DOI:10.7544/issn1000-1239.2020.20200181
57(8): - , 2020

基于双路注意力机制的学生成绩预测模型

李梦莹[1] 王晓东[1] 阮书岚[2] 张 琨[2] 刘 淇[2]
[1](河南师范大学计算机与信息工程学院 河南新乡 453000)
[2](中国科学技术大学计算机科学与技术学院 合肥 230027)
(lmengying@yeah.net)

Student Performance Prediction Model Based on Two-Way Attention Mechanism

Li Mengying[1], Wang Xiaodong[1], Ruan Shulan[2], Zhang Kun[2], and Liu Qi[2]
[1](*College of Computer and Information Engineering, Henan Normal University, Xinxiang, Henan* 453000)
[2](*College of Computer Science and Technology, University of Science and Technology of China, Hefei* 230027)

Abstract The prediction and analysis of student performance aims to achieve personalized guidance to students, improve students' performance and teachers' teaching effectiveness. Student performance is affected by many factors such as family environment, learning conditions and personal performance. The traditional performance prediction methods either treat all the factors equally, or treat all students equally, which cannot achieve personalized analysis and guidance for students. Therefore, we propose a two-way attention (TWA) based students' performance prediction model, which can assign different weights to different influence factors, and pay more attention to the important ones. Besides, we also take the individual features of students into account. Firstly, we calculate the attention scores of the attributes on the first-stage performance and the second-stage performance. Then we consider a variety of feature fusion approaches. Finally, we made better predictions of student performance based on the integrated features. We conduct extensive experiments on two public education datasets, and visualize the prediction results. The result shows that the proposed model can predict student performance accurately and have good interpretability.

Key words performance prediction; attention mechanism; attribute characteristics; feature fusion; personalized analysis

摘 要 学生成绩的预测与分析旨在实现对学生的个性化指导,提升学生成绩及教师的教学成果.学生成绩受家庭环境、学习条件以及个人表现等多种因素的影响.传统的成绩预测方法往往忽视了不同因素对同一学生成绩的影响程度不同,而且不同学生受同一因素的影响程度也不同,所构建的模型无法实现对学生的个性化分析与指导.因此提出一种基于双路注意力机制的学生成绩预测模型(two-way attention, TWA),该方法不仅有区别地对待了这些因素对成绩的影响程度,而且考虑到了学生的个体差异性.该方法通过两次注意力计算分别得到各属性特征在第1阶段成绩和第2阶段成绩上的注意力得分,并考虑了多种特征融合方式,最后基于融合后的特征对期末成绩进行更好地预测.分别在2个公开数据集上对模型进行了验证,并根据各属性特征在期末成绩上的概率分布对预测结果进行可视化分析.结果显示,所构建模型能够更准确地预测出学生成绩,并且具有良好的可解释性.

关键词 成绩预测;注意力机制;属性特征;特征融合;个性化分析

中图法分类号 TP399

收稿日期:2020-03-19;**修回日期**:2020-05-13

通信作者:王晓东(wxd@htu.edu.cn)

教育数据挖掘旨在从海量的教育数据中发现隐藏在其中的内在联系与规律,为学生学习、教师教学以及教育管理者的管理提供一些帮助[1].作为教育数据挖掘领域的一个重要研究分支,学生成绩预测有助于教师对学生的学习过程进行及时有效的干预和指导,例如识别出有风险的学生,以便及时提供干预措施[2].此外,还可用于在线测评[3]、认知诊断[4]、学生画像构建[5]和推荐系统[6],具有重要的研究意义与应用价值.

目前,对学生成绩进行预测分析及其成绩关键影响因素挖掘研究已引起国内外学者的关注.在学生成绩预测方面,蒋卓轩等人[7]通过从 MOOC 学习者的诸多行为特征中选择出若干典型学习行为特征,并利用所选择的特征对学习者能否成功完成学习任务获得通过证书进行预测,从中找出潜在的认真学习者.Pandey 等人[8]在影响学生成绩的 18 个属性特征中通过计算各个属性特征的信息增益率挑选出 8 个重要属性,并利用所挑选的 8 个重要属性构建决策树对学生成绩进行预测.在学生成绩影响因素挖掘方面,Bhardwaj 等人[9]通过对印度某大学 300 名学生成绩进行研究发现,学生成绩受家庭住址、家庭年收入、母亲受教育情况、生活习惯及学生历史成绩等因素影响比较大.Thiele 等人[10]提出学生的社会人口学特征(如种族、性别和经济地位)和学业特征(如学校类型和在校表现)与他们的学业表现联系紧密.

虽然以上工作已经取得了比较好的表现,但仍然存在 2 个方面问题:1)当前工作仅考虑已挑选的特征对学生成绩的影响,而忽略了未挑选特征的影响.例如 Pandey 在构建决策树对学生成绩进行预测时仅选用信息增益率较高的 8 个特征,而忽略了剩余 10 个特征对学生成绩的影响.2)当前工作假设关键因素对所有学生的影响程度是相同的,忽略了学生的个体差异.事实上不同因素对同一学生成绩的影响程度是不同的,并且不同学生受同一因素的影响程度也是不同的.如何更全面准确地分析利用这些属性特征对学生成绩进行预测,同时挖掘出影响不同成绩学生的关键因素,实现对学生的个性化分析与指导是目前学生成绩预测研究所面临的一项重大挑战.

为了解决以上挑战,文本提出了一种基于双路注意力机制的学生成绩预测方法(two-way attention, TWA).该方法通过双路注意力机制为不同的属性特征赋予不同的注意力权重,实现了学生属性更全面准确的利用,进而保证了学生成绩的准确预测.具体而言,首先,TWA 模型通过两次注意力计算分别得到各属性特征在第 1 阶段成绩和第 2 阶段成绩上的注意力得分.然后,在此基础上进行双路特征融合并对期末成绩进行预测.最后在 2 个公开数据集上进行大量实验,实验结果证明了本文所提出方法的有效性.

本文的主要贡献有 3 个方面:

1) 通过对数据进行挖掘分析发现学生个体之间存在差异性,不同成绩类别学生所受关键影响因素不同,并且学生期末成绩与第 1 阶段和第 2 阶段历史成绩有很大关联;

2) 提出的双路注意力机制可以让模型充分学习各属性特征与成绩间的关系信息,有效标识不同属性特征对成绩的重要程度,同时可以弥补普通注意力机制的不足,提升模型的预测能力;

3) 在 2 个公开数据集上验证了模型的准确性和有效性,同时模型也具有良好的可解释性.

1 相关工作

本节从学生成绩预测分析、注意力机制研究 2 个方面介绍相关工作.

