doi:10.3969/j.issn.1001-4616.2023.03.012

基于 K-XGBoost 融合模型的高校学生学情预测研究

赵宇奔[1],王鑫宁[2],李　崇[1]

(1.中国海洋大学工程学院,山东 青岛 266100)
(2.中国海洋大学基础教学中心,山东 青岛 266100)

[**摘要**]　高精准的学情预测是提升高校教学水平促进教学改革的重要技术手段. 目前学情预测存在数据维度单一和数据结构不平衡等问题,降低了预测模型的准确性与泛化能力. 为此,本文提出了 K-XGBoost 学情预测融合模型. 首先,该模型通过精准特征提取与重构,构建基于高校教务处数据库的多维度学情特征集;其次,设计基于最小 2-范数的聚类算法,创新性地建立无监督数据平衡化机制;最后,基于损失函数优化的 XGBoost 集成学习方法设计学情预测模块,构建高准确性和高泛化能力的 K-XGBoost 学情预测融合算法. 实验结果表明,K-XGBoost 多个子类模型的预测值均较好地逼近真实值,可将成绩预测结果的平均绝对误差(MAE)和均方根误差(RMSE)相较基线 XGBoost 模型分别降低了 76.19%、85.33%,显著提升了学情预测的准确性和泛化能力.

[**关键词**]　K-XGBoost,学情预测,数据挖掘,机器学习,集成学习
[**中图分类号**]TP181　[**文献标志码**]A　[**文章编号**]1001-4616(2023)03-0089-09

Research on Undergraduate Academic Prediction Based on K-XGBoost Fusion Model

Zhao Yuben[1],Wang Xingning[2],Li Chong[1]

(1.College of Engineering,Ocean University of China,Qingdao 266100,China)
(2.Teaching Center of Fundamental Courses,Ocean University of China,Qingdao 266100,China)

Abstract:High-precision prediction of academic conditions is an important technical means to improve the teaching level of colleges and promote teaching reform. At present,there are problems such as single data dimension and unbalanced data structure in academic prediction,which reduces the accuracy and generalization ability of the prediction model. To the end,this paper proposes a K-XGBoost academic situation prediction fusion model. Firstly,through accurate feature extraction and reconstruction,the model constructs a multi-dimensional set of academic features based on the database of the Academic Affairs Office of the University. Secondly,the clustering algorithm based on the minimum 2-norm is designed,and the unsupervised data balancing mechanism is innovatively established. Finally,the XGBoost integrated learning method based on loss function optimization designs the academic situation prediction module,and constructs a K-XGBoost learning situation prediction fusion algorithm with high accuracy and high generalization ability. The experimental results show that the predicted values of K-XGBoost models can well approximate the real values,and the MAE and RMSE of performance prediction results are reduced by 76.19% and 85.33% respectively compared with XGBoost models,which significantly improves the accuracy and generalization ability of the academic performance prediction.

Key words:K-XGBoost,academic performance prediction,data mining,machine learning,ensemble learning

目前,大数据、人工智能等新兴技术在教育领域取得众多进展,引领了高校教育教学模式的智能化变革[1-2]. 其中,学情预测作为教与学的关键环节,是提升学生自我认知和加强教师教学干预的重要技术手

收稿日期:2022-07-29.
基金项目:中央高校基本科研业务费专项(202213016)、山东省自然科学基金项目(ZR201910230031)、2022 年度青岛市社会科学规划研究项目(QDSKL2201014).
通讯作者:王鑫宁,博士,讲师,研究方向:智能信息处理,大数据分析与可视化. E-mail:wangxinning@ ouc.edu.cn

段,成为教育与计算机领域学科交叉融合的前沿研究热点[3]. 学情预测是指利用相关学情数据,依靠一定手段预测学生未来表现[4]. 然而,学情数据表现出多维度、不平衡的特点,以往的学情预测研究鲜有建立兼顾学情数据特点的预测模型,难以满足精准预测的需求.

早期的学情预测研究主要基于数理统计方法,通过考试成绩、课后作业等预测未来成绩[5]. 例如,Hussain 等人[6]使用来自教育系统中学生做练习的时间、次数等预测成绩,表明支持向量回归(support vector regression,SVR)的准确率高于线性回归(linear regression,LR)和朴素贝叶斯(Naive Bayes). Mueen 等[7]收集了学生课程数据,并采用 SMOTE 方法平衡数据集,表明 Naive Bayes 分类器较好的预测准确率. 除单一统计模型外,有研究提出了融合模型[8]. 例如,Marbouti 等[9]使用成绩预测不及格风险,表明了包含 SVR、K 近邻(K-nearest neighbor,KNN)和 Naive Bayes 的融合模型在预测成绩时较单一模型有更好的准确率. 以上数理统计方法易于理解,但其通常使用在特征维数较小的数据集上,难以保证学习到多特征之间的深度联系.

近年来,神经网络方法为大数据背景下的学情预测提供了重要手段[10]. KIM B H 等[11]提出了一种基于双向长短时记忆(bidirectional long short term memory,BLSTM)的深度学习算法 GritNet,利用学生的课堂参与度预测其能否毕业,关注于序列中最相关的部分,一定程度上克服了样本不平衡的影响. Liu 等[12]基于学生练习记录,提出了一种基于马尔科夫原理和注意力机制的 BLSTM 框架 EKT,跟踪学生学习情况,在时间序列上预测学生成绩. Waheed 等[13]利用深度人工神经网络(deep artificial neural network,Deep ANN)预测 22 437 名学生退课风险. 相对于数理统计方法,深度神经网络可以建立掌握学情数据更深层特征的复杂模型[14].

