DOI: 10.18721/JCSTCS.13106
УДК 004.85, 004.62, 378.147

THE ASSESSMENT OF THE RESULTS OF A MASSIVE OPEN ONLINE COURSE USING DATA MINING METHODS

S.A. Nesterov, E.M. Smolina

Peter the Great St. Petersburg Polytechnic University,
St. Petersburg, Russian Federation

The paper presents the results of a grade reports analysis for five sessions of a massive open online course "Data Management" at openedu.ru. For our research, we used clustering and classification in the R programming environment. Clustering showed the presence of four groups of course participants with nearly similar course results. These clusters were similar for all five sessions of the course we analyzed. We also showed it is possible to predict whether a participant completes the course or drops out, based on the test results during the first half of the course. The course lecturers can use the results to plan measures for keeping the students in the course. Also, such a type of analysis helps to understand the reasons why the students drop out of the course. The lecturers can take them into account to modify the course structure and learning content. This new knowledge about the course participants can be used during the next course sessions. We expect that for other courses with a similar structure, the clustering results will be also similar. The approach to predict whether a student drops out or completes the course used in the paper is applicable for other courses as well.

Keywords: MOOC, learning management systems, Data Mining, clustering, classification.

Citation: Nesterov S.A., Smolina E.M. The assessment of the results of a massive open online course using Data Mining methods. Computing, Telecommunications and Control, 2020, Vol. 13, No. 1, Pp. 65-78. DOI: 10.18721/JCSTCS.13106

ОЦЕНКА РЕЗУЛЬТАТОВ ПРОВЕДЕНИЯ МАССОВОГО ОТКРЫТОГО ОНЛАЙН КУРСА С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ

С.А. Нестеров, Е.М. Смолина

Санкт-Петербургский политехнический университет Петра Великого,
Санкт-Петербург, Российская Федерация

Представлены результаты исследования отчетов об оценках пяти сессий дистанционного массового онлайн курса «Управление данными» на портале Открытого образования openedu.ru. В ходе исследования решались задачи кластеризации и классификации. Исследование проводилось с использованием языка программирования R. Кластеризация показала наличие четырех групп слушателей курса, сходных по результатам прохождения курса. Характеристики этих групп близки для всех рассмотренных сессий курса. Показано, что на основании результатов прохождения тестов в первой половине курса можно с высокой точностью предсказать, бросит ли слушатель изучение курса или будет учиться до его окончания. Полученные результаты можно использовать при планировании мероприятий с целью удержания слушателей на курсе. Подобный анализ помогает понять причины, по которым студенты бросают изучение курса, и учесть это при корректировке его структуры.

Ключевые слова: MOOK, системы дистанционного обучения, интеллектуальный анализ данных, кластеризация, классификация.

Ссылка при цитировании: Нестеров С.А., Смолина Е.М. Оценка результатов проведения массового открытого онлайн курса с использованием методов интеллектуального анализа данных // Информатика, телекоммуникации и управление. 2020. Т. 13. № 1. С. 65-78. DOI: 10.18721/JCSTCS.13106

Introduction

As e-learning is currently developing at an accelerating pace, massive open online courses (MOOCs) providing simultaneous education to thousands of students are becoming more and more popular. At the same time, this type of e-learning has a common drawback of only a small percent of students completing the courses [1, 2].

Information systems of distance online education accumulate large amounts of data on course participants and the results of their studies. We can analyze these data and give recommendations to increase the quality of the courses. For example, some e-learning platforms such as Moodle provide inbuilt tools for statistics analyses of completed tests [3–5]. In other instances, the data in the form of students' reports or event log files get downloaded from the education monitoring system and analyzed by additional external means, including those using Data Mining algorithms.

Nowadays, an emerging direction of data analysis called Educational Data Mining is developing new methods of data analysis in education [6–8]. This type of analysis helps to identify characteristic groups of students [8, 9], predict students' course results [10–12], as well whether or not the participants will finish the course [11, 13], determine the most difficult tasks [4] and general behavioral patterns the participants display [1, 14].

This paper analyses grade reports of the participants of "Data management" course on an open education platform openedu.ru [15]. We engaged in clustering (to identify characteristic groups of course participants) and binary classification (to predict whether or not the participants will finish the course) tasks.

