# 基于特征工程的学习分析研究

汪红霞

(安徽新华学院 大数据与人工智能学院,安徽 合肥 230088)

**摘 要**:针对当前在线教育产生的大量数据,首先,使用特征工程技术对在线学习者的学习数据进行深层次的行为提取;其次,采用逻辑回归分类算法对学习数据进行建模;最后,结合机器学习作进一步训练和验证,得到性能更优的预测模型。实验表明,使用该模型对学习效果进行分析,能够为老师的教学和学生的学习提供较好的指导。

**关键词**:学习行为分析;特征工程;逻辑回归

**中图分类号**:TP391 **文献标识码**:A **文章编号**:2095-7726(2021)12-0040-04

为了应对新冠疫情,社会对教育信息化的要求更高,利用多种信息化手段进行在线学习已经成为全球教育教学的一种主要方式[1]。相比于传统的线下学习,在线教育和学习软件会产生海量的教育数据,如何利用特征工程分析技术去发掘这些教育数据的潜在价值是国内外高等教育的一个重要研究方向。

通过学习分析技术,研究学习者在学习过程中所产生的客观数据,能发掘出教育大数据潜在的有效价值,这对教育信息化的发展非常有利[2-4]。当前,大多数算法和模型在预测能力上已经取得了不错的效果,但在算法准确度上并没有实质的突破,其主要原因是底层数据提取的行为特征不完善,导致相应的算法或模型难以进一步优化[5]。针对上述问题,相关研究人员提出了特征工程分析技术,其特点是重视数据本身所处的情境特性,强调将得到的数据转换成具有物理意义、适合用于建模的特征数据,以提升数据分析和建模的可操作性和可解释性,从而为后期模型的优化打下坚实基础[6]。

## 1 特征工程的应用学习分析

蓝墨云班课是一款免费线上教学软件,它可以记录师生教学过程中的行为足迹[7]。本文以该平台上收集的“算法与数据结构”课程数据为基础,利用特征工程技术进行案例研究。首先概述了实验数据的来源;然后在特征提取前对数据进行预处理;之后结合学习者的行为特征构建思路,从特征集中提取较为高效的特征数据;再以学生们的成绩是否会有不及格的风险为分类目标,分析学生的行为特征,构造出特征选择策略;最后在经特征工程筛选的特征数据的基础上,使用逻辑回归分类算法根据蓝墨云班课学习者的学习成绩风险进行建模实践。

### 1.1 实验数据集概述

本研究的数据来源于使用蓝墨云班课的某班级的 84 位学生的“算法与数据结构”课程期末汇总数据和明细数据。

### 1.2 学生行为特征的提取和预处理

本研究根据学生行为特征,从特征数据集中提取相关特征。采用的数据集为 2020 级学生“算法与数据结构”课程的数据,原始数据较为完整。

#### 1.2.1 特征提取

学生行为特征的提取以特征构建为基础,从原始数据中提取学生的学习行为为基础特征,然后把期末总成绩不及格的学生定义为风险者,把期末总成绩及格的学生定义为无风险者,提供给预测学习风险使用。

(1)基础特征提取

本研究的数据来源于同一个班级同一门课程,

**收稿日期**:2021-08-02

**基金项目**:安徽省高等学校省级质量工程项目(2020kcszjxtd37,2020kfkc261,2019jyxm0505);安徽新华学院校级质量工程智慧课堂项目(2019zhktx03)

**作者简介**:汪红霞(1979—),女,安徽宣城人,副教授,硕士,研究方向:计算机视觉。

数据所针对的对象仅为学生,提取出的16项学生相关的学习行为特征,见表1。

(2)风险者标记

本研究以每位学生的期末最后总成绩为标准,将不及格学生标记为0,及格学生标记为1。使用Python提取学生数据特征,共有84份样本,以供学习风险预测研究。

(3)特征向量

研究设置$P$作为是否有学习风险的变量,$P=\sum_{i=1}^{16}a_iX_i$,其中,$X_i$为表1中的第$i$($i$=1~16)个特征,$a_i$为第$i$个特征所对应的特征向量的值。阈值设置为0.5,如果$P\leqslant 0.5$,则标记为$P$=0;反之,则标记为$P$=1。