除了数理统计和神经网络,集成学习方法日益成为学情预测领域的研究热点[15-16]. BATOOL S 等[17]采用随机森林(random forest,FR)算法探究学生个人行为、家庭成员状态等人口统计学特征对成绩的影响. Ahmed 等[18]选择了包括 450 名学生、20 维家庭特征的数据集,使用梯度提升决策树(gradient boosting decision tree,GBDT)来预测学生在期末考试中的表现,并证明了其相对 SVR、逻辑回归(logistic regression,LR)的优势. Duan 等[19]根据 10 000 名学生的高考成绩和课程情况预测成绩,验证了极度梯度提升(eXtreme gradient boosting,XGBoost)算法模型的整体准确性、稳定性和有效性. 由此可见,集成学习方法在学情预测研究中展现出显著的优越性.

以往的研究侧重于预测模型的选择与构建,忽略了学情数据本身对预测结果的影响. 首先,多维综合“涌现”的复杂性特征是学情预测具有科学性和高准确性的保障,将学生以往的成绩相关数据、行为数据、任课教师行为数据等多方面相结合是必要的. 其次,学情数据存在样本不平衡问题,尤其是以考试成绩为特征时,而以往的研究通常忽略了样本不平衡问题,即使利用了欠采样、重采样[20]或 SMOTE 方法[21],也可能会导致模型欠拟合、过度拟合或过度泛化. 为此,本文提出了 K-XGBoost 学情预测融合模型,突出贡献如下:

(1)精准提取并重构高校教务处数据库数据,建立包含成绩和行为在内的多维度学情数据集,作为预测模型挖掘更深层特征的前提;

(2)首次提出基于最小 2-范数的 K-means++算法,以无监督聚类方式匹配与划分样本子类,实现学情数据平衡化机制;

(3)创新性地将聚类算法与损失函数优化的 XGBoost 集成方法融合并应用于学情预测,构建了高准确性、高泛化能力的预测模型.

学生课程成绩和平均绩点的预测结果评估了 K-XGBoost 相较于基线模型的优越性能,验证了 K-XGBoost 融合模型对学情预测的有效性. 该模型的建立为克服数据不平衡问题提供了新思路,为利用融合模型开展学情预测提供了新方法.

1 方法选择与模型构建

本文构建了 K-XGBoost 学情预测融合模型,包括特征提取与重构、基于优化 K-means++的数据平衡化机制和融合预测模型构建,如图 1 所示.

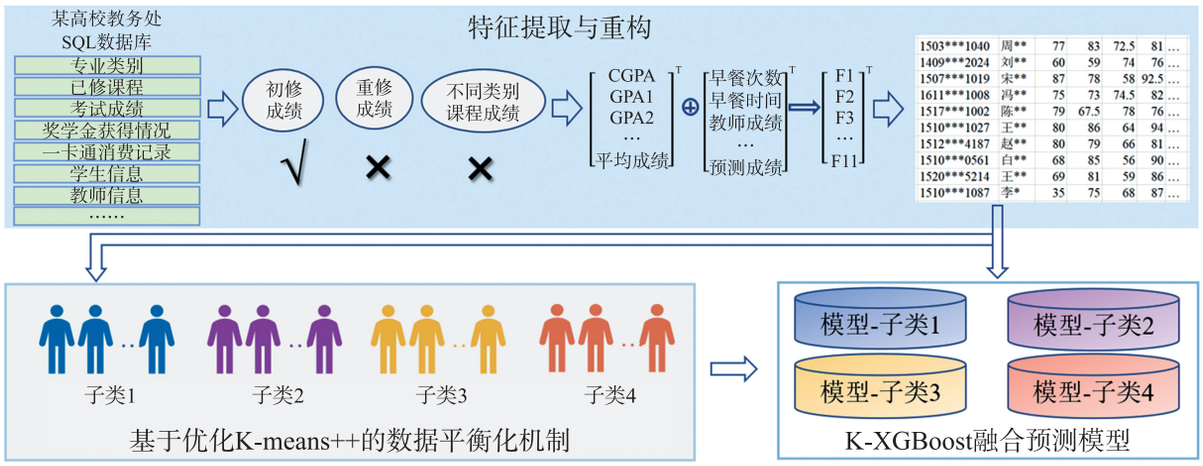

图 1　K-XGBoost 学情预测融合模型

Fig. 1　K-XGBoost fusion model of academic performance prediction

1.1　特征提取与重构

本文以某高校教务管理系统 SQL 数据库为数据来源，获取多维度、综合性学情数据. 本研究选取该数据库中 6 个数据集，主要包含 2003 年~2018 年 10 万余名学生的 108 034 条基本个人信息、2006 年~2018 年考试成绩记录 3 206 314 条、获得奖学金数据 25 589 条、校园卡消费记录 4 800 多万条、在职教师基本个人信息 5 337 条. 以此构建多维度学情特征数据集，保证研究的可靠性和实际应用价值. 但是，原始学情数据为非结构化数据，存在数据类型无法被机器学习算法识别、数据冲突、数据信息密度低等问题，影响模型对学情信息的挖掘和预测. 因此，本文首先将原始数据进行转换、筛选和重构.

