MOOC 中学生流失现象分析与预警

陈云帆，张铭
（北京大学信息科学技术学院，北京 100871）

摘要：慕课（Massive Open Online Course，简称 MOOCs，音译为“慕课”）是一种兴起于 2012 年的新型的在线教育模式。经过一段时间的发展和观察，学生参与 MOOC 课程往往很难坚持到最后，大约只有平均而言 5% 左右的学生可以完成课业任务。学生的严重流失现象制约了 MOOC 的发展。本文通过观察了《数据结构与算法》 MOOC 课程学生完成课程的情况，对影响学生的因素进行了分析，并结合课程设计进行了深入讨论。此外，本文提出了一种通用的在线预警学生流失系统，以期帮助降低学生流失率。该系统由行为采样器、分类器构成，对不同的课程，提供可选的差分器和衰减器，通过无人工监督的方式，在课程进行中对学生流失进行预警。该系统基于《数据结构与算法》课程进行了深入测试。为了验证其可扩展性，对于《计算概论》 MOOC 课程进行了重复实验，结果表明系统具有较好的可扩展性，但在扩展使用时，需要分析课程特点，选择适当的可选组件。基于以上研究，结合通过对行为特征的分析，提出了部分引导学生学习的可行建议。

关键词：慕课；学生流失预警；学生引导
中图分类号：TP311

Student Churn Analysis and Prediction on MOOC

CHEN Yunfan, ZHANG Ming
(School of EECS, Peking University, Beijing 100871)

Abstract: The campaign of Massive Open Online Courses (MOOCs) in the area of e-learning and distant education gains significant popularity among both students and educators.As observation of MOOC gets deeper, studies pointed out that student, however, can hardly finish a course on MOOCs. Less than average 5% of students could follow course. Such a low retention rate will restrict development of MOOCs in future. This paper shows observation of data on MOOCs. A factor analysis is applied to students' behavioral data. The result is discussed with course design on MOOC platform. A general online system for predicting students' churn is developed. The system consists of sampler, classifier, optional differentiator and optional attenuato. With unsupervised learning, this system can fit on different on-going courses. A discussion of this system with the course Data Structures and Algorithms is conducted. To test the scalability of the system, the system also applied on the course Introduction to Computing. Scalability of the system is fine but an analysis of course and to choose optional components based on the analysis is required. Additionally, based on research above, some suggestions for instructing students on MOOC is given.

Key words: MOOC; Prediction of Churn; Instructions for students

0 引言

从 2012 年开始，大规模开放在线课程（Massive Open Online Course，简称 MOOC，中译“慕课”）飞速发展。北京大学从 2012 年底酝酿投入这个很可能影响高等教育未来生态的战略方向，2013 年 9－10 月间 14 名来自理工和人文社科的教师开出了 11 门慕课。北京大学张铭教授主讲的《数据结构与算法》课程[1]于 2013 年 10 月 20 日正式上线 Coursera 平台，

基金项目：自然科学基金项目（项目编号：61472006）；博士点基金（项目编号：20130001110032）
作者简介：陈云帆（1992-），男，北京大学信息科学技术学院智能科学系本科生
通信联系人：张铭（1966-），女，教授，主要研究方向：文本挖掘、社会网络分析. E-mail: mzhang@net.pku.edu.cn
[1] https://www.coursera.org/course/dsalgo

在课程期间，共有来自 112 个不同国家和地区的 13,683 名学生注册并参与了课程学习。其中有 9,088 名学生来自于发展中国家。

课程共持续 14 周，后两周内容为选修课程，不参与课程评分。课程的评分由平时作业和期中考试两部分构成，平时作业包括客观题目测试（Quiz）和编程作业（Programming Assignment）两部分。分数组成为：论坛讨论 10%，课程小测 30%，POJ 编程作业 20%，期中 15%，期末 25%。

虽然《数据结构与算法》课程吸引了上万名学生注册，然而最后完成课程的学生仅有 55 人。若放宽标准，以访问课程来衡量，最终也仅有 1037 人完整跟下了课程。保留率不足 10%，流失情况如图 1 所示。

类似的情况也可以在其他课程中见到，例如美国麻省理工学院在 edX 上开设的《电路与电子》课程吸引了 154,763 名学生选修，最终只有 7,157 名学生完成了课程[1]。

