ТРАЕКТОРИИ ИСПОЛЬЗОВАНИЯ ОБРАЗОВАТЕЛЬНЫХ ОНЛАЙН РЕСУРСОВ СТУДЕНТАМИ СМЕШАННОГО КУРСА

А.А. Бахитова
Национальный Исследовательский Университет «Высшая Школа Экономики»
Санкт-Петербург

С внедрением технологий в процесс обучения и появлением массовых онлайн-курсов начала активно развиваться сфера образовательной аналитики. Под образовательной аналитикой подразумевается «измерение, сбор и анализ данных об обучающихся и о среде обучении с целью понимания и оптимизации процесса обучения» [1]. Развитию этой области способствует появление все большего объема данных о действиях студентов в виртуальных образовательных окружениях (VLE) в сочетании с совершенствованием компьютерных методов анализа данных.

Значительная часть работ в области образовательной аналитики посвящена анализу данных об активности студентов в онлайн-курсах или в виртуальных образовательных окружениях с целью предсказывания итоговой оценки, как, например, в [2, 3]. Это включает работы по выделению студенческих траекторий в смешанных курсах, то есть курсах, которые включают онлайн-компоненту.

Вопрос о студенческой мотивации в смешанных курсах требует особого внимания из-за минимизации контакта с преподавателями, будь то из-за большого числа студентов или из-за дистанционного формата общения. Поэтому, для поддержки студенческой вовлеченности, и разработки программ, способных учитывать особенности обучения студентов, важно изучить как студенты используют образовательные ресурсы в условиях смешанных курсов.

В докладе представлены образовательные траектории студентов, поступивших на специализацию «Обработка и анализ данных» Санкт-Петербургского кампуса Высшей Школы Экономики в 2016 году. Специализация междисциплинарная; студенты осваивают навыки анализа данных, машинного обучения и основы языка программирования R. Данные представлены за период с 1 сентября 2016 года до (конец 3 модуля) и содержит информацию о студентах набора 2016 года.

Специализация представляет собой смешанный курс, включающий дополнительные задания на платформе для онлайн-обучения Stepik.org. Выполнение заданий на платформе носят необязательный характер, но сами задания тесно связаны с материалами курса и призваны укрепить приобретенные знания через практические задания.

Таблица 1. Описание переменных

Название переменной	Описание переменной	Min	Med	Mean	SD	Max
grade	общее количество правильных ответов по результатам 3 контрольных работ	0.0	36.0	33.8	9.4	47.0
r_logs	количество строчек кода на сервере RStudio	42.0	3847.0	4139.4	2866.7	18924.0
r_not_tuesday	количество активных дней в RStudio не в дни проведения аудиторных занятий (по вторникам) поделенное на общее количество активных дней	21.4	55.3	54.5	11.8	83.3
stepik_percent	процент выполненных заданий на платформе Stepik	7.1	88.7	84.0	15.5	100.0
stepik_share_wrong	количество неправильных ответов на платформе Stepik поделенное на правильные ответы	0.1	0.7	0.8	0.4	2.6
forum_posts	количество постов на форуме	0.0	0.0	0.9	1.6	8.0
forum_answers	количество ответов и комментариев под постами на форуме	0.0	0.0	1.3	2.9	18.0
attendance	посещенные занятия	1.0	12.0	10.7	3.9	16.0

Аудиторные занятия курса проводятся с использованием сервера со встроенной средой разработки RStudio, в которую студент может заходить в любое время, в том числе и для дополнительной подготовки дома. Кроме того, для студентов действует форум вопросов и ответов, активное участие на котором поощряется дополнительными баллами к оценке. Это вызвано результатами предыдущего исследования, в котором была обнаружена связь активности на форуме, то есть коммуникации со сверстниками, обсуждения вопросов, связанных с анализом данных и высокими оценками за курс [4].

Чтобы выделить типичные образовательные траектории студентов применялся модельный метод (model-based) кластеризации данных с помощью библиотеки Mclust для языка программирования R. Для нахождения оптимального числа групп использовался Байесовский информационный критерий (BIC). Выбранные для кластеризации переменные перечислены в Таблице 1. В качестве дополнительной переменной для сравнения характеристик получившихся групп, но не участвовавшей при кластеризации, использовалась переменная *grade* — сумма оценок за контрольные, как показатель усвоенных навыков за курс.