The purpose of this research is to enhance MOOC efficiency using Data Mining results accessible for lecturers from standard reports. The research tasks include:

- obtaining new knowledge on the course participants based on the data found in the course reports;
- analyzing possibilities of applying the obtained results to the future sessions of the course, as well as other courses at openedu.ru and similar platforms.

Preliminary data analysis

As we mentioned before, the research presents an analysis of grade reports of the participants of "Data management" MOOC at openedu.ru. The course takes up one semester and starts twice a year, in fall and spring. We analyzed five sessions of the course: the fall of 2016, the spring and fall of 2017 and 2018.

The course lasts for 16 weeks, each week presenting a new topic to study. The course content includes video lectures, lecture notes, workshops, weekly tests (Homework in reports). The students have a Midterm Exam after the 8th week and a Final Exam after the 16th week. The final course grade consists of the homework results (an average for all weeks), the midterm and final exams combined. During the first session of the course these grades are summed up with weighting factors: 0.3 for the homework results, 0.35 for the midterm exam and 0.35 for the final exam. The subsequent sessions were subject to some changes: the final exam required only participant's identity authentication and its contribution to the final grade increased significantly. The new weighting factors now were 0.2, 0.2 and 0.6 respectively.

We chose R programming language for analysis since it presents a variety of tools for statistical data processing, visualization and machine learning [16–18]. We downloaded delimited text files of reports from openedu.ru and imported them into R, where they were displayed as data frames. The absent grade data were replaced by zero. Thus, we assumed the difference between the case a student failed to complete the task and a case a student failed the test scoring zero was insignificant. Table 1 presents a fragment of a grade report after the above-mentioned replacement.

Table 1

Fragment of the report under consideration

id	Grade	Homework 1	...	Homework 16	Midterm Exam	Final Exam
217782	0.0	0.0	...	0.0	0.0	0.0
181077	0.05	0.8	...	0.0	0.0	0.0
180553	0.94	1.0	...	1.0	0.933	0.9

For each session of the course, the authors calculated a percentage of the participants enrolling in the course, but failing to complete any tasks. We used the final course grade for the purpose (defined as Grade at openedu.ru). If a participant's Grade amounts to zero, the participant never commenced performing the tasks. It is noteworthy that the Grade is presented with values between 0 and 1 rounded to two decimal places in the report, so a roundoff error may occur: the students completing only a small part of a task of only one week may fall into the group of the students who never commenced to perform the tasks. Table 2 shows the results.

Table 2 demonstrates that the largest number of participants enrolled in the first session of the course. This was probably due to the interest in the new course at the website. In the fall of 2018 the enrollment deadlines were prolonged significantly which also had a positive impact on the number of the applicants. For five sessions of the course, only 31, 32, 23, 19 and 23% of the participants (respectively) commenced to perform the tasks.

Table 2

Number of students failing to perform any task

The course starting period	Enrolled in the course	Commenced performing the tasks	Failed to perform any task, %
class 2016	2547	798	69
class 2017 (spring)	1572	499	68
class 2017 (fall)	1823	427	77
class 2018 (spring)	1504	279	81
class 2018 (fall)	2346	529	77

A bar chart in Fig. 1 demonstrates a number of students of the first course session who commenced performing the tasks. The x-axis shows the number of the task, while the y-axis presents the number of students. 798 students performed the tasks of the first week, then a steep drop occurs with only 435 proceeding to perform the second week tasks. In the course of the subsequent weeks, the number of active participants continues to fall gradually, however we observe a slight increase in numbers on the midterm exam week.

