DOI:10.19651/j.cnki.emt.2107179

基于 LGB-FFM-LR 算法的在线课程评分预测方法研究*

刘昱萌 刘 斌
(陕西科技大学 电子信息与人工智能学院 西安 710021)

摘 要:针对在线教育课程客观评价较差的问题,设计了基于决策树算法的梯度提升算法-场感知因式分解机-逻辑回归(LightGBM-FFM-LR)算法的评分预测模型。该模型采集在线课程观看历史数据,提取用户的通用特征、时间特征等特征值,并着重考虑特征值的高维特征和低维特征关系来实现多维特征组合,改善数据稀疏性,从而提升评分预测性能。通过对某在线课程网站的脱敏数据实验表明,该模型的评分预测值与评分实际值的决定系数为 0.87,平均均方误差为 0.42,提升了模型的泛化能力,对在线课程的预测评分结果更加客观真实。

关键词:在线课程;评分预测;LightGBM;FFM;LR
中图分类号:TP391;TP18 **文献标识码**:A **国家标准学科分类代码**:510.1050

Prediction methods of online course grading based on LGB-FFM-LR

Liu Yumeng Liu Bin
(School of Electronic Information and Artificial Intelligence, Shaanxi University of Science & Technology, Xi'an 710021, China)

Abstract: For the problem of poor objective evaluation of online education courses, a prediction model based on light gradient boosting machine-field-aware factorization machine-logistic regression is designed. The model collects online course viewing history data, extracts users' generic features, temporal features and other feature values, and focuses on the relationship between high-dimensional features and low-dimensional features of feature values to achieve multi-dimensional feature combinations and improve data sparsity, so as to improve rating prediction performance. Based on masked data test on an online course website, the determination coefficient between predicted grading and actual one in this model is 0.87, with 0.42 of average mean square error, improved model generalization capability, this model serves more objective and realistic predicted grading to online course.

Keywords: online course;grading prediction;LightGBM;FFM;LR

0 引 言

在线教育资源从文字、图片逐渐转变为音频、视频、在线直播等多媒体形式。疫情期间,在线课程占据主要教学地位,各大教育机构网课销量直线上升[1-3]。越来越多的用户开始关注课程的用户评分,用户评分是对课程属性特征及满意度的综合评估,用户首先会关注评分较高的课程[4-5]。

随着在线课程的迅速发展,各种刷分、刷评论现象层出不穷,用户评分不能真实反映课程质量。学者开始关注用户的隐式反馈信息,观看课程的有效时间反映真实观看行为,可以更加具体、准确地表达出用户的偏好。Hong 等[6]提出流式数据特征,采用 K 最近邻(K-nearst neighbors)算法评估课程质量,准确度达到 0.72;Brinton 等[7]提出时间特征,采用随机森林、决策树算法训练分类器,评估特征有效性,准确率达到 0.73。在评分预测算法研究方面,徐日等[8]基于自适应增强算法(adaptive boosting,Adaboost)算法框架将评分预测问题转化为二分类问题,基于矩阵分解模型得到的评分预测框架在精准度上有明显提高;陆君之[9]将随机森林回归算法应用于评分预测模型,以电影实际评分为参考数据,随机森林回归算法得到的评分预测模型在性能上明显优于其他预测算法,预测时的相对误差也远远低于其他算法。杨贵军等[10]基于潜在狄利克雷分配模型(latent dirichlet allocation,LDA)模型,量化用户评论为主题特征向量作为解释变量,将用户评分作为被解释变量,采用分布式梯度增强算法(eXtreme gradient boosting,

收稿日期:2021-07-06
* 基金项目:国家自然科学基金(61871260)项目资助

XGBoost)，并加入样本扰动和属性扰动生成多个模型进行集成，构建用户评分预测模型。丁勇等[11]提取并融合元数据和评分数据的相似性权重，构建同质关系网络，提出一种融合网络表示学习与 XGBoost 算法的评分预测模型。综上，当前学者对于评分的研究侧重于关注用户的隐式反馈特征或显示反馈特征，模型建立较为单一，准确率和真实性远远不够。

