enhong, Huang Zhenya, et al. Cognitive ability analysis of students for personalized learning [J]. Communications of the CCF, 2017, 13(4): 28-35 (in Chinese)
(刘淇, 陈恩红, 黄振亚, 等. 面向个性化学习的学生认知能力分析[J]. 计算机学会通讯, 2017, 13(4): 28-35)

[4] Fan Xitao. Item response theory and classical test theory: An empirical comparison of their item/person statistics [J]. Educational and Psychological Measurement, 1998, 58(3): 357-381

[5] De La Torre J. DINA model and parameter estimation: A didactic [J]. Journal of Educational and Behavioral Statistics, 2009, 34(1): 115-130

[6] Dong Shenghong, Qi Shuqing, Dai Haiqi, et al. Research on manual assignment methods for difficulty and distinction parameters [J]. Testing Research, 2005, 1(1): 25-32 (in Chinese)
(董圣鸿, 漆书青, 戴海琦, 等. 题目难度, 区分度参数人工赋值方法的研究[J]. 考试研究, 2005, 1(1): 25-32)

[7] Beck J, Stern M, Woolf B P. Using the student model to control problem difficulty [C] //Proc of the 6th Int Conf on User Modeling. Berlin: Springer, 1997: 277-288

[8] Kubinger K D, Gottschall C H. Item difficulty of multiple choice tests dependant on different item response formats--An experiment in fundamental research on psychological assessment [J]. Psychology Science, 2007, 49(4): 361-374

[9] Wu Runze, Liu Qi, Liu Yuping, et al. Cognitive modelling for predicting examinee performance [C] //Proc of the 24th Int Joint Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2015: 1017-1024

[10] Zhu Tianyu, Huang Zhenya, Chen Enhong, et al. Cognitive diagnosis based personalized question recommendation [J]. Chinese Journal of Computers, 2017, 40(1): 176-191 (in Chinese)
(朱天宇, 黄振亚, 陈恩红, 等. 基于认知诊断的个性化试题推荐方法[J]. 计算机学报, 2017, 40(1): 176-191)

[11] DiBello L V, Roussos L A, Stout W. 31a review of cognitively diagnostic assessment and a summary of psychometric models [J]. Handbook of Statistics, 2007, 26: 979-1030

[12] Zhang Xiao, Sha Ruxue. Research advance in DINA model of cognitive diagnosis [J]. China Examinations, 2013(1): 32-37 (in Chinese)
(张潇, 沙如雪. 认知诊断 DINA 模型研究进展[J]. 中国考试, 2013(1): 32-37)

[13] Maris E. Estimating multiple classification latent class models [J]. Psychometrika, 1999, 64(2): 187-212

[14] Wu Ruize, Xu Guandong, Chen Enhong, et al. Knowledge or gaming?: Cognitive modelling based on multiple-attempt response [C] //Proc of the 26th Int Conf on World Wide Web Companion. New York: ACM, 2017: 321-329

[15] Chen Yuying, Liu Qi, Huang Zhenya, et al. Tracking knowledge proficiency of students with educational priors [C] //Proc of the 26th ACM Conf on Information and Knowledge Management. New York: ACM, 2017: 989-998

[16] Huang Zhenya, Liu Qi, Chen Enhong, et al. Question difficulty prediction for reading problems in standard tests [C] //Proc of the 31st AAAI Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2017: 1352-1359

[17] Wang Weiqiang, Gao Wen, Duan Lijuan. Text mining on the Internet [J]. Computer Science, 2000, 27(4): 32-36 (in Chinese)
(王伟强, 高文, 段立娟. Internet 上的文本数据挖掘[J]. 计算机科学, 2000, 27(4): 32-36)

[18] Wei Shunping. Learning analysis technology: Mining educational data's value in the era of big data [J]. Modern Educational Technology, 2013, 2(23): 5-11 (in Chinese)
(魏顺平. 学习分析技术: 挖掘大数据时代下教育数据的价值[J]. 现代教育技术, 2013, 2(23): 5-11)

[19] Yang Pei, Yang Zhihao, Luo Ling, et al. An attention-based approach for chemical compound and drug named entity recognition [J]. Journal of Computer Research and Development, 2018, 55(7): 1548-1556 (in Chinese)
(杨培, 杨志豪, 罗凌, 等. 基于注意机制的化学药物命名实体识别[J]. 计算机研究与发展, 2018, 55(7): 1548-1556)