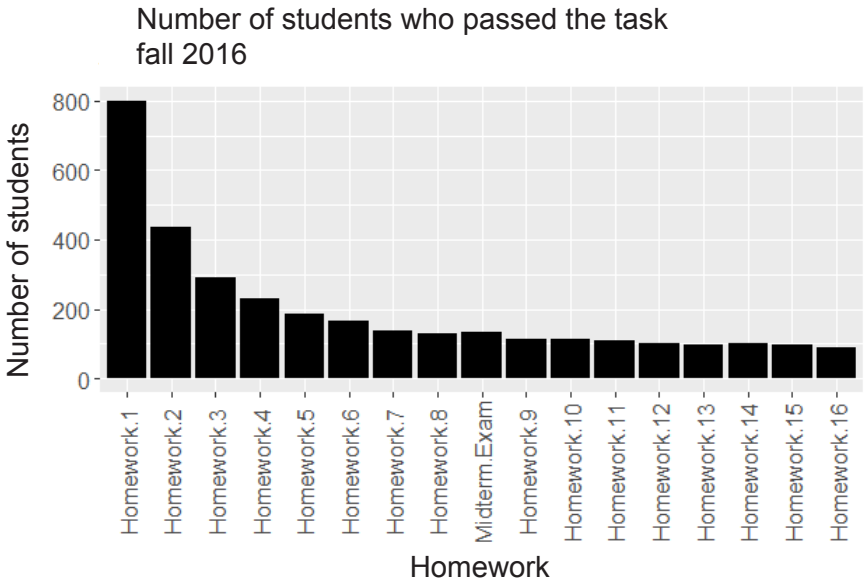

Fig. 1. Number of students completing the tasks (fall 2016)

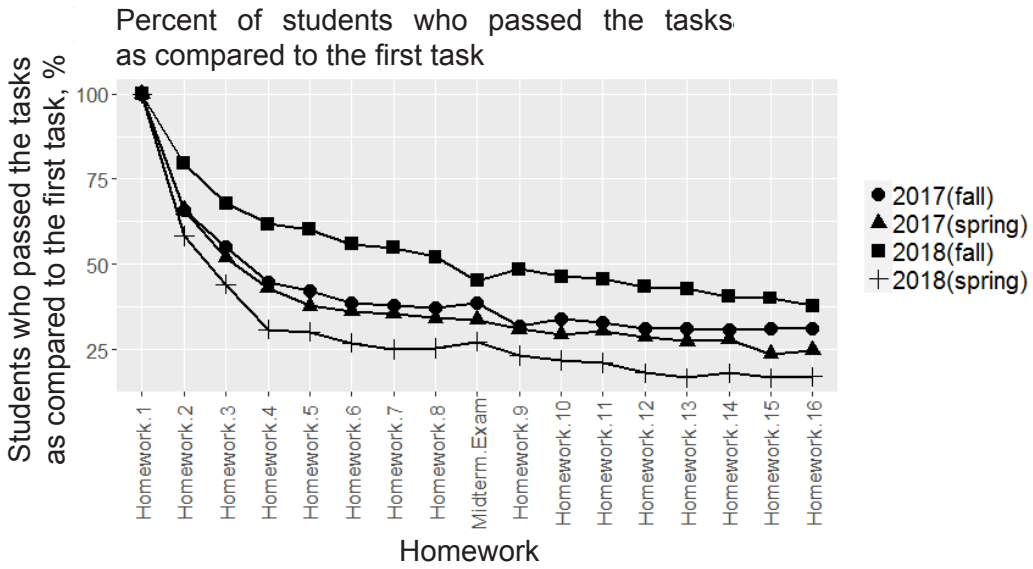

Fig. 2. Percentage of the students performing the task

The spring start of 2017 had 499 participants commencing to perform the first task. Similarly to the first start, the second week exhibited a dramatic fall down to 329 participants. The following weeks saw a smooth drop in the number of participants. The subsequent sessions showed a similar dependency. We should note that the midterm exam marks the moment after which the number of active participants remains virtually the same.

Fig. 2 presents graphs for all five course sessions under consideration with the number of students taking tests weekly displayed as percentage of the students performing the first week tasks. Figure shows that all the course sessions exhibit a sharp fall in the number of active participants after the first week. This can be due to a number of reasons, for example:

- the content turned out to be uninspiring or too difficult;
- the participants realized the course required a lot of time;
- the participants had no intention to study and wanted to "have a look" only;
- the course is too extensive and long.

The next task of this research was to find groups of participants similar in their level of activity in the course.

Clustering

Based on the reports of the progress in the course we can divide the participants into groups according to their results. This is a clustering task we can describe in the following manner. Let I be a multitude of the course participants:

$$I = \{ i_1 , i_2 , \ldots, i_n \},$$

where each of the participants possesses a set of attributes:

$$i_j = \{ x_1 , x_2 , \ldots, x_m \},$$

x_k is an independent variable which can assume values from a certain multitude (usually numerical values).