首先，转换成绩、课程和奖学金数据中的学期（如“2018 秋”）以及校园卡消费记录中的消费时间（如“12:33”）等数据为数值型. 然后，筛选成绩数据集中部分学生课程重修、人工输入数据等导致的信息冗余或冲突的数据. 最后，重构具有高信息密度的多维度学情特征集，例如：利用课程所属类别标签，提取所预测科目的同类课程成绩；计算学生获得奖学金金额、平均成绩、累积平均绩点 CGPA、前三学年成绩绩点 GPA1、GPA2、GPA3（不足三学年即为空，算法可识别）、当前学期的平均绩点 GPA 以及任课教师该课程的平均成绩；获取学生近 100 d 内的早餐次数和平均早餐时间等. 本文以预测概率统计成绩和当前学期的 GPA 为例，通过特征提取与重构，获得包含学生已有知识储备、本学期状态和教师教课情况等多方面信息的 12 维学情特征集，如表 1 所示. 该特征集具有真实性、综合性和可靠性，为实现学情预测优越性能提供数据保障.

表 1　学情特征描述

Table 1　Academic features description

特征	取值范围	特征	取值范围	特征	取值范围
学号	030000~183099	平均成绩	0~100	任课教师平均成绩	0~100
CGPA	0~5	同类别平均成绩	0~100	概率统计成绩	0~100
GPA1	0~5	奖学金金额	0~50 000	当前学期的 GPA	0~5
GPA2	0~5	早餐次数	0~100		
GPA3	0~5	早餐时间	6~10		

1.2　学情数据平衡化

以上的学情特征存在数据段分布不平衡的问题，以 4 420 名学生的概率统计成绩为例，0~60 分的数据为 892 条，占比 20.18%；60~100 分的数据为 3 528，占比 79.82%. 样本量小的数据段易受样本量大的数据段影响，其预测值较真实值偏高；同时，样本量大的数据段也会受个别小数据量样本的影响. 因此，本节基于优化的聚类算法在无监督情况下划分学生类别，获得反映不同学生学情的 K 个子类群体，在建立数据平衡化机制的同时，保留了原始数据的分布特征，以助于预测模型具有更高的准确性和泛化能力.

原始的 K-means++算法随机选取样本点作为第一个初始聚类中心 c_1，容易造成初始类别分布不均匀. 为克服这一问题，本文计算并选择训练集中 2-范数最小的样本为 c_1. 此时的 c_1 分布在所有数据样本的边缘，以此进一步选择其他初始中心，保证初始聚类中心的分散性以获得更佳的类别划分结果，计算公

式如下:

$$c_1=\min\|x_i\|_2=\min(|x_{i1}|^2+|x_{i2}|^2)^{1/2},\tag{1}$$

式中,x_i 表示第 i 个样本,x_{i1} 和 x_{i2} 分别表示第 i 个样本的两维学情特征. 然后,根据 K-means++算法[22],依次选取第 2 至第 K 个聚类中心. 具体地,计算每个样本与当前已有聚类中心之间的最短欧式距离:

$$D(x_i)=\min[\sqrt{(x_i-c_1)^2},\cdots,\sqrt{(x_i-c_K)^2}],\tag{2}$$

式中,c_1、c_K 分别表示第 1、K 个聚类中心. 然后,计算每个样本点被选为下一个聚类中心的概率:

$$P(x_i)=D(x_i)^2\sum_{i=1}^{m}D(x_i)^2,\tag{3}$$

式中,m 为样本数量. 选择最大概率值所对应的样本作为初始聚类中心. 最后,计算各个类簇的平均值,并将其更新为聚类中心:

$$c_j=\frac{1}{N_j}\sum_{i=1}^{N_j}x_i.\tag{4}$$

式中,N_j 表示第 j 个簇的样本数量. 在此基础上,计算每个样本与新聚类中心的距离以更新聚类,并重复此操作直到中心不再发生变化. 以上迭代过程示意如图 2 所示:图 2(a)为二维空间上样本点,图 2(b)中虚线圆为基于最小 2-范数 K-means++算法获取的初始聚类中心,图 2(c)-(d)为第一次迭代过程,图 2(e)-(f)为第二次迭代过程.

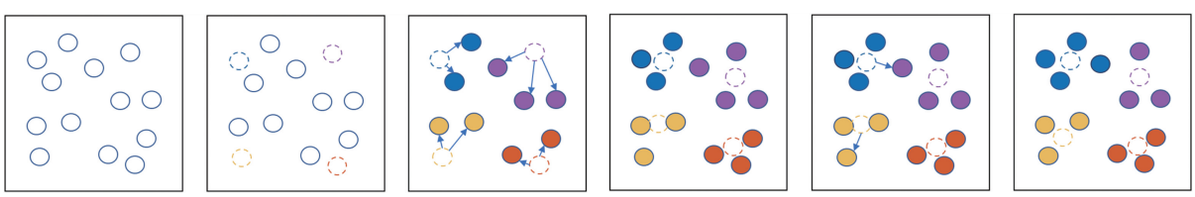

图 2　聚类迭代过程示意图

Fig. 2　Diagram of iterative process of clustering

以上利用优化的 K-means++算法,在保持原来真实数据值和数据量的基础上,依据特征分布建立灵活的数据平衡化机制,实现不同子类内部的学情数据平衡化,以助于提升学情预测的准确性和泛化能力.

