**MOOC大数据分析**

基于中心矩与n-gram对用户行为进行建模

**光电与信息工程学院　 通信工程闽台合作专业**

**136102017014（学号）　郑义恒（姓名）　　指导教师　徐哲鑫**

# 简介

## 论文简述

MOOC（Massive Open Online Course）旨在免费提供大学在线课程，使全世界都能接受高等教育，因此吸引了来自不同年龄段、教育背景和国籍的多元化生源。尽管如此，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。并且，随着诸如Coursera，edX和Udacity之类的大规模开放在线课程（MOOC）的迅速普及，MOOC引起了教育界和计算机科学界以及公众的关注。 以墨尔本大学于2013年首次提供的Coursera MOOC 平台上的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％[2]。我对来自XuetangX 在线教育平台的用户行为数据集进行了类似的统计, 在69823名用户中，平均通过率仅为24.31%，平均辍学率达到75.69%。因此，对提高课程通过率来说，对有辍学倾向的用户的识别成为了十分重要的一个环节。

对于线下课堂的教授者来说，判断一个学生能否通过这门课程往往根据xxx等主观因素与xxx等客观因素以下指标：课堂表现、课程的难度、作业情况、身边同学的成绩、家庭背景等等。对于在线教育也是如此，但在庞大的用户基数面前，让教师人工识别高风险学生成为了一件几乎不可能的任务。

因此，我们可以让机器来识别高风险用户，具体实现可以借鉴人工识别的思路，但课堂表现这类指标十分抽象，除此之外的指标，大都可以很轻松的进行量化的描述，唯独上课的表现、状态，在传统课堂中这依靠教师的经验进行主观判断，但在MOOC中我们要如何让计算机明白，学生处于什么样的状态呢？常见的方法是统计开课时间内，行为序列中各类型学习行为的频次、每周的页面点击次数、视频观看次数等指标，但现实中用户的真实学习时长各不相同，以固定时间段内的行为频次来量化学生的状态，这种方案忽视了序列中隐含特征的尺度不变性和平移不变性。

为了满足实际场景中对不定长序列的特征提取需求，我基于序列中的行为间隔分布情况和行为转移方式，设计了相应的**行为转移间隔模型（BTIM,Behavioral Transition Interval Model）**以量化本任务中学生的学习状态。实验表明，该模型对序列长度不敏感的同时能够提取有效特征。并将BTIM与传统机器学习算法结合，应用在Xuetang MOOC平台提供的用户行为数据集上，在识别辍学用户的任务中展现出了和目前其他深度学习方案高度接近的预测性能，并且不同于深度学习的黑盒特性，**BTIM**与机器学习结合的解决方案具有更好的解释性，能够根据拟合结果对辍学行为进行更具因果性的分析与推理，并且本方案结合了十分成熟的机器学习算法，在计算方面具有更快的收敛速度。

## 相关研究

应用认知模型评分

Understanding Dropouts in MOOCs[2]：

简述：

对XuetangX平台xx时间内的数据进行分析，在人口统计学的维度上进行了初步的分析，并通过选课行为以及点击情况对用户进行分类，提出了一种自适应的特征提取模型，其能够基于用户与课程的背景信息自动选择有价值的特征，并在预测辍学用户的任务上的性能优于机器学习模型。

该论文的团队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A/B测试，结果显示其干预策略显著的提升了用户的活跃程度。

在我的论文的辍学预测任务中使用了和此团队相同的训练数据与验证数据，并且达到了近似的预测性能，我的论文中提取了课程与用户人数以及辍学率特征，此点参考了此论文中对和特征的提取思路。