当前课程评分存在用户评分不真实、主观性较强，特征提取单一，评分方法准确度较低等问题。为了解决上述问题，本文在回顾已有方法的基础上，融合用户的隐式反馈特征和显示反馈特征，充分挖掘用户特征，基于 LightGBM-FFM-LR 模型预测课程评分，结合用户真实观看数据分析评分预测的真实性，并与其余学者的模型对比模型泛化能力和预测性能，证实评分结果更加真实客观，具有更高的参考价值。

1　解决方案

1.1　算法研究

在阅读大量文献后，借鉴电影评分算法和广告点击率（click through rate，CTR）以及广告转化率（click conversion rate，CVR）算法，结合本文数据集特点，提出基于决策树算法的梯度提升算法-场感知因式分解机-逻辑回归（light-gradient boosting machine-field-aware factorization machine-logistic regression，Light GBM-FFM-LR）融合算法的评分模型。

轻度提升机[12-17]（light gradient boosting machine，LightGBM）和特征域感知因子分解机模型[18-22]（field-aware factorization machines，FFM）在广告点击率和广告转化率预测模型中得到了很好的预测效果。李雄飞等[23]考虑到当前被转化的广告数据较少，现有融合模型侧重于分析高维组合特征，少有模型考虑到低阶特征间的关系，且难以达到平衡。因此提出了 LightGBM-FFM 融合模型，侧重于考虑高维特征和低维特征之间的关系，利用 FFM 模型实现多维特征组合，提升 CVR 和 CTR 的预估性能。刘金梅等[24]提出评分填充和时间算法，改善数据稀疏性，提升评分预测性能。评分预测问题是推荐系统的核心问题，数据稀疏性和冷启动问题首当其冲。当前评分预测问题多使用文本特征及用户评分数据，但存在多数用户未评分情况，数据稀疏性问题难以解决，在综合考虑当前问题之后，借鉴 CTR 和 CVR 相关模型以及当前评分预测模型，提出 LightGBM-FFM-LR 算法模型。

1.2　LightGBM-FFM-LR 模型

如图 1 所示，本文采用的模型由 LightGBM 模型、FFM 模型和 LR 模型组成。考虑到特征选择对于最终评分结果有较大的影响，LightGBM 模型使用了基于直方图的决策树算法，将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。遍历数据时根据离散化后的值作为索引在直方图中累积统计量，然后根据直方图的离散值，遍历寻找最优的分割点，大大加快了模型的训练速度。因此选用 LightGBM 模型对样本进行特征选择和特征排序将特征重要性高的样本构成特征矩阵，将其与其余特征一起送入 FFM 模型进行特征组合。FFM 模型将高度相关的特征进行特征组合，构建特征工程，提升模型预测能力。逻辑回归模型(logistic regression，LR)训练特征数据得到逻辑回归方程，计算评分概率。

