УДК 378.4:004.81

МОДЕЛЬ ПОВЕДЕНИЯ ОБУЧАЮЩЕГОСЯ, ПОЛУЧЕННАЯ С ИСПОЛЬЗОВАНИЕМ ТЕХНОЛОГИИ МАШИННОГО ОБУЧЕНИЯ

И.Н. Цырельчук
Декан факультета инновационного непрерывного образования, кандидат технических наук, доцент

Е.Н. Шнейдеров
Зам. декана факультета инновационного непрерывного образования

А.С. Терешкова
Магистрант факультета компьютерного проектирования

Белорусский государственный университет информатики и радиоэлектроники, Республика Беларусь, E-mail: shneiderov@bsuir.by

И.Н. Цырельчук
Декан факультета инновационного непрерывного образования. Основным направлением научных исследований является моделирование, оптимизация и надёжность приборов СВЧ, а также организация учебного и научно-исследовательского процессов в БГУИР.

Е.Н. Шнейдеров
Заместитель декана инновационного непрерывного образования. Занимается исследованием методов обеспечения надёжности систем и устройств, а также технической организацией дистанционных образовательных технологий.

А.С. Терешкова
Обучается в магистратуре по направлению автоматизации использования математических методов с использованием машинного обучения для обработки данных.

.

Аннотация. Построение модели поведения обучающихся предполагает сбор и обработку большого количества информации. Популярность массовых открытых онлайн-курсов (МООК) позволила накапливать информацию о поведении студентов на протяжении освоения дисциплины, а использование машинного обучения – обрабатывать большие наборы данных и получать более точные прогнозы. В данной статье описан один из возможных методов получения модели поведения обучающегося и прогнозирования оценки на его основе с использованием технологий машинного обучения.

Ключевые слова: МООК, онлайн-обучение, модель поведения обучающегося, машинное обучение.

На сегодняшний день значительное количество учебных занятий учреждений образования являются аудиторными. Однако известно, что в современной образовательной Internet-среде за последние десятилетие большое распространение получили массовые открытые онлайн-курсы – МООК, являющиеся автономными образовательными модулями, которые включают в себя учебный мультимедиа-материал и задания для автоматического контроля знаний обучающегося. С точки зрения образовательного процесса МООК является развитием бумажного учебника с заданиями для решения. Однако ключевым отличием МООК и учебника является интерактивность и сбор статистики об обучающихся. Поэтому, с точки зрения науки, МООК представляет обширную область для количественного и качественного анализа поведенческих процессов обучающихся и разработки новых моделей обучения.

Разработка модели обучения является сложной исследовательской задачей, предполагающей формализацию предметной области, проведение эксперимента и выполнение численного анализа промежуточных результатов. Задача осложняется тем, что как правило специалисты в области численного анализа далеки от педагогики, а квалифицированные педагоги не знакомы с современными методами обработки данных и построения моделей. Наиболее трудоёмким этапом в этой задаче является проведение эксперимента. Этот этап решается с помощью анализа статистики МООК, позволяя оперировать входными данными в виде информации об обучающихся и получать результаты их обучения: время и интенсивность работы с учебными материалами, результат выполнения заданий, активность на форумах и др. При таком взгляде на образовательный процесс интерес представляет модель, которую будем называть моделью поведения обучающегося, позволяющая с определённой достоверностью спрогнозировать результаты обучения конкретного обучающегося по его начальным действиям в МООК в процессе первых недель обучения. Получение такой модели позволит обозначить те ключевые факторы в процессе формирования МООК, которые будут определять действия обучающихся и учёт которых позволит повысить процент обучающихся, успешно завершивших курс.

Существующие эксперименты в области образования на постсоветстком пространстве, как правило, проводились на небольших группах обучающихся, часто из одного и того же учебного заведения, имеющих примерно одинаковый опыт обучения. Использованием МООК позволяет не только увидеть полную картину, включающую студентов со всего мира, но и даёт возможность индивидуально рассмотреть каждого.

В классическом подходе, например, с 50 обучающимися, достаточно применить методы теории вероятности (например, дисперсионный анализ). В случае малого объёма данных, используемые методы позволяют выявить только те первичные факторы, влияние которых наиболее значимо. В случае большого объёма выборки обучающихся имеет место обратный эффект, когда значимое влияние фактора не относится к исследуемой области из-за неоднородности входных данных. В данной статье приводится пример построения модели поведения обучающегося с использованием технологии машинного обучения по алгоритму, изображённому на рисунке 1.

