基于集成学习的 MOOC 学习数据分析及辍课预测

光电与信息工程学院　通信工程（闽台合作）工程专业
136102017014（学号）　郑义恒　指导老师 徐哲鑫

【摘　要】近年来，大规模的开放式在线课程（MOOC）迅速发展，并吸引了数百万的在线用户，然而，高居不下的辍学率是一个一直存在的问题。为了降低辍学率，通过分析用户的行为来发现辍学倾向并进行干预是本本文研究的问题。针对这个问题，本文中设计了行为转移间隔模型（BTIM）用于量化用户的状态，并结合机器学习模型预测辍学倾向。在 BTIM 模型分为行为间隔分布与行为转移矩阵两个模块，行为间隔分布模块采集学习行为间距的统计学度量指标，行为转移矩阵模块从频次上描述学习行为间的转移模式。出于预测结论的可解释性要求，预测模型选择决策树与线性回归模型，并应用模型组合方法提高泛化能力。本研究最终实现了对辍学用户的高效预测 AUC 85.2% F1 90.15%，该结果与领域内较为优秀的研究十分接近，且本研究中良好的模型解释性，对分析辍学行为的成因以及高风险用户的推断起到一定作用。

【关键字】在线教育；数据分析；特征构造；机器学习；时间序列

目录

1 绪论 3
1.1 研究背景与意义 3
1.2 研究现状与趋势 3
1.3 本文研究内容与组织结构 4
2 行为转移间隔模型 4
2.1 模型设计思路 4
2.2 探索性数据分析 4
2.3 数据清洗 5
2.4 行为间隔特征模块设计 6
2.5 行为转移特征模块设计 7
2.6 背景特征模块设计 8
2.7 特征组合 8
3 辍学预测算法 9
3.1 预测算法设计思路 9
3.2 决策树算法原理阐述 9
3.3 XGBoost 算法简述 10
3.4 随机森林算法阐述 10
3.5 线性回归算法原理阐述 11
3.6 模型组合方法 11
4 模型测试与分析 11
4.1 模型性能评价指标 11
4.2 模型验证 12
4.3 验证结果分析 12
5 总结与展望 14
5.1 总结 14
5.2 展望 14

1 绪论

1.1 研究背景与意义

2020 年 COVID-19 阻断了传统的线下教学路径，教育领域经历了一场深层次变革，在线学习在这一年成为主流，MOOC（Massive Open Online Course）开始被更多人所知，并且随着诸如 XuetangX，edX 和 Udacity 等的大规模开放在线课程（MOOC）的迅速普及，在线教育引起了教育界和计算机科学界以及公众的关注。但在 MOOC 上学习的用户仍然面临着课程通过率低这一主要问题。以墨尔本大学于 2013 年在 Coursera MOOC 平台上提供的 Discrete Optimization 课程的完成率数据为例，在 51,306 名学生中，只有 795 名学生完成课程，完成率 1.5%。只有 27,679 名学生（约占 54%）曾经参加过讲座和测验/作业，即使在参与过课程活动的学生这一组，结业率也仅为 2.9%，远低于传统学习方式的课程通过率。

在解决在线学习通过率低的问题上，众多研究都聚焦于对辍学用户的提早预测这一问题。原因在于，在传统教育方式中，在用户展现出辍学倾向时对其加以干预，对提升学习表现有显著的作用。而对于线下课堂的教授者来说，判断一个学生能否通过课程往往根据课堂表现、学习状态等主观因素以及学习成绩、作业情况等客观因素。对于在线教育也是如此，但在庞大的用户基数下，有限的教师资源令人工识别高风险学生这一工作变得难以开展。因此，现有方案是训练人工智能模型以识别高风险用户，但诸如课堂表现这类由教师主观判断的指标其抽象程度高无法被机器直接处理。

所以，对这类在传统课堂中这依靠教师的经验进行主观判断的指标进行量化，建立高效的辍学预测模型，成为了在线教育的重点问题，对该问题的探索有助于了解辍学问题的成因，也能从一种新的维度对在线教育进行深入了解与解构。

1.2 研究现状与趋势

目前，对辍学预测问题的研究可以分为三个主要方向，数据集与知识图谱构建、行为数据挖掘、深度预测模型。

数据集与知识图谱[11]是对在线教育领域进行研究的基础，其中涵盖了行为数据与课程数据等背景信息的整合以及以及知识概念关系的提取等研究工作。清华大学的 Jifan Yu,Gan Luo 团队基于 XuetangX 在线教育平台 2015 至 2017 年内所以有的用户数据建立了一个大规模公开数据库，其中包含 706 门真实在线课程，38181 个教学视频，114563 个概念，199199 名 MOOC 用户的选课、视频观看记录。并根据维基百科等资源建立了课程知识图谱，此研究成果为对数据要求高的深度模型在 MOOC 研究中的应用起到了推动作用 [7]。

在数据挖掘方向的研究中，清理冗余数据抑制数据中的噪声对于进一步的分析工作十分重要，卡耐基梅隆大学的 Tanmay Sinha 从认知科学的角度对用户行为数据进行分析，其基于“有限能力信息处理方法”构建学生的信息处理指标，该方法断言人独立地分配有限的认知资源来完成任务，并以此对学习行为（点击流组合[10]）根据学习状态与学习的信息量进行打分。并应用 N-gram（常用在 NLP 领域）方法对点击流序列进行提取，其研究结果显示，选择合适的采样宽度能在保留特征的同时抑制噪声，对数据清洗与挖掘做出了贡献，并且该研究成果也展示出认知心理学研究在 MOOC 领域不可忽视的潜力[2]。

