第 59 卷 第 5 期
2023 年 9 月

南京大学学报(自然科学)
JOURNAL OF NANJING UNIVERSITY
(NATURAL SCIENCE)

Vol. 59, No. 5
Sept., 2023

DOI:10.13232/j.cnki.jnju.2023.05.011

基于Lightgbm和XGBoost的优化深度森林算法

谢军飞[1],张海清[1,3],李代伟[1,3*],于 曦[2],邓钧予[1,3]

(1.成都信息工程大学软件工程学院,成都,610225;2.成都大学斯特灵学院,成都,610106;
3.四川省信息化应用支撑软件工程技术研究中心,成都,610225)

摘 要:教育规模不断扩大,高校在校生人数持续上升,导致学生的能力参差不齐.为了提升教育水平,教师需掌握学生在校期间的学习状态,预测学生期末成绩是教师掌握学生学习状态的重要途径之一.目前的研究工作主要采用传统的机器学习算法进行成绩预测,如随机森林、贝叶斯、深度森林等,但精度不高;也有利用深度学习算法进行预测,但模型缺少可解释性.Lightgbm(Light Gradient Boosting Machine)算法内存消耗低,时间复杂度低,而XGBoost(eXtreme Gradient Boosting)算法精度高.因此,基于提高精度与降低模型内存消耗的策略,将深度森林中的随机森林与极限随机森林模块分别替换为Lightgbm和XGbBoost,提出一种基于Lightgbm和XGBoost算法的优化深度森林算法LIGHT-XDF.在八个数据集上与其他模型进行对比实验,结果表明,LIGHT-XDF算法的综合性能最好.

关键词: Lightgbm算法,XGBoost算法,深度森林,综合性能

中图分类号:TP181 **文献标志码:**A

Optimized deep forest algorithm based on Lightgbm and XGBoost

Xie Junfei[1], Zhang Haiqing[1,3], Li Daiwei[1,3*], Yu Xi[2], Deng Junyu[1,3]

(1. School of Software Engineering, Chengdu University of Information Technology, Chengdu, 610225, China;
2. Stirling College, Chengdu University, Chengdu, 610106, China; 3. Sichuan Province Engineering Technology Research Center of Support Software of Informatization Application, Chengdu, 610225, China)

Abstract: The continuous expansion of education scale and the continuous increase in the number of students in universities lead to uneven abilities of students. To improve education level, during their school years, teachers need to grasp the learning status of students, one of the important ways for which is to predicting students′ final grades. Current researches mainly use traditional machine learning algorithms to predict results, such as Random forest, Bayesian, deep forest, etc., but the accuracy is not high. Deep learning algorithms are also used for prediction, but the model lacks interpretability. In terms of model comprehensive performance, Lightgbm (Light Gradient Boosting Machine) algorithm has low memory consumption and time complexity, while XGBoost (eXtreme Gradient Boosting) algorithm has high precision. Therefore, to improve accuracy and reduce model memory consumption, we replace the Random Forest module and the Extreme Random Forest module in Deep Forest with Lightgbm and XGBoost algorithms, and propose an optimized Deep Forest algorithm LIGHT-XDF based on Lightgbm and XGBoost algorithms. Compared with other models on eight datasets, experimental results show that our LIGHT-XDF algorithm proves the best comprehensive performance.

Key words: Lightgbm, XGBoost, Deep Forest, comprehensive performance

基金项目:欧盟项目(598649-EPP-1-2018-1-FR-EPPKA2-CBHE-JP),国家自然科学基金(61602064),四川省科技厅项目(2021YFH0107,2022YFS0544,2022NSFSC0571)

收稿日期:2023-08-14

*通讯联系人,E-mail:ldwcuit@cuit.edu.cn

目前的高校教育中,学生能力的衡量指标主要是学生的成绩,同时,学生成绩也是衡量教师教学水平的重要依据.在专业课程的教学中,教师可通过学生的成绩对教学行为和方案进行临时调整,对不同情况的学生进行不同的教学干预,保证学生的学习质量,可以通过挖掘学生相关的信息来建立学生成绩预测模型,对其未来的学习表现进行预判.由于教师无法在教学期间通过获得学生的成绩来掌握学生的学习情况,所以对学生的课程成绩的预测显得尤为重要.