亮点：

可能存在的不足：

在该团队提出的CFIN模型中，对一个行为序列样本进行预测时，对模型有价值的特征来自三个方面：

* 该样本对应的课程的历史信息
* 该样本对应的用户的历史信息
* 样本本身的信息

也就是说，面对一名新用户、一门新课程所产生的行为序列时，由于历史信息的缺乏，对新样本的预测性能可能会降低。

1. 基于用户学习的时间模式对用户进行聚类

2. 发现了用户在选择主课时还会选择一门替补课程这一做法

3. 根据用户的选课情况，将课表重合度高的用户视为同伴，分析得出辍学同伴多的用户自身的辍学可能性高这一结论

4. 基于用户的学力有限这一原则，在模型中引入了注意力机制，以衡量每门课程在学习中的受重视程度。

在本文中，我使用和该团队相同的数据集进行辍学预测，获得了相近的结果（F1 score差距2%）

Modeling and Predicting Learning Behavior in MOOCs[9]：

1. 得出努力的学生不一定更有可能完成课程这一结论。

2. 学生完成课程的可能性与完成课程的好友数量呈正相关。

Identifying At-Risk Students in Massive Open Online Courses[10]：

1. 根据用户的先期行为（课程开始的数周内）识别可能辍学的用户。

Your click decides your fate: Inferring Information Processing andAttrition Behavior from MOOC Video Clickstream Interactions：

1. 将研究重心放在学生的思考方式上，应用了一种认知模型来解释MOOC视频流点击交互中涉及的动态认知过程，可以有效地将其用作对重要学习者的行为进行预测。基于“有限能力信息处理方法”构建学生的信息处理指标，该方法断言人们独立地分配有限的认知资源来完成任务，并分配权重来量化不同学习行为中获取的信息量。

2. 发现在报名的学生中，75％有参与学习：其中50％的人主要是看视频讲座，其余25％的人则主要参加家庭作业和作业。因此，大多数学生将观看视频讲座作为其主要的MOOC活动。

3. 对点击行为序列用不同的粒度进行进行建模，使用较高级别的类别而不是原始点击来表示点击行为，使得高级序列比原始的点击流序列具有更好的抗噪性，以及更低的向量维数。

4. 使用n-gram对行为序列进行统计，发现多种出现频次较高的行为组合。

MOOCs for Universities and Learners An analysis of motivating factors：

1. 提出不应该假设所有学习者都打算完成整个课程这一想法，许多学习者可能只对课程的一部分感兴趣，或是时间上不允许，他们知道在某个时间点后将无法完成课程。不过由于MOOC是免费的，不会因未完成而受到惩罚，这使得学习者能够灵活的参加（和退出）课程。

根据我目前的调研情况，还未有论文使用行为间隔的分布情况与行为转移矩阵来衡量用户的行为模式，在本文中我提取了这两种特征以分析用户的辍学行为。

## 数据集介绍

在这项工作中的分析是对来自XuetangX的数据集进行的。 XuetangX于2013年10月推出，现已成为中国最大的MOOC平台之一。它提供了1000多门课程，吸引了超过1000万注册用户。XuetangX有十二类课程：艺术，生物学，计算机科学，经济学，工程学，外语，历史，文学，数学，哲学，物理学和社会科学。 XuetangX中的用户可以选择学习模式：讲师监督模式（IPM）和自定义模式（SPM）。 IPM遵循与常规教室相同的课程安排，而在SPM中，可以有更加灵活的时间表来自己进行在线学习。通常，在XuetangX中IPM课程的时间跨度超过16周，而SPM课程的时间跨度较长。每个用户可以报名一门或多门课程。当学习一门课程时，系统会记录多种类型的活动：视频观看（观看，停止等），论坛讨论（问题提出、答复等），提交作业（正确/错误答案、重置答案）以及单击网页（单击、关闭课程页面）等活动。