由于 MOOC 数据集中涵盖的特征众多，数据维度较高，应用拟合能力强大的深度学习算法在辍学预测这一方向上成为了一些研究团队的选择[14]，清华大学的冯文政,唐杰团队基于选课与行为数据应用 DeepWalk[8]算法对用户进行分类，并进一步提出了一种特征提取模型，该模型基于带注意力机制的深度神经网络，能够基于用户与课程的信息自动选择有价值的特征，研究结果显示在预测辍学用户的任务上的性能优于传统机器学习模型。该团队还基于对用户的分类结果，对各类用户进行分析，发现不同类别的用户对各类学习活动的关注度在统计学上有显著差异，并针对用户的特性提出了初步的辍学干预策略，并在线进行了双盲测试，结果显示其干预策略显著的提升了用户的活跃程度。目前，该团队已经将此研究成果部署在 XuetangX 平台上，起到显著的辍学干预效果[3]，并且该研究成果被人工智能领域顶

级会议 AAAI 收录。在本论文中使用与该研究相同的数据集，取得了十分接近的预测性能。

1.3 本文研究内容与组织结构

本文研究的内容是如何建立学生行为状态量化模型并识别可能辍学的学生，进而根据模型推测辍学行为的影响因素并提出干预措施。

在本文的组织结构上，首先在绪论中介绍辍学预测的研究背景与相关领域的研究进展，其中涉及到 MOOC 在近年的发展趋势与相关领域中较为出色的研究成果。在第二章中介绍行为特征模型的设计思路与细节以及数据集的基本情况。在第三章中详细叙述预测模型的原理以及模型组合方法。在第四章中分析模型的预测结果以及辍学行为的相关原因。第五章为研究总结与展望。

2 行为转移间隔模型

2.1 模型设计思路

建立学习行为转移间隔模型（BTIM，Behavioral Transition Interval Model）的目的是，从原始数据中提取能够描述学习行为特质的指标。基于这一目的，在建立模型前对数据集进行初步的了解与探索，进而对数据进行针对性的清洗，再对清洗完成的数据提取时间间隔特征、行为转移特征以及背景特征，特征提取完成后，将特征组合成训练数据集，以训练预测算法。

2.2 探索性数据分析

探索性数据分析的目的是了解数据的概况并对数据进行简要的描述，为数据清洗、特征提取以及预测模型的选择与调整提供理论支持[12]。

首先对数据量以及样本比例进行分析，可以观察到训练样本约为测试样本数量的 2.3 倍，正负样本比例在 1:3 左右，存在一定程度的样本不均衡问题，在相同的迭代次数下，预测模型会对样本更多的类别产生过拟合，或对样本更少的类别产生欠拟合问题。因此，在预测模型中应该对非辍学样本的判断予以适当的倾斜。

表 2-1 ： 有监督数据集数据量概览

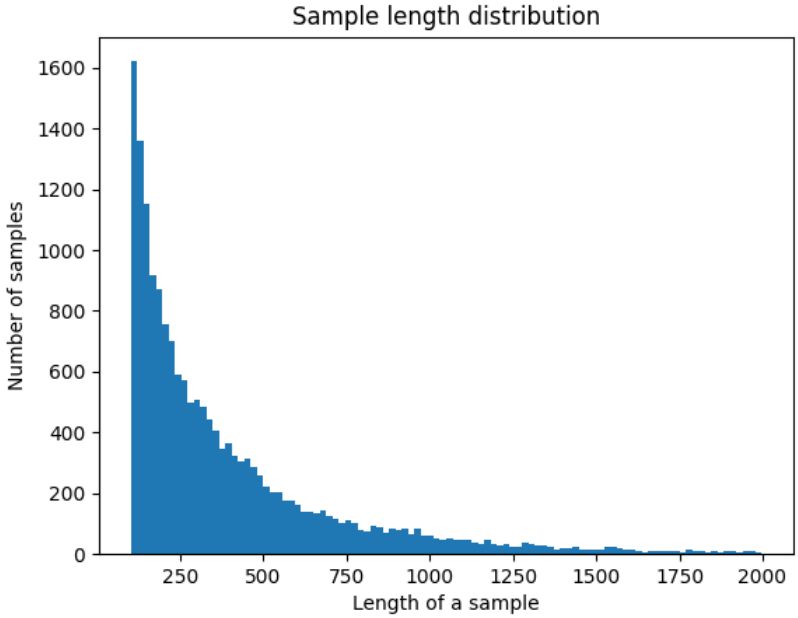
	文件名	数据文件大小	注册样本数量	数据维度（列）
训练集	trian_log.csv	3.92GB	157943	7
	train_truth.csv	1352KB	157943	2
测试集	test_log.csv	1.74GB	67700	7
	test_truth.csv	580KB	67700	2
用户信息	user_info.csv	111MB	9627149	4
课程信息	course_info.csv	534KB	6411	6