**表1 蓝墨云班课学生的相关特征**

| 特征 | 名称 | 特征 | 名称 |
|---|---|---|---|
| $X_1$ | 签到数 | $X_9$ | 讨论答疑参与次数 |
| $X_2$ | 观看视频个数 | $X_{10}$ | 讨论答疑发言次数 |
| $X_3$ | 观看视频时长 | $X_{11}$ | 讨论答疑解答总次数 |
| $X_4$ | 非视频资源查看数 | $X_{12}$ | 讨论答疑被老师点赞次数 |
| $X_5$ | 课堂表现参与次数 | $X_{13}$ | 测试参与次数 |
| $X_6$ | 投票问卷参与次数 | $X_{14}$ | 测试平均分 |
| $X_7$ | 头脑风暴参与次数 | $X_{15}$ | 作业任务参与次数 |
| $X_8$ | 头脑风暴被老师点赞次数 | $X_{16}$ | 出勤率 |

1.2.2 特征处理

特征提取之后,采用无量纲化将特征标准化,分析解决有风险和无风险的样本数量不平衡问题。

(1)特征无量纲化

本研究提取了16项特征,这些特征的数据取值范围差别很大,量纲也不同,如个数单位为“个”,实看视频时长为“分钟”等,没有办法进行比较。由于本研究的原始数据集体量较小,并且希望保留原始数据中潜在权重关系和数值意义,本文采用归一化中的最值区间缩放法和标准化中的Z-score法进行特征无量纲化处理。

1)最值区间缩放法

该方法基于原始数据分布对数据进行线性变换,将其映射到[0,1]上,该方法的计算公式为

$$X'=\frac{X-\mathrm{Min}}{\mathrm{Max}-\mathrm{Min}}。\tag{1}$$

式中,$X$为样本数据的值,Min为样本数据中的最小值,Max为样本数据中的最大值,$X'$为样本数据缩放到[0,1]区间上所对应的值。

2)Z-score法(标准化法)

该方法会将原始数据集标准化成均值为0、方差为1的新数据集,标准化后的值会被缩放到[-1,1]。该方法的计算公式为

$$X'=\frac{X-u}{\sigma}。\tag{2}$$

式中,$u$为样本数据均值,$\sigma$为样本数据的标准差,$X'$为样本数据缩放到[0,1]区间的值。

(2)样本数量不平衡处理

本研究样本中及格的学生占总样本的85%左右,剩余为不及格的学生,不及格学生和及格学生样本数量差距明显,这种数据量的不平衡会影响模型的准确性。为解决此不平衡问题,采用采样法和留一叉验证法进行处理。采样法适合于处理“样本总量少、样本数据类别差距明显”问题,留一叉验证法适合于处理“总样本数量小、样本利用率高能对模型进行验证”的问题。在特征选择中使用SMOTE方法平衡数据,在预测建模时使用留一叉验证法平衡样本训练模型并验证模型有效性。

### 1.3 特征选择

机器学习建模过程中,特征数量不是越多越好,还需要排除无关或影响甚微的特征,以提高模型的泛化能力,降低模型运算成本。本研究分如下两个步骤进行。

(1)剔除“低质量”的特征。查找缺失或低质量的特征,进行剔除。

(2)选择“相关”的特征。使用Python相关算法探究各项特征与学习“是否有风险”这一变量有关的特征。

1.3.1 剔除低质量特征

在数据存储阶段,将缺失的数据进行默认0值保存,所以此处查找缺失特征时,使用max函数查找最大值等于0的特征。而在样本数据中,只有投票问卷参与次数这一特征为缺失特征。

经查看数据后发现,可能存在某些特征数据在班级所有学生中是等值的,这类数据不具有研究意义,所以采用max函数与min函数做数据比对进行筛选。最后得到观看视频个数这一特征为等值数据。

经研究,从数据特征集中剔除投票问卷参与次数和观看视频个数这两个特征。

1.3.2 选择相关特征

首先依据风险者标记将样本划分为风险者和无风险者两个子样本，然后采用 SMOTE 采样方法平衡样本，获得有风险者和无风险者各 71 人次样本数据集；通过使用 ttest_ind 函数对两类学习者的特征数据进行 T 检验，研究特征数据和成绩风险的相关性，T 检验结果如表 2 所示。