1.1 学生成绩预测分析

学生成绩与学生的表现以及所处的环境息息相关,利用教育数据挖掘技术在诸多潜在影响因素中挖掘出影响不同学生成绩的关键因素,并对学生成绩做出早期准确预测,对于实现学生的个性化指导以及提升教学成果具有重大意义.在过去的研究中,学生成绩预测分析方法主要是基于统计和机器学习的传统方法.Zhang 等人[11]利用学生历史成绩和在校行为信息,运用朴素贝叶斯、决策树、多层感知器和支持向量机等预测模型分别对学生成绩进行预测,发现多层感知器模型的预测效果更好.Mueen 等人[12]根据学生的历史学习成绩和论坛参与度,运用朴素贝叶斯、神经网络和决策树等数据挖掘技术来预测学生的学习成绩,结果显示朴素贝叶斯模型在此数据集上效果最好,预测准确度可达 86%.Francis 等人[13]将影响学生成绩的特征因素划分为人口统计特征、学术特征、行为特征以及额外特征 4 类,并提出一种将分类与聚类相结合的方法对成绩进行预测,结果显示综合考虑学术特征、行为特征以及额外特征时得到的预测结果最好.谢娟英等人[14]通过对葡萄牙学生数据挖掘发现,学生的成绩

与学生所在学校、家庭住址、母亲学历、家庭有无网络有极大相关性,与父亲受教育程度、上学路上花费时间、想上大学、是否恋爱也具有一定的相关性.

以上研究在建模的过程中要么平等对待了所有因素对学生成绩的影响程度,要么平等对待了不同成绩层次的学生,并将所挖掘出的关键影响因素视为对所有学生的影响程度相同,忽略了学生的个体差异性,所构建的预测模型解释性不强,且无法实现对学生的个性化分析与指导.

1.2 注意力机制

注意力机制最早应用于图像处理领域[15],旨在使模型在训练的过程中能够高度关注指定的目标.注意力机制主要是模拟人的注意力[16],可以用人类生物系统来解释[17].例如我们的视觉处理系统会根据我们的需求有选择地聚焦于图像中我们所感兴趣的部分,而忽略其他不相关的信息,从而有助于我们感知到一些关键信息.深度学习中的注意力机制核心思想就是从众多信息中选取对于当前任务目标更为关键的信息,根据其重要度不同对其赋予不同的权重.近年来,注意力机制在图像处理[18]、自然语言处理[19]、语音识别[20]等领域广泛应用.黄友文等人[21]提出了一种基于卷积注意力机制和长短期记忆网络的图像描述生成模型,解决了现有的基于卷积神经网络和循环神经网络搭建的图像描述模型在提取图像关键信息时精度不高而且训练速度缓慢等问题.Cheng 等人[22]将注意力机制应用到机器翻译任务中,并提出了全局注意力和局部注意力 2 种机制,奠定了注意力机制在自然语言处理中的应用基础.于重重等人[23]提出一种基于注意力机制的检索式匹配问答方法,针对输入的中文词向量信息建立实体关注层模型并采用注意力机制算法,很好地解决了检索式匹配问答模型对中文语料适应性弱和句子语义信息被忽略的问题.注意力机制在以上领域中的成功应用也为其在教育数据挖掘领域的研究提供了新的思路.

2 基于双路注意力机制的成绩预测模型

2.1 问题定义

给定一个学生特征集合 M,M 可由一系列的属性特征 $attributes=\{x_1,x_2,\cdots,x_n\}$ 以及第 1 阶段历史成绩 G_1 及第 2 阶段历史成绩 G_2 表示,即 $M=\{attributes,G_1,G_2\}$,其中 n 为属性特征的数量.对于该学生,他的期末成绩为 y_i,$y=\{y_1,y_2,\cdots,y_c\}$ 为学生成绩所划分的类别集合.学生成绩预测任务的目标就是根据给定的学生特征 M,判定 M 的成绩类别 y_i.

2.2 模型结构

本文旨在通过对教育数据的分析和挖掘,实现对学生期末成绩的准确预测,并找出影响不同成绩类别学生的关键因素,对学生进行个性化分析和指导.通过对数据进行统计分析,结果显示学生前 2 个阶段的历史成绩和期末成绩的关联性很大.结合学生属性特征及两阶段历史成绩提出一种基于双路注意力的成绩预测模型(two-way attention, TWA),模型框架如图 1 所示,TWA 共包含 3 层:

1) 输入编码层.首先对各属性值及历史成绩进行预处理,包括数值转换、归一化、分组等.在此基础上,将离散的属性值映射到高维的特征空间,生成各属性的特征表示和历史成绩的特征表示.

2) 双路注意力层.根据学生的各属性特征和两阶段历史成绩特征,进行双路注意力机制计算.分别得到各属性基于第 1 阶段成绩的注意力得分 $\boldsymbol{\beta}$ 以及基于第 2 阶段成绩的注意力得分 $\boldsymbol{\gamma}$,利用两路注意力得分进行属性特征加权求和,得到能体现重要性程度的学生属性特征 $\boldsymbol{f}_1$ 与 $\boldsymbol{f}_2$.

3) 标签预测层.对学生属性特征 $\boldsymbol{f}_1$ 与 $\boldsymbol{f}_2$ 进行特征融合,得到具有更丰富信息的最终特征 $\boldsymbol{f}$,从而进行更好的学生成绩预测.

2.2.1 输入编码层(input embedding layer)

输入编码层主要是对各属性值以及历史成绩进行预处理,并将其转成向量表示.具体包括:

1) 数据预处理.为了便于模型的处理,根据属性特征值的特点对数据进行预处理操作,包括对二元数据进行数字编码转换、数值归一化、对成绩进行分组等.年龄属性和缺课次数相对于其他属性取值比较大,为避免其对实验结果的干扰,采用式(1) min-max 归一化方法将年龄和缺课次数标准化到 $[0,1]$.

$$x^*=\frac{x-x_{\min}}{x_{\max}-x_{\min}},\tag{1}$$

其中,x^* 是年龄属性或缺课次数标准化后的属性取值,x 为年龄属性或缺课次数的原始取值,$x_{\min}$ 为样本数据的最小值,$x_{\max}$ 为样本数据的最大值.