随着机器学习技术的快速发展,其已经广泛运用于教育数据挖掘领域.神经网络[1]、Lightgbm[2]、支持向量机(Support Vector Machine, SVM)[3]、随机森林(Random Forest, RF)[4]、XGBoost[5]等机器学习方法已被广泛运用于成绩预测、GPA(Grade Point Average)预测等多种教育数据挖掘.目前研究还存在以下不足:首先,大部分工作采用SVM和决策树等传统机器学习算法进行建模,模型精度还有提升空间;其次,部分工作采用深度学习方法构建预测模型,虽然预测结果较好,但模型缺少可解释性;模型综合性能方面,Lightgbm算法内存消耗低,时间复杂度低,XGBoost算法精度高.针对上述问题,基于提高精度与降低模型内存消耗的策略,本文将深度森林中的随机森林与极限随机森林模块分别替换为Lightgbm和XGBoost算法,提出一种基于Lightgbm和XGBoost的优化深度森林算法LIGHT-XDF,以学生的社会数据和行为数据作为输入构建预测模型,并对模型的结果进行分析评估.

1 成绩预测的研究现状

对学生成绩预测的研究可以追溯到2000年,Harackiewicz et al[6]在心理学入门课程中考察了学生的掌握目标和绩效目标,经过三个学期的观察,分析获得的数据,认为掌握目标能预测随后的心理学课程的入学率,而绩效目标可以预测长期的学习成绩.2007年Nghe et al[7]在预测两名大学生学习成绩方面比较了决策树和贝叶斯算法,数据来源是人口数据、入学数据、过往表现数据,结果发现20%的方差,并且决策树的性能优于贝叶斯.Zimmermann et al[8]基于用本科成绩预测研究生阶段成功的有效性分析,数据来源是学生过去的表现和入学数据,目的是找到最能解释学生在第三年的学习成绩的预测变量.Romero et al[9]利用学生的网上论坛活动来预测学生的成绩,数据来源是讨论板日志.早期预测的准确性较低,而聚类和关联规具有较好的预测精度.2014年Marbouti et al[10]将变量定义为指定学生是否通过课程,使用RF、决策树、KNN(*K* Nearest Neighbor)和SVM开发了不同的预测模型,但实验结果表明,没有一个单一模型能在所有方面都取得令人满意的结果,因此,他们进一步利用集成学习来开发预测模型,并通过特征选择和增加训练集的规模来优化模型,最终其在所有模型中表现最好.

2017年Asif et al[11]用大一和大二的考试成绩来预测毕业生的表现,在预测结果基础上研究典型学生的学习过程,数据集是学生之前的成绩,目的是区分成绩好和成绩差的学生.Helal et al[12]利用学生的异质性属性,采用不同的机器学习方法对学习成绩进行预测,数据为社会数据、过去的表现和学生学习活动,目的是预测最容易失败的学生,发现单一方法不优于其他方法.Polyzou and Karypis[13]通过建立特征子集的模型找出下学期可能表现较差的学生和影响学生表现的因素,数据来源是过去表现、学生课程细节和课程特点.

2020年Haridas et al[14]在智能辅导系统中预测学生成绩、学校风险学生和阅读困难学生,数据来源是过去表现,发现形成性评估数据与过去的总和分数能更好地预测学生表现.

2023年,张政庭等[15]构建了基于k-means聚类与BP神经网络的预测方法,利用学生的公共基础课和专业基础课成绩,找出专业核心课与公共基础课、专业基础课的潜在联系,从而进行预测,实验表明其所提出的算法泛化能力较好且精度更高.许欢和夏道明[16]将Bootstrap方法、递归特征消除、支持向量回归模型和变邻域搜索算法相结合,提出一种针对点、区间的成绩预测模型,采用学生的课堂内外表现为数据,并与传统的Bootstrap支持向量模型作对比.王洪亮和赵圆圆[17]基于校园一卡通数据,利用改进的离群点检测算法从多角度对学生校园行为数据进行挖掘,发现每天用餐的规律性、进入图书馆次数、借阅图

书次数及吃早饭次数之和与学业成绩具有强相关性，最后基于决策树构建了成绩预测模型.

2　基础知识

深度森林(Deep Forest，DF)是2017年Zhou and Feng[18]提出的一种深度学习方法. 和深度神经网络(Deep Neural Networks，DNN)相比，深度森林容易训练，计算开销小，超参数少，无须进行复杂调参，能适应各种大小的数据集，泛化性较好. 目前，DF被广泛运用于许多领域，证明了其在分类和预测方面的鲁棒性[19−20]. DF主要由两部分组成，分别是多粒度扫描(Multi-Grained Scanning)和级联森林(Cascade Forest).

