探索 MOOC 数据背后的学习行为分析

地球与空间科学学院 李柯毅 2100012618

摘要：MOOCs（大规模开放在线课程）已经成为当代教育领域的关键组成部分，通过其巨大的学习资源和灵活性吸引了全球范围内的学习者。本研究旨在探索 MOOC 学习行为，从而深入了解学生的学术成就和参与模式之间的关系。利用关联性分析，聚类分析和回归分析，我们对 MOOC 数据进行了全面的探索性分析。首先，我们使用关联性分析确定了不同变量之间的相关关系，揭示了各项行为指标之间的潜在联系。随后，通过聚类分析，我们将学生划分为不同的群体，从而揭示出不同学习特征的群体模式。最后，通过回归分析，我们尝试建立模型来预测学术成就并探索影响学生成绩的关键因素。本研究对于深入理解 MOOC 学习行为、改善教学方法以及提高学习者成绩具有重要的指导意义。

关键字：MOOC，学习行为，关联性分析，聚类分析，回归分析，教学模式。

1. 背景和数据介绍

我们收集了来自 MOOC 平台的学员数据，包括注册时间、地理位置、活动记录等。数据预处理包括缺失值处理和标准化，以确保数据质量和一致性。数据具体变量名称及涵义如下：

session_user_id: 用户会话 ID	signature_track_register_time: 账号注册时间
locale: 用户所在地区	duration: 持续时间
timezone: 用户所在时区	begin_time: 开始时间
access_group_id: 访问组 ID	total_actions: 总行动次数
registration_time: 注册时间	assignment_score_sum: 作业分数总和
last_access_time: 最后访问时间	lecture_item_views: 讲座查看次数
email_announcement: 是否接收电子邮件通知	quiz_score_sum: 测验分数总和
email_forum: 是否接收论坛电子邮件	forum_posts_count: 论坛帖子数量

in_signature_track: 是否有账号	forum_posts_votes: 论坛帖子投票数
wishes_proctored_exam: 是否希望监考考试	forum_comments_count: 论坛评论数量
email_review: 电子邮件审核	forum_comments_votes: 论坛评论投票数
deleted: 是否已删除	forum_reputation: 论坛声望
grade: 成绩	video_views: 视频观看次数
	page_views: 页面访问次数

2. 关联性分析

本研究中我们分析了成绩（grade）与其他各数值变量之间的相关系数（图 1）并绘制了散点图（具体实现的 python 代码附在文末）。

直接对原始数据绘制散点图发现，很多变量受到离群值感受，相关性表现不明显。故先对数据做清洗工作，剔除每个变量最大的三个值所在的行，之后以 grade 为纵轴，以其他变量为

	Correlation with Grade
last_access_time	0.26231234500996126
duration	0.42451778995698863
begin_time	-0.16419644345423884
total_actions	0.6549388940143915
assignment_score_sum	0.980114476026383
lecture_item_views	0.3214152856749906
quiz_score_sum	0.853687677157159
forum_posts_count	0.21528690019920338
forum_posts_votes	0.11375347586269119
forum_comments_count	0.21612798815658907
video_views	0.3615077916700441
page_views	0.7931244438882151

图 1：grade 与其他各参数的相关系数（这里保留 0.1 以上的值）

横轴绘制散点图。结果表明，学生总成绩与学习任务成绩、小测成绩、页面访问次数三项呈现较为明显的正相关关系，与注册时间、最后访问时间、总行动次数、论坛帖子数量、视频观看次数五项呈现较弱的正相关，与其他项无明显相关性（图 2）。