表 2-2 ： 样本数量与比例

样本数量	非辍学样本	辍学样本	非辍学样本占与辍学样本比例
测试数据集	16383	51316	32%
训练数据集	38126	119817	32%

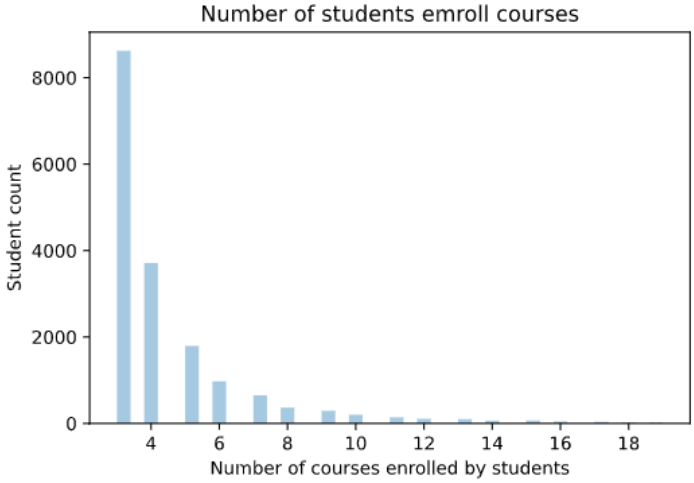
其次,在对训练集中单一样本的长度进行统计的过程中，发现行为序列样本长度分布高度不均衡，长度不均衡的样本需要更繁琐的特征工程，以发现具有尺度不变性与平移不变性的特征，并且，

使用时间序列模型如 ARIMA 对此类序列进行预测时需要填充或截断至各序列等长，在长度高度不均衡的情况下会导致有效数据损失或数据稀疏的问题。因此在特征工程中提取的特征应当对序列长度不敏感。



（图 2-1 样本序列长度分布情况）

通过观察课程的学生数量分布与学生报名的课程数量分布，可以发现有 8000 名以上用户报名课程的数量在两门以下，会对个人辍学率的计算方式造成影响，若以辍学课程数与个人总课程数的比例作为个人辍学率，在报名课程数越少的用户中，个人辍学率越容易收到影响，以致于辍学率特征不能如实反映用户的情况。因此，应对个人辍学率的计算设定课程数阈值，在阈值之下的用户采用均值填充等方式降低不确定的干扰。



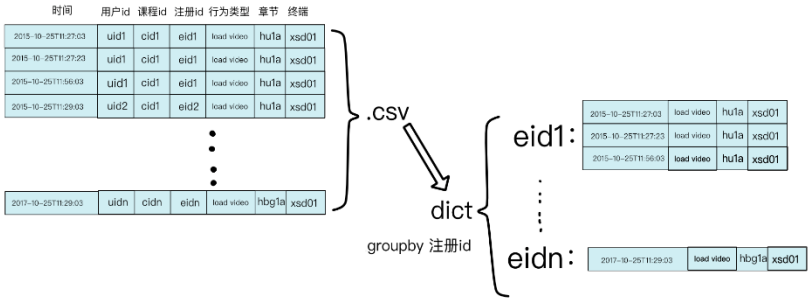
（图 2-2 学生报名的课程数量分布情况）

2.3 数据清洗

学习行为原始数据为字符串类型存储的逗号分隔符文件（csv），清理冗余数据抑制数据中的噪声能

够有效提升特征提取模型的性能。首先，遍历整个日志文件，对其中的每一行日志，根据该行中注册码进行分类，并以注册码为索引存入字典数据结构中（Python dict），在此存储过程中将字符串类型的特征值以无符号整型数值替换，并使用字典结构存储字符值与整型值之间的映射关系。在对日志文件的遍历过程中，得到“课程识别码（C）—注册码（E）—用户识别码（U）”之间的唯一映射关系，并存储在字典格式中。

此步骤是为了降低查询某课程或某用户下的所有数据时的时间开销而设计的，只需存储“C-E-U”这一映射关系，并配合字典结构，即可高效的进行查询筛选等操作。



（图 2-3 数据归类示意图）

2.4 行为间隔特征模块设计

行为间隔特征设计的目的是对用户的学习行为操作频率以及在操作频率的变化趋势进行衡量。学习序列的不同阶段，依据对课程的兴趣、课程难度、个人因素等原因，用户会表现出不同的学习频率，对于用户学习频率的基准值与波动情况，在此模块中选择计算间隔序列的均值与方差进行量化。课程与用户的背景多样性会导致多种不同的行为模式，例如，在课程作业提交截止日前，会出现学习行为的波峰，反映在时间间隔序列中就表现为波谷。为了量化用户学习行为的聚集程度与聚集位置，在模块中选择计算间隔序列的峰度与偏斜度来进行衡量。

由于在线学习中存在大量视频观看任务，所以在本模块的设计中加入时长阈值参数，将行为间隔分为长操作间隔与短操作间隔，以分割视频观看任务，在模型实际运行中分别对长序列与短序列计算均值$\overline{X}$、方差S^2、峰度[4]G_1、偏斜度[4]G_2，最终行为间隔特征模块输出八个统计特征。