We need to form a multitude of clusters

$$C = \{ c_1 , c_2 , \ldots, c_g \},$$

where each cluster includes similar objects from I multitude of participants under consideration:

$$c_h = \{ i_j , i_p \mid i_j , i_p \in I , d (i_j , i_p) < \sigma \}.$$

Here $d(i_j , i_p)$ is a measure of closeness between objects (distance), σ is a boundary value of distance to include objects in one cluster [19].

The task was to divide the multitude of the course participants into groups with similar attributes (clusters) and compare the clusters obtained for different course sessions.

We based the choice of clusters on the research of the dependency of the change of the total mean squared deviation (squared distance between each element and the cluster center) on the number of clusters [16, 20]. This approach uses a number of clusters corresponding to the elbow of the curve (the so-called "elbow method"). According to this criterion each course session had a value of 4 clusters. We used the k-means clustering algorithm. Our clustering algorithm did not take the results of the participants who failed to commence performing any task at all into account.

To define each cluster, the authors constructed graphs describing average cluster grades for the weekly tests and exams. Fig. 3 demonstrates clustering for the course session starting in the fall of 2018. The graphs for the other course sessions are visually very similar [9].

Thus, in the course of the study we defined four major groups of active course participants present in every session under consideration:

1. students with a stable performance (Fig. 3, cluster 3);
2. students with high performance in the first half of the course, low performance in the second half who still completed the course (cluster 4);
3. students who attended the first two weeks with occasional attendance in the following weeks (cluster 2);
4. students with high performance in the first weeks and low further *attendance* who dropped out of the course after the midterm (cluster 1).

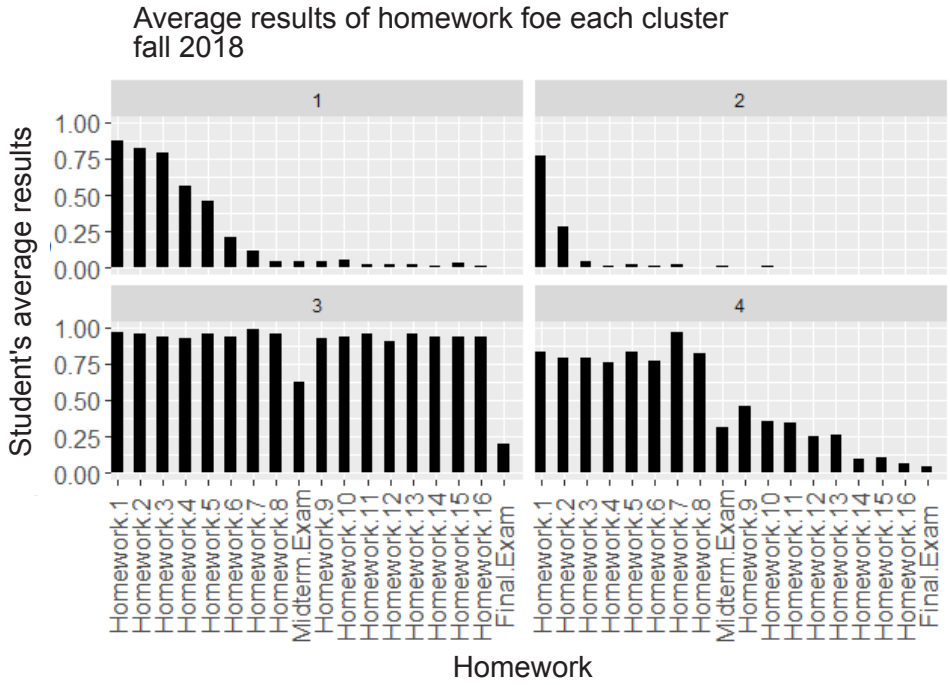

Fig. 3. Average cluster grades for the course session of the fall of 2018

Classification

The most interesting and essential analysis task is obtaining a predicative model which uses the results of the previous courses to benefit the upcoming sessions. In this research, we used the results of students participating in 4 sessions of the course (2016, the spring and fall of 2017, spring of 2018) as a training sample, while applying the classification model obtained to the fifth course session (the fall of 2018) to predict the completion rate for the course. The purpose of this prediction was to determine whether a participant drops out or continues to study in the course until the end. Thus, we reduced the task to a binary classification.