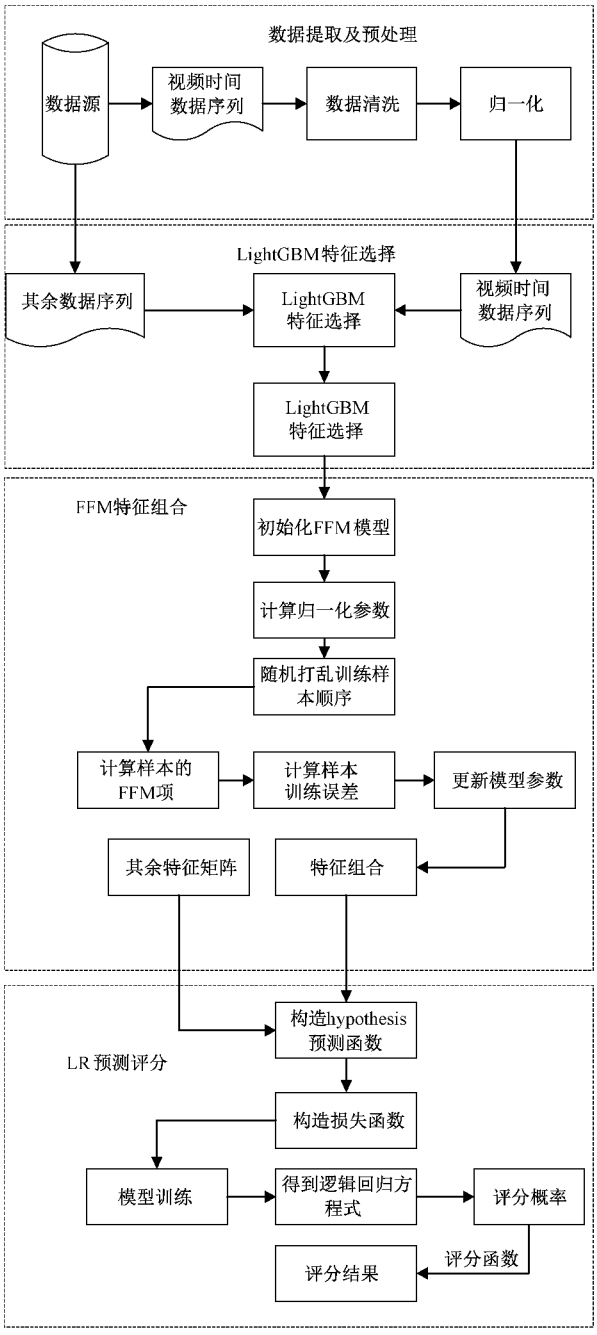

图 1　模型架构

1）特征排序和特征选择。对 49 门课程共计 10 356 条有效数据进行特征排序和特征选择。

采用单边梯度采样(gradient-based one-side sampling)算法,排除大部分梯度小的样本,进行特征降维处理。采用直方图算法和带深度限制的 Leaf-Wise-Learning 算法进行特征重要性排序和特征选择。最终选取对模型贡献度较高的 17 个特征,如图 2 所示。

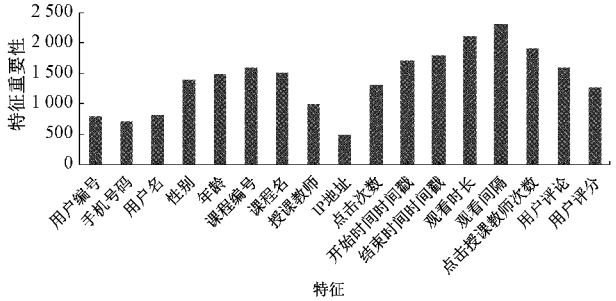

图 2 特征重要性排序

借鉴 CVR 和 CTR 算法思想,高质量特征有助于提高模型准确性和性能,特征工程采用组合一阶离散特征的方法得到高阶组合特征,提升复杂关系的拟合能力。比如:将年龄、性别、地区、时间特征等进行组合,可以体现出不同地区用户观看课程的类型和观看难度以及性别年龄偏好。组合特征可以反映出更多的观看特征和社会现象。相较而言,组合后的特征维度高、可解释性强,如表 1、2 所示。

表 1 组合前特征

组合前单一特征
年龄
性别
地区
课程名
观看时长
……

表 2 组合后特征

组合后多样特征		
集中学习,学习状况较好,课程难度较大场景	兴趣学习,学习自觉性适中场景	集中学习,计算机偏好,课程难度较大场景
6～14 岁;中小学生;河南地区;奥数视频;观看时长>4 h。	18～25 岁;女大学生;陕西地区;Flash 动画制作;观看间隔<5 h。	30～45 岁;男性从业人员;北京地区;C++语言高阶学习;观看时长>10 h,观看间隔>3 h。