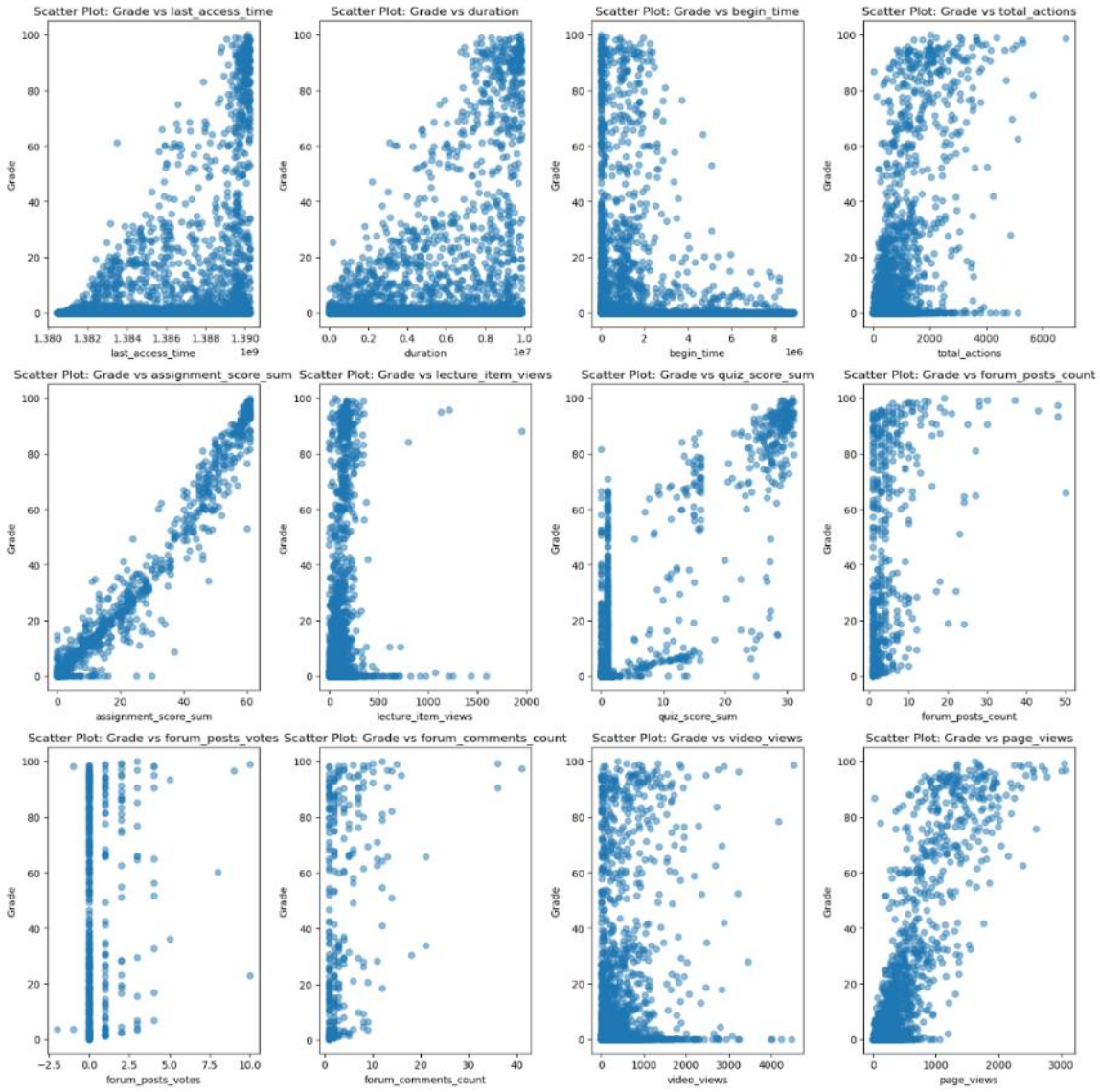

图 2：grade 与其他各参数绘制散点图

3. 聚类分析

本研究利用 WEKA 对 12690 名学生做聚类分析，采用 K 均值聚类算法，该算法能够将学员分成不同群体，以相似性进行分类。经过试验我们发现，将分组参数设置为 5 时结果可解释性最强，该数目可以最大效率利用各变量信息，同时有效避免过度分类的问题。我们使用欧氏距离作为相似性度量，其他参数均保持默认设置（图 1）。