XuetangX数据集包含了四类文件：

**行为日志（无辍学标签） ：** 2015年至2017年间在XuetangX平台上所有用户的学习行为日志。

**行为日志（有辍学标签）**：本论文中用于辍学预测任务的训练集和测试集，同时也是**Understanding Dropouts in MOOCs[2]**论文中评价模型预测性能所使用的数据集。

**用户信息文件** ：年龄、性别、学历。

**课程信息文件** ：开课时间、结课时间、课程属性（IPM 或 SPM）、课程类别（经济学、工程学等）。

#HERE

辍学预测数据集中含247门课程，均为IPM课程， 课程的平均辍学率是 78.73% ，每门课程平均有639人参与。

图1.课程的参与人数直方图

数据集中含69823名用户，平均辍学率是75.69%，平均每位用户参与2.6门课程，24.8%的用户参与了两门以上的课程，在这24.8%的用户中平均每人参与4.6门课程。

图2.学生报名课程数量的直方图

（出于对样本质量的考虑，将选课数量在三门以下的学生以及学生数量在三人以下的课程被排除在统计数据之外）

（1-1）

（1-2）

基于以上结果，我假设：

辍学的学生与不辍学学生之间存在不同的学习习惯。

学生个人的学习习惯与其历史信息存在关联。

课程的类别等属性不同会导致不同的学习习惯。

基于对数据的探索性分析以及以上假设，本文中我以历史信息与行为特征来分析学生辍学的可能性。

历史信息采用数据集中的course\_info.csv 与user\_info.csv文件。

行为特征从train\_log.csv中提取，特征提取过程分为两个部分，分别是单步行为间隔信息和行为转移矩阵。单步行为间隔指的是行为日志中时间相邻的两条记录之间的时间差。行为转移指的是，用户依次进行的两次行为之间的转移关系。再将提取到的特征作为训练数据，训练机器学习模型对辍学样本与非辍学样本进行分类。

## 探索性数据分析

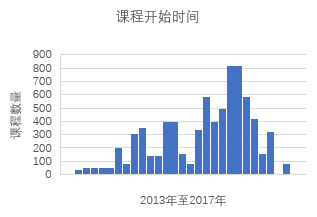
行为日志（无辍学标签）数据概览[2]：

|  |  |  |
| --- | --- | --- |
| 类型 | IPM | SPM |
| 视频活动 | 50,678,849 | 38,225,417 |
| 57% | 43% |
| 论坛活动 | 443,554 | 90,815 |
| 83% | 17% |
| 作业活动 | 7,773,245 | 3,139,558 |
| 71% | 29% |
| 网页活动 | 9,231,061 | 5,496,287 |
| 63% | 37% |

表1-1 ：行为日志内各类行为的计数值，IPM与SPM分别代表讲师监督模式与自定义模式这两类不同课程的统计值。

辍学预测数据集概览：

我根据日志数据集中的索引找到在用户信息以及课程信息中与日志索引相对应的数据，以分析数据集内样本的分布情况。分布情况如下：



样本均衡情况:

在辍学预测数据集中，辍学样本占比76%，非辍学样本占24%。

# 特征工程

## 数据清洗

辍学预测数据是以字符串类型存储的逗号分隔符文件（csv），在有限的硬件条件下，为了降低在特征工程中查询数据时的时间与内存开销，有必剔除冗余数据，并选用更高效的数据结构。

数据清洗流程中我使用Python语言进行实际操作，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，以便后续分析。

在对整个日志文件的遍历过程中，同时得到“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。

（groupby示意图）

## 时间间隔特征

2.2.1 阐述：

2.2.2提取方式

## 行为转移特征­­

## 历史信息

## 量化方式

## 历史信息

# 预测模型

目前我已经通过特征工程量化了用户的行为特征，并将其与历史信息进行组合，余下的的工作就是基于这些特征，选择合适的模型拟合特征和辍学标签间的关系。

我希望可以通过模型拟合的过程了解用户辍学的成因，所以选用了易于解释的决策树模型以及线性回归模型。

# 回归

# 线性

# 非线性

## 模型1

回归模型

## 模型2

分类模型

# 总结

# 参考文献