2) 离散值向量化.为实现对离散的属性特征值以及历史成绩的初始编码,本文随机生成一个高维的参数矩阵,将离散的属性值与两阶段成绩值经过随机化生成的参数矩阵后转化成属性矩阵 $\boldsymbol{A}\in$

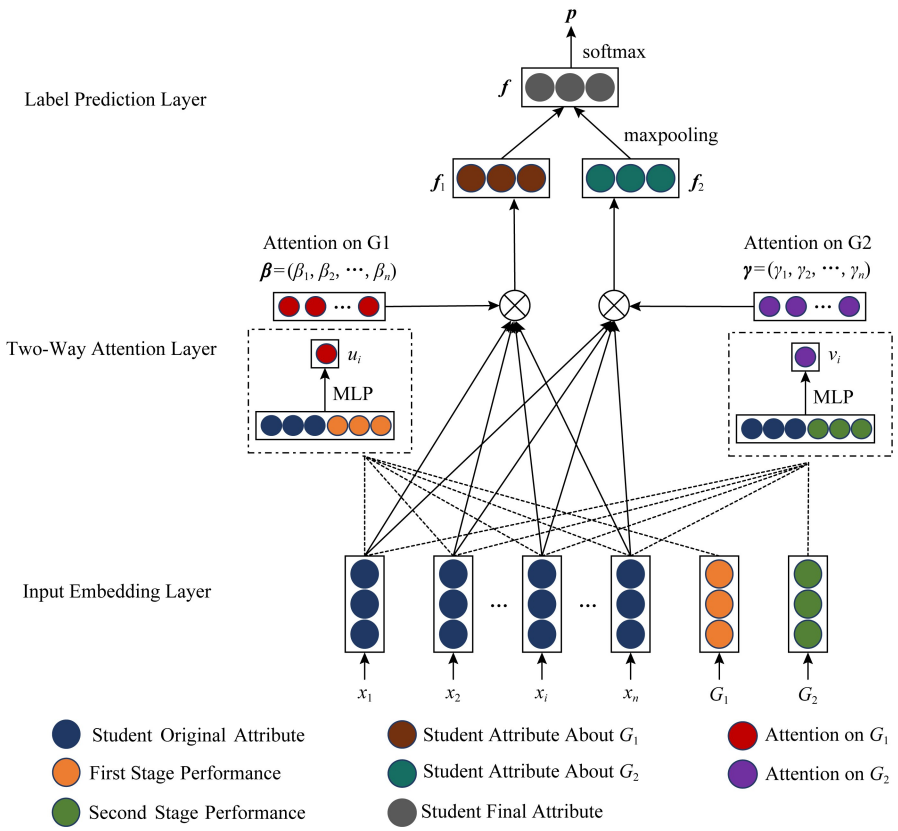

Fig. 1 Framework of TWA

图 1 Two-Way Attention 模型框架图

$\mathbb{R}^{k\times n}$ 以及成绩向量 $\boldsymbol{g}_1\in\mathbb{R}^k$ 和 $\boldsymbol{g}_2\in\mathbb{R}^k$，其中 n 为属性特征数量，k 为各属性特征及历史成绩向量维度.

$$\boldsymbol{A}=\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{k1} & a_{k2} & \cdots & a_{kn} \end{bmatrix}, \tag{2}$$

$$\boldsymbol{g}_1=(\boldsymbol{g}_{11},\boldsymbol{g}_{12},\cdots,\boldsymbol{g}_{1k})^{\mathrm{T}}, \tag{3}$$

$$\boldsymbol{g}_2=(\boldsymbol{g}_{21},\boldsymbol{g}_{22},\cdots,\boldsymbol{g}_{2k})^{\mathrm{T}}. \tag{4}$$

2.2.2 双路注意力层(two-way attention layer)

通过对各属性特征和历史成绩进行编码后，可以得到特征矩阵 $\boldsymbol{A}$ 和成绩向量 $\boldsymbol{g}_1$ 与 $\boldsymbol{g}_2$.考虑到不同属性特征对成绩影响程度不同，因此所构建的学生成绩预测模型需要学习不同属性特征对于成绩预测结果所起的关键性作用.同时，考虑到学生期末成绩与前两阶段历史成绩有很强的关联性，模型应能够自动挖掘期末成绩和历史成绩之间的内在联系，从而进一步增强对学生期末成绩预测的能力.因此，本文设计了双路注意力机制，对学生期末成绩的重要影响因素进行了建模.

双路注意力层的目标是根据各属性特征与前两阶段历史成绩间的关系信息，分别利用注意力机制为各属性特征分配合适的注意力权重，从而解决了不同因素对学生成绩的影响程度不同以及不同成绩学生所受的关键影响因素也不同的问题.通过利用双路注意力机制，可以实现更全面准确地利用这些属性特征对学生成绩进行预测.具体来说，本文采用多层感知器(multi-layer perceptron, MLP)操作来进行注意力权重计算，视特征矩阵 $\boldsymbol{A}$ 中的每一列向量 $\boldsymbol{A}_i$ 为对应位置属性特征向量，即 $\boldsymbol{A}_i=(a_{1i},a_{2i},\cdots,a_{ki})^{\mathrm{T}}$，基于第 1 阶段历史成绩向量 $\boldsymbol{g}_1$，可以得到任意属性特征 $\boldsymbol{A}_i$ 的注意力权重 u_i，具体计算过程为

$$u_i=MLP([\boldsymbol{g}_1;\boldsymbol{A}_i]),i=1,2,\cdots,n. \tag{5}$$

同理，基于第 2 阶段历史成绩向量 $\boldsymbol{g}_2$，可以得到属性特征向量 $\boldsymbol{A}_i$ 的对应注意力权重 v_i：

$$v_i=MLP([\boldsymbol{g}_2;\boldsymbol{A}_i]),i=1,2,\cdots,n, \tag{6}$$

无特殊说明全文中[.;.]皆表示特征拼接操作.

用 softmax 函数对所得权重进行归一化处理，分别得到各属性特征在第 1 阶段历史成绩上的注意力得分 $\boldsymbol{\beta}=(\beta_1,\beta_2\cdots,\beta_n)$ 以及第 2 阶段历史成绩上的注意力得分 $\boldsymbol{\gamma}=(\gamma_1,\gamma_2,\cdots,\gamma_n)$，该过程可形式化表示为

$$\boldsymbol{\beta}_i=\mathrm{softmax}(u_i)=\frac{\exp(u_i)}{\sum_{i=1}^{n}\exp(u_i)},\tag{7}$$

$$\boldsymbol{\gamma}_i=\mathrm{softmax}(v_i)=\frac{\exp(v_i)}{\sum_{i=1}^{n}\exp(v_i)},\tag{8}$$

其中,$\boldsymbol{\beta}_i$ 指第 i 个属性特征在第 1 阶段历史成绩上的注意力得分,$\boldsymbol{\gamma}_i$ 指第 i 个属性特征在第 2 阶段历史成绩上的注意力得分.

将注意力得分向量 $\boldsymbol{\beta}$ 和 $\boldsymbol{\gamma}$ 分别与特征矩阵 $\boldsymbol{A}$ 中对应位置的属性特征值进行加权求和,得到基于第 1 阶段成绩的学生属性特征 $\boldsymbol{f}_1$ 以及基于第 2 阶段成绩的学生属性特征 $\boldsymbol{f}_2$,具体计算过程为

$$\boldsymbol{f}_1=\sum_{i=1}^{n}\boldsymbol{\beta}_i\boldsymbol{A}_i,\tag{9}$$

$$\boldsymbol{f}_2=\sum_{i=1}^{n}\boldsymbol{\gamma}_i\boldsymbol{A}_i.\tag{10}$$

2.2.3 标签预测层(label prediction layer)

标签预测层的主要任务是根据在双路注意力层所得到的基于第 1 阶段历史成绩的学生属性特征 $\boldsymbol{f}_1$ 以及基于第 2 阶段历史成绩的学生属性特征 $\boldsymbol{f}_2$ 预测目标学生的成绩类别.鉴于 $\boldsymbol{f}_1$ 与 $\boldsymbol{f}_2$ 的信息互补性,首先对两者进行特征融合,以便更全面准确地利用这些属性特征对学生成绩进行预测.特别地,本文考虑了 3 种特征融合方式,分别是 maxpooling,avgpooling 和 concatenation.