1.3　预测模型构建

在以上特征平衡化的基础上,基于以往研究中 XGBoost 集成学习方法[23]表现出的优越性[19],本文将其作为学情预测研究的基线模型,构建 K-XGBoost 学情预测融合模型. 在预测模型中,预测值为已建立的树对该样本的预测值之和,表示为:

$$\hat{s}_i=\sum_t f_t(x_i),\tag{5}$$

式中,i 表示第 i 个样本,s_i 为第 i 个样本的真实成绩. 目标函数如下:

$$\mathcal{L}(\phi)=\sum_i l(s_i,\hat{s}_i)+\sum_t \Omega(f_t),\tag{6}$$

式中,$\sum_i l(s_i,\hat{s}_i)$ 为 m 个样本的损失函数之和;t 表示第 t 棵树,为正则化项,即前 k 棵树的复杂度. 为提升梯度收敛速度,将损失函数二阶泰勒展开为:

$$\mathcal{L}^{(t)}=l(s_i,\hat{s}_i^{(t-1)}+f_t(x))\approx l(s_i,\hat{s}_i^{(t-1)})+g_i f_t(x_i)+\frac{h_i}{2}f_t^2(x_i),\tag{7}$$

式中,$g_i=\partial_{\hat{s}_i}l(s_i,\hat{s}_i)$ 为损失函数的一阶导数,$h_i=\partial^2_{\hat{s}_i}l(s_i,\hat{s}_i)$ 为损失函数的二阶导数. 然后,为使正则项能够被表示为树的参数,将其展开为:

$$\Omega(f_t)=\gamma M+\frac{1}{2}\lambda\sum_{m=1}^{M}(w_m)^2,\tag{8}$$

式中,γ 和 λ 为控制模型复杂性的超参数. M 表示叶节点数量,表示叶节点权重. 在上述基础上,为了用贪心算法遍历所有特征来完成节点分割,将式(6)转化为:

$$\mathcal{L}^{*}=-\frac{1}{2}\sum_{j=1}^{M}\frac{\left(\sum_{i\in I_j}g_I\right)^2}{\frac{1}{2}\sum_{i\in I_j}h_i+\lambda}+\gamma M, \tag{9}$$

式中，I_j 表示第 j 个节点包含的样本. 目标函数 $\mathcal{L}^{*}$ 的值越小，模型中树的结构越好. 然而，在模型训练过程中，平均绝对误差损失在预测值和真实值的误差很小时给定模型过小的惩罚，而均方根误差损失中平方项的存在导致误差较大时给定模型过大的惩罚. 因此，为保证损失函数给定小误差点和异常数据点合适的惩罚并保证模型较快的收敛速度，本文改进 K-XGBoost 模型中的损失函数为：

$$l_{\delta,h}=\delta^2\left(\sqrt{1+\left((s_i-\hat{s}_i)/\delta\right)^2}-1\right)+h(s_i-\hat{s}_i)^2. \tag{10}$$

该式可通过调节超参数 δ 降低损失函数的敏感度，同时加入可有超参数 h 控制的误差平方项以加快收敛速度. 以上过程形成的集成树模型如图 3 所示，算法在已建立树的基础上基于损失函数构建新树，从而使预测结果不断逼近真实值.

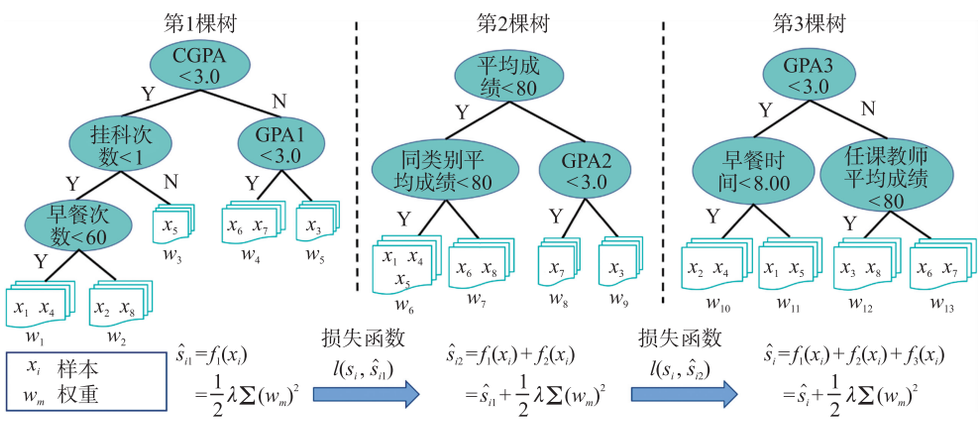

图 3　K-XGBoost 集成树模型

Fig. 3　Tree ensemble model of K-XGBoos

2　实验设计

2.1　模型训练

采用前述聚类机制将学生划分为不同的子类. 特征选择为累积平均绩点 CGPA 和早餐次数，分别代表学生已有的认知和学生在本学期的状态. 由于 K-XGBoost 预测效果受类簇数量 K 的影响，本研究设置 $K=4,5,6$ 进行对比. 子类划分结果如图 4 所示，可以看出，随着 K 值的增加，学生被划分为内部分布更紧密的子类.

在以上基础上，将数据集按 8∶2 随机划分为训练集和测试集，利用 K-XGBoost 算法为训练集中的各个子类群体分别训练模型，并将测试集样本根据其所属的类别输入至对应的子类模型进行成绩预测.