由于 MOOC 课程位于开放、自由的互联网上，课程又往往伴随着严格的时间要求，所以导致学生流失率在 MOOC 上格外严重。由于 MOOC 数据规模大、实时性高，使用计算机的手段进行分析和研究是很有必要的。然而目前尚未见到基于计算机的手段进行流失率分析和研究的报告。

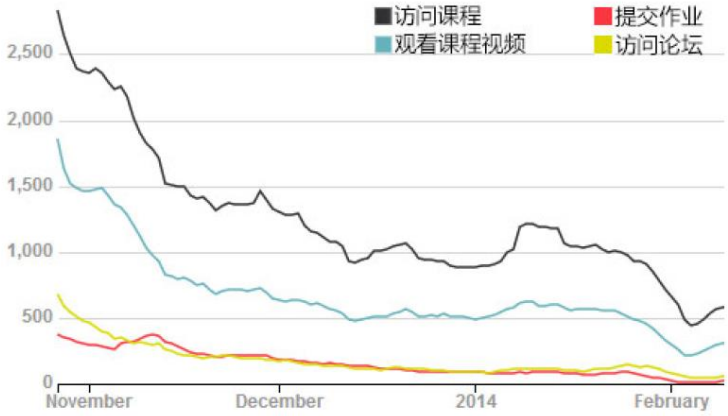

图 1 数据结构与算法学生流失情况

Fig.1 Student Churn of Data Structures and Algorithms Course

在相关研究部分，本文讨论了近年来关于 MOOCs 的研究，重点关注了关于学生行为数据的分析结论。在影响流失的学生行为数据分析部分，通过统计分析，对影响学生流失的因素进行了分析，并对结果进行了讨论。研究了未完成数据结构与算法课程的学生行为表现，并提出了一些改进建议。

本文介绍了一个学生流失预警系统。作者基于机器学习的方法，开发了一个 MOOC 学生流失预警系统。系统依赖前面对行为数据分析的结果，建立了有效的诊断机制，对学生流失情况能起到很好的预警效果。利用该系统，教师在课程进程中，可以有针对性地调整课程、进行引导，以帮助学生完成课程学习。该系统除了可以运行在《数据结构与算法课程》上，也在《计算概论》MOOC 课程上进行了测试。测试结果显示了系统的通用性，也指出了部分不足。最后，作者针对上述研究存在的问题，以及相关的研究课题，提出进一步研究的方向和具体方法建议。

1 相关研究

在教育学、社会学以及心理学中，关于学生流失的研究始于 1938 年[2]。由于那时美国

大学的辍学率高达 45%，引发了这个横跨 25 所学校的调查研究。随后，通过分析学生参与学校学习的过程[3]和退学动机[4]建立了解释学生退学行为的整合模型。也有研究者研究大学入学新生引导课程的完成与学生流失之间的关系[5]。

MOOC 作为新生事物，目前对其的研究方向较为分散。但是已经有部分研究者指出，高流失率影响了 MOOC 的发展，应当作为研究重点进行深入研究。Ramesh 通过建模[6]将用户分为积极和消极两类，同时对用户参与情况进行建模，以预测成绩[7]。Kizilcec 等人借助三门 MOOC 课的数据对用户进行聚类，预测他们是否会流失[8]。上述相关研究表明，MOOC 中的用户流失问题已经成为了一个非常重要的研究课题，国内外相关研究人员已从多角度展开了研究和分析。但是，目前还没有对用户流失问题进行一个系统和全方位的分析。而且，很多已发表的研究方法还局限于小规模的数据，并没有很好地利用 MOOC 大规模数据的特点。

由于 MOOC 是以一门课程而非一系列课程为单位的，同时用户规模与之前相比有了显著增加，而此前的研究主要研究学生从校园中退学的流失，所采用的研究方法和结论多数并不适用于 MOOC，为此需要寻求新的研究手段和研究工具。

2 学生行为数据分析

在 MOOC 上学习的学生在流失前，会存在一些可观察到的预兆，例如访问课程次数减少、论坛活跃度下降、不再重视作业和考试等。通过观察，作者列举出 4 类 15 条学生行为数据。学生流失前，除了可以观察到行为数据的绝对数字以外，应当也可以观察到行为数据的改变。共有 12 类行为数据可以计算以周为单位的差分。对参与了计分测试用户行为变量及行为变量的差分与是否保留做相关性分析，得到表 1。