以 maxpooling 方式进行特征融合时,取相应位置最大值,该过程可以形式化表示为

$$\boldsymbol{f}=\max(\boldsymbol{f}_{1i},\boldsymbol{f}_{2i}),i=1,2,\cdots,k.\tag{11}$$

以 avgpooling 方式进行特征融合时,对各属性特征对应位置的 2 个值求平均,该过程可以形式化表示为

$$\boldsymbol{f}=\frac{1}{2}(\boldsymbol{f}_{1i}+\boldsymbol{f}_{2i}),i=1,2,\cdots,k.\tag{12}$$

以 concatenation 方式进行特征融合时,直接将学生属性特征 $\boldsymbol{f}_1$ 与 $\boldsymbol{f}_2$ 进行拼接,该过程可以形式化表示为

$$\boldsymbol{f}=[\boldsymbol{f}_1;\boldsymbol{f}_2],\tag{13}$$

式(11)~(13)中 $\boldsymbol{f}_{1i}$ 指学生属性特征 $\boldsymbol{f}_1$ 中第 i 个元素,$\boldsymbol{f}_{2i}$ 指学生属性特征 $\boldsymbol{f}_2$ 中第 i 个元素,$\boldsymbol{f}$ 为 $\boldsymbol{f}_1$ 与 $\boldsymbol{f}_2$ 进行特征融合后输入分类层的最终特征.

随后将融合后的特征 $\boldsymbol{f}$ 输入 MLP 中得到分类结果.本文所使用的是一个 3 层全连接网络,在 2 个隐含层中使用 ReLU 激活函数,输出层使用 softmax 函数得到各成绩类别的分类预测得分 $\boldsymbol{p}$.

$$\boldsymbol{p}=MLP(\boldsymbol{f}).\tag{14}$$

2.2.4 模型训练

本文使用反向传播算法来训练网络模型,用交叉熵作为分类损失,通过迭代求解损失值和随机梯度下降来优化模型,使得损失函数的值收敛到最小.考虑到模型的复杂性,避免模型在训练的过程中出现过拟合,本文引用了 L2 正则项对参数进行约束,故模型的最终损失函数为

$$L=-\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{y}_i\log\boldsymbol{p}_i+\lambda\|\theta\|^2,\tag{15}$$

其中,N 为训练集数据量,$\boldsymbol{y}_i$ 为第 i 个学生样本的标签,$\boldsymbol{p}_i$ 为第 i 个学生样本的预测概率,$\lambda\|\theta\|^2$ 为 L2 正则项,θ 为模型的所有参数集合.

3 实 验

3.1 数据集

本文在 student performance[24] 中的葡萄牙语成绩数据集(portuguese)以及数学成绩数据集(math)中展开实验.其中葡萄牙语成绩数据集中的有效数据为 649 条,数学成绩数据集中有效数据 357 条.2 个数据集都包含有 30 个维度属性特征信息,前两阶段历史成绩 G_1 和 G_2 以及期末成绩 G_3,涉及 13 种二元数据,4 种标称数据以及 16 种数值数据.其中 30 维属性特征信息以及前两阶段历史成绩作为输入,期末成绩类别为最终输出目标,关于数据集描述如表 1 所示.(其中编号 1~33 分别指学校、性别、年龄、…、上课缺席次数、第 1 阶段历史成绩、第 2 阶段历史成绩和期末成绩.)

Table 1 The Description of student performance Dataset

表 1 student performance 数据集描述

ID	Attributes	Characteristics
1	School	GP=0,MS=1
2	Sex	F=0,M=1
3	Age	[15,20]
⋮	⋮	⋮
30	Absences	Number of school absences (from 0 to 93)
31	G_1	First period grade (from 0 to 20)
32	G_2	Second period grade (from 0 to 20)
33	G_3	Final grade (from 0 to 20)(output target)

针对上述各属性特征包含信息的差异性,本文进一步对其进行类别划分.如包含学生性别在内的学生基本信息、家庭主要监护人以及父母教育水平等家庭因素信息、学生的社交及消费情况、学生学习

地址以及是否使用网络在内的学习条件信息等,具体划分结果如表 2 所示.

本文也探究了学生期末成绩与历史成绩之间的关联关系,结果如表 3 所示.通过对数据集进行数据统计分析,发现学生期末成绩 G_3 与前 2 个阶段的历史成绩 G_1 和 G_2 具有很强的相关性.在葡萄牙语成绩数据集(portuguese)中,学生期末成绩与第 1 阶段历史成绩保持一致的占 67.35%,与第 2 阶段历史成绩保持一致的占 75.08%,与第 1 阶段历史成绩和第 2 阶段历史成绩至少有一个保持一致的占 85.65%.由此可见,在葡萄牙语成绩数据集中,学生期末成绩与前 2 个阶段的历史成绩具有很强的关联性.对于数学成绩数据集(math)也可以得到同样的结论.

Table 2 Category of Each Attribute

表 2 各属性特征分类

Attribute Types	The Specific Attributes
Basic Information	Sex, Age, Health
Family Factors	Famsize, Guardian, Pstatus, Famrel, Famsup, Mjob, Fjob, Medu, Fedu
Student Performance	Studytime, Freetime, Romantic, Failures, Higher, Dalc, Walc, Reason, Absences, Goout, Activities
Learning Conditions	School, Address, Internet, Schoolsup, Traveltime, Nursery, Paid

Table 3 Statistical Results on student performance Dataset

表 3 student performance 数据集数据统计结果

Dataset	Total Number	$G_3=G_1$	$G_3=G_2$	$G_3=G_1$ or $G_3=G_2$
portuguese	634	427	476	543
math	357	232	296	318

本文所有实验均按照 8:2比例划分成训练集和测试集,每次实验用训练集训练模型并选择最优参数,用测试集计算各项指标.