表 2-3:行为间隔特征表

长间隔序列特征:	长间隔方差	长间隔均值	长间隔峰度	长间隔偏度
短间隔序列特征:	短间隔方差	短间隔均值	短间隔峰度	短间隔偏度

X_i：间隔序列中的单个元素

n：间隔序列长度

$$\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i \qquad （公式 2-1）$$

$$u_k=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^k \qquad （公式 2-2）$$

$$S^2=u_2 \qquad （公式 2-3）$$

$$G_1=\frac{n^2u_3}{(n-1)(n-2)S^3} \qquad （公式 2-4）$$

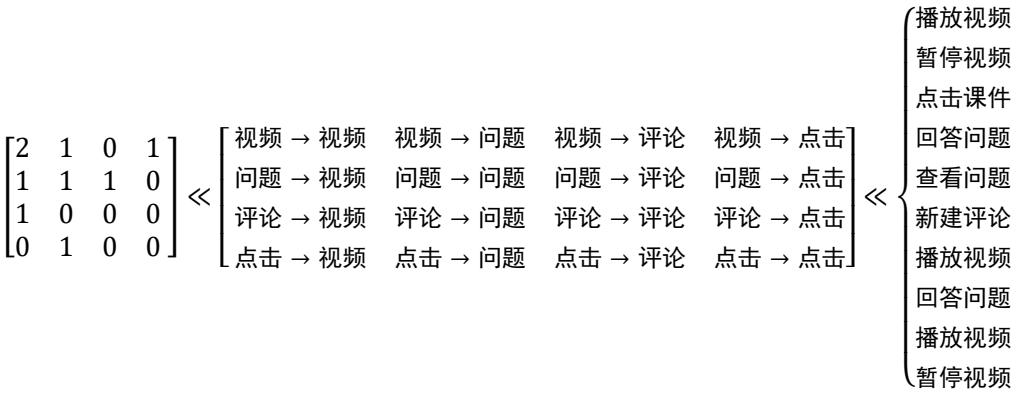
$$G_2=\frac{n^2(n+1)u_4}{(n-1)(n-2)(n-3)S^4}-\frac{3(n-1)^2}{(n-2)(n-3)} \qquad （公式 2-5）$$

2.5 行为转移特征模块设计

设计行为转移特征的目的是，从转移频次的角度提取用户的学习模式。根据行为模式的不同，用户选择的下一次学习行为的类别概率分布也不同，为了量化行为转移概率，在此模块基于图结构的邻阶矩阵设计了行为转移矩阵。将每一类学习行为视作图结构中的一个结点，结点间的关系用邻接矩阵存储，结点间边的权值表示经过这条边的次数，即该日志中此类转移行为的发生次数。

表 2-4 学习行为分类表

视频类	查找、加载、播放、暂停、停止
提问类	获取问题、保存、确认、重置、回答正确、回答错误
评论类	新增评论、新增讨论、删除评论、删除讨论
点击类	点击课件、点击课程简介、点击论坛、点击 about、点击教学计划、关闭课件、关闭课程简介



（图 2-4 行为转移特征提取示意图）

2.6 背景特征模块设计

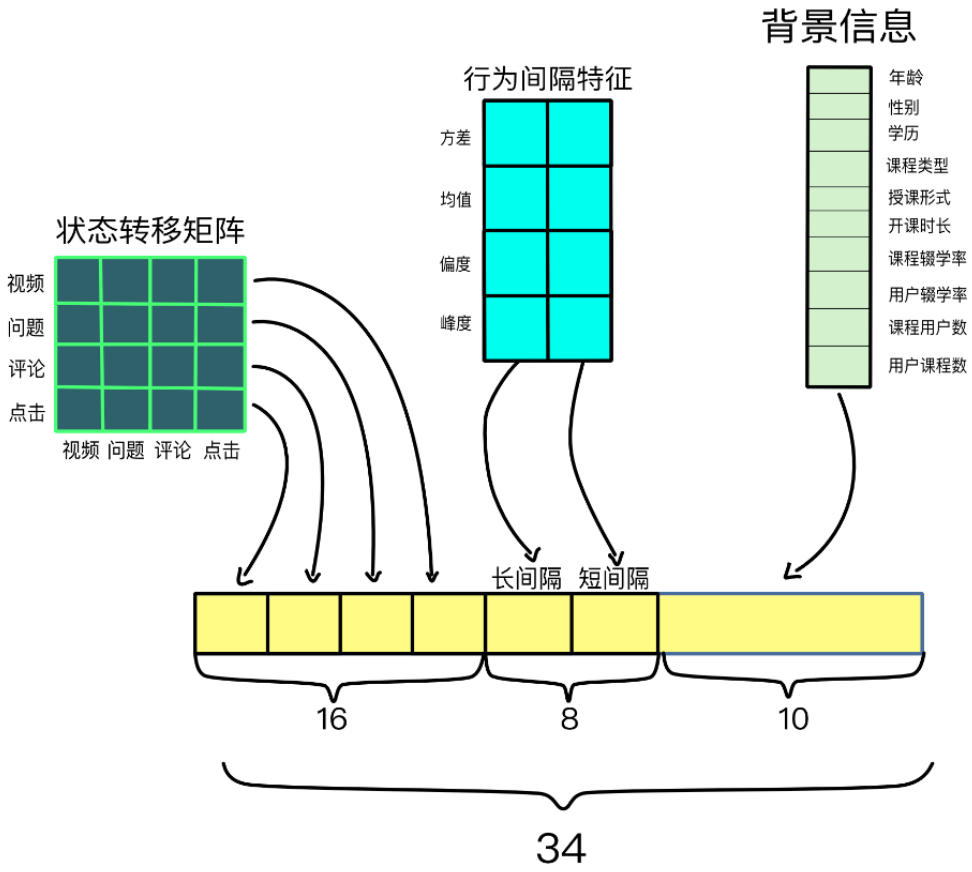
背景特征指的是从行为序列之外的用户与课程信息数据集中，提取的有效特征，提取背景特征的目的是量化用户与课程的既有历史信息，其独立于行为特征之外，通过训练样本的注册识别码调用相关历史信息与辍学标签组合为背景特征。在背景特征提取完成后，每个样本共含有 10 项背景特征如特征表所示。