2.1　多粒度扫描　多粒度扫描是对输入的特征进行分析，以挖掘特征之间的顺序关系为目的. 如图1所示，多粒度扫描采用多种滑动窗口来扫描输入特征，多种窗口在输入的特征向量上滑动，之后RF会针对滑动窗口提取的特征进行信息提取. 具体流程：首先，输入含有p维特征的数据，再采用长度为k的滑动窗口提取特征，设置步长为n，通过下式得到s个k维特征片段：

$$s=(p-k)/n+1$$

之后，将特征片段分别输入RF和CRTF(Completely-Random Tree Forests)模型，计算后输出类概率向量，再把所有森林输出的类概率向量进行拼接，最终生成转换特征向量，作为级联森林的输入.

2.2　级联森林　级联森林由多个级联层组成，每个级联层都包含两个RF和两个CRTF，每个RF和CRTF都包含m棵树，如图2所示. 针对每个级联层的分类，在RF中，每个节点随机选择特征，再使用基尼指数最大的特征作为该分裂节点划分的条件；同时，在CRTF中，每个节点随机选择特征，再使用该特征进行划分并生成子节点，直到每个叶子节点中只包含同一类样本. 每个森林通过叶子节点上不同种类样本的比例来估算类的分布，再针对森林中全部的树的输出结果取均值，得到每个森林的类分布向量，共得到四个森林的类分布向量. 因此，假设有m种分类的数据集便会得到$4m$维特征向量. 对上一层和当前层的结果进行拼接，作为下一层的输入. 在最后一层针对所有向量取均值，得到最终的类向量，将概率最大的类作为样本的预测结果. 为了避免模型过拟合，训练选择了10折交叉验证法，如果性能没有显著提升，训练过程会自动停止.

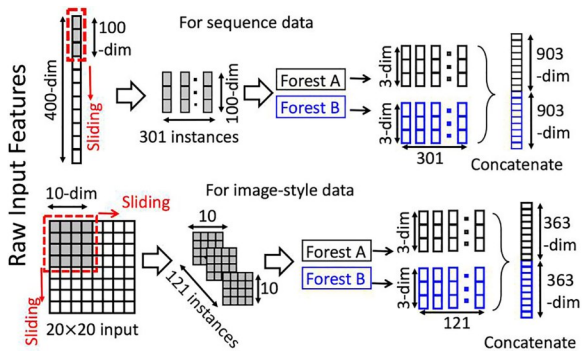

图1　多粒度扫描

Fig. 1　Multi granularity scanning

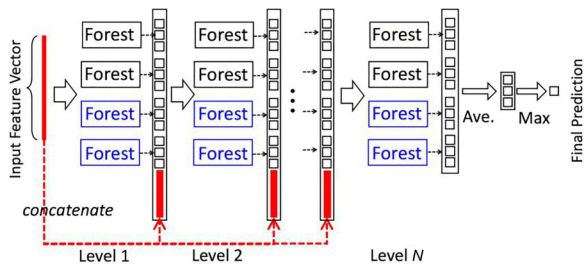

图2　级联森林的结构

Fig. 2　The structure of Cascading Forest

3　基于Lightgbm和XGBoost算法的优化深度森林算法

DF算法在一些小规模数据集上的精准度还有提升空间，而采用多粒度扫描也增加了模型的复杂度，DF模型中RF的决策树很多时，训练需要的空间会加大，时间会加长. 针对以上问题，本文采用基于决策树的Lightgbm和XGBoost算法对深度森林模型的级联森林结构进行了优化.

Lightgbm采用单边梯度采样，减少大量只有小梯度的数据实例，在计算信息增益时只利用有高梯度的数据，和遍历所有特征值相比，节省了时间和空间的开销. 同时，采用互斥特征捆绑算法，将不会同时为非零值的特性进行融合绑定，降低特征数量，降低模型的时间复杂度. 最后，采用直方图算法，把连续的浮点特征值离散化成K个整数，同时构造一个宽度为K的直方图. 在遍历数

据的时候,将离散化后的值作为索引在直方图中累积统计量,由于只需保存特征离散化后的值,相比保存离散化之前的特征值,降低了内存消耗.