3.2 数据预处理

3.2.1 成绩分组

在本文所选用的数据集中,G_1,G_2 和 G_3 分别表示第 1 阶段历史成绩,第 2 阶段历史成绩和期末成绩,并且都是一种 0～20 的数值数据.由于本文中样本数量的限制,通过对各个成绩上的样本数量进行统计分析发现,在某些成绩上的样本数量分布过少.通过观察分析成绩的数据分布并结合目前常用的成绩等级划分方法,将学生成绩划分为 A,B,C,D 这 4 个组别,用来区分不同的学生个体.其中,A 组优秀:16～20 分;B 组良好:13～15 分;C 组中等:10～12 分;D 组不及格:<10 分.分别对 2 个数据集中期末成绩分布情况进行统计,统计结果如表 4 所示:

Table 4 The Statistical Results of Final Grade Distribution

表 4 期末成绩分布统计结果

Dataset	Group A	Group B	Group C	Group D
portuguese	82	194	273	85
math	40	91	134	92

3.2.2 异常数据处理

通过对数据进行统计分析得知,2 个数据集中的数据均没有缺失值,但存在学生期末成绩为 0 的情况.对此类数据进行进一步分析发现,当该生期末成绩为 0 时,其缺席次数并不高而且前 2 个阶段历史成绩也均处于正常水平,故将此种情况视为该生未参加期末考试,对其结果的预测也失去意义.因此,将期末成绩为 0 的数据视为异常数据并对其进行删除处理,保留剩余的 634 条葡萄牙语成绩数据和 357 条数学成绩数据,进行更进一步的数据挖掘分析.

3.2.3 评价指标

本文中除了预测准确率(*Accuracy*)外,还采用精确率(*Precision*)、召回率(*Recall*)和 *F1-Measure* 进行模型分类预测性能度量.*Accuracy* 表示的是正确分类的样本个数占整个样本的比例,准确率越高表明预测越准确.*Precision* 表示正确分类的正例个数占预测为正例总数的比例.*Recall* 表示正确分类的正例个数占预测为正数的比例.*F1-Measure* 是 *Precision* 和 *Recall* 的折中,*F1-Measure* 值越高,分类效果越好[25].

$$Precision=\frac{TP}{TP+FP}, \tag{16}$$

$$Recall=\frac{TP}{TP+FN}, \tag{17}$$

$$F1\text{-}Measure=\frac{2\times Precision\times Recall}{Precision+Recall}, \tag{18}$$

其中,TP 表示真实标签为正例也被正确判定为正例;FP 表示真实标签为负例但是被错误地判定为

正例;FN 表示真实标签为正例但未被正确地判定为正例;TN 表示真实标签为负例的未被判定为正例.

3.3 实验参数设置

本文所提模型基于深度学习框架 PyTorch 展开实验,优化器为随机梯度下降 SGD,其中 $batchsize=16$,初始的学习率为 0.01,模型迭代次数 $epoch=2\,000$,属性类别数为 30,初始化属性特征维度为 128 维,即在式(2)中 $n=30$,$k=128$.式(5)(6)中所用的是两层全连接网络,拼接特征进行 256,32,1 的特征维度变换.对于式(14)中的 3 层全连接网络,当采用 concatenation 的方式进行特征融合时,特征进行 256,128,64,4 的特征维度变换,当采用 maxpooling 或者 avgpooling 时,进行 128,128,64,4 的特征维度变换,上述各全连接网络隐含层之间采用的激活函数为 ReLU.

3.4 实验结果与分析

3.4.1 对比实验

将本文所提出的基于双路注意力机制的学生成绩预测方法(TWA)同支持向量机[26](support vector machine, SVM)、逻辑回归[27](logistic regression, LR)、高斯朴素贝叶斯[28](gaussion naive bayes, GaussionNB)、决策树[29](decision tree, DT)等 4 种传统的分类预测方法分别在 student performance 中的葡萄牙语成绩以及数学成绩这 2 个公开教育数据集中进行对比实验,验证本文提出方法的有效性.实验结果如表 5 和表 6 所示.

Table 5 Performance on Portuguese

表 5 葡萄牙语数据集上的预测结果 %

Classifier	*Accuracy*	*Precision*	*Recall*	*F1-Measure*
GaussionNB	53.54	51.20	59.22	54.92
LR	72.44	70.33	68.22	69.26
SVM	88.98	90.35	85.11	87.65
DecisionTree	90.55	88.73	90.62	89.67
Our TWA	**96.06**	**95.51**	**95.20**	**95.36**

Note: The best performance is in bold.

Table 6 Performance on Math

表 6 数学数据集上的预测结果 %

Classifier	*Accuracy*	*Precision*	*Recall*	*F1-Measure*
GaussionNB	61.97	47.51	58.91	52.60
LR	74.65	76.62	74.01	75.29
SVM	80.28	81.91	77.60	79.69
DecisionTree	92.96	93.23	92.90	93.06
Our TWA	**95.77**	**95.27**	**96.49**	**95.88**

Note: The best performance is in bold.

从表 5 和表 6 的实验结果可以看出,相比其他 4 种传统的成绩预测方法(GaussionNB,LR,SVM 和 DecisionTree),本文基于 two-way attention 的方法在 2 个公开教育数据集上均取得了最好的预测效果.在葡萄牙语成绩数据集和数学成绩数据集上的预测准确率可分别达到 96.06%和 95.77%,相比于最好的传统方法 Decision Tree 分别提升了 5.51%和2.81%.此外,在查准率(*Precision*)、查全率(*Recall*)以及 *F1-Measure* 这 3 个指标上也均有显著性的提高.

对比实验中的 4 种传统机器学习方法预测准确率普遍不高,分析其原因可能为:传统方法没有针对特定的成绩目标提取更多的特征信息,而是将各属性特征直接作为分类特征输入模型进行学习训练,平等地对待了各属性特征对期末成绩的影响程度.而本文引入注意力机制,可以有区别性地对待各属性特征的重要性.除此之外,与普通注意力机制不同的是,鉴于期末成绩与前两阶段历史成绩的强关联性,在模型中通过设计双路注意力来挖掘出更多的隐藏信息并进行信息互补,有效弥补普通注意力机制的不足,从而大大提升了模型的预测能力,取得了较好的预测效果,实验结果也证明了本文所提方法的有效性.

3.4.2 双路注意力机制的消融研究

为了进一步验证本文所提方法的有效性,在葡萄牙语成绩数据集上进行了双路注意力机制的消融研究,所有实验均在融合方式为 maxpooling 下进行.实验结果如表 7 所示,其中 No_attention 指的是模型完全不采用 attention 机制而平等对待所有属性特征,即图 1 中去掉 Two-Way Attention Layer 模块,将 Input Embedding Layer 之后的属性特征求平均,再与成绩向量 $\boldsymbol{g}_1$ 和 $\boldsymbol{g}_2$ 求和,得到输入分类预测层的特征 $\boldsymbol{f}$.G1_attention 与 G2_attention 分别指的是单路 attention 的结果,即图 1 中变为 One-Way Attention 而不考虑另一路的影响.可以得出:

1) 相较于 G1_attention 和 G2_attention 这样的单路注意力机制以及 TWA 这样的双路注意力机制,No_attention 的结果有了大幅度的下降,其中准确率相较 TWA 下降 6.3%,这说明注意力机制能有效地学习不同属性的相对重要性,相较于平等对待所有属性特征的影响,能更好地提升模型的预测能力.