2) 构建特征工程。LightGBM 模型输出的特征矩阵均为离散型数据,导致数据变得很稀疏。FFM 模型引入域(ficld)概念,同一特征针对不同 field 使用不同隐向量,模型建模更加准确。本文中,将用户性别、年龄、观看地区划分为同一个 field,同一门课程的观看时间、观看间隔化为同一个 field。不同 field 的特征交义,可以学习到不同的隐含含义,从而挖掘到隐含的特征关系,构建高阶特征工程,提升模型准确性。

$$y(x)=w_0+\sum_{i=1}^{n}w_ix_i+\sum_{i=1}^{n}\sum_{i=j+1}^{n}<V_{i,f_j},V_{j,f_i}>x_ix_j \tag{1}$$

其中,n 表示特征个数;f_j 表示第 j 个特征对应的特征域;V_{i,f_j} 表示特征 x_i 在特征域 f_j 中挑出的隐向量。

3) 构建 LR 模型,进行评分预测。考虑到 LR[25-28]模型在数据缺失或特征空间较大时表现效果较差,因此前两步着重对特征进行处理。LR 模型在时间和内存需求上十分高效,而且对于数据中的噪声鲁棒性较好,因此采用 LR 模型进行评分。将 FFM 模型的特征组合及其余特征矩阵输入至 LR 模型,并对其进行训练得到逻辑回归方程。

$$\log\left(\frac{p}{1-p}\right)=\beta_0+\beta_1x_1+\cdots+\beta_nx_n \tag{2}$$

其中,n 为输入样本数量;$\beta_0,\beta_1,\cdots,\beta_n$ 为回归系数;$x_0,x_1,\cdots,x_n$ 为样本指标变量;p 为时间-视频概率,进行加权平均得出评分。课程的评分可以体现出不同地区、不同类型课程的受众人群和适应程度,有利于课程的进一步改革和推广。

2 数据采集和预处理

2.1 数据采集

近年来,各大教育机构积极打造在线课程学习网站,形成了学而思、新东方、慕课(MOOC)、沪江网校等在线学习平台。本实验选取某在线课程网站脱敏数据作为数据源,部分数据通过网络爬虫获取,选取用户的基本信息、课程信息、观看课程日志、位置时间、教师信息等关键指标。基于 Python 语言编写爬虫程序,采用 Scrapy 框架抓取数据。Scrapy 是一个爬取网站数据、提取结构性数据的应用程序框架。在抓取过程中更易伪装为普通用户访问行为,Scrapy 采用异步式处理请求,可以灵活调节并发量,抓取速度快。数据采集的爬虫流程如图 3 所示。

2.2 数据预处理

1)异常值处理

每个用户在数据库中都具有唯一标识符和其相对应的视频观看日志,将其匹配组成用户视频对。异常值分为如下情况:

(1) 清除没有课程评分的用户视频对,删除有评分但观看轨迹不完整的用户视频对。

(2) 删除至少有一个特征为“空”的用户视频对。

(3) 对重复值进行删除,缺失值不作处理。

(4) 用户未观看完课程、观看时长过短所产生的数据,设置阈值检测当前用户观看的时长是否满足视频总时长的 2/3。

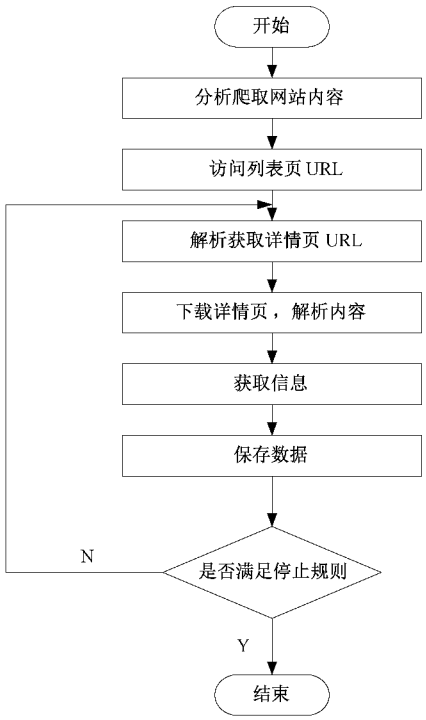

图 3　数据采集爬虫流程

(5) 由于电脑设备故障产生重复点击的问题，视频点击频率在 30～60 min 内的点击次数，视为有效数据。

(6) 使用 boxplot 工具处理异常值。

本文采用箱线图进行异常值检测，箱线图可以根据数据分布情况辨别出异常值，一般将小于下边界的数值与大于上边界的数值称为异常值。异常值对于评分结果会产生较大影响，因此将其移除，可用作分析课程情况。

2)归一化处理

从观看日志特征中提取观看时长、观看频率和观看间隔 3 种数据，这些记录体现出用户观看视频的真实时间行为。3 种特征的具体信息如表 3 所示。

表 3　特征信息

序号	特征维度		
	特征名	特征描述	长度
1	视频观看时长 $T_i, i \in [1,24]$	整个学习期间观看视频时间长度	24
2	视频观看频率 $F_i, i \in [1,24]$	整个学习期间，第 i 小时内点击视频的频率	24
3	视频观看间隔 $W_i, i \in [1,8]$	真实观看时间内，第 i 次课和第 $i+1$ 次课的间隔时间	8

3 个特征值之间相差较大，数值较高的特征会在综合分析中占据主导地位，削弱数值较低的特征值的作用。为了保证结果的可靠性，需要对原始数据进行标准化处理。采用最小-最大规范化(Min-Max Normilization)方法对数据进行标准化，如式(3)所示，再从经过预处理和清洗之后的数据集中提取用户学习行为特征。

$$z_i = \frac{\max(x) - x_i}{\max(x) - \min(x)} \tag{3}$$

其中，$\max(x)$为样本数据最大值；$\min(x)$为样本数据最小值；x_i 表示待处理的第 i 个数据。Min-Max Normilization 是对原始数据的线性变换，保留数据关系，将数值映射到[0,1]区间，消除量纲和数据取值范围影响，避免了数值过高/低的特征引发的数值问题，经过标准化的数据可以提升模型预测评分的性能。

3　实验结果及分析

3.1　模型训练

以量化后的 10 356 条有效数据进行实验，利用多种预测评分算法构造模型，本次实验主要选择了随机森林算法、XGBoost 算法、LightGBM 算法、LightGBM-LR 算法这 4 种算法作为对比实验。

3.2　评价指标

本文选择了 3 种评价指标评判模型性能，分别是平均绝对误差(MAE)、均方根误差($RMSE$)、决定系数(R^2)。

$$MAE = \frac{1}{m}\sum_{i=1}^{m} | \hat{y}_i - y_i | \tag{4}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{m_1}} \tag{5}$$

$$R^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{6}$$

其中，m 为样本容量；y_i 代表第 i 个样本的真实值；$\hat{y}_i$ 代表第 i 个样本的预测值；$\bar{y}$ 代表真实值的平均值。MAE 反映模型整体误差情况；$RMSE$ 反映模型预测精度，两者值越小，说明模型效果越好；R^2 说明预测值与实际数据的相关程度，值越接近 1，说明模型效果越好。

3.3　结果分析

1)算法对比

本实验将数据以 7∶3 的比例划分为训练集和测试集，采用 5 种预测评分算法对训练数据集进行模型训练，并通过测试集验证模型的有效性。

根据表 4 可知，对比其他算法，本文所采用的方法在 3 种评价指标中均有不同程度的提升，模型预测效果更好。本文的预测模型，旨在挖掘特征关联性，融合显示反馈特征和隐式反馈特征，通过集成算法提升模型预测性能，相比传统预测算法，模型误差降低了 3.1%，模型性能提升了 1.8%。本文选用时间特征结合通用特征进行训练，更加符合用户的真实观看行为，进一步提升了预测结果的可信度。