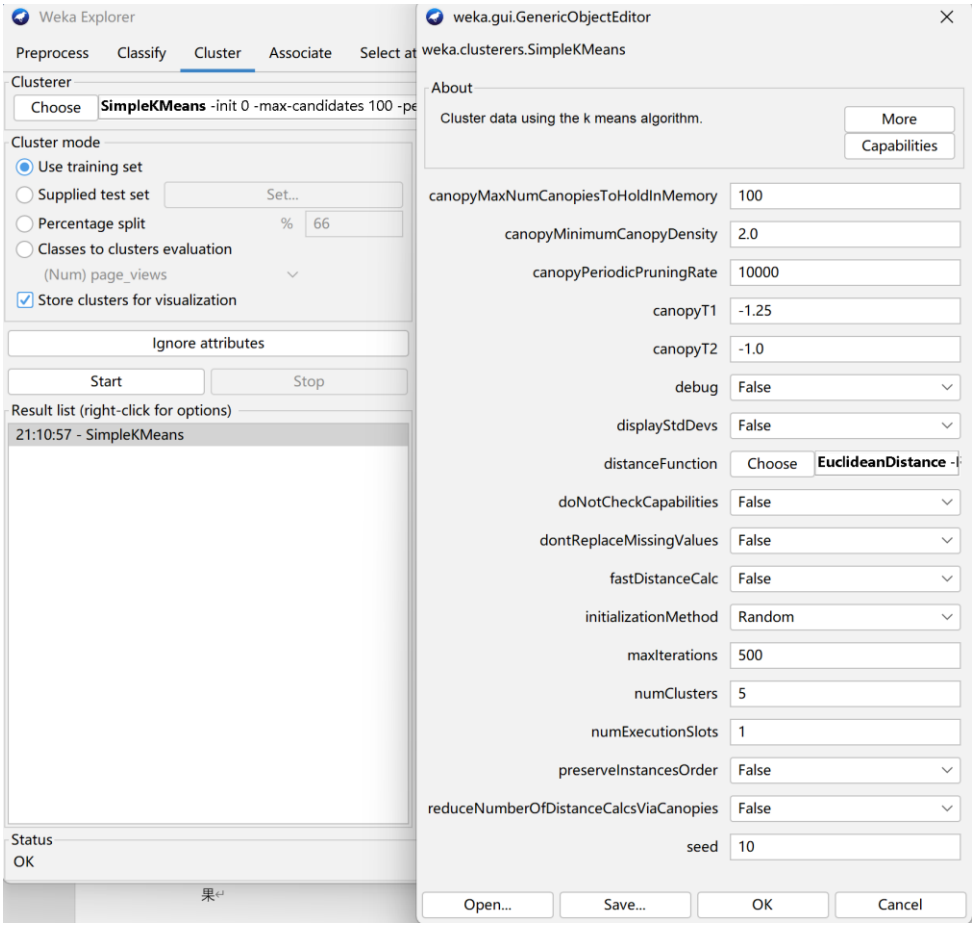

图 3: 聚类模型及参数设置

操作得到聚类结果，该模型误差平方和为 16579.46，误差控制良好。利用 WEKA 绘制散点图，但由于数据数量较多，集中度较高，此处散点图展示效果欠佳。对于每类数据特点作进一步分析，发现聚类群体特征主要体现在三个方面：