2) 通过比较 G1_attention 与 G2_attention,可以发现后者预测能力更强,这说明近期的历史成绩

与期末成绩更相关,因此对期末成绩预测更具备参考性,该现象也与表 3 所示的统计规律相一致.

3) 相较于单路注意力机制,本文所提的双路注意力机制在各项评价指标上都有明显的提升,就准确率而言,TWA 相较 G1_attention 提升 3.93%,相较 G2_attention 提升 1.57%,这说明双路具有一定的信息互补性,当进行双路融合时,能进一步提升模型的预测性能.

Table 7 Ablation Study on Two-Way Attention Mechanism

表 7 双路注意力机制的消融研究 %

Methods	*Accuracy*	*Precision*	*Recall*	*F1-Measure*
No_attention	89.76	87.51	87.86	87.69
G1_attention	92.13	91.84	90.24	91.03
G2_attention	94.49	93.56	93.10	93.33
Our TWA	**96.06**	**95.51**	**95.20**	**95.36**

Note: The best performance is in bold.

3.4.3 双路特征融合实验

本文在葡萄牙语数据集上进行了模型变种实验,比较不同特征融合方式(avgpooling, concatenation 和 maxpooling)对模型实验结果的影响.实验结果如表 8 所示:可以看出,按照 maxpooling 方式进行特征融合时的效果最好,预测准确率可达到 96.06%.按照 concatenation 方式进行特征融合的效果次之,其预测准确率为 95.28%,但较 maxpooling 融合方式下降了 0.78%.按照 avgpooling 方式进行特征融合的预测准确率为93.70%,较 maxpooling 融合方式下降了 2.36%.

Table 8 Results of Different Feature Fusion Ways on Portuguese

表 8 葡萄牙语中不同特征融合方式下的预测结果 %

Methods	*Accuracy*	*Precision*	*Recall*	*F1-Measure*
avgpooling	93.70	93.37	94.05	93.71
concatenation	95.28	93.54	**95.67**	94.60
maxpooling	**96.06**	**95.51**	95.20	**95.36**

Note: The best performance is in bold.

由于特征融合方式的不同使得预测结果有些差异,分析其原因可能为:按照 maxpooling 方式进行特征融合时,各属性特征中的更高注意力得分被保留,使得各属性特征所表征的信息更加准确和全面,其预测效果最好.按照 concatenation 方式进行融合后的特征涵盖了特征融合前的所有信息,其预测效果次之.而按照 avgpooling 方式进行特征融合时对各属性特征所对应的注意力得分取平均,可能导致某些关键属性特征的显著影响力下降,使得该方式下的预测效果相对较差.此外,数据集的数据量也可能会对实验结果造成一些影响.

3.5 可视化分析

为了挖掘出影响不同成绩类别学生的具体因素,实现对学生的个性化指导,我们对不同成绩类别中的各属性特征进行了注意力结果可视化,更直观地显示出每个特征对成绩预测结果的影响.通过对不同成绩类别的学生进行分析来反映学生个体差异的情况.本文对学生葡萄牙语期末成绩按照不同成绩分组进行各属性特征的注意力分布可视化分析.以 avgpooling 方式进行特征融合为例,可视化结果如图 2 所示.其中横坐标表示属性特征编号且与表 1 数据集描述中的编号保持一致,纵坐标表示各属性特征对应的注意力权重,图中虚线表示注意力权重值为 0.1.由于属性特征的数量较多,根据各属性特征的概率分布情况,本文将注意力权重大于 0.1 的属性特征视为学生期末成绩的关键影响因素.

分析图 2 可知,对于期末成绩类别为 A 的学生而言,属性特征 24(家庭关系)和 8(父亲受教育程度)是影响他们成绩的关键因素,其中家庭关系的影响最为显著,其所占比重已经超出 50%,说明良好的家庭关系是取得优异成绩的关键.其次,属性特征 26(和朋友外出次数)以及 29(自身的健康状况)也对该类别的学生成绩有一定的影响.

对于期末成绩类别为 B 的学生而言,其成绩所受的影响因素种类比较多.其中,属性特征 14(一周内学习时长)以及 19(是否参加课外活动)是影响他们成绩的关键因素,说明学习时间的投入以及适当的课外活动对他们是很有必要的.此外,属性特征 7(母亲受教育程度)、8(父亲受教育程度)、11(选择学校原因)、12(监护人)、13(上学路上花费时间)、16(学校对教育的额外支持)、18(是否补课)、24(家庭关系)、25(课余时间)、26(和朋友外出次数)、27(工作日是否饮酒)、28(周末是否饮酒)、29(自身健康状况)等也对其成绩有不同程度的影响.

对于期末成绩类别为 C 的学生而言,属性特征 7(母亲受教育程度)、10(父亲工作)、25(课余时间长短)、以及 29(自身健康状况)是影响他们成绩的关键因素.其次,属性特征 8(父亲受教育程度)、9(母亲工作)、11(选择学校原因)、12(监护人)、13(上学路上花费时间)、14(一周内学习时长)、24(家庭关系)、

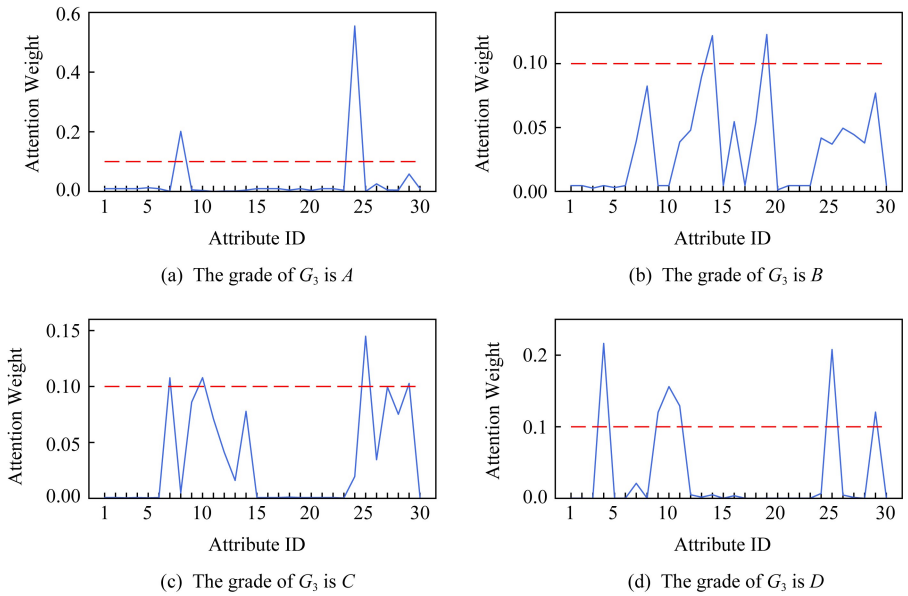

Fig. 2 The attribute probability distribution on each grade group

图 2 各成绩分组上的属性概率分布

26(和朋友外出次数)、27(工作日饮酒情况)、28(休息日饮酒情况)等对成绩也有一定的影响.