可以观察到，对于《数据结构与算法》课程，行为数据的一阶差分并不单独具有相关性。也就是说学生是否会流失取决于行为的绝对数字，而非变化量。这可能是由于《数据结构与算法》课程每周的模块性较强，相互之间关系并不非常大，因此学生很有可能会有选择性地参与某些周的学习，或按照自己的节奏调整学习顺序。同时，作者观察到，学生并不热衷于在论坛上对其他同学给出负性评价，导致这一行为的数据量极少，无法进行分析。

有多种行为数据特征和是否流失直接相关，可以根据这些相关性，对学生参与课程学习给出一定的指导：

(1) 访问课程：学生访问课程即表明学生不会流失，为了帮助学生跟上课程，应当在每周发送通知，提醒学生访问课程查看信息内容；

(2) 页面浏览次数：学生访问课程页面的次数越多表明学生越不会流失，所以应当伴随每周发送的通知，对本周发布的内容进行简介，并对学生应当浏览哪些课程材料给出指导；

(3) 观看视频次数：每周的引导页面应当直接给出视频链接，引导学生观看视频；

(4) 视频暂停次数：暂停行为可能意味着学生使用零散时间进行学习，为了给学生这样学习提供方便，应当尽量缩短单次视频的时间，并在视频中穿插视频内练习，以缩短单次学习时间；

(5) 测试行为：学生参与计分测试次数越多越不容易流失，而不是一次成功最好，这可能是因为反复尝试直至成功给学生极大的成就感，鼓励学生进一步学习，因此应当在合理范围内尽量提供多次尝试的机会；

(6) 互动行为：几乎所有的互动行为都让学生更容易坚持学习课程，因此，老师和助教应当尽量创造活跃、轻松的论坛氛围，鼓励学生参与论坛讨论。

表 1 学生行为数据

Tab.1 Students' Behavior Data

类别	数据	可差分	与保留与否的相关性	差分及其相关性
访问行为	是否访问课程	否	0.300**	N/A
	页面浏览次数	是	0.390**	-0.002
学习行为	视频观看次数	是	0.146**	0.040
	平均暂停次数	否	0.149**	N/A
	平均播放速率	否	0.035	N/A
测试行为	尝试作业次数	是	0.198**	0.021
	尝试编程次数	是	0.357**	0.013
互动行为	浏览论坛次数	是	0.248**	-0.063*
	浏览帖子次数	是	0.245**	-0.039
	发表帖子次数	是	0.161**	0.031
	发表评论次数	是	0.106**	0.009
	点赞同次数	是	0.083**	0.006
	点反对次数	是	0.042	-0.042
	增加标签次数	是	0.042	0
	删除标签次数	是	b	B

**. 在 0.01 水平（双侧）上显著相关

*. 在 0.05 水平（双侧）上显著相关

b. 数据差异太小，无法回归

3 学生流失预警系统

学生流失预警系统是本文作者提出的可以在线运行的无人工监督系统，可以适应不同的课程特性，对学生的流失进行合理预警。

3.1 系统模型

系统由每一周用户行为采样器，可选的差分器、可选的衰减器、分类器四部分构成，接受用户行为数据，输出用户是否会流失的预测结果，如图 2。对于其中的可选部分，当不使用衰减器时,分类器输入将选用最后一个差分器的结果；当不使用差分器时，衰减器输入为各周采样器结果；当两者均不使用时，分类器输入为各周采样器结果。

运行时流失，是流失的一种近似，描述的是在课程运行过程中如何判定用户是否已经流失。为了给出合理的定义，本文对学生连续未访问课程周数与是否最终流失进行了统计，如图 3 所示。数据选取参与了课程评分项目的课程用户共 3714 名。可以观察到，如果用户连续不少于三周没有访问课程，保留的可能性则不大。因此，本文定义运行时流失为：截至目前，连续三周没有访问课程。