表 2-5 背景特征表

用户背景特征	年龄、性别、学历、 用户课程数、课程辍学率
课程背景特征	开课时长、授课形式、课程类型、 课程学生数、学生辍学率

2.7 特征组合

至此，特征的提取已经完成，需要将特征组合成训练样本输入模型，此处使用较为简单的组合方式，对行为转移矩阵逐行读取，与其它特征首尾拼接，组成一个含有 34 个元素的样本。



（图 2-5 特征组合示意图）

3 辍学预测算法

3.1 预测算法设计思路

辍学预测算法的实质是将训练数据作为自变量,将辍学标签作为因变量，拟合自变量与因变量之间的关系，并根据此种关系判断测试数据的类别归属。出于对模型预测结果解释性的要求，所以选择线性回归与决策树模型构建辍学预测算法，并在模型验证过程中运用模型组合方法进一步提高预测性能。

3.2 决策树算法原理阐述

决策树模型基于对特征的选择与特征值的分割，来构造规则以判断样本的归属。对特征以及分割点的选择由其对模型收敛过程的贡献度决定，对贡献度的计算是决策树模型的核心，不同的决策树模型在贡献度的定义上不同，常用的贡献度有信息增益（ID3 算法）与信息增益率（C4.5 算法）。贡献度定义如下：

D: 数据集, $|D|$ 数据集中样本数量

C_k: 样本所属的标签类别 k = 1,2,3 ... K

$|C_k|$ 为属于类 C_k 的样本个数

A_n: 数据集中的特征 $n=1,2,3...N$

由此可知，特征 A_n 有 V 个不同的取值 $\{A_n^1, A_n^2, A_n^3, ..., A_n^V, \}$，可以根据 A_n 的取值将 D 划分为 V 个子集，可知 $|D|=\sum_{i=1}^{v}|D_i|$，由此,在数据集 D 已知，并选择其中一类特征 A 的情况下，可以计算数据集 D 的熵 H，以及 A 对 D 的条件熵：

$$H(D)=-\sum_{i=1}^{k}\frac{C_k}{D}\log_2\frac{C_k}{D} \qquad \text{（公式 3-1）}$$

$$H(D|A)=\sum_{i=1}^{V}\frac{|D_i|}{|D|}*H(D_i) \qquad \text{（公式 3-2）}$$

信息增益即特征 A 与 D 之间的互信息 G 定义如下：

$$G(D,A)=H(D)-H(D|A) \qquad \text{（公式 3-3）}$$

信息增益率 G_{Rate} 定义如下：

$$G_{Rate}(D,A)=\frac{G(D,A)}{H(D)} \qquad \text{（公式 3-4）}$$

在决策树结构中节点分裂的目的是使得叶子节点内的样本尽可能的相同，使树结构向有序的方向演进，即决策树收敛的过程是熵值不断减小的过程。在使决策树的整体熵值减小的过程中使用贪心算法，首先选择一个使决策树的熵值下降最多的特征，将其用作根节点进行分裂，在余下的节点中重复这一过程，最终会得到一棵局部最优的树。局部最优的树结构预测性能在实际应用中表现并不优秀，所以衍生出了集成树模型，通过多棵树多套规则共同判断样本的归属。

在辍学预测任务中，决策树模型的拟合过程被视为，依据行为特征的分类和取值总结出样本的划分规则，该划分规则阐明了对样本进行判断的根据与推理过程，其对预测结果具有解释作用，并且对分析辍学行为的成因可以起到指导。

（表 3-1 决策树划分规则示意图）

特征	取值	特征	取值	特征	取值	标签	
年龄	>20	学历	低于初中	课程类型	数学	辍学	规则 1
					文学	不辍学	规则 2
			高于初中	课程时长	>9000	辍学	规则 3
					<=9000	不辍学	规则 4
	<=20	峰度	>10.9	性别	男，NaN	辍学	规则 5
					女	不辍学	规则 6
			<=10.9 或 NaN	序列偏度	>11.2	不辍学	规则 7
					<=11.2	辍学	规则 8

3.3 XGBoost 算法简述

XGBoost 算法是对梯度提升树（GBDT）算法的改进版本，能够自主处理数据中的缺失值。该算法属于集成学习模型的一种，Boost 指的是以串行的方式训练基分类器，各分类器之间有依赖。每次训练时，对前一层基分类器分错的样本给与更高的权重，其本质是多个决策树算法的组合，利用下一个决策树模型对目前模型的残差进行拟合。并且，XGBoost 不采用以熵值度量的 ID3 与 C4.5 算法，XGBoost 采用 CART 算法，也就是基尼系数（Gini Index），在计算过程中比 ID3 与 C4.5 快速，并且 XGBoost 使用二阶泰勒展开近似计算基尼系数。

3.4 随机森林算法阐述

随机森林算法和 XGBoost 同为决策树模型中的集成学习模型，并且使用相同的损失函数 CART，不同之处在于随机森林算法各分类器之间无强依赖，可以并行计算。其名字中的随机一词源于对结点的分裂的步骤中，并非采用二分查找，而是随机选择分裂位置，这一随机的特性导致随机森林算法具有更好的泛化能力，并且在面对高维度数据时收敛更稳定。