表 4　算法对比

模型	*MAE*	*RMSE*	R^2
随机森林	0.66	0.73	0.69
XGboost	0.60	0.69	0.72
LightGBM	0.53	0.65	0.76
LightGBM-LR	0.47	0.57	0.80
LightGBM-FFM-LR	0.35	0.42	0.87

2)视频预测评分

(1) 根据提取的入模指标,经模型训练并通过式(2)计算时间-视频概率 $p_1,p_2,\cdots,p_n$。

(2) 根据式(7)计算预测的加权评分,其中 R 的取值范围为 0～5,R 值越大,代表时间特征对视频最终评分影响越大。如表 4 所示为部分课程的预测评分和用户评分对比。

$$R=(w_1p_1+w_2p_2+\cdots+w_np_n)\times 5 \tag{7}$$

表 5 所示为部分课程的相关信息及预测评分,评分范围为 0～5 分。根据表中所列数据,用户评分与预测评分存在较大差异,与用户观看行为不完全相符。Flash 动画制作观看时长达 13 h,视频观看间隔较短,反映用户对该门课程的兴趣度较高,持续性学习意愿较强,课程质量达标,用户评分 2.3 分与用户实际观看行为不相符。表中预测评分较高的课程,经研究发现该门课程的教授教师关注度较高,课程的点击次数较为频繁。实验结果表明,单纯用户评分无法体现出课程的真实情况,参考度较低。本文所得评分更加贴近用户观看实况,对课程质量及用户选择具有更大的价值。

表 5　部分课程预测评分

课程名	观看时长/h	观看间隔/h	用户评分	预测评分
Flash 动画制作	13	2	2.3	4.5
计算机组成原理	2	5	4.6	2.6
一元二次函数	17	0.5	3.5	4.8
高等数学	8	1	3.2	4.3
人工神经网络	3	2	4.8	3.5
大学物理	6	3	4.1	3.2

3.4　对比实验

为了验证本文模型的有效性,采用公开数据集 MOOC 数据集做对比实验。如图 4 所示,对比文献[6]所采用的支持向量机(support vector machine, SVM)、随机森林和 C50 决策树算法,分别以 *MAE*、*RMSE*、R^2 作为评价标准。

由图 4 可知,本文算法对比其他 3 大算法,在 *RMSE*、*MAE* 误差方面均有所降低,R^2 的值更加接近 1。本文所选算法相比文献[4]中的算法,更能挖掘数据隐藏的规律,在特征选择和特征融合时可以较好地挖掘特征之间的关联性,在训练时模型的效果更好,预测结果更加接近真实评分。

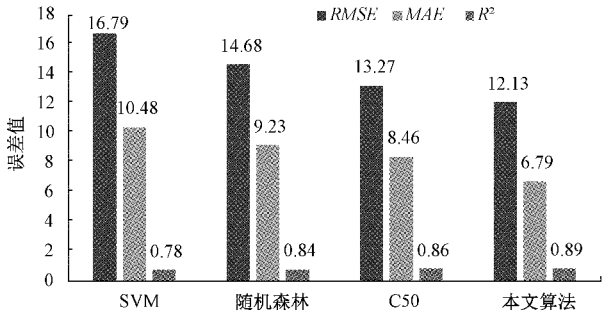

图 4　算法改进对比(%)

4　参数优化

4.1　LightGBM 模型参数

对于 LightGBM 参数调优,本文主要采用 Scikit-learn 提供的网格搜索(GridSearch)对模型的学习率、叶子节点总数、树的深度、特征建树比例 4 个主要参数进行调整。