对于期末成绩类别为 D 的学生而言,属性特征 4(家庭住址)、9(母亲工作)、10(父亲工作)、11(选择学校原因)、25(课余时间)以及 29(自身的健康状况)均为影响期末成绩的关键因素,属性特征 7(母亲受教育程度)也对其成绩有些许影响.

通过对不同成绩类别学生所受的影响因素进行挖掘后可知,在 30 种属性特征中,有 15 种属性特征对学生期末成绩存在影响,其所属类别如表 9 所示:

Table 9 Category of the Key Attributes

表 9 重要属性特征分类

Attribute Types	The Specific Attributes
Family Factors	Medu, Fedu, Mjob, Fjob, Famrel, Guardian
Student Performance	Reason, Studytime, Activities, Freetime, Goout
Learning Conditions	Address, Traveltime, Paid
Basic Information	Health

对表 9 中各属性类别所包含的属性特征数量进行统计分析可知,在 15 种关键影响因素中,家庭因素类别中占有 6 种,所占比重可达 40%.学生表现类别占 33.3%,学习条件类别占 20%,基本信息类别占 6.7%.由此可见,家庭因素信息以及学生表现是影响学生成绩的重要因素,不容忽视.

为实现对学生的个性化分析与指导,本文对所挖掘出的对学生期末成绩存在影响的 15 种属性特征进行注意力得分可视化,可视化结果如图 3 所示.图 3 中横坐标表示各属性特征,纵坐标表示各属性特征所对应的注意力权重.

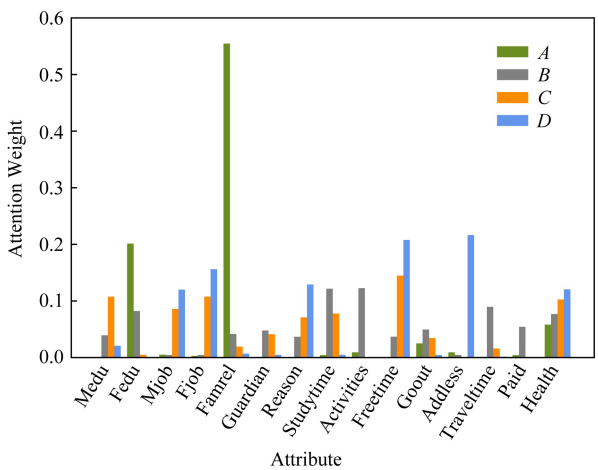

Fig. 3 The influence of each attribute on different grades

图 3 各属性特征对不同成绩的影响

通过对不同成绩类别学生进行个性化分析,从中发现的主要现象和结论为:

1) 从图 3 的结果可以观察到,在家庭因素中,成绩类别为 A 和 B 的学生受父亲所受教育程度的影响比较大,而成绩类别为 C 和 D 的学生受父母亲工作影响比较大.我们推测出现这样的结果是因为父亲受教育程度不同,其教育理念有很大差异,对

孩子学习有着直接的影响.父母作为孩子的第一任老师,由于工作原因对孩子陪伴及教育的缺失也会影响孩子成绩.家庭关系对成绩类别为 A 的学生影响尤其显著,对成绩类别为 B,C,D 的同学影响不明显.可能是因为 A 类学生已经具有丰富的知识储备和良好的学习习惯,家庭关系的好坏对其学习情绪和学习状态有直接的影响,相对于其他因素而言,该因素更为重要.

2) 在学生自身表现方面,学生选择学校的原因与学生成绩关联性比较大,成绩越差,其所占比重越大,说明学习动机对学生成绩有着直接影响.一周内学习时长对成绩类别为 B 和 C 的学生影响比较大,而对成绩类别为 A 和 D 的学生影响甚微,说明在学习上的时间投入是很有必要的.是否参加课外活动对成绩为 B 的学生影响较大,而对其他学生影响不明显.课后自由时间对成绩类别为 A 的学生几乎无影响,而对另外 3 种类别学生的影响程度呈现出上升趋势.我们推测可能对于成绩好的学生而言,课余时间对他们成绩的提升并不是特别重要,相比较而言,其他的行为特征属性可能更重要一点.

3) 就学习条件对不同成绩的影响而言,家庭住址对成绩类别为 D 的学生影响非常明显,我们推断出现这样的结果是因为家庭经济条件的影响.上学路上花费时间对成绩类别为 B 的学生影响较大,而对其他成绩类别的学生影响甚微.补课只对成绩类别为 B 的学生有较为显著的影响,这也说明补课并不是对所有学生都是必要的.

4) 此外,学生成绩越差,自身健康状况对其成绩的影响程度越大.我们推断出现这样的结果可能是因为身体不舒服而导致参加考试时不能够正常发挥.

4 总 结

学生成绩预测是近年教育数据挖掘领域的一个研究热点,也是进行学习分析的重要目标之一.本文针对目前相关研究中没有考虑到不同因素对同一学生成绩的影响程度不同,而且不同学生受同一因素的影响程度也不同等问题,提出了一种基于双路注意力机制的学生成绩预测模型.首先,该模型可以实现对离散属性特征变量输入的处理.其次,模型设计了双路注意力机制有效地学习不同属性特征的相对重要性并通过特征融合进行信息互补,使得模型预测能力更强.最后,在葡萄牙语成绩和数学成绩这 2 个公开教育数据集上的大量实验结果表明,本文所提出的基于双路注意力机制的学生成绩预测模型均取得了最好的预测效果,充分证明了模型在学生成绩预测问题上的有效性.

未来的研究工作中,可以对于不同特征之间的组合或者更高阶的特征对成绩预测结果的影响上进行更多地考虑和设计.