В результате кластеризации алгоритм выделил 5 групп. Для более удобного сравнения групп представлена таблицы со средними по каждой из используемых переменных (Таблица 2).

Таблица 2. Результаты кластеризации

Переменная/Номер кластера(кол-во студентов)	**1(33)**	**2(33)**	**3(28)**	**4(93)**	**5(6)**
r_logs	4499.50	3301.37	6575.91	3164.25	10212.00
r_not_tuesday	57.34	55.05	61.82	50.21	66.86
stepik_percent	83.43	75.80	91.33	84.60	90.06
stepik_share_wrong	0.64	1.02	0.93	0.72	0.87
forum_posts	2.30	0.59	2.08	0.00	3.83
attendance	11.63	8.98	12.31	10.28	14.83
forum_answers	1.93	0.00	5.21	0.00	6.67
grade	36.44	28.58	34.50	34.03	40.50

Для сравнения характеристик активности студентов в каждом из кластеров использовался непараметрический тест Краскела-Уоллиса, показавший значимую разницу между группами для всех переменных. Для апостериорного анализа переменных далее применялся критерий Данна с поправкой Бонферрони.

Как видно из Таблицы 2 во второй кластер попали наименее активные студенты с низкими показателями активности на сервере RStudio, на форуме вопрос и ответов и плохой посещаемостью. Оценки за контрольные у этих студентов значимо ниже оценок пятого кластера ($p<.05$). Пятый кластер является противоположностью первому: показатели активности в R самые высокие, значимо выше, чем во всех остальных кластерах ($p<.05$), кроме третьего. Студенты также чаще используют R не в дни занятий, что может говорить о высокой заинтересованности в предмете и программировании. Также эти студенты чаще задают вопросы, отвечают и комментируют на форуме, посещают больше занятий и выполняют много заданий на платформе Stepik. Однако этот кластер представлен менее всего — 6 студентов.

Три оставшихся кластера имеют схожие оценки (разница не значима), при этом в третьем кластере активнее используют R и выполняют больше заданий на Stepik и в целом много комментируют и отвечают на вопросы на форуме — разница значима для всех переменных ($p<.05$). Активность студентов в третьем кластере сравнима по показателям с пятым кластером наиболее успешных студентов, за исключением количества логов в R и результатов контрольных. Четвертый кластер составляет почти половину студентов на курсе — 93, студенты отличаются невовлеченностью в коммуникацию на форуме. В первом кластере низкая доля ошибок на Stepik — это может быть связано как с более ответственным подходом к решению заданий, так и со списыванием.

Таким образом, в работе были представлены результаты кластеризации активности студентов в смешанном курсе, которые показывают существование различных стратегий использования доступных ресурсов среди студентов. Выделение таких стратегий важно, потому что это поможет создавать и проектировать системы, которые учитывали бы особенности образовательной активности студентов для улучшения качества учебных курсов. Это также важный шаг в создании адаптивных курсов, подстраивающих сложность материала в процессе обучения под студентов.

Статья подготовлена в ходе проведения исследования 17-05-0024 в рамках Программы «Научный фонд Национального исследовательского университета „Высшая школа экономики“ (НИУ ВШЭ)» в 2016 – 2017 гг. и в рамках государственной поддержки ведущих университетов Российской Федерации «5-100».

ЛИТЕРАТУРА

1. Call for Papers of the 1st International Conference on Learning Analytics & Knowledge (LAK 2011). https://tekri.athabascau.ca/analytics.
2. Brooks C., Thompson C., Teasley S. A Time Series Interaction Analysis Method for Building Predictive Models of Learners Using Log Data LAK ’15 / New York, NY, USA: ACM, 2015. Pp. 126–135.
3. Jovanović J. [и др.]. Learning analytics to unveil learning strategies in a flipped classroom // The Internet and Higher Education. 2017. (33). Pp. 74–85.
4. Musabirov I., Okopny P., Pozdniakov S. Enabling Information Access in Virtual Learning Environment: The Case of Data Science Minor ACM, 2016. 119 p.