在模型预估准确度方面,通常迭代次数越多模型的预估准确度越高,学习率越高模型准确度越高。但学习率过高,很容易造成模型过拟合。因此将训练集拆分,70%作为新的训练集,30%作为测试集。依据新训练集对模型进行参数调整,最终确定 4 个主要参数设置为:学习率 $learning\_rate=0.03$,叶子节点总数 $num\_leaves<=31$,树的深度 $max\_depth=6$,特征建树比例 $feature\_fraction=0.6$。

4.2　FFM 模型参数优化

FFM 模型对稀疏样本数据具有很好的效果,本文中,FFM 将低维特征进行组合同时结合其余低维特征,挖掘其关联关系,使模型训练效果更优。如式(8)所示,FFM 模型采用逻辑损失(logistic loss)函数作为损失函数,并加入 L2 正则项。

$$\sum_{i=1}^{n}\log(1+\exp(-y_i\varnothing(w,x_i)))+\frac{\lambda}{2}\|w\|^2 \tag{8}$$

$$\varnothing(w,x_i)=\sum_{i=1}^{n}\sum_{j=i+1}^{n}<w_{i,f_j},w_{j,f_i}>x_ix_j \tag{9}$$

其中,式(8)中 y_i 是第 i 个样本的 label;n 是训练集数量;λ 是 L2 正则项系数。式(9) $\varnothing(w,x_i)$ 为 FFM 去掉常数项和一次项,其中 w_{i,f_j} 是特征 i 对 field f_j 的隐向量,与上文中的 V_{i,f_j} 含义相同。

FFM 模型的参数估计,通常采用如下算法:随机梯度下降法(stochastic gradient descent,SGD)、交替最小二乘法(alternating least square,ALS)、马尔科夫链蒙特卡洛法由蒙特卡洛方法(Monte Carlo simulation,MC)和马尔科夫链(Markov chain,MC)组成。因 SGD 算法训练速度较快且复杂度低,因此选用 SGD 算法求解 FFM 模型参数,SGD 训练时采用单样本的损失函数计算梯度更新模型。模型主要分为两步,第 1 步梯度分步计算,提升计算效率;第 2 步

采用自适应更新学习率，随着迭代次数增加，参数的学习率随着历史梯度的累加而减少。各个参数可以较快达到最优，测试误差不会出现太大的波动。

5　结　　论

本文通过挖掘用户观看视频真实行为，融合用户显示反馈特征和隐式反馈特征，使用集成算法深度挖掘影响评分的特征，并使用多种评分预测算法构建模型进行对比实验，利用公开 MOOC 数据集对模型进行检验，结果说明采用 LightGBM-FFM-LR 算法得到的预测评分效果最好，在训练速度和训练精度方面也有所提升，得到的评分结果更具现实意义。从模型预测效果来看，本文的方法以及在参数优化方法上还有可以提升的空间，在今后的工作中将更加深入研究该评分预测算法，力求达到更好的预测效果。

参考文献

[1] 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.

[2] 郑庆华，董博，钱步月，等. 智慧教育研究现状与发展趋势[J]. 计算机研究与发展，2019，56(1)：209-224.

[3] 王逸凡，李国平. 基于语义相似度及命名实体识别的主观题自动评分方法[J]. 电子测量技术，2019，42(2)：84-87.

[4] 马志丽. 中小学校外在线教育的现状及教学模式研究[D]. 北京：北京邮电大学，2019.

[5] 詹玉广. 大型客运站人流量在线预测模型研究[J]. 国外电子测量技术，2020，39(11)：94-97.

[6] HONG B，WEI Z，YANG Y. Online education performance prediction via time-related features[C]. IEEE/ACIS International Conference on Computer & Information Science，IEEE，2017(5)：95-100.

[7] BRINTON C G，CHIANG M. MOOC performance prediction via clickstream data and social learning networks [C]. Computer Communications，IEEE，2015(10)：2299-2307.