参 考 文 献

[1] Wang Sheng. Educational data mining promotes the analysis of personalized learning approaches for college students [J]. Exam Weekly, 2014, (34): 176-176 (in Chinese)
(王盛. 教育数据挖掘促进高校学生个性化学习途径分析[J]. 考试周刊, 2014, (34): 176-176)

[2] Bienkowski M, Feng M, Means B. Enhancing teaching and learning through educational data mining and learning analytics: An issue brief [R]. Washington: US Department of Education, Office of Educational Technology, 2012, 1: 1-57

[3] Lykourentzou I, Giannoukos I, Mpardis G, et al. Early and dynamic student achievement prediction in e-learning courses using neural networks [J]. Journal of the Association for Information Science and Technology, 2009, 60(2): 372-380

[4] 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 Press, 2015: 1017-1024

[5] Ren Changlin, Wang Xuelong. Research on students' performance portrait method based on multidimensional data [J]. Journal of Physics Conf, 2019, 1237(3): No.032043

[6] Li Hui, Li Haining, Zhang Shu, et al. Intelligent learning system based on personalized recommendation technology [J]. Neural Computing and Applications, 2019, 31(9): 4455-4462

[7] Jiang Zhuoxuan, Zhang Yan, Li Xiaoming. Analysis and prediction of learning behavior based on MOOC data [J]. Journal of Computer Research and Development, 2015, 52(3): 614-628 (in Chinese)
(蒋卓轩, 张岩, 李晓明. 基于 MOOC 数据的学习行为分析与预测[J]. 计算机研究与发展, 2015, 52(3): 614-628)

[8] Pandey M, Sharma V K. A decision tree algorithm pertaining to the student performance analysis and prediction [J]. International Journal of Computer Applications, 2013, 61(13): 1-5

[9] Bhardwaj B K, Pal S. Data Mining: A prediction for performance improvement using classification [J]. arXiv preprint arXiv:1201.3418, 2012

[10] Thiele T, Singleton A, Pope D, et al. Predicting students' academic performance based on school and socio-demographic characteristics [J]. Studies in Higher Education, 2016, 41(8): 1424-1446

[11] Zhang Xu, Xue Ruojuan, Liu Bin, et al. Grade prediction of student academic performance with multiple classification models [C] //Proc of the 14th Int Conf on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD). Piscataway, NJ: IEEE, 2018: 1086-1090

[12] Mueen A, Zafar B, Manzoor U, et al. Modeling and predicting students' academic performance using data mining techniques [J]. International Journal of Modern Education and Computer Science, 2016, 8(11): 36-42

[13] Francis B K, Babu S S. Predicting academic performance of students using a hybrid data mining approach [J]. Journal of Medical Systems, 2019, 43(6): 1-15

[14] Xie Juanying, Zhang Yi, Chen Enhong. Key factors mining and performance prediction of students' performance [J]. Journal of Nanjing University of Information Science and Technology: Natural Science Edition, 2019, 11(3): 316-325 (in Chinese)
(谢娟英, 张宜, 陈恩红. 学生成绩关键因素挖掘与成绩预测[J]. 南京信息工程大学学报: 自然科学版, 2019,11(3): 316-325)

[15] Sun Xiaowan, Wang Ying, Wang Xin, et al. Aspect-based sentiment analysis model based on dual-attention networks [J]. Journal of Computer Research and Development, 2019, 56(11): 2384-2395 (in Chinese)
(孙小婉, 王英, 王鑫, 等. 面向双注意力网络的特定方面情感分析模型[J]. 计算机研究与发展, 2019, 56(11): 2384-2395)

[16] Mnih V, Heess N, Graves A. Recurrent models of visual attention [C] //Advances in Neural Information Processing Systems. Massachusetts: MIT Press, 2014: 2204-2212

[17] Xu K, Ba J, Kiros R, et al. Tell: Neural Image Caption Generation with Visual Attention [J]. IEEE Transactions on Neural Networks, 2015, 5: 157-166

[18] Chen Long, Zhang Hanwang, Xiao Jun, et al. Sca-CNN: Spatial and channel-wise attention in convolutional networks for image captioning [C] //Proc of the 30th IEEE Conf on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2017: 5659-5667

[19] Rush A M, Chopra S, Weston J, et al. A neural attention model for abstractive sentence summarization [J]. arxiv:org/abs/1509.00685, 2015

[20] Bahdanau D, Chorowski J, Serdyuk D, et al. End-to-end attention-based large vocabulary speech recognition [C] // Proc of the 2016 IEEE Int Conf on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ: IEEE, 2016: 4945-4949

[21] Huang Youwen, You Yadong, Zhao Peng. Image caption generation model with convolutional attention mechanism [J]. Journal of Computer Application, 2020, 40(1): 23-27 (in Chinese)
(黄友文, 游亚东, 赵朋. 融合卷积注意力机制的图像描述生成模型[J]. 计算机应用, 2020, 40(1): 23-27)

[22] Cheng Yong, Shen Shiqi, He Zhongjun, et al. Agreement-based joint training for bidirectional attention-based neural machine translation [C] //Proc of the 25th Int Joint Conf on Artificial Intelligence. Menlo Park, CA: AAAI Press, 2016: 2761-2767

[23] Yu Chongchong, Cao Shuai, Pan Bo, et al. Retrieval matching question and answer method based on improved CLSM with attention mechanism [J]. Journal of Computer Application, 2019, 39(4): 972-976 (in Chinese)
(于重重, 曹帅, 潘博, 等. 基于注意力机制的改进 CLSM 检索式匹配问答方法[J]. 计算机应用, 2019,39(4): 972-976)

[24] Cortez P, Silva A M G. Using data mining to predict secondary school student performance [C] //Proc of the 15th Conf on European Concurrent Engineering. Ostend, Belgium: EUROSIS, 2008: 5-12

[25] Wu Bei. Research and application of grade prediction model based on decision tree algorithm [D]. Xi'an: Xi'an University of Technology, 2019 (in Chinese)
(吴蓓. 基于决策树算法的成绩预测模型研究及应用[D]. 西安: 西安理工大学, 2019)

[26] Burman I, Som S. Predicting students academic performance using support vector machine [C] //Proc of the 2019 Conf on Amity Int Conf on Artificial Intelligence (AICAI). Piscataway, NJ: IEEE, 2019: 756-759

[27] Bahadir E. Using neural network and logistic regression analysis to predict prospective mathematics teachers' academic success upon entering graduate education [J]. Educational Sciences: Theory and Practice, 2016, 16(3): 943-964

[28] Pujianto U, Azizah E N, Damayanti A S. Naive Bayes using to predict students' academic performance at faculty of literature [C] //Proc of the 5th Int Conf on Electrical, Electronics and Information Engineering (ICEEIE). Piscataway, NJ: IEEE, 2017: 163-169

[29] Zhu Jianlin, Kang Yilin, Zhu Rongbo, et al. College academic achievement early warning prediction based on decision tree model [C] //Proc of the 7th CCF Conf on Big Data. Berlin: Springer, 2019: 351-365

Li Mengying, born in 1994. Master candidate. Her main research interest is data mining.

Wang Xiaodong, born in 1963. PhD, professor and PhD supervisor. His main research interests include ontology engineering and data mining.

Zhang Kun, born in 1990. PhD. His main research interests include natural language processing and deep learning.

Ruan Shulan, born in 1996. Master candidate. His main research interests include data mining, sentiment analysis, and multimedia modeling.

Liu Qi, born in 1986. PhD, professor and PhD supervisor. His main research interests include data science, data mining, machine learning, social network analysis and recommender systems.