[20] Zhang Ying, Wang Chao, Guo Wenya, et al. Multi-source emotion tagging for online news comments using bi-directional hierarchical semantic representation model [J]. Journal of Computer Research and Development, 2018, 55(5): 933-944 (in Chinese)
(张莹, 王超, 郭文雅, 等. 基于双向分层语义模型的多源新闻评论情绪预测[J]. 计算机研究与发展, 2018, 55(5): 933-944)

[21] Gao Yunlong, Zuo Wanli, Wang Ying, et al. Sentence classification model based on sparse and self-taught convolutional neural networks [J]. Journal of Computer Research and Development, 2018, 55(1): 179-187 (in Chinese)
(高云龙, 左万利, 王英, 等. 基于稀疏自学习卷积神经网络的句子分类模型[J]. 计算机研究与发展, 2018, 55(1): 179-187)

[22] Chen Ke, Liang Bin, Ke Wende, et al. Chinese micro-blog sentiment analysis based on multi-channels convolutional neural networks [J]. Journal of Computer Research and Development, 2018, 55(5): 945-957 (in Chinese)

(陈珂,梁斌,柯文德,等.基于多通道卷积神经网络的中文微博情感分析[J].计算机研究与发展,2018,55(5):945-957)

[23] Miner G, Elder IV J, Hill T. Practical Text Mining and Statistical Analysis for Non-Structured Text Data Applications [M]. Amsterdam, Netherlands: Academic Press, 2012

[24] Mikolov T, Sutskever I, Chen Kai, et al. Distributed representations of words and phrases and their compositionality [C] //Proc of the 26th Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2013: 3111-3119

[25] Le Q, Mikolov T. Distributed representations of sentences and documents [C] //Proc of the 31st Int Conf on Machine Learning. New York: ACM, 2014: 1188-1196

[26] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition [J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324

[27] Graves A, Liwicki M, Fernández S, et al. A novel connectionist system for unconstrained handwriting recognition [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(5): 855-868

[28] Cambria E, Gastaldo P, Bisio F, et al. An ELM-based model for affective analogical reasoning [J]. Neurocomputing, 2015, 149: 443-455

[29] Bowman S R, Angeli G, Potts C, et al. A large annotated corpus for learning natural language inference [EB/OL]. (2015-08-21) [2018-08-17]. https://arxiv. org/abs/1508.05326

[30] Yin Wenpeng, Ebert S, Schütze H. Attention-based convolutional neural network for machine comprehension [EB/OL]. (2016-02-13) [2018-08-17]. https://arxiv. org/abs/1602.04341

[31] Hochreiter S, Schmidhuber J. Long short-term memory [J]. Neural Computation, 1997, 9(8): 1735-1780

[32] Gers F A, Schmidhuber J, Cummins F. Learning to forget: Continual prediction with LSTM [J]. Neural Computation, 2000, 12(10): 2451-2471

[33] Benesty J, Chen Jingdong, Huang Yiteng, et al. Noise Reduction in Speech Processing [M]. Berlin: Springer, 2009: 37-40

[34] Liu Qi, Chen Enhong, Xiong Hui, et al. Enhancing collaborative filtering by user interest expansion via personalized ranking [J]. IEEE Transactions on Systems, Man, and Cybernetics: Part B, 2012, 42(1): 218-233

[35] Cox D R. The regression analysis of binary sequences [J]. Journal of the Royal Statistical Society: Series B, 1958, 20(2): 215-242

[36] Drucker H, Burges C J C, Kaufman L, et al. Support vector regression machines [C] //Proc of the 26th Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 1997: 155-161

[37] Breiman L. Random forests [J]. Machine Learning, 2001, 45(1): 5-32

Tong Wei, born in 1984. PhD candidate. His main research interests include statistics, data analysis in education database.

Wang Fei, born in 1997. Master candidate. His main research interests include data mining and deep learning.

Liu Qi, born in 1986. PhD, associate professor. His main research interests include data mining and knowledge discovery in database, machine learning method and application.

Chen Enhong, born in 1968. PhD, professor and PhD supervisor. His main research interests include data mining and machine learning, social network analysis, and recommender systems.