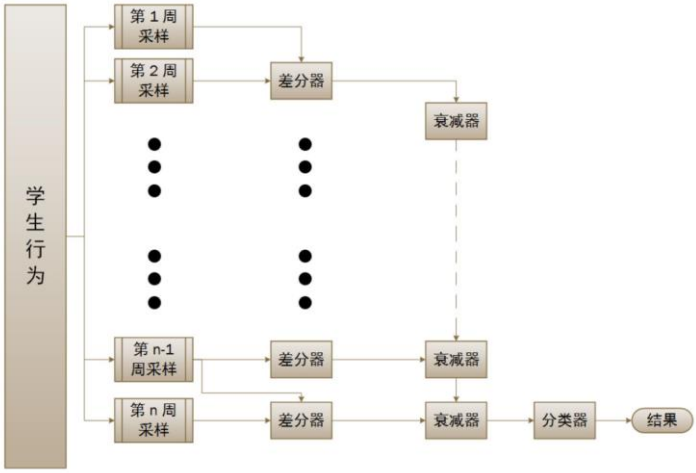

图 2 流失预警系统逻辑

Fig.2 Churn Prediction System Outline

3.2 参数训练

本系统为实时在线系统，拥有完整的时序逻辑。在任何时刻 n（第 n 周），系统以 n−3 及之前的数据（图 4 中的红色部分）作为系统输入，并以 n–2 到 n 周数据给出用户是否在运行中流失，作为用户流失与否的标签（图 4 中黄色部分），进行分类器训练。

训练完成后，可以以所有 n 周的数据作为输入，预测每个用户是否会流失。系统在 n > 4 的条件下可以运行。

在行为感知到分类器之间的一个可选部分为差分器。由于用户行为变化可能是重要的因素，例如用户可能在流失前减少访问课程的次数，因此可以考虑添加本层。处于系统复杂度的考虑，仅采用一阶逆向差分器。

第 n 个差分器的输入为第 n−1 周的采样X_{n-1} 及第 n 周的采样X_n，输出为Y_{n-1}，差分结果和原向量构成新的特征向量。当 n ≥ 2 时，变量间关系满足公式 1。

$$Y_n = \begin{pmatrix} X_n \\ \nabla X_n \end{pmatrix} = \begin{pmatrix} X_n \\ X_n - X_{n-1} \end{pmatrix} \quad \text{(公式 1)}$$

当 n=1 或不选择使用可选差分器时，Yn=Xn。在后续的实验中发现，对于连续性较强，课程内容耦合度较高的课程，应当使用差分器。

可以合理推测，用户越近的行为越可能有预测力，但较远的行为也可能具有效力，因此可以采用指数衰减的方法进行建模，可以添加衰减器。衰减系数定义为 α (0−0.5)，数值越大表明越多地保留往周数据，为 0 时与不选择衰减器相同。定义可选衰减器的输出为 Zn，变量间关系满足公式 2。

$$Z_n = \alpha^{n-1} Y_1 + \sum_{k=2}^{n} (1-\alpha)\alpha^{n-k} Y_k = \begin{cases} Y_n & n = 1 \\ \alpha Z_{n-1} + (1-\alpha) Y_n & else \end{cases} \quad \text{(公式 2)}$$