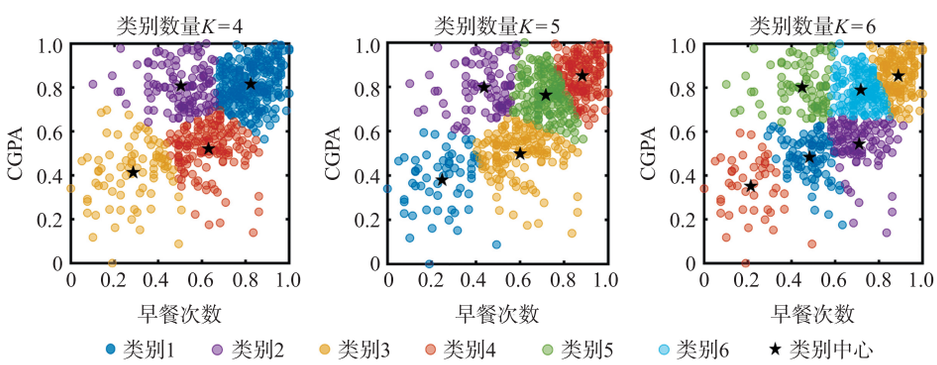

图 4　$K=4,5,6$ 时的聚类结果

Fig. 4　Cluster results when $K=4,5,6$

2.2　模型评价指标

本文选取以下 4 个模型评价指标:平均绝对误差(mean absolute error,MAE)反映成绩预测误差的大小;均方根误差(root mean squared error,RMSE),与 MAE 相结合反映样本误差的离散程度;可决系数(coefficient of determination,记为 R^2)反映预测成绩对真实成绩的解释的程度;纳什效率系数(Nash-sutcliffe efficiency coefficient,NASH)进一步考察模型的可信度. 数学表达式如下:

$$MAE=\frac{1}{n}\sum_{i=1}^{n}|s_i-\hat{s}_i|, \tag{11}$$

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(s_i-\hat{s}_i)^2}, \tag{12}$$

$$R^2=1-\frac{\sum_{i=1}^{n}(s_i-\hat{s}_i)^2}{\sum_{i=1}^{n}(s_i-\bar{s}_i)^2}, \tag{13}$$

$$NASH=1-\left[\frac{\frac{1}{n}\sum_{i=1}^{n}(s_i-\hat{s}_i)^2}{\frac{1}{n}\sum_{i=1}^{n}(\hat{s}_i-s_{i,avr})^2}\right]. \tag{14}$$

式中,*MAE* 和 *RMSE* 越接近于 0、R^2 和 *NASH* 越接近于 1,模型拟合得越好.

3　结果与讨论

表 2 展示了不同模型在测试集上的单科成绩预测结果. 对于 MAE 和 RMSE,XGBoost 为 9.328 0、13.261 9,K-XGBoost 在 $K=4$ 时为 3.213 9、6.572 2,分别降低了 76.19%、85.33%,并且随着 K 值的增大而继续减小. 值得注意的是,XGBoost 和 K-XGBoost 在 $K=4$ 时的 RMSE 远大于 MAE,可以得知不同样本的误差差别较大,即存在预测误差很大的样本点,而 K-XGBoost 在 $K=5,6$ 时的 RMSE 与 MAE 均相差较小. 同时,K-XGBoost 模型的 R^2 和 NASH 较基线 XGBoost 有显著提升.

表 2　不同模型的单科成绩预测结果指标比较

Table 2　Comparison of f individual subject prediction indicators under different models

模型	平均绝对误差(*MAE*)	均方根误差(*RMSE*)	决定系数(R^2)	纳什效率系数(*NASH*)
XGBoost	9.328 0	13.261 9	0.606 1	0.528 2
K-XGBoost($K=4$)	3.213 9	6.572 2	0.905 2	0.784 5
K-XGBoost($K=5$)	1.463 7	1.968 5	0.975 0	0.970 3
K-XGBoost($K=6$)	0.991 4	1.378 7	0.984 1	0.980 6

为进一步验证 K-XGBoost 对不同类型预测目标的性能,除预测学生的单科成绩以外,本文同时预测学生在当前学期的平均绩点 GPA,不同模型的预测结果指标如表 3 所示. 从表中可以看出,K-XGBoost 各项指标均显著优于 XGBoost,且随着 K 值增大各项指标持续改善:MAE 和 RMSE 由基线 XGBoost 的 0.327 9、0.370 7 可分别减小至 0.133 1、0.177 4,降低了 59.41%、52.14%,且两者之间的差值逐渐减小;R^2 和 NASH 分别达到 0.970 2 和 0.967 5,表明预测值与真实值的强相关性.

表 3　不同模型的 GPA 预测结果指标比较

Table 3　Comparison of GPA prediction indicators under different models

模型	平均绝对误差(*MAE*)	均方根误差(*RMSE*)	决定系数(R^2)	纳什效率系数(*NASH*)
XGBoost	0.327 9	0.370 7	0.883 5	0.663 9
K-XGBoost($K=4$)	0.217 0	0.306 5	0.922 0	0.785 8
K-XGBoost($K=5$)	0.159 7	0.213 4	0.962 5	0.952 9
K-XGBoost($K=6$)	0.133 1	0.177 4	0.970 2	0.967 5

本文随机选取 200 个样本点可视化预测值与真实值的误差,如图 5~图 8 所示. 图中点由真实成绩和预测成绩构成,黑色线为最佳拟合线,即预测成绩等于真实成绩;蓝色虚线之间为真实成绩与预测成绩误

差在 10 分以内的样本；红色虚线之间为真实成绩与预测成绩误差在 20 分以内的样本.