XGBoost算法通过不断地进行特征分裂来产生一棵树,而生成的树就是一个新函数,然后利用新函数拟合前一次函数产生的残差,达到提高精度的目的.除了拥有高精度外,XGBoost模型还具有泛化能力强、性能良好、不易过拟合等特点.

本文在级联森林结构中引入Lightgbm和XGBoost,将Lightgbm模型中的叶子数量设为100,单个叶子上数据的最小数量设为300,学习率设为0.05,每棵树的最大深度设为6,重新构建级联森林,如图3所示.算法的伪代码如下所示.

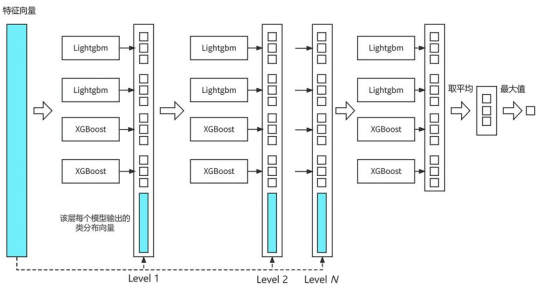

图3 LIGHT-XDF算法的示意图

Fig. 3 The schematic diagram of LIGHT-XDF

算法 优化的级联森林算法

Input: 训练集 $D=\{(X_1,Y_1),(X_2,Y_2),\cdots,(X_m,Y_m)\}$,每个样本$X_i=(X_1,X_2,\cdots,X_{id})$的$d$个特征,每个样本的标签$Y_i\in Y=\{1,2,\cdots,k\}$

Output: 级联森林模型

1.Begin

2. For $i=1$ to L do # L为4

3. If$(i==1\|i==2)$

4. 构建一个Lightgbm F_i,其中每棵树的叶子节点数为$n_1,n_2,\cdots,n_{300}$

5. 在F_i上进行训练,得到森林中每棵树的阈值$\{\theta_1,\theta_2,\cdots,\theta_t\}$

6. 在训练集D上进行预测,得到每个样本的得分S_i,用得分排序来选择前$k\%$的样本作为正样本D_{i+}

7. 从D_{i+}中有放回地随机采样m'个样本和相同数量的负样本D_{i-},形成新样本集D_i

8. Else

9. 构建一个XGBoost F_i,其中每棵树的叶子节点数为$n_1,n_2,\cdots,n_t$

10. 在F_i上进行训练,得到森林中每棵树的阈值$\{\theta_1,\theta_2,\cdots,\theta_t\}$

11. 在训练集D上进行预测,得到每个样本的得分S_i,用得分排序来选择前$k\%$的样本作为正样本D_{i+}

12. 从D_{i+}中有放回地随机采样m'个样本和相同数量的负样本D_{i-},形成新的样本集Di

13. End if

14. End for

15. 构建一个级联分类器G,其中每个分类器为F_i中的一个模型

16. 在训练集D上进行训练,得到G中每个分类器的阈值$\{\gamma_1,\gamma_1,\cdots,\gamma_L\}$

17. 输出级联森林模型,包含模型集合$\{F_1,F_2,\cdots,F_L\}$和分类器阈值集合$\{\gamma_1,\gamma_1,\cdots,\gamma_L\}$

18.End

基于LIGHT-XDF的成绩预测模型的工作流程如图4所示.首先对原始数据进行清洗,将数据按7:3的比例分为训练集和验证集,将训练集输入模型中的多粒度扫描进行特征选择;将选择后的数据输入优化的级联森林,数据会经过两个Lightgbm和两个XGBoost模型进行训练,同时进