A、地理位置差异：各群体的地理位置和所在时区各不相同，包括来自不同地区的学员，比如亚洲和美洲的不同城市。

B、学习持续时间：不同群体的学习持续时间也有所不同，从注册到最后一次访问的时间跨度存在差异。

C、行为模式差异：不同群体的行为模式存在显著差异。有些学员可能更活跃于特定功能，比如在论坛上交流、观看视频、参与测试，有些学员涉及的活动较少。

根据以上特征解读聚类得到的五组数据，每组数据以中心数据的特点表示，可以得到以下结果：

①Cluster 0（3%）：

用户 ID: dd7d6a83cda141c356ae23bbc5a974b444bc060b

地区: Asia/Chongqing

注册时间和最后访问时间: 1381331413, 1382603233

特点：该群体可能是亚洲尤其是重庆地区的学员，其注册到最后访问的时间较短。

行为模式：参与度较低，涉及的活动比较少，但在特定功能上可能有一定的行为。

②Cluster 1（42%）：

用户 ID: ff95fcf5cbd98a6e28554720ef5f173be9a46020

地区: America/Los_Angeles

注册时间和最后访问时间: 1386680615, 1389930389

特点：美国洛杉矶地区的学员，注册到最后访问的时间较长。

行为模式：参与度较高，在多个功能上有较多的活动，包括论坛交流、视频观看、测验参与等。

③Cluster 2（7%）：

用户 ID: 5ec3ece9d43f88ab511fb688a13cecc1a54dc723

地区: Asia/Shanghai

注册时间和最后访问时间: 1380609645, 1385135849

特点：亚洲上海地区的学员，注册到最后访问的时间较长。

行为模式:在多个功能上有着活跃的参与,尤其视频观看、论坛数量显著高于其他组别。

④Cluster 3（13%）：

用户 ID: c9b38f020612aff9f433c07ff53af3c9b7d214c9

地区: America/Los_Angeles

注册时间和最后访问时间: 1379424411, 1381562196

特点：美国洛杉矶地区的学员，注册到最后访问的时间较短。

行为模式：相对较少的参与度，在少数功能上有行为表现。

⑤Cluster 4（35%）：

用户 ID: 7bf55820d11be58487610d8a8e5ef064eebf3005

地区: America/Los_Angeles

注册时间和最后访问时间: 1385967043, 1385967102

特点：美国洛杉矶地区的学员，注册到最后访问的时间很短，有较低的活跃度。

行为模式：在极少功能上有行为表现，几乎没有参与各类课堂活动。

图 4：聚类结果及可视化

通过本项聚类研究，我们发现 MOOC 上学生可根据地区、学习时长和行为模式分为 5 类，每种类别的特点与学习成绩也呈现出一定的相关性：高活跃度全面参与的类别学习成绩最高，中等活跃度较为被动的类别成绩次之，活跃度低的类别成绩最低；

令人意外的是，高活跃度视频与论坛类别（Cluster 2）成绩只是中等，或许说明特定功能的活跃度对学术成绩影响不大。当然该结果很可能是受到地域的影响，该类别主要为中国尤其是上海等大城市的样本，学生可能在老师的高压下完成慕课各项任务，或者只是盲目刷高各项任务的参与度来换取好的成绩。这也反映在样本中，视频观看和论坛交流等没有很多硬性要求且可以按时长或次数计分的项目参与数据很好，但其他要求较高或者不方便计分的项目则参与度则很低。因而，他们实际的学习效果或者学习意愿可能并不如各项任务数据一样优秀，这可能也间接反映在他们最终的成绩上。相比之下，美国的学生的竞争压力相对较小，没有如此“内卷”的环境，其任务完成度可以更好反映他们的学习兴趣和学习投入，因而美国地区高参与度的类别成绩也位居首位。以上分析基于现实经验推断而得，具体结论仍需更多数据和调研以进一步验证。

针对不同群体的行为模式和兴趣点，教育者可以提供个性化支持和指导，使得学习体验更加贴合学员需求，并满足不同学员的学习偏好。对于数据反映出来的部分地区盲目竞争的情况，教育者也应善于引导，制定合理的评分规则，减少无效内卷的发生。此外，对于流失率较高的群体，可以实施措施鼓励更多参与和持续学习，以提供更好的学习效果。