图 5 为基线 XGBoost 模型的预测结果可视化. 从图中可以看出，预测值与真实成绩普遍存在较大偏差，甚至部分样本的偏差大于 20 分. 同时，低分样本的预测值普遍大于其真实值，这说明低分数段的样本预测结果受到了样本量大的高分数段样本的影响，模型难以准确识别出有不及格风险的学生.

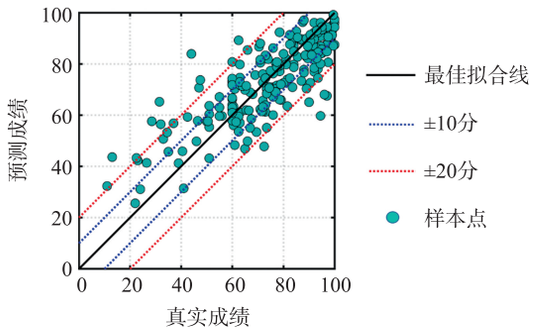

图 5　XGBoost 预测结果可视化

Fig. 5　Visualization of prediction results of XGBoost

图 6 为 K-XGBoost 在 $K=4$ 时预测结果可视化. K-XGBoost 4 个模型的样本点表现出集中在不同的分数段，且与图 4 中聚类结果吻合. 例如，图 4 中当 $K=4$ 时，类别 1 的样本点同时具有较高的 CGPA 和早餐次数，对应图 6 中模型−子类 1 中的样本点都具有较高的真实成绩和预测成绩. 从整体看，K-XGBoost 对高分数段样本的预测值较为准确，对于低分数段样本的预测值不再普遍偏高，4 个模型的样本点均更集中在最佳拟合线附近，预测结果显著优于 XGBoost.

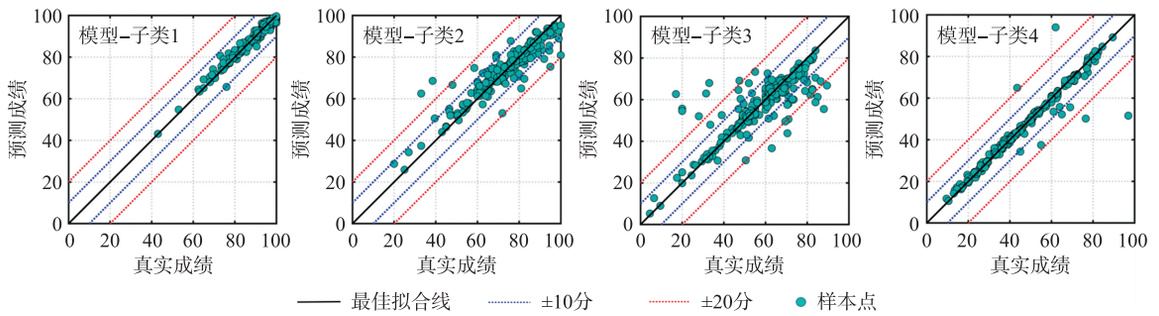

图 6　K-XGBoost 在 $K=4$ 时预测结果可视化

Fig. 6　Visualization of prediction results of K-XGBoost when $K=4$

图 7 为 K-XGBoost 模型在 $K=5$ 时的预测结果可视化. K-XGBoost 在 $K=5$ 时的样本点整体偏差较 $K=4$ 时进一步减小，每个模型中样本的预测值和真实值的偏差均在 10 分以内，只有极小部分样本的偏差在 10~20 分. 这说明，随着 K 值的增加，学生被划分成更多不同的类别，每个模型中的样本点更加相似，学习所得的模型更具有针对性，高低分数段数据不平衡的问题也得以克服.