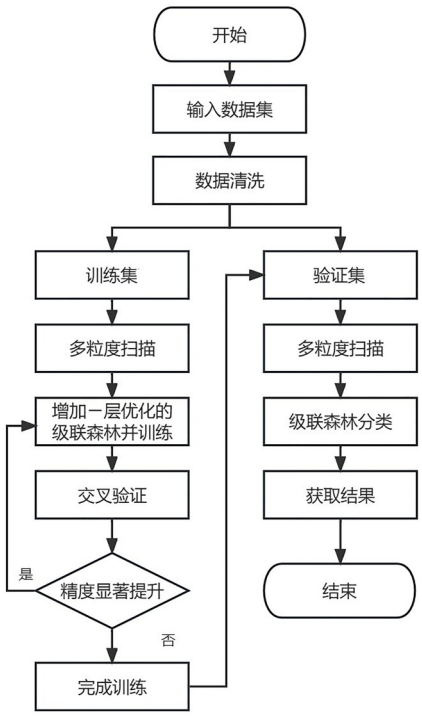

图4 基于LIGHT-XDF的成绩预测模型的工作流程

Fig. 4 Performance prediction model based on LIGHT-XDF

行10折交叉验证,此时,每个Lightgbm和XGBoost都生成不同比例的不同种类的样本;最后,对两个Lightgbm和两个XGBoost的输出结果取均值,将概率最大的类作为样本的预测结果;如果精度有显著提升,则再生成一个级联森林层,对上一层的结果和当前层的结果进行拼接,作为下一层的输入继续训练,直到模型精度无显著提升,结束训练;再将验证集输入模型,得到结果之后计算模型的各项指标.

4　实验分析

4.1　实验数据　基于UCI公开数据集进行实验,选择Dataset for Empirical Evaluation of Entry Requirements包含的Chemical Engineering,Electrical and Electronics Engineering(EEE),Mechanical Engineering以及Adult,Dry_Bean,Bank Marketing,Winequality-red和Winequality-white共计八个数据集,其中,Chemical Engineering,Electrical and Electronics Engineering,Mechanical Engineering为学生成绩数据集,其余为普通分类数据集,如表1所示.针对数据集中含有的文本数据进行one-hot编码处理.

4.2　评价指标及对比模型　为了验证模型的有效性,通过10折交叉验证进行实验,并采用精度、召回率、$F1$和AUC作为模型的评估指标.在选择对比模型时,考虑到DF模型是2017年提出的,故加入DF的优化模型ada-mdf[21]与lgb-df[22],在算法上类似深度学习,同时又加入了优化的深度学习模型ASTGNN[23]进行对比.最终,选择了RF,NB(Naive Bayes),KNN,SVM,Lightgbm,XGBoost,ASTGNN,DF,ada-mdf和lgb-df模型进行对比实验.

表1　实验使用的数据集

Table 1　Datasets used in experiments

数据集	样本数量	特征数	类别
Adult	48842	14	2
Dry_Bean	13611	17	7
Bank Marketing	45211	17	2
Winequality-red	1599	12	6
Winequality-white	4898	12	7
Chemical Engineering	188	8	4
EEE	348	8	4
Mechanical Engineering	161	8	4

4.3　实验结果与分析　为了验证LIGHT-XDF算法的有效性,将LIGHT-XDF与RF,NB,KNN,SVM,Gcforest,ada-mdf和lgb-df在八个数据集上做对比实验,实验结果如表2~6所示,黑体字表示结果最优,下画线表示结果次优.

算法精度,LIGHT-XDF在三个数据集上表现最优,在同一数据集上和表现第二的算法相比,最高提升1%;在三个数据集上的表现为第二,与表现最优的算法平均相差仅1.5%.

召回率,LIGHT-XDF在两个数据集上表现最优,在同一数据集上和表现第二的算法相比,最高提升0.011;在四个数据集上的表现为第二,与表现第一的算法平均相差仅0.05.

$F1$,LIGHT-XDF在三个数据集上表现最优;在另外三个数据集上的表现为第二,与表现第一的算法平均相差在0.01内.

表2　LIGHT-XDF和十种对比算法在八个数据集上的算法精度比较

Table 2　Classification accuracy of LIGHT-XDF and other ten algorithms on eight datasets

数据集	RF	NB	KNN	SVM	Lightgbm	XGBoost	ASTGNN	ada-mdf	lgb-df	Gcforest	LIGHT-XDF
Chemical Engineering	**98.25%**	92.98%	50.88%	54.39%	95.74%	**98.25%**	97.92%	95.75%	**98.25%**	96.49%	**98.25%**
EEE	95.24%	92.38%	56.19%	53.33%	94.29%	97.14%	96.25%	96.55%	94.29%	98.10%	**99.05%**
Mechanical Engineering	93.88%	87.76%	44.90%	42.86%	91.84%	95.92%	91.07%	87.81%	**96.42%**	93.88%	95.92%
Adult	84.33%	81.10%	84.33%	79.36%	86.94%	87.14%	**94.25%**	85.40%	86.91%	86.29%	86.95%
Dry_Bean	92.63%	76.35%	71.79%	63.37%	92.87%	92.61%	91.75%	90.74%	**93.05%**	91.31%	**93.05%**
Bank Marketing	90.68%	84.70%	88.37%	88.56%	91.04%	90.71%	**92.61%**	90.59%	**91.07%**	90.75%	90.87%
Winequality-red	66.46%	55.21%	51.25%	51.67%	65.21%	66.25%	**72.36%**	64.00%	67.71%	63.33%	68.54%
Winequality-white	66.12%	45.58%	46.94%	43.27%	**67.01%**	66.05%	63.74%	58.94%	65.03%	57.21%	66.40%