[8] 徐日，张谧. 基于 Adaboost 算法的推荐系统评分预测框架[J]. 计算机系统应用，2017，26(8)：107-113.

[9] 陆君之. 基于随机森林回归算法的电影评分预测模型[J]. 江苏通信，2018，34(1)：75-78.

[10] 杨贵军，徐雪，赵富强. 基于 XGBoost 算法的用户评分预测模型及应用[J]. 数据分析与知识发现，2019，3(1)：118-126.

[11] 丁勇，陈夕，蒋翠清，等. 一种融合网络表示学习与 XGBoost 的评分预测模型[J]. 数据分析与知识发现，2020，4(11)：52-62.

[12] ZHANG Y Y，ZHU C F，WANG Q R. LightGBM-based model for metro passenger volume forecasting[J]. IET Intelligent Transport Systems，2021，14(13)：1815-1823.

[13] 顾桐，许国良，李万林，等. 基于集成 LightGBM 和贝叶斯优化策略的房价智能评估模型[J]. 计算机应用，2020，40(9)：2762-2767.

[14] 佐磊，胡小敏，何怡刚，等. 小样本数据处理的加速寿命预测方法[J]. 电子测量与仪器学报，2020，34(11)：26-32.

[15] WANG F J，CHENG H L，DAI H L，et al. Freeway short-term travel time prediction based on LightGBM algorithm [J]. IOP Conference Series Earth and Environmental Science，2021，2489(1)：97-104.

[16] GU J S. A novel credit risk assessment model based on LightGBM[J]. Journal of Simulation，2020，8(3)：71-73.

[17] 国强强，朱振方. 基于 LightGBM 算法的移动用户信用评分研究[J]. 计算机技术与发展，2020，30(9)：210-215.

[18] 邓秀勤，谢伟欢，刘富春，等. 基于特征工程的广告点击转化率预测模型[J]. 数据采集与处理，2020，35(5)：842-849.

[19] 王昊，刘震. 基于信息感知权重和误差预测的时间序列在线预测方法[J]. 仪器仪表学报，2020，41(11)：31-41.

[20] 孙晓燕，聂鑫，暴琳，等. 基于改进 Wide&Deep 交互特征提取的移动 APP 转化率预估[J]. 郑州大学学报(工学版)，2020，41(6)：26-32.

[21] ZHANG L，SHEN W，LI S，et al. Field-aware neural factorization machine for click-through rate prediction[J]. IEEE Access，2019，7：75032-75040.

[22] CHEN J H，LI X Y，ZHAO Z Q，et al. A CTR prediction method based on feature engineering and online learning [C]. International Symposium on Communications & Information Technologies，IEEE，2017：1-6.

[23] 李雄飞，周晋男，张小利. 基于混合模型的广告转化率问题研究[J]. 东北大学学报(自然科学版)，2019，40(7)：942-947.

[24] 刘金梅，舒远仲，张尚田. 基于评分填充和时间的加权 Slope One 算法[J]. 计算机技术与发展，2021，31(1)：35-42.

[25] 边玉宁，陆利坤，李业丽，等. 基于逻辑回归的金融风投评分卡模型实现[J]. 计算机科学，2020，47(z2)：116-118.

[26] 周围. 基于 LightGBM-Logistic 回归的网贷个人信用评分模型研究[D]. 杭州：浙江工商大学，2018.

[27] LU P，WANG H，TOLLIVER D，et al. Prediction of bridge component ratings using ordinal logistic regression model [J]. Mathematical Problems in Engineering，2019，2019：1-11.

[28] 刘明昌. 豆瓣网站电影在线评分的混合预测模型研究[D]. 河北：河北大学，2017.

作者简介

刘昱萌，硕士研究生，主要研究方向为数据挖掘、大数据分析等。

E-mail：1660475552@qq.com

刘斌，硕士，副教授，主要研究方向为数据挖掘、大数据分析、人工智能等。

E-mail：Liubin@sust.edu.cn