当不选择使用可选衰减器时，Zn=Yn。在后续的实验中发现，对于连续性较弱，课程内容关联度较低的课程，应当使用衰减器。

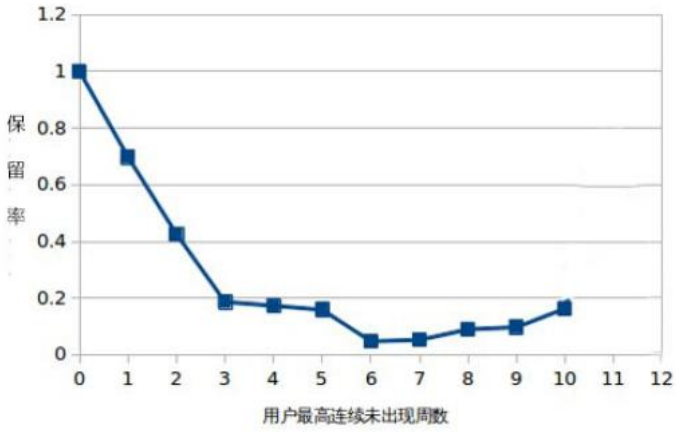

图 3 运行时流失定义依据

Fig.3 Definition of Runtime Churn

图 4 系统的时序逻辑
Fig.4 Timeline of Churn Prediction System

3.3 实验评测

实验数据选取《数据结构与算法》课程在 Coursera 平台上的第一期课程，只采用计分的前 12 周进行实验。由于系统特性，实验从第 4 周开始，至第 9 周止。为了选取衰减器合适的参数，只选取衰减模块进行系统性能测试，测试指标为 F1-Measure，随着课程的推进，测试结果如下。综合而言，取值为 0.5 时具有较好的效果，因此后续测试采用 0.5 的取值。

本文对可能的四种系统结构 进行了实验，各种条件下的 F1-Measure 可见图 6。可以见得，在 n≤7 的条件下，只加入衰减器为最优选择，当 n>7 的时候，只加入差分器最优，但此时四种条件相差无几。综合而言，系统只使用衰减器为最佳选择。

从实验结果可以发现，课程初期的用户动机差异较大。系统在运行的初期阶段效果不佳，这主要是由于第一周访问课程的用户较多。而这部分用户存在大量不具有课程所需基础只是访问课程观看内容的用户，和对 MOOC 形式感兴趣想要观察课程的用户。因此，他们的行为特征与主要用户不同，在数据中构成了噪音，导致在最初阶段系统的性能不佳。为了提升这部分效果，可以考虑对用户动机进行研究。

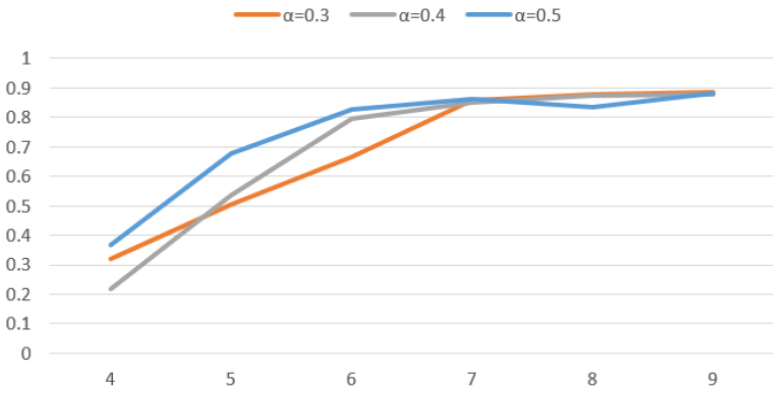

图 5 不同 α 参数下的系统效果
Fig.5 System test result with different α

同时可以发现，的确如假设的一样，用户连续多周的低频行为预示着用户流失，单独一周的预测力不如多周结合并进行合理衰减的结果。另外，使用差分器的效果不佳，可能存在两个原因。一方面是因为需要一定的绝对数量的行为才能跟上课程，而用户跟不上课程可能会流失。这部分底线对于不同水平的用户均存在。为了提升系统的效果，应当考虑研究用户的知识背景，并借此提升系统效果。另一方面是由于课程的耦合度较低，模块化较强，用户可以自己安排课程学习顺序。从第七周—图的内容开始，使用差分器的效果有所提升，这也从另一方面说明了差分器的效果取决于课程内容的耦合度。因为从“图”这一章的内容开始，需要此前的内容作为基础知识进行学习，不能够自由地安排学习顺序。

通过差分器的效果差异可以得出，为了帮助学生跟上课程，应当在课程时间表中注明各

部分内容的依赖关系，以便学生安排学习。同时，应该引导学生每周访问课程内容或参与论坛讨论，鼓励学生在零散时间进行学习。通过合理设计课程视频、学习材料，使学生可以有更高的学习自由度。