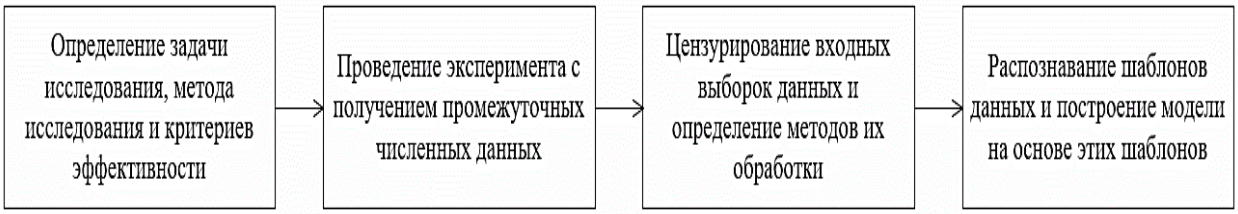

Рисунок 1. Последовательность этапов исследований для получения модели поведения обучающегося

В качестве исходных данных используется информация, полученная на основе изучения данных известных МООК «Введение в C++» и «Введение в Java», опубликованных EPFL осенью 2013 года. Данные содержат информацию о 13787 обучающихся на курсе «Введение в C++» и 17716 обучающихся на курсе «Введение в Java».

Модель поведения обучающегося будет строиться на основе данных о поведении студентов в системе электронного обучения с течением времени. Гипотеза заключается в том, что модель не зависит от курса, так как оба курса имеют схожую структуру с точки зрения продолжительности, содержания, количества и качества заданий. Предполагается также, что поведение студентов будет схожим.

В ходе проведения эксперимента осуществлялся сбор данных об активности студентов в системе электронного обучения. Также можно собирать информацию о движении и нажатиях мыши, последовательности нажатий клавиш на клавиатуре и взаимодействиях с видео. На основе этих данных строятся временны́е ряды прохождения студентами курсов для последующего их анализа. Для каждого студента есть временны́е метки событий следующих типов: регистрация, просмотр форума, подписка на форум, просмотр темы, подписка на тему, публикация в теме, просмотр лекции, повторный просмотр лекции, скачивание лекции, просмотр видео, повторный просмотр видео, сдача теста, повторная сдача теста, сдача итогового задания, повторная сдача итогового задания.

После сбора необходимых данных происходит их цензурирование с целью выявления важных для исследования характеристик. В первую очередь извлекаются простые характеристики, например, количество просмотренных видео, количество попыток сдачи теста и т. д., а затем – более сложные, например, взаимодействия с видео (паузы, перемотки и т. д.) [2].

Визуализация полученных данных позволяет получить представление о том, какие переменные являются более информативными для последующего анализа. На рисунке 2 представлены временны́е ряды двух студентов. Первый студент (слева) получил 87 %, а второй – 66 %. На рисунке видно, что активность первого студента гораздо выше активности второго, хотя они оба выполнили все задания курса.

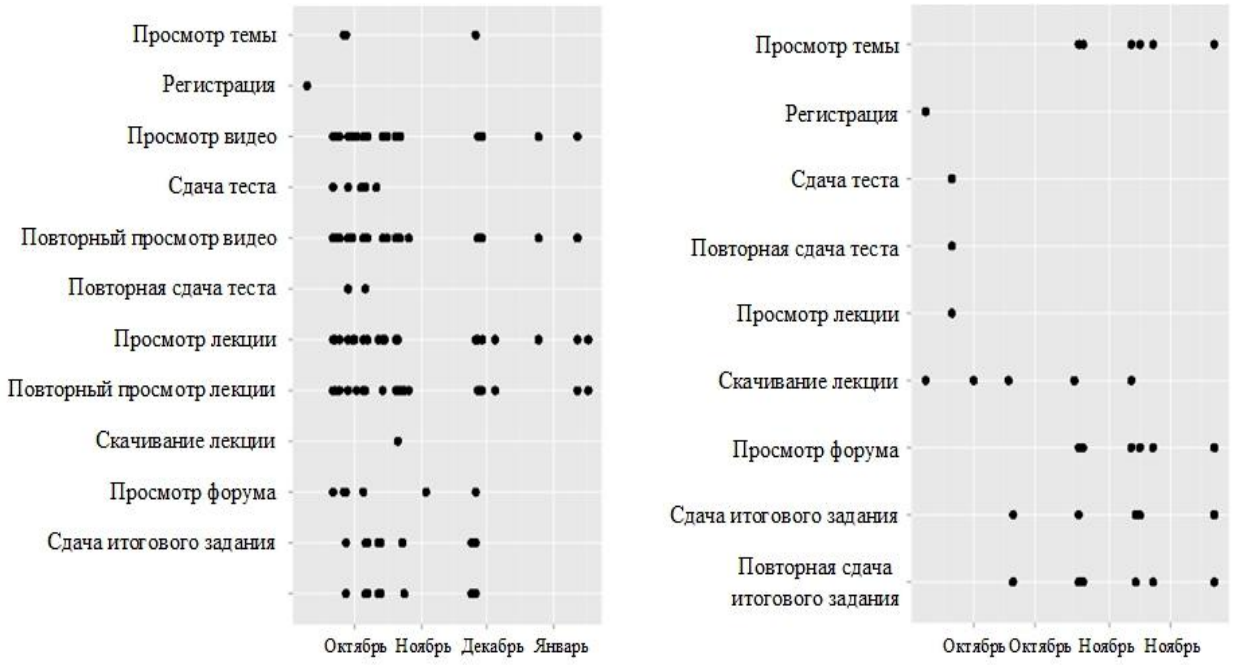

Рисунок 2. Визуализация временны́х рядов двух случайных студентов

По данным активности в системе можно выявить такие характеристики студента, как, например, прокрастинация (процент сданных незадолго до крайнего срока заданий) или регулярность (дисперсия времени, прошедшего между двумя входами в систему).