**表 2　各项行为特征 T 检验的结果**

| 特征 | 名称 | T 检验结果 $p$ 值 |
|---|---|---|
| $X_1$ | 签到数 | 0.002 7 |
| $X_2$ | 出勤率 | 0.003 9 |
| $X_3$ | 观看视频时长 | 0.000 01 |
| $X_4$ | 非视频资源查看数 | 0.0000 01 |
| $X_5$ | 课堂参与次数 | 0.056 3 |
| $X_6$ | 头脑风暴参与次数 | 0.000 01 |
| $X_7$ | 头脑风暴被老师点赞次数 | 0.003 9 |
| $X_8$ | 讨论答疑参与次数 | 0.000 01 |
| $X_9$ | 讨论答疑发言次数 | 0.000 01 |
| $X_{10}$ | 讨论答疑解答总次数 | 0.002 2 |
| $X_{11}$ | 讨论答疑被老师点赞次数 | 0.009 2 |
| $X_{12}$ | 测试参与次数 | 0.000 01 |
| $X_{13}$ | 测试平均分 | 0.000 01 |
| $X_{14}$ | 作业任务参与次数 | 0.587 6 |

研究以 T 检验的检验水准 $\alpha$=0.05 作为标准，认为 $p$<0.05 的特征有显著差异。经过多次 T 检验后发现作业任务参与次数无显著差异($p$>0.05)；而课堂参与次数和出勤率的 $p$ 值在 0.05 波动比较大，可能受采样方法的影响比较大；其他的行为特征均保持在 $p$<0.05，有些较为显著的差异。于是剔除作业任务参与次数、课堂参与次数和出勤率 3 个差异性较小的行为特征。

### 1.4　重要特征选择和评估

为提高模型预测效果和建模的效率，进一步采取递归特征消除法筛选出更少相关特征。递归特征消除法通过递归减少特征集的数量来选择特征。设置特征集阈值个数(Number of feature sets，简称 nfts)，对每个特征的确定一个权重模型使用原始特征集训练，训练后，将重要性低的特征删除。重复训练，并找到特征集数量达到 *nfts*。但递归特征消除法得到的特征集会受到 *nfts* 的影响，由于 *nfts* 的设置具有一定的盲目性，过小可能会导致相关特征被移除特征集，过大则会导致信息冗余，所以采取 RFECV 方法进行交叉验证并寻找最佳的 *nfts*。

使用 RFECV 函数进行交叉验证得到精度系列得分，计算其系列得分均值。以模型的精度得分平均值为标准，探寻 RFECV 的交叉验证的最佳次数，平均精度值和交叉验证的次数的关系如图 1 所示。由图 1 知，当交叉次数为 8 次时，平均精度达到了最佳，因此采用交叉次数为 8 的 RFECV 方法实现特征选择。

使用交叉验证 8 次的 RFECV 函数去获取最佳的 *nfts*，精度均值的变化曲线如图 2 所示。由图 2 知，当 *nfts* 取 6~8 时，预测建模的效果归好，“性价比”最高。

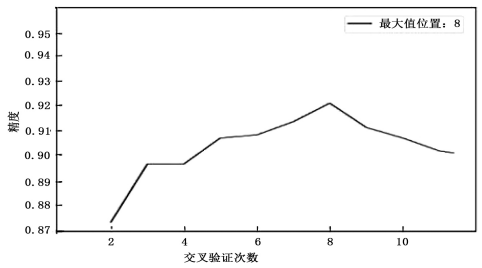


**图 1　精度与交叉验证次数的关系**

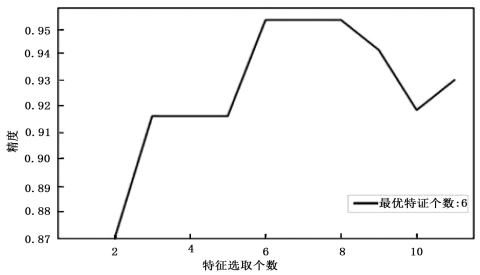


**图 2　选择特征个数对预测准确率的影响**

根据实验结果，选取最佳 *nfts* 为 8，使用 RFECV 方法，筛选出优先级从高到低的 8 项重要特征集为：非视频资源查看数、签到数、测试参与次数、讨论答疑解答总次数、头脑风暴参与次数、讨论答疑被老师点赞次数、观看视频时长、测试平均分。