AUC,LIGHT-XDF 在四个数据集上表现最优,在同一数据集上和表现第二的算法相比,最高提升 0.011;在两个数据集上的表现为第二,与表现第一的算法平均相差在 0.01 内.

时间效率,LIGHT-XDF 与其他模型相比还有一定的提升空间,如表 6 所示,这也为后续改进指明了方向.

综上,在精度、召回率、*F*1、*AUC* 四个指标中,LIGHT-XDF 模型在一半以上的数据集上表现最好,在剩余的大部分数据集中的表现也是第二名,并且和第一名的差距很小,在 0.01 之内. 证明本文提出的 LIGHT-XDF 模型的综合性能优于其他模型.

表 3 LIGHT-XDF 和十种对比算法在八个数据集上的算法召回率比较

Table 3 Model recall rates of LIGHT-XDF and other ten algorithms on eight datasets

数据集	RF	NB	KNN	SVM	Lightgbm	XGBoost	ASTGNN	ada-mdf	lgb-df	Gcforest	LIGHT-XDF
Chemical Engineering	0.875	0.808	0.249	0.251	0.753	0.875	0.861	**0.991**	0.875	0.866	0.881
EEE	0.75	0.783	0.353	0.253	0.721	0.901	0.9	0.911	0.729	0.833	**0.922**
Mechanical Engineering	0.75	0.714	0.327	0.25	0.66	0.833	0.796	0.75	**0.842**	0.75	0.833
Adult	0.769	0.798	0.777	0.62	0.793	0.798	**0.884**	0.775	0.792	0.783	0.798
Dry_Bean	0.936	0.762	0.716	0.624	0.934	0.936	0.91	0.913	0.939	0.92	**0.94**
Bank Marketing	0.683	0.715	0.621	0.73	**0.732**	0.73	0.697	0.685	0.725	0.695	0.714
Winequality-red	0.367	0.371	0.254	0.205	0.36	0.36	0.343	0.284	**0.379**	0.285	0.373
Winequality-white	0.363	0.312	0.223	0.145	0.378	0.393	**0.413**	0.309	0.362	0.254	0.354

表 4 LIGHT-XDF 和十种对比算法在八个数据集上的算法的 *F*1 比较

Table 4 *F*1 of LIGHT-XDF and other ten algorithms on eight datasets

数据集	RF	NB	KNN	SVM	Lightgbm	XGBoost	ASTGNN	ada-mdf	lgb-df	Gcforest	LIGHT-XDF
Chemical Engineering	**0.980**	0.927	0.432	0.383	0.937	0.979	0.963	0.938	**0.980**	0.963	**0.980**
EEE	0.938	0.920	0.531	0.371	0.942	0.973	0.988	0.959	0.939	0.976	**0.990**
Mechanical Engineering	0.910	0.851	0.442	0.257	0.899	0.949	0.934	0.825	**0.958**	0.910	0.950
Adult	0.839	0.819	0.841	0.756	0.865	0.866	**0.924**	0.848	0.864	0.857	0.864
Dry_Bean	0.926	0.761	0.716	0.600	0.928	0.926	**0.942**	0.908	0.931	0.913	0.931
Bank Marketing	0.897	0.858	0.870	0.834	**0.905**	0.900	0.887	0.894	**0.905**	0.899	0.901
Winequality-red	0.646	0.539	0.490	0.471	0.638	0.647	0.625	0.616	0.663	0.606	**0.668**
Winequality-white	0.647	0.446	0.455	0.309	**0.661**	0.652	0.612	0.567	0.641	0.547	0.651