4. 回归分析

对学生成绩和其他各项指标做回归分析，发现学生总成绩与学习任务成绩、小测成绩和页面访问次数之间出现明显的正相关关系，R^2等于 0.924，学习任务成绩、小测成绩和页面访问次数三项均在 0.001 条件下显著。其原因推测如下：①学习任务成绩通常是课程中重要的组成部分，它们涵盖了学生在课程中完成的作业、项目或考试成绩。如果课程设计合理，学习任务成绩与课程总成绩可能有较高的相关性。②小测成绩可能是课程中对学生知识点掌握情况的测验，最终成绩与小测成绩正相关，可能暗示了学生对课程内容的理解和掌握程度。③页面访问次数可能代表了学生在课程平台上的活跃程度和学习投入。较高的页面访问次数可能意味着学生更加勤奋地学习、复习或参与课程讨论，与此同时，学习投入可能带来更好的学习成果。综上，学生倾向于在学习任务成绩高、小测成绩好的情况下更加努力学习。他们可能更倾向于参与更多的在线活动和课程资源，这种积极的学习态度可能会带来更好的学习结果。

OLS Regression Results

Dep. Variable:	grade	R-squared:	0.924
Model:	OLS	Adj. R-squared:	0.924
Method:	Least Squares	F-statistic:	6436.
Date:	Sun, 12 Nov 2023	Prob (F-statistic):	0.00
Time:	02:12:55	Log-Likelihood:	-5635.4
No. Observations:	1597	AIC:	1.128e+04
Df Residuals:	1593	BIC:	1.130e+04
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-8.8043	0.337	-26.113	0.000	-9.466	-8.143
assignment_score_sum	0.8903	0.020	44.942	0.000	0.851	0.929
quiz_score_sum	0.7333	0.036	20.363	0.000	0.663	0.804
page_views	0.0169	0.001	25.684	0.000	0.016	0.018

Omnibus:	31.516	Durbin-Watson:	2.024
Prob(Omnibus):	0.000	Jarque-Bera (JB):	21.210
Skew:	-0.156	Prob(JB):	2.48e-05
Kurtosis:	2.530	Cond. No.	1.14e+03

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.14e+03. This might indicate that there are
strong multicollinearity or other numerical problems.

图 5：学习任务成绩、小测成绩和页面访问次数线性回归结果显著

此外对关联性分析中呈现若相关性的变量视频访问时间和总行动次数和总成绩做回归分析，发现结果也在 0.001 水平上显著，但 R^2略小。（图 6）。

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  grade   R-squared:                       0.741
Model:                            OLS   Adj. R-squared:                  0.741
Method:                 Least Squares   F-statistic:                     2281.
Date:                Sun, 12 Nov 2023   Prob (F-statistic):               0.00
Time:                        02:17:16   Log-Likelihood:                -6612.0
No. Observations:                1597   AIC:                         1.323e+04
Df Residuals:                    1594   BIC:                         1.325e+04
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          1.0172      0.523      1.944      0.052      -0.009       2.044
total_actions  0.0539      0.001     62.618      0.000       0.052       0.056
video_views   -0.0653      0.002    -41.951      0.000      -0.068      -0.062
==============================================================================
Omnibus:                      381.694   Durbin-Watson:                   2.017
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1412.166
Skew:                           1.129   Prob(JB):                    2.25e-307
Kurtosis:                       7.016   Cond. No.                     1.86e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.86e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

图 6：视频访问时间和总行动次数线性回归结果显著

5. 总结：

①学生总成绩与学习任务成绩、小测成绩、页面访问次数三项呈现较为明显的正相关关系，与注册时间、最后访问时间、总行动次数、论坛帖子数量、视频观看次数五项呈现较弱的正相关，与其他项无明显相关性。

②通过本项聚类研究，发现聚类群体特征主要体现在三个方面：地理位置、学习时间、行为模式。针对不同群体的行为模式和兴趣点，教育者可以提供个性化支持和指导，使得学习体验更加满足学员兴趣，实现因材施教。

③回归分析得到，总成绩与学习任务成绩、小测成绩和页面访问次数、变量视频访问时间和总行动次数均在 0.001 条件下显著，呈现明显的正相关关系。