3.5 线性回归算法原理阐述

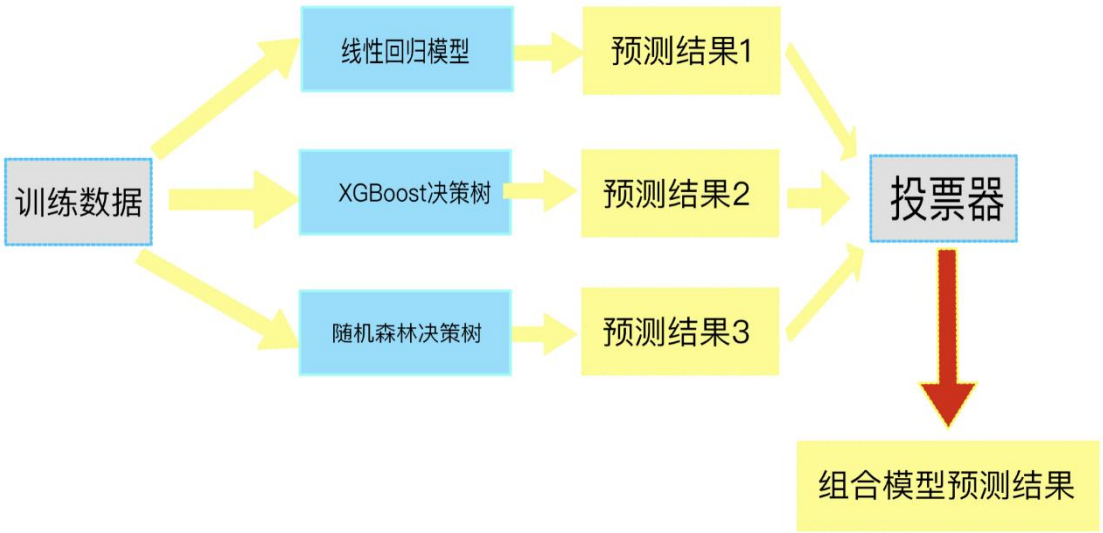
线性回归模型是较为原始的机器学习回归模型，其假设模型输入值与预测值之前存在线性关系，模型通过求解线性关系参数拟合数据，其线性关系参数的数值表明该特征与模型输出概率间的线性关系。线性模型输出结果为 0 至 1 之间的概率值，但本文中将回归模型应用于分类任务，所以通过设定**样本划分阈值**划分概率值，将大于阈值的样本划分为阳性，小于等于阈值的样本划分为阴性。[15]

在实际预测中，由于传统线性回归只能拟合线性关系，难以拟合非线性关系，所以在本研究中对线性回归模型进行了改进，特征组合后得到的 34 维的样本（原始样本）数据输入线性模型，可以视作计算含有 34 个变量的线性回归，对原始样本进行二次方展开，即可得到包含了原始样本一次项与二次项的复杂样本，对复杂样本进行线性回归，等价于同时使用非线性与线性关系拟合样本标签。

3.6 模型组合方法

模型组合即组合多个模型的预测结果。由于单一模型受限于其模型结构难以对全体样本均有良好的拟合结果，为了获得更好的预测性能，使模型具有更强的泛化能力[9]，使用相同的数据对所有的单一模型训练完成后，在对测试数据进行预测的过程中，使用投票机制均衡的考虑多个模型的判断结果。

投票机制设计为，在任意一个单一模型判断样本为辍学时，则组合模型判断样本辍学。该判定偏向的根据是，训练数据集中辍学样本的数量数倍于非辍学样本，因此模型对辍学样本的拟合程度优于对非辍学样本的拟合。



（图 3-1 模型组合方法示意图）

4 模型测试与分析

4.1 模型性能评价指标

对模型预测性能的评价指标采用领域内常用的 AUC 值与 F1 分数。AUC 值表示将一个辍学样本预测为辍学的概率值比预测为非辍学的概率值大的可能性，其数值介于 0.5 到 1 之间，越接近 1 分类性能越优秀。F1 分数（F1 Score），是统计学中用来衡量二分类模型精确度的一种指标。 它同时兼顾了分类模型的精确率和召回率。 F1 分数可以看作是模型精确率和召回率的一种调和平均，其数值介于 0 至 1

之间，越接近 1 分类性能越优秀。

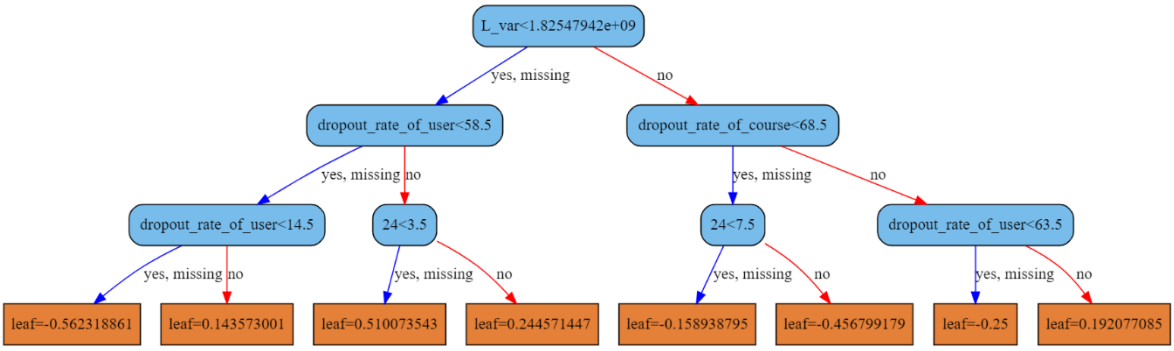
4.2 模型验证

对模型的验证分别采用一次项与二次项特征对单一模型与组合模型进行辍学样本预测实验，结果表明，在使用两种不同的特征配置的情况下，组合模型的 AUC 性能均显著超过单一模型预测方案，其 F1 性能与单一方案较为接近。并且与相关研究中使用深度神经网络的预测方案性能十分接近，可见本研究中对模型的组合以及特征的提取有一定的成效。