表 5 LIGHT-XDF 和十种对比算法在八个数据集上的算法 *AUC* 比较

Table 5 *AUC* of LIGHT-XDF and other ten algorithms on eight datasets

数据集	RF	NB	KNN	SVM	Lightgbm	XGBoost	ASTGNN	ada-mdf	lgb-df	Gcforest	LIGHT-XDF
Chemical Engineering	**0.988**	0.953	0.673	0.696	0.971	**0.988**	0.946	0.972	**0.988**	0.977	**0.988**
EEE	0.968	0.949	0.708	0.689	0.961	0.980	0.975	0.977	0.962	0.987	**0.994**
Mechanical Engineering	0.959	0.918	0.633	0.619	0.945	0.972	**0.982**	0.919	0.972	0.959	0.973
Adult	0.766	**0.809**	0.777	0.620	0.797	0.798	0.742	0.769	0.793	0.784	0.790
Dry_Bean	0.957	0.862	0.835	0.786	0.958	0.956	0.892	0.946	**0.959**	0.949	**0.959**
Bank Marketing	0.684	0.716	0.622	0.505	**0.734**	0.730	0.674	0.672	0.726	0.695	0.708
Winequality-red	0.799	0.731	0.707	0.710	0.791	0.797	0.751	0.784	0.806	0.780	**0.811**
Winequality-white	0.802	0.683	0.690	0.669	**0.807**	0.801	0.751	0.760	0.796	0.750	0.804

表6　LIGHT-XDF和十种对比算法在八个数据集上的算法执行时间的比较(s)

Table 6　Execution time (s) of LIGHT-XDF and other ten algorithms on eight datasets

数据集	RF	NB	KNN	SVM	Lightgbm	XGBoost	ASTGNN	ada-mdf	lgb-df	Gcforest	LIGHT-XDF
Chemical Engineering	0.604	**0.063**	0.443	0.072	0.484	0.531	13.250	3.083	1.450	3.060	1.785
EEE	0.545	**0.074**	0.538	0.080	0.547	0.566	15.274	3.262	1.860	3.186	1.944
Mechanical Engineering	0.509	**0.059**	0.471	0.068	0.568	0.595	14.723	3.046	1.615	3.077	1.783
Adult	13.557	**12.636**	14.910	26.48	13.963	13.104	308.820	71.795	16.002	89.210	39.752
Dry_Bean	3.848	4.074	2.615	4.262	**2.544**	4.106	254.230	28.299	19.409	61.214	60.437
Bank Marketing	15.775	12.358	17.868	12.32	**12.280**	15.303	325.780	115.705	15.180	70.722	36.411
Winequality-red	0.719	**0.147**	0.567	0.215	0.872	0.974	43.152	5.672	5.539	5.475	8.712
Winequality-white	1.212	**0.374**	0.839	0.887	1.111	1.205	68.345	16.897	8.460	15.932	24.595

5　结论

在学期中对学生成绩进行预测具有重要意义,但是目前的预测模型比较单一,且精度不高.本文提出一个基于Lightgbm和XGBoost的优化深度森林算法LIGHT-XDF,在级联森林中引入Lightgbm和XGBoost,利用XGBoost提高精度,利用Lightgbm降低模型复杂度.使用LIGHT-XDF算法,以学生的行为数据和社会数据为输入,预测学生的期末成绩等级,实验结果表明,与其他模型相比,LIGHT-XDF在绝大部分数据集上的综合性能最好.

未来将进行特征选择研究,进一步提高模型的精度和时间效率.

参考文献

[1] Zhang Y Z, Xiong R, He H W, et al. Lithium-ion battery remaining useful life prediction with Box-Cox transformation and Monte Carlo simulation. IEEE Transactions on Industrial Electronics, 2019, 66(2): 1585－1597.

[2] 肖迁,焦志鹏,穆云飞,等.基于Lightgbm的电动汽车行驶工况下电池剩余使用寿命预测.电工技术学报,2021,36(24):5176－5185. (Xiao Q, Jiao Z P, Mu Y F, et al. Lightgbm based remaining useful life prediction of electric vehicle lithium-ion battery under driving conditions. Transactions of China Electro-technical Society, 2021, 36(24): 5176－5185.)

[3] Platt J C. Sequential minimal optimization: A fast algorithm for training support vector machines. Microsoft Research, 1998, DOI:US4558132 A.