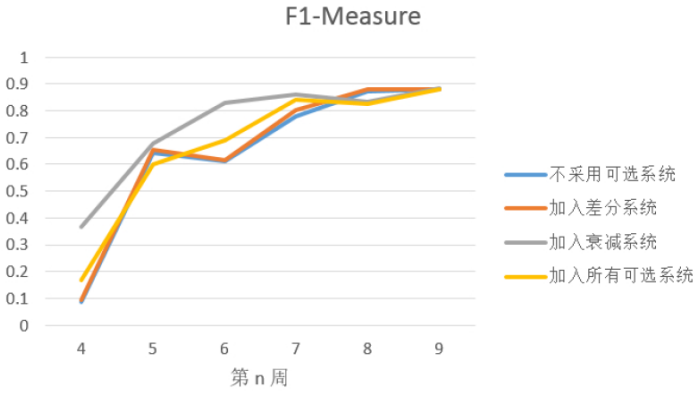

图 6 四种系统结构下的实验结果
Fig.6 Test result of 4 system structures

3.4 可扩展性实验

作者选用了北京大学同期开设的《计算概论》Coursera 慕课课程[2]来检验系统的可扩展性。实验结果证实了系统具有一定的可扩展性，也验证了部分实验结果讨论中的假设。

作者对可能的四种系统结构进行了实验，各种条件下的 F1-Measure 可见图 7。可以见得，只加入差分器为最优选择。虽然看起来在最初的两周不加入任何可选部分最优，但在第四周时，若不加入任何可选部分，事实上系统无法运行。综合而言，系统只使用衰减器为最佳选择。本文将两门课程最好的实验结果进行对比，对比结果可见图 8。可以见到，虽然实验结果趋势相似，但系统对于《计算概论》课程的整体性能不如《数据结构与算法》课程。

作者发现，《计算概论》课程选用差分器效果最佳，而采用衰减器效果不佳。这是因为《计算概论》的连续性较强，一周的缺课可能导致后续课程无法跟进，而《数据结构与算法》课程则是一段时间缺课才会导致无法跟上。这与上一实验结论中的模块化决定差分器及衰减器性能一致。本系统针对不同的课程特性，构架中的可选部分应当有调整。

《计算概论》课程的预测结果不如《数据结构与算法》课程的另一个原因，是《计算概论》课程自始至终没有关闭课程注册，导致始终有不以学习为动机的用户涌入课程，因此，系统采样器可以考虑添加用户过滤模块。当然也有可能是由于用户行为特征提取不适用于计算概论课程，这有待于进一步实验进行讨论和验证。