（表 4-1 本文模型预测性能）

本文模型预测性能			
	模型	AUC（%）	F1(%)
一次项特征	线性回归	83.8	89.49
	XGBoost	83.4	90.04
	随机森林	84.6	90.16
添加二次项特征	线性回归	84.8	90.05
	XGBoost	83.7	90.12
	随机森林	83.2	90.10
一次项特征	组合模型	85.1	89.71
添加二次项特征	组合模型	85.2	90.07

（表 4-2 使用相同数据集的深度神经网络预测方案性能）



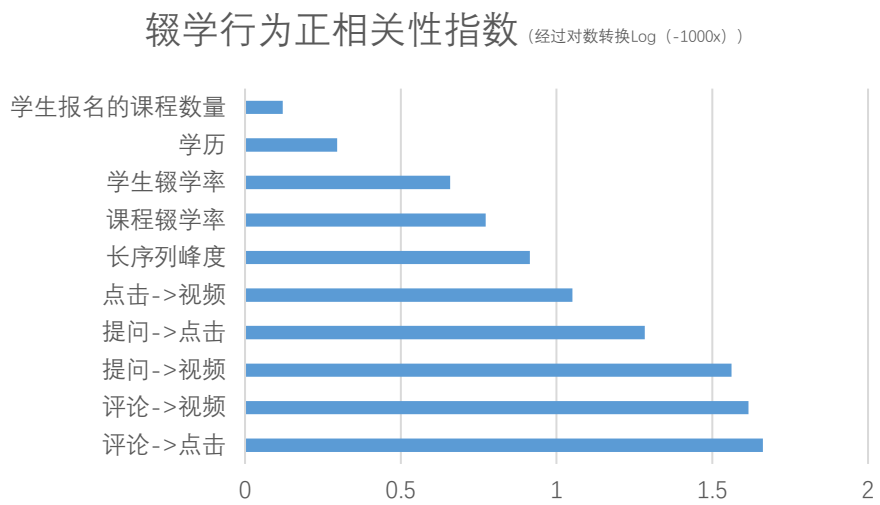
Understanding Dropouts in MOOCs[3]预测性能		
模型	AUC(%)	F1（%）
CFIN	86.40	90.92
CFIN-en	86.71	90.95

4.3 验证结果分析

(图 4-1 拟合训练数据所构建的决策树，树深度设置为 3)

基于前文对决策树与线性回归模型的原理阐述，对拟合后模型的参数进行分析，发现这两类模型对特征的选取存在较大差异，决策树模型仅使用 4 种特征就达到了与线性回归相当的预测性能，但必须说明的是，决策树模型选取的四种特征中用户与课程的辍学率这两种特征是基于用户的在其它课程上的

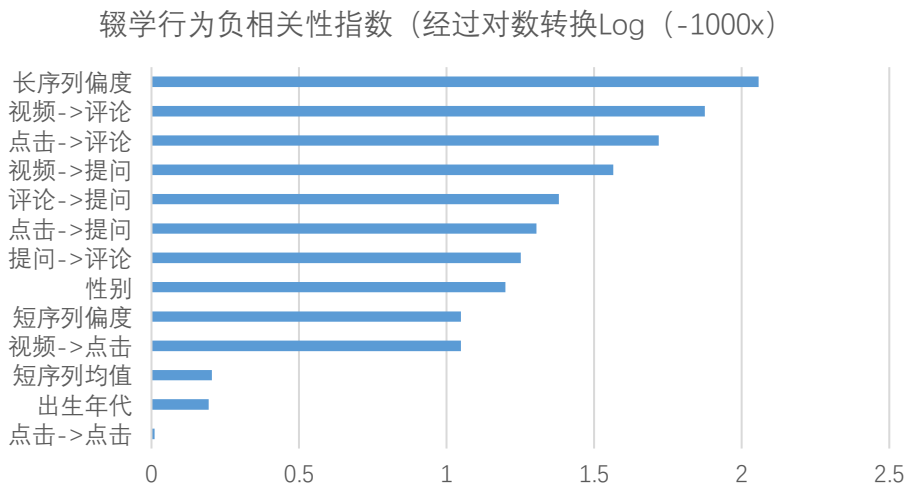
学习记录提取的，由于对此类特征的依赖，容易在实际应用中遇到模型冷启动问题。但对线性回归模型



的系数进行对比发现其对历史信息的依赖性相对决策树模型显著的低，由此可以推测在发生冷启动问题的场景下，使用线性回归模型是一种值得尝试的解决方案。

（图 4-2 与辍学行为呈正相关关系的特征)

并且，对线性回归模型的参数进行比较可以发现，行为转移特征对线性回归模型起到了显著的影响，从侧面说明了对行为转移特征的提取是有效的。长序列偏度与辍学行为的负相关性最为显著，根据偏度的定义可知，偏度越大则序列中的长间隔更多的分布在行为序列的前段，可以据此推测，随着学习过程的进行，倾向于提高学习频率的用户更不易辍学。对回归相关指数仔细观察可以发现，将四类活动作为转移的目标节点时，转移的目标是评论和提问类结点时，其频次与辍学行为呈现负相关关系，转移的目



标是点击和视频类结点时，其频次与辍学行为呈现正相关关系。对此现象，可以推测，以评论与提问类行为为转移目标，表示用户倾向于在学习之后进行提问等活动，而以点击和视频类行为为目标的用户则倾向于在学习前进行提问等活动。