图 8 为 K-XGBoost 模型在 $K=6$ 时的预测结果可视化. 相较于 $K=5$ 时，K-XGBoost 在 $K=6$ 时的样本点更

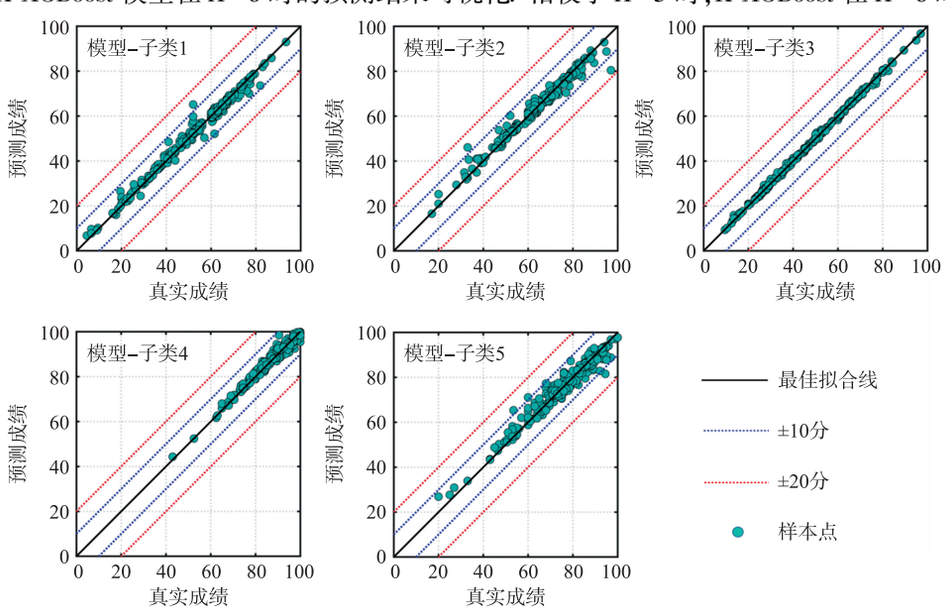

图 7　K-XGBoost 在 $K=5$ 时预测结果可视化

Fig. 7　Visualization of prediction results of K-XGBoost when $K=5$

加集中于最佳拟合线附近,所有样本的预测值和真实值的偏差均在 10 分以内. 这进一步证明 K 值的增加对克服数据不平衡、提高预测准确性和模型泛化能力有显著作用,验证了 K-XGBoost 融合模型的有效性. 然而,这种有效性必然会随 K 值的不断增加而破坏,因为当 K 值过大时,每个类别的样本点过少,无法满足模型训练对数据量的需求. 对于本文的数据,在 $K=6$ 时 K-XGBoost 的预测准确性能够基本满足实际需求.

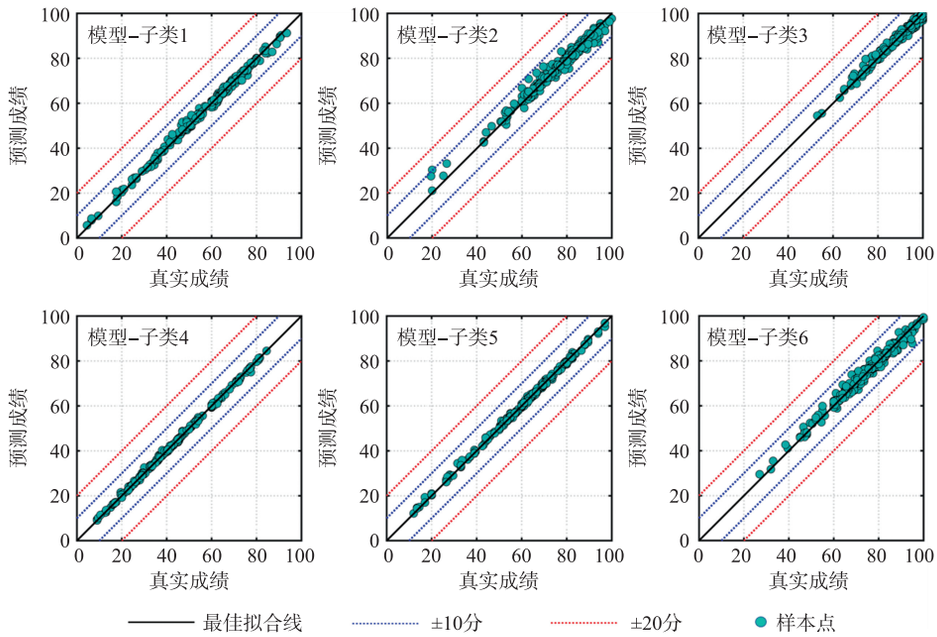

图 8　K-XGBoost 在 $K=6$ 时预测结果可视化

Fig. 8　Visualization of prediction results of K-XGBoost when $K=6$

图 9 为 XGBoost 基线模型和不同 K 值下 K-XGBoost 融合模型对 GPA 的预测结果可视化. 从图中可以看出,XGBoost 的预测结果中只有少数样本点分布于最佳拟合线上,多数样本的真实值与预测值误差在 0.5 左右. 而在 K-XGBoost 模型中:当 $K=4$ 时,样本点的分布情况相较于 XGBoost 模型有显著改善;当 $K=5$ 时,所有样本点的误差均在 0.5 以内;当 $K=6$ 时,样本更紧密地分布在最佳拟合线附近. 这一现象与预测单科成绩的结果相符合,进一步验证了 K-XGBoost 优越的学情预测性能,并且模型性能在一定范围内随着 K 值的增加而提升.