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

[5] Chen T Q, Guestrin C. XGBoost: A scalable tree boosting system//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, CA, USA: ACM, 2016: 785－794.

[6] Harackiewicz J M, Barron K E, Tauer J M, et al. Short-term and long-term consequences of achievement goals: Predicting interest and performance over time. Journal of Educational Psychology, 2000, 92(2): 316－330.

[7] Nghe N T, Janecek P, Haddawy P. A comparative analysis of techniques for predicting academic performance//2007 37th Annual Frontiers in Education Conference, Global Engineering: Knowledge without Borders, Opportunities without Passports. Milwaukee, WI, USA: IEEE, 2007: T2G-7－T2G-12.

[8] Zimmermann J, Brodersen K H, Heinimann H R, et al. A model-based approach to predicting graduate-level performance using indicators of undergraduate-level performance. Journal of Educational Data Mining, 2015, 7(3): 151－176.

[9] Romero C, López M I, Luna J M, et al. Predicting students′ final performance from participation in on-line discussion forums. Computers & Education, 2013(68): 458－472.

[10] Marbouti F, Diefes-Dux H A, Madhavan K. Models for early prediction of at-risk students in a course using standards-based grading. Computers & Education, 2016(103): 1－15.

[11] Asif R, Merceron A, Ali S A, et al. Analyzing

undergraduate students′ performance using educational data mining. Computers & Education, 2017(113):177－194.

[12] Helal S, Li J Y, Liu L, et al. Predicting academic performance by considering student heterogeneity. Knowledge-Based Systems, 2018(161):134－146.

[13] Polyzou A, Karypis G. Feature extraction for next-term prediction of poor student performance. IEEE Transactions on Learning Technologies, 2019, 12(2): 237－248.

[14] Haridas M, Gutjahr G, Raman R, et al. Predicting school performance and early risk of failure from an intelligent tutoring system. Education and Information Technologies, 2020, 25(5):3995－4013.

[15] 张政庭,周恒宇,崔璀,等. 基于K-means-BP神经网络的高校专业核心课程成绩预测. 中国医学教育技术, 2023, 37(2): 212－217, 228. (Zhang Z T, Zhou H Y, Cui C, et al. Performance prediction of professional core course based on K-means and BP neural network. China Medical Education Technology, 2023, 37(2):212－217, 228.)

[16] 许欢,夏道明. 基于机器学习方法的学生成绩的点预测和区间预测. 信息工程大学学报, 2023, 24(2): 177－182. (Xu H, Xia D M. Point prediction and interval prediction of student achievement based on machine learning. Journal of Information Engineering University, 2023, 24(2):177－182.)

[17] 王洪亮,赵圆圆. 学生行为数据与学业成绩的关系研究——基于离群点检测算法. 石家庄职业技术学院学报, 2023, 35(2):35－40. (Wang H L, Zhao Y Y. A correlational study on the student behavior data and academic performance: Based on outlier detection algorithm. Journal of Shijiazhuang Vocational Technology Institute, 2023, 35(2):35－40.)

[18] Zhou Z H, Feng, J. Deep forest . National Science Review, 2019, 6(1):74－86.

[19] AlJame M, Imtiaz A, Ahmad I, et al. Deep forest model for diagnosing COVID-19 from routine blood tests. Scientific Reports, 2021, 11(1):16682.

[20] Dong L, Qi J F, Yin B S, et al. Reconstruction of subsurface salinity structure in the south China sea using satellite observations: A Lightgbm-based deep forest method. Remote Sensing, 2022, 14(14):3494.

[21] Su R, Liu X Y, Wei L Y, et al. Deep-Resp-Forest: A deep forest model to predict anti - cancer drug response. Methods, 2019(166):91－102.

[22] 滕玲,施三支,张梦菲,等. 基于深度森林的高校贫困生认定模型研究. 长春理工大学学报(自然科学版), 2022, 45(3): 131－137. (Teng L, Shi S Z, Zhang M F, et al. Research on the identification model of poor college students based on deep forest. Journal of Changchun University of Science and Technology (Natural Science Edition), 2022, 45(3):131－137.)

[23] Guo S N, Lin Y F, Wan H Y, et al. Learning dynamics and heterogeneity of spatial-temporal graph data for traffic forecasting. IEEE Transactions on Knowledge and Data Engineering, 2022, 34(11): 5415－5428.

(责任编辑 杨可盛)