We could not use the final exam results as the target attribute, because the exam with remote identity authentication and proctoring requires payment at openedu.ru, so only a small percentage of participants engages in it. Therefore, to form the target attribute (hereinafter referred to as *targetAttr*) we used grade values for the homework of the last (15^{th} and 16^{th}) weeks. If a student has a grade exceeding zero for a test of the 15^{th} or 16^{th} week, *targetAttr* acquires a value of 1 (completed the course), otherwise 0 (dropped out). Then we had to determine the weeks most suitable for predicting. For a pooled sample of the participants of the first four course sessions, we calculated a number of n week participants as well as their percentage in relation to the number of students who managed to complete the course (Table 3).

Table 3 shows results up to the midterm. As we can see, the majority of the participants who reached the midterm continued their studies. Thus, we decided it makes sense to predict "whether a participant finishes the course or drops out" before the midterm. Then we trained the classification models using the data of the first 4–7 and 8 weeks of the course.

Using an R tool, *sample()*, we randomly formed a training dataset (with 75 % of the initial data) and a test dataset (including the rest 25 %). It is important to note, that the training dataset included approximately 20 % of the *targetAttr* values equalled 1, while the rest 80 % equalled 0, thus making the dataset unbalanced. Table 4 displays the results of solving the classification task using three methods: k-nearest neighbors, Naive Bayes and decision trees. We used R packages class, naivebayes and rpart, respectively.

Table 3

Number of *n* week participants

Homework	Number of *n* week participants	Percentage of the students who completed the course, %
1	1869	22
2	1199	35
3	901	46
4	719	58
5	638	66
6	585	72
7	544	77
8	528	80
Midterm Exam	543	77
Completed the course	423	

Table 4

Classification results (unbalanced sample)

Algorithm	Week	Characteristics			
		accuracy	*precision*	*recall*	*f*1
k-nearest neighbors	4	0.79	0.75	0.73	0.74
	5	0.8	0.73	0.77	0.75
	6	0.81	0.75	0.77	0.76
	7	0.81	0.75	0.83	0.79
	8	0.82	0.76	0.81	0.79
Naive Bayes	4	0.79	0.68	0.9	0.78
	5	0.82	0.71	0.94	0.81
	6	0.82	0.71	0.94	0.81
	7	0.84	0.73	0.95	0.82
	8	0.85	0.75	0.95	0.84
Decision trees	4	0.79	0.74	0.73	0.73
	5	0.79	0.7	0.82	0.75
	6	0.83	0.78	0.79	0.78
	7	0.81	0.74	0.86	0.8
	8	0.84	0.77	0.87	0.82

The quality characteristics of the classification model presented in Table 4 are defined in the following way [20]. Let's assume that after the training dataset the binary classifier showed:

- TP – a number of true positive predictions;
- TN – a number of true negative predictions;
- FP – a number of false positive predictions;
- FN – a number of false negative predictions.

Then we can calculate the characteristics as:

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}};$$

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}};$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}};$$

$$f1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}.$$

To improve the classification results we balanced the training dataset using random undersampling [16]. The obtained sample had 40 % of the *targetAttr* values equalling 1. Table 5 contains the classification results.

Table 5

Classification results (balanced sample)

Algorithm	Week	Characteristics			
		accuracy	*precision*	*recall*	*f*1
k-nearest neighbors	4	0.79	0.67	0.9	0.77
	5	0.77	0.66	0.88	0.75
	6	0.81	0.7	0.89	0.78
	7	0.82	0.7	0.93	0.8
	8	0.86	0.76	0.94	0.84
Naive Bayes	4	0.78	0.67	0.9	0.77
	5	0.81	0.7	0.94	0.8
	6	0.82	0.71	0.94	0.81
	7	0.84	0.73	0.96	0.83
	8	0.86	0.75	0.96	0.84
Decision trees	4	0.78	0.66	0.89	0.76
	5	0.79	0.68	0.88	0.76
	6	0.81	0.7	0.91	0.79
	7	0.83	0.71	0.95	0.82
	8	0.85	0.75	0.93	0.83

Tables 4 and 5 show that the balanced sample does not significantly improve the results. Moreover, *precision* characteristic demonstrates a rise of false responses, meaning the share of true positive objects is reduced. Thus, we cannot deem the balanced training feasible for the current dataset.