В результате цензурирования полученных данных получаем таблицу, содержащую в себе информацию о каждом студенте по выбранным переменным, которая впоследствии используется для обучения модели. Так как мы имеем выборку входных данных (метки поведения студента в системе) и полученный в ходе эксперимента результат (оценка студента), в данном случае используется контролируемое машинное обучение для получения модели поведения.

При получении данной модели были использованы следующие алгоритмы машинного обучения: метод опорных векторов (англ. Support Vector Machine, SVM), случайные леса (англ. – Random Forests, RF) и бустинг (англ. Generalised Boosted Regression, GBM). Каждый обучающийся представляется в виде вектора его характеристик, полученных в результате цензурирования собранных данных. В качестве показателя точности модели используется среднеквадратическое отклонение, выражающая среднее расстояние между экспериментальным и прогнозным значениями.

Для реализации алгоритмов машинного обучения использовался язык программирования R и фреймворк CARET [3]. На рисунке 3 представлен листинг кода для построения модели с помощью метода опорных векторов.

```
library(caret)
# Build the model
control <- trainControl(method="repeatedcv", number=10, repeats
    =3)
model.svm <- train(Grade ~ ., data=students.tr, method="
    svmRadial", trControl=control, tuneLength=5)
# Predict the grades
grades = predict(model, students.ts)
```

Рисунок 3. Листинг программного кода построения модели SVM с использованием пакета CARET языка программирования R

Аналогично строятся модели на основе случайных лесов (RF) и бустинга (GBM), что позволит быстро создавать прототипы и сравнивать точность различных моделей.

Для оценки точности каждой модели необходимо вычислить среднеквадратическое отклонение результатов прогнозирования (рисунок 4). Для удобства также выполняется визуализация полученных значений (рисунок 5).

```
results <- resamples(list(svm = model.svm, rf = model.rf, gbm =
    model.gbm))
# boxplots of results
bwplot(results)
```

Рисунок 4. Сравнение точности различных моделей

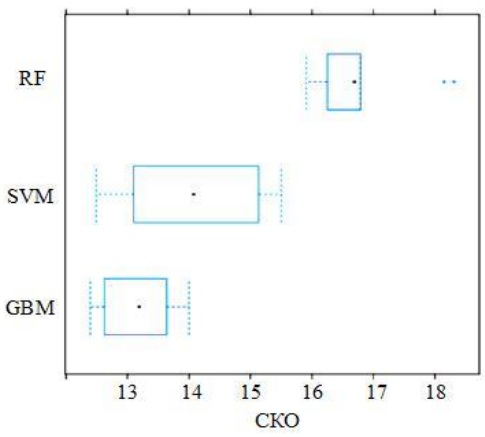

Рисунок 5. Ошибки каждой модели с точки зрения СКО

На рисунке 5 видно, что наибольшую точность обеспечивает модель GBM, так как она обеспечивает наименьшую ошибку прогнозирования.

Разработанный подход можно рассматривать как один из способов получения модели поведения обучающихся в частности и использования технологий машинного обучения в процессе образования в целом. Данный подход не претендует на оптимальность и не был апробирован в процессе образования, однако он может позволить получить представление о количестве студентов, которые успешно окончат курс, ещё задолго до его завершения.

Литература

[1] Kidzinski, L. A tutorial on machine learning in educational science / L. Kidzinski, M. Giannakos, D. G. Sampson, P. Dillenbourg, Pierre // Proceedings of the 2nd International Conference on Smart Learning Environments / Sinaia, Romania, 2015.

[2] N. Li. How do in-video interactions reflect perceived video difficulty? / Li, N., Kidzinski, L., Jermann, P., Dillenbourg, P. // Proceedings of the European MOOCs Stakeholder Summit 2015 / Lausanne, Switzerland, 2015.

[3] Kuhn, M. The CARET Package. [Электронный ресурс] – Режим доступа : http://topepo.github.io/caret/index.html.

THE MODEL OF STUDENT BEHAVIOR, OBTAINED USING MACHINE LEARNING TECHNOLOGY

I.N. TSYRELCHUK, PhD
Dean of the Faculty of Innovative and Lifelong Learning

E.N. SHNEIDEROV
Vice-dean of the Faculty of Innovative and Lifelong Learning

A.S. TERESHKOVA
Master student of the Faculty of Computer-Aided Design

Belarusian State University of Informatics and Radioelectronics, Belarus
E-mail: shneiderov@bsuir.by

Abstract: Building a model of student behavior involves the collection and processing of a large amount of information. The popularity of massive open online courses (MOOCs) allowed to accumulate information about students behavior throughout the course of the discipline, and the use of machine learning - to process large data sets and get more accurate predictions. This article describes one of the possible methods for obtaining student behavior model and predicting assessment based on it using machine learning technologies.

Keywords: MOOC, online education, student behavior model, machine learning.