同时作者观察到，对于两门课程，在第八周都有算法性能的下降，可能是由于两门课程都在第八周附近布置了期中考试导致的，学生面对期中考试的行为特征可能与平时有所不同。

[2] https://www.coursera.org/course/pkuic

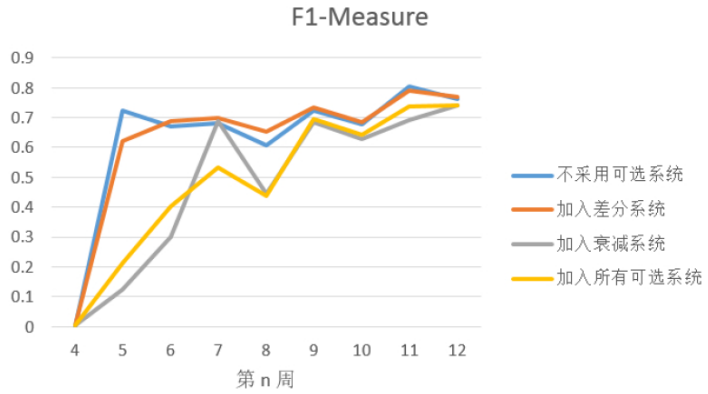

图 7 四种系统结构下的计算概论课程实验结果
Fig.7 Test result of Intro. to Computation course

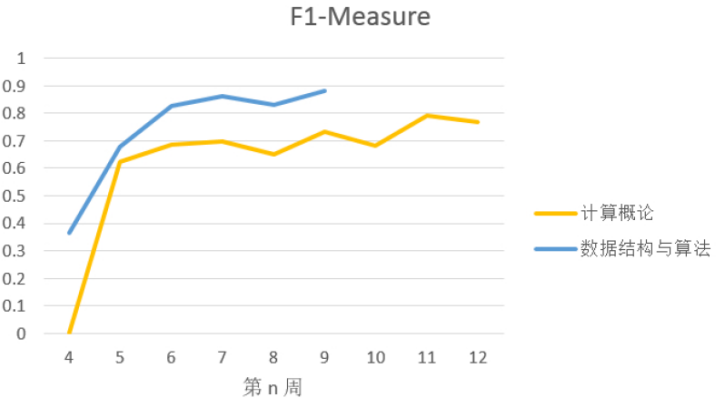

图 8 两门课程实验结果对比
Fig.8 Test result of 2 courses

4 结论

MOOC 的出现及其发展是近几年中重要的互联网事件，MOOC 在未来的教育中也将占据一席之地。合理设计 MOOC 课程，增强学生体验，利用 MOOC 改善课内课程的学习是重要的研究方向。目前，MOOC 的流失是对 MOOC 发展的重要阻碍。MOOC 的大规模特性导致使用计算机的方法进行研究和分析必不可少。

通过对学生行为数据的统计分析，本文发现，学生访问课程、学习课程内容、参与课程测试、参与论坛互动行为都可以有效地保留学生。因此课程应当在此方面对学生进行足够的引导和帮助。

由采样器、差分器、衰减器、分类器构成的无人工监督在线学生流失预警系统可以有效对学生流失进行预警。对于耦合性较低的课程，应当选用衰减器，对于耦合性较高的课程，应当选用差分器，本系统有较好的可扩展性。为了提升系统性能，还应当进一步考虑在采样器环节添加用户过滤算法，结合用户学习动机，优化系统性能。针对系统展现出的结果，课程中教师应当给学生足够的学习指导，例如指出本周内容包括哪些，某些课程内容依赖前几周的知识，具体涉及的章节是什么等等。

本研究相关内容还具有大量值得进一步研究的方面，除了在系统上尝试更多类型的衰减、分类模型以求最好的效果以外，还应当建立短时预测模型，并类似相关研究中提到的教育学研究，提出学生行为模型。

作者注意到，对于两门课程的实验，系统在初期都具有较明显的“冷启动”现象，在最初

的几周运行情况较差。除了可以添加采样器的用户过滤算法以外，应当建立合理的系统结构，针对短时间课程进行预警。由于MOOC上大多数课程的持续时间为6到8周，如果系统可以对这部分课程进行有效预警，则具有较大的实用价值。

为了辅助建立短时预警模型，解释学生行为与流失现象，建立合理的学生行为模型是有必要的。学生行为模型可以对每个学生个体进行训练，弥补现在的模型对于学生整体进行统计学习的不足。这样的模型可以适应不同的学生情况，并且解释学生的行为原因。

可以考虑通过学生行为动机为基础，建立动机与行为的概率图，并依据次进行学习和训练，通过图关系解释学生行为动机和行为数据。

[参考文献] (References)

[1] Lori B, David E P, Jennifer D, Glenda S S, Andrew D H, DT S. Studying learning in the worldwide classroom: Research into edx's first mooc[J]. Research & Practice in Assessment, 2013, 8:13-25

[2] John H M. College student mortality[M]. Washington: US Government Printing Office, 1938

[3] Vincent T. Dropout from higher education: A theoretical synthesis of recent research[J]. Review of educational research. 1975, 45(1):89-125

[4] Tinto V. Leaving college: Rethinking the causes and cures of student attrition[M]. Chicago: University of Chicago Press, 1987.

[5] Charles A B, Jeffrey D Kromrey. A longitudinal study of the retention and academic performance of participants in freshmen orientation course.[J]. Journal of College Student Development, 1994, 35(6): 444-449.

[6] Doug C. MOOCs and the funnel of participation[A]. Dan S, Katrien V, Erik D, Xavier O. Proceedings of the Third International Conference on Learning Analytics and Knowledge[C]. New York: ACM, 2013. 185-189.

[7] Arti R, Dan G, Bert H, Hal D III, Lise G. Modeling Learner Engagement in MOOCs using Probabilistic Soft Logic[A]. Jonathan H, Sumit B, Kalyan V. NIPS Workshop on Data Driven Education[C]. 2013.

[8] Kizilcec R F, Piech C, Schneider E. Deconstructing disengagement: analyzing learner subpopulations in massive open online courses[A]. Dan SKatrien V, Erik D, Xavier O. Proceedings of the third international conference on learning analytics and knowledge[C]. New York: ACM, 2013: 170-179.