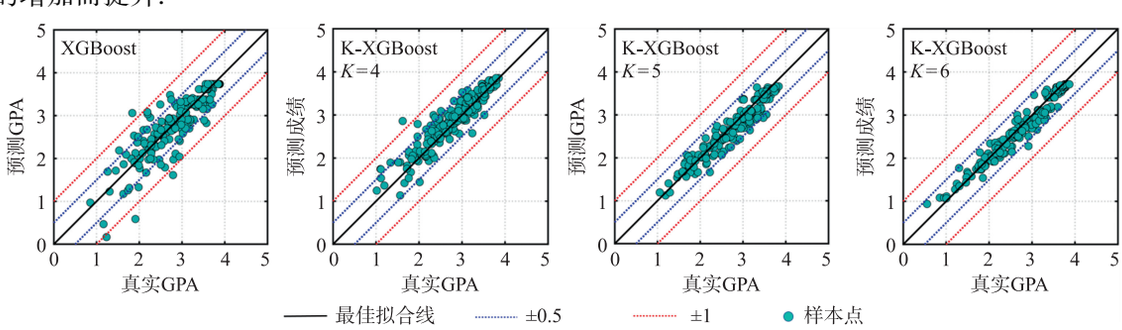

图 9　GPA 预测结果可视化

Fig. 9　Visualization of prediction results of GPA

4　结论

本文围绕学情预测中数据维度单一、数据结构不平衡导致的预测结果准确性低、泛化性差的问题,设计了 K-XGBoost 学情预测融合模型. 该模型首先提取并重构某高校教务处数据库的多维度学情特征,然后通过最小 2-范数聚类算法创新性地建立有效的数据结构平衡化机制,最后基于损失函数改进的 XGBoost 集成方法构建 K-XGBoost 学情预测算法. 实验结果表明,K-XGBoost 对百分制课程成绩和五分制 GPA 预测结果的平均绝对误差(MAE)和均方根误差(RMSE)分别可降低至 3.213 9 分、6.572 2 分和 0.217 0、0.306 5 以下,显著提升了学情预测的准确性和泛化能力,实现了理想的学情预测性能,论证了基于优化聚类算法的数据平衡机

制与集成学习方法在学情预测中的有效性.

在未来的工作中,将力求寻找更全面的学情数据,并进一步从算法的结构、算法与数据的适配度、模型中 K 值的最优选择机制等方面优化学情预测模型,以获得更精准、更稳定的学情预测性能. 同时,将扩展学情预测模型的适用范围,将其应用于学生心理健康预测、身体素质预测以及职业发展预测等更广泛的领域.

[参考文献]

[1] 陈桂香. 大数据对我国高校教育管理的影响及对策研究[D]. 武汉:武汉大学,2017.

[2] RAY S,SAEED M. Applications of educational data mining and learning analytics tools in handling big data in higher education[M]//Applications of big data analytics. Switzerland:Springer,Cham,2018:135-160.

[3] YANG F,LI F W B. Study on student performance estimation,student progress analysis,and student potential prediction based on data mining[J]. Computers & education,2018,123:97-108.

[4] SALAL Y K,ABDULLAEV S M,KUMAR M. Educational data mining:Student performance prediction in academic[J]. International journal of engineering and advanced technology,2019,8(4C):54-59.

[5] 聂秀山,马玉玲,乔慧妍,等. 任务粒度视角下的学生成绩预测研究综述[J]. 山东大学学报(工学版),2022,52(2):1-14.

[6] HUSSAIN M,ZHU W,ZHANG W,et al. Using machine learning to predict student difficulties from learning session data[J]. Artificial intelligence review,2019,52(1):381-407.

[7] MUEEN A,ZAFAR B,MANZOOR U. Modeling and predicting students' academic performance using data mining techniques[J]. International journal of modern education & computer science,2016,8(11):36-42.

[8] 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.

[9] MARBOUTI F,DIEFES-DUX H A,MADHAVAN K. Models for early prediction of at-risk students in a course using standards-based grading[J]. Computers & education,2016,103:1-15.

[10] TOMASEVIC N,GVOZDENOVIC N,VRANES S. An overview and comparison of supervised data mining techniques for student exam performance prediction[J]. Computers & education,2020,143:103676.

[11] KIM B H,VIZITEI E,GANAPATHI V. GritNet:Student performance prediction with deep learning[J/OL]. arXiv Preprint,2018. 10.48550/arXiv:1804.07405.

[12] LIU Q,HUANG Z,YIN Y,et al. Ekt:Exercise-aware knowledge tracing for student performance prediction[J]. IEEE transactions on knowledge and data engineering,2019,33(1):100-115.

[13] WAHEED H,HASSAN S U,ALJOHANI N R,et al. Predicting academic performance of students from VLE big data using deep learning models[J]. Computers in human behavior,2020,104:106189.

[14] MONTAVON G,SAMEK W,MüLLER K R. Methods for interpreting and understanding deep neural networks[J]. Digital signal processing,2018,73:1-15.

[15] PANDEY M,TARUNA S. A comparative study of ensemble methods for students' performance modeling[J]. International journal of computer applications,2014,103(8):26-32.

[16] AMRIEH E A,HAMTINI T,ALJARAH I. Mining educational data to predict student's academic performance using ensemble methods[J]. International journal of database theory and application,2016,9(8):119-136.

[17] BATOOL S,RASHID J,NISAR M W,et al. A random forest students' performance prediction(rfspp) model based on students' demographic features[C]//2021 Mohammad Ali Jinnah University International Conference on Computing(MAJICC). USA:IEEE,2021:1-4.

[18] AHMED D M,ABDULAZEEZ A M,ZEEBAREE D Q,et al. Predicting university's students performance based on machine learning techniques[C]//2021 IEEE International Conference on Automatic Control & Intelligent Systems(I2CACIS). New York,USA:IEEE,2021:276-281.

[19] DUAN D,DAI C,TU R. Research on the prediction of students' academic performance based on XGBoost[C]//2021 Tenth International Conference of Educational Innovation through Technology(EITT). New York,USA:IEEE,2021:316-319.

[20] 张新玉. 类不平衡数据分类关键技术研究[D]. 武汉:武汉大学,2021.

[21] GHORBANI R,GHOUSI R. Comparing different resampling methods in predicting students' performance using machine learning techniques[J]. IEEE access,2020,8:67899-67911.

[22] ARTHUR D,VASSILVITSKII S. k-means++:The advantages of careful seeding[C]//Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithm. New Orleans:SIAM,2006:1027-1035.

[23] CHEN T,GUESTRIN C. Xgboost:A scalable tree boosting system[C]//Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining. California. USA:ACM,2016:785-794.

[责任编辑:黄　敏]