## 2　学习风险预测模型的实现和验证

### 2.1　学习风险预测模型实现

2.1.1 学习风险预测模型

逻辑回归模型(Logistic Regression，简称 LR)是一种广义的线性回归分析模型，它经常被用来解决二分类的问题。本研究采用该模型对学生学习是否存在风险进行预测。线性回归产生的预测值 $Z=\omega^T X+b$ 是实值，将它转换为(0，1)值。采用 sigmoid 函数进行学习风险预测模型的构建，所构建模型的数学表达式为

$$H_\theta = \frac{1}{1+X_\theta} 。 \quad (3)$$

式中：$X_\theta$ 为样本的输入值；$H_\theta(X)$ 为模型的输出值，$H_\theta(X)$ 的值在[0,1]之间。

2.1.2 模型参数

本研究使用网格搜索交叉验证法进行模型调参，选用 SMOTE 进行样本数量平衡，采用网格搜索方法进行逻辑回归参数调优，将逻辑回归的可能参数集合输入到网格搜索，并以模型的曲线下方的面积大小(Area Under Curve,简称 AUC)值作为评分指标，以获取最优的 LR 参数。调参实验运行得到的结果表明，调优算法中 solver 参数选择为 newton-cg，参数最大迭代次数在此优化算法下有效。因最大迭代次数对 LR 的运行有着较大的影响，最大迭代次数若过小，则运算结果可能没有收敛；若过大，则梯度下降迭代的次数就会过大，模型运算缓慢。所以择优设置为 200，既防止不收敛又防止迭代次数过大。最终选择的 LR 参数如表 3 所示。

表 3 选择的 LR 参数

| 参数 | 描述 | 设置结果 |
|---|---|---|
| penalty | 用于指定惩罚项中使用的规范 | 12 |
| dual | 对偶或原始方法 | False |
| tol | 停止求解的标准 | 0.000 5 |
| c | 正则化系数λ的倒数 | 0.1 |
| fit_intercept | 是否存在截距或偏差 | True |
| intercept_scaling | 人造特征被考虑时，降低其影响 | 0.5 |
| class_weight | 用于表示分类模型中各种类型的权重 | None |
| random_state | 随机数种子 | None |
| solver | 优化算法选择参数 | newton-cg |
| max_iter | 算法收敛最大迭代次数 | 200 |
| multi_class | 分类方式选择参数 | ovr |

## 2.2 学习风险预测模型的验证

2.2.1 学习风险预测模型的验证

为验证学习风险预测的正确性，本文首先使用 REF 包裹法以样本选择的前 6 项重要特征为基础，再以得到的 LR 模型参数进行建模，并使用留一交叉验证法来验证建构的模型有效性，以接受者操作特征曲线（receiver operating characteristic curve，简称 ROC 曲线，如图 3 和图 4 所示）以及 AUC 值作为评价指标，评价该模型的稳定性，其中 ROC 曲线的横坐标为假阳性率，纵坐标为正阳性率，并使用 joblib 函数保存模型。代码运行实验结果如图 3 所示，从图中可以看出，最终得到的模型的准确率为 96.42%，AUC 值达到 0.978。

2.2.2 学习风险预测模型的测试

利用平行班级相同课程的蓝墨云班课教学数据，按照上述建模思路建立新的模型，并再次进行验证测试。这组测试数据共有学生信息数据 106 条，其中有 36 个学生成绩有风险，70 个学生成绩无风险。将此组数据运行于所构建的模型，并设置参数为调优参数进行测试，实验结果如图 4 所示。实验结果表明模型的准确率为 97.1%，而模型的 AUC 值为 0.996，预示着模型的效果非常好。

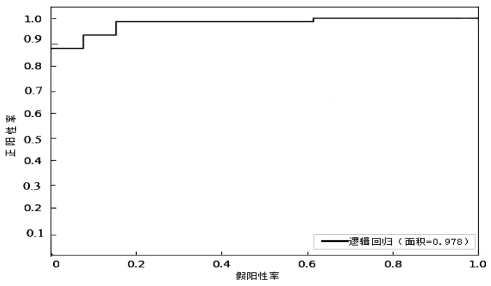


图 3 模型的 ROC 曲线

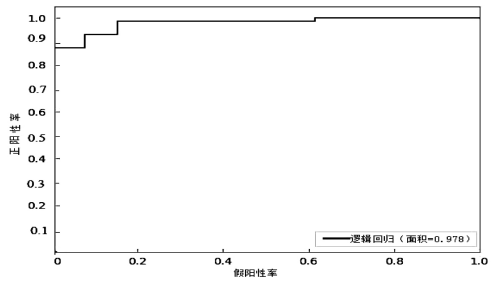


图 4 新模型的 ROC 曲线

## 3 结语

本研究以学生期末成绩的风险预测建模为核心，深入探究了基于特征工程的学习分析的应用以及预测模型的构建。通过运用特征工程在蓝墨云班课真实数据上提取影响学生成绩的特征属性，并在此基础上构建了学习成绩风险预测模型，最后利用真实数据进行验证，得出模型能够以 95%左右的准确率预测学生成绩和 AUC 得分在 0.95 以上，这说明模型的预测效果非常好，教师可以凭此判断学习者的学习进度或学习行为是否存在异常，同时学习者也可