(图 4-3 与辍学行为呈负相关关系的特征)

根据决策树的构建原理分析，越靠近树顶部的特征，其样本分割情况越接近真实标签，而长序列方差值被作为首要的划分特征，首先可以肯定 BTIM 中对行为间隔信息的提取具有现实意义，其次在后续的工作中可以细化对此类特征的提取，可能对进一步提升模型表现会有帮助。并且 24 特征也就是提问

行为向课件点击行为转移的频次，被放置在第三层可见其在决策树中的重要性，提升对此类行为转移特征的提取精度可以作为一个优化的方向。

5 总结与展望

5.1 总结

本文从用户行为间隔与行为转移方式两个较为新颖的角度对行为特征进行量化建立了 BTIM 模式，此模型减轻了行为序列长度对特征采集的影响。并将 BTIM 与机器学习模型结合对辍学用户进行预测，其间采用模型组合方法对单一模型的泛化性能进行提升，最终模型预测性能接近在行业中部署的深度学习解决方案[3]，并根据模型拟合的结果对辍学行为的重大影响因素进行推测，并且分析了不同机器学习模型对数据的倾向，对模型冷启动问题提出可能的解决方法。

5.2 展望

在本文中对辍学预测问题进行了系统的研究，但仍存在诸多细节尚未完善，如对长短间隔序列分割阈值的推导以及更详尽的行为转移矩阵与间隔分析。伴随着 MOOC 的发展，用户数量的激增，课程教学平台的性质或许会发生改变，也许会向知识科普平台转型，辍学或许不再被人提起[1]，但对时间序列的分析与提取在金融、制造业等多个领域都有应用，产业一直在变化，但分析的实质未曾改变。

参考文献:

[1]. Davis H C, Dickens K, Leon Urrutia M, et al. MOOCs for Universities and Learners an analysis of motivating factors[J]. 2014.

[2]. Sinha T, Jermann P, Li N, et al. Your click decides your fate: Inferring information processing and attrition behavior from mooc video clickstream interactions[J].arXiv:1407.7131, 2014.

[3]. Feng W, Tang J, Liu T X. Understanding dropouts in MOOCs[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2019, 33(01): 517-524.

[4]. Zwillinger, D. and Kokoska, S. CRC Standard Probability and Statistics Tables and Formulae Chapman & Hall: NewYork [M],2000, Section 2.2.24.1

[5]. Arindam Banerjee and Joydeep Ghosh.Concept-based Clustering of Clickstream Data.

[6]. Chen T, He T, Benesty M, et al. Xgboost: extreme gradient boosting[J]. R package version 0.4-2, 2015, 1(4).

[7]. Yu J, Luo G, Xiao T, et al. MOOCCube: A Large-scale Data Repository for NLP Applications in MOOCs[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 3135-3142.

[8]. Perozzi B, Al-Rfou R, Skiena S. Deepwalk: Online learning of social representations[C]//Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. 2014: 701-710.

[9]. Xiao Y, Wu J, Lin Z, et al. A deep learning-based multi-model ensemble method for cancer prediction[J]. Computer methods and programs in biomedicine, 2018, 153: 1-9.

[10]. Wang G, Zhang X, Tang S, et al. Unsupervised clickstream clustering for user behavior analysis[C]//Proceedings of the 2016 CHI conference on human factors in computing systems. 2016: 225-236.

[11]. Wang X, Huang T, Wang D, et al. Learning Intents behind Interactions with Knowledge Graph for Recommendation[J]. arXiv preprint arXiv:2102.07057, 2021.

[12]. Cox V. Exploratory data analysis[M]//Translating Statistics to Make Decisions. Apress, Berkeley, CA, 2017: 47-74.

[13]. Rousseeuw P J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis[J].

Journal of computational and applied mathematics, 1987, 20: 53-65.

[14]. Qiu J, Tang J, Liu T X, et al. Modeling and predicting learning behavior in MOOCs[C]//Proceedings of the ninth ACM international conference on web search and data mining. 2016: 93-102.

[15]. Wang H Y, Yang M, Stufken J. Information-based optimal subdata selection for big data linear regression[J]. Journal of the American Statistical Association, 2019, 114(525): 393-405.

[16]. Probst P, Wright M N, Boulesteix A L. Hyperparameters and tuning strategies for random forest[J]. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2019, 9(3): e1301.

MOOC learning data analysis and dropout prediction based on integrated learning

College of Photonic and Electronic Engineering Communication Engineering
1362017014 ZhengYiheng Advisor XuZhexin

【Abstract】massive open online courses (MOOCs) have developed rapidly and have attracted millions of online users. However, the high dropout rate is a persistent problem. In order to reduce the dropout rate, this problem is addressed by analyzing user behaviors . Here, a Behavior Transfer Interval Model (BTIM) is designed to convert the user's state and combined with a machine learning model to predict the tendency to drop out. In the BTIM model division, there are two modules of behavior interval distribution and behavior transition matrix. The behavior interval distribution module collects statistical indicators of learning behavior changes, and the behavior transition matrix module describes the transfer mode between learning behaviors in terms of frequency. This study finally achieved an efficient prediction of AUC 85.2% F1 90.15% for dropout users . This result is very close to a large number of excellent studies in the field, and the good model interpretation in this study is useful for analyzing the causes of dropout behavior and high-risk users.

【Keyword】online education; data analysis; feature construction; machine learning; time series