（下转第 57 页）

航空制造技术，2014，39(5)：52-57.
[5] 孙家坤，杨超英.基于MBD的三维CAPP系统分析与实施[J].山东建筑大学学报，2013，27(4)：363-366.
[6] 王占富，丁来军.MBD模式下机加工艺研究[J].机械设计与制造，2014，35(3)：145-147.
[7] 王虓，王细洋.基于MBD技术的飞机结构件三维工艺规程卡定义[J].机械工程师，2013，41(2)：60-62.
[8] 胡保华，闻立波，杨根军，等.基于MBD的三维数字化装配工艺设计及现场可视化技术应用[J].航空制造技术，2011，36(22)：81-85.
[9] 潘康华.基于MBD的机械产品三维设计标准关键技术与研究[D].北京：机械科学研究总院，2012：50-53.
[10] 唐远志，向雄方.汽车车身制造工艺[M].北京：化学工业出版社，2009：103-105.
[11] 王境宇，邓立营.基于CATIA的产品定义信息三维表达及组织方法[J].制造业自动化，2011，33(11)：130-133.
[12] 周秋忠，查浩宇.基于三维标注技术的数字化产品定义方法[J].机械设计，2011，29(1)：33-36.

【责任编辑　刘建华】

# Development of 3D Annotation and Management System for BIW Welding Information

**WANG Pengfei，WANG Lin，YE Hongling，CAI Feiyang**

(School of Mechanical and Vehicular Engineering, Bengbu University, Bengbu 233030, China)

**Abstract:** In order to reduce the errors in the transmission process of BIW Welding information, using CATIA secondary development technology, combined with VB language and ACCESS database, the information management system of BIW Welding is developed，and an example of an assembly is used to verify the application. The results show that 3D annotation method and its management system can effectively reduce welding errors and improve welding efficiency.

**Keywords:** BIW welding; 3D annotation; management information system

(上接第 43 页)

以借此探寻学习上存在的问题，进而调整自身的学习行为。

**参考文献：**

[1] 高瑞洁.基于特征工程的大学生在线学习习惯挖掘与分析[D].哈尔滨：哈尔滨师范大学，2020：14-36.
[2] 王莎莎，王梅. 学习分析技术研究现状综述[J].中国教育技术装备，2019(8)：7-11.
[3] 陈佑清.论有效教学的分析模型[J].课程·教材·教法，2012，32(11)：3-9.
[4] 谢玉华，胡永生. 教育大数据支持下学习分析技术研究综述[J].当代教育实践与教学研究，2018(9)：33-34.
[5] 欧阳嘉煜，范逸洲，罗淑芳，等. 特征工程：学习分析中识别行为模式的重要方法[J]. 现代教育技术，2018，28(4)：13-19.
[6] 李若晨.基于特征工程的MOOC学习者行为分析和辍课预测[D].上海：华东师范大学，2019：4-42.
[7] 过珺，曹路舟，潘新征. 基于蓝墨云班课的课堂大数据分析在教学中的应用初探：以学生学习行为数据的分析为例[J]. 电脑知识与技术，2018，14(19)：130-131.

【责任编辑　刘建华】

# Research of Learning Analysis Based on Feature Engineering

**WANG Hongxia**

(College of Big Data and Artificial Intelligence, Anhui Xinhua University, Hefei 230088, China)

**Abstract:** In view of the large amount of data generated by current online education, feature engineering technology is firstly used to extract the in-depth behavior of online learners' learning data. Then the logistic regression classification algorithm is used to model the learning data. Finally, a better prediction model is obtained by further training and verification with machine learning. The experimental results show that the model can provide better guidance for teachers and students to analyze the learning effect.

**Keywords:** learning behavior analysis; feature engineering; logistic regression