Combining the classification and clustering results poses a more interesting task. Fig. 3 shows results of dividing the participants into characteristic groups (clusters) for the fall of 2018. The groups displayed in the Figure and named in the list below it include the following numbers of students: 201 in group 1; 84 in group 2; 98 in group 3; 146 in group 4.

For the prediction purposes we chose models based on the prediction results for 8 weeks of learning as the values of the generalized metric $f1$ (see Table 4) are the highest specifically by the midterm. Table 6 shows confusion matrices for each group.

Table 6

Confusion matrices for each group

Method	Group															
	1				2				3				4			
k-nearest neighbors		fact				fact				fact				fact		
	pred		0	1	pred		0	1	pred		0	1	pred		0	1
		0	0	32		0	29	7		0	98	0		0	141	2
		1	11	158		1	40	8		1	0	0		1	2	1
Naive Bayes		fact				fact				fact				fact		
	pred		0	1	pred		0	1	pred		0	1	pred		0	1
		0	0	1		0	14	7		0	98	0		0	143	3
		1	11	189		1	55	8		1	0	0		1	0	0
Decision trees		fact				fact				fact				fact		
	pred		0	1	pred		0	1	pred		0	1	pred		0	1
		0	0	16		0	29	9		0	98	0		0	139	2
		1	11	174		1	40	6		1	0	0		1	4	1

The classifier demonstrates acceptable precision for the first group. That is probably connected with the fact the group includes the students with the highest grades on a permanent basis: 190 out of 201 students completed the course. Interestingly enough, responses of each algorithm to the same dataset were different. The *k*-nearest neighbors determined a smaller share of positive objects than it should have. The Naive Bayes classifier, on the contrary, defined more values as positive falsely. The decision trees algorithm demonstrated the best prediction results for the first group: almost all the responses were correct.

We can see, that any classifier formed using any of the chosen algorithms is underperforming in terms of predictions for the second group showing a high rate of false positive results. 84 students of the group displayed high academic performance in the first half of the session, but then it dropped: only 15 participants completed the course. At the same time, the models predict that more than a half of the students would complete it. By the end of the 8th week most of the students had sufficiently high grades. We can attribute the classification errors to this reason.

The results for the third group turned out to be the most accurate. This is probably due to the fact the cluster includes the students who dropped out exclusively.

While predicting the reply to the question of the fourth group students dropping out the *k*-nearest neighbors algorithm committed no errors. The Naive Bayes algorithm ignored the values equaling 1, and the decision trees predicted more positive results than there were in reality. This cluster included 146 students, 3 of them completed the course.

Analysis of the results

While analyzing the obtained results it is important to note that the grade reports are the main source of information on students' progress for the course lecturer at openedu.ru. The platform currently fails to provide any inbuilt analytical means for lecturers. For this reason, analyses of grade

reports similar to the one described in this article may be vital for any course at the website, and possibly for other courses using the same Open edX platform. This feature distinguishes this system from LMS Moodle wide-spread in universities, as the latter one offers e-learning analysis tools [3–5].

We analyze the obtained results starting with the participants clustering. Firstly, solving this problem gave us a better understanding of the peculiarities of the course participants' behavior. Dividing the participants into 4 described groups was not obvious, this result can be useful for the subsequent course sessions. Secondly, the repetitive nature of the clustering results shows that small changes to the course procedures (similar to the prolongation of the enrollment deadlines in the fall of 2018 we previously described) had little impact. This may testify that all the sessions of the course are equal. If we choose to analyze the data of other courses, the number and characteristics of the clusters may be different, but the most important factor is the repetitive nature for different sessions, provided the course content was subject to no major changes.

In other cases, we might face a reverse problem: if the course content is significantly revised, we need to understand whether the participants exhibit different behavior. We can expect the number and characteristics of clusters into which the participants fall to change as well.

Let us compare the clusters obtained with the results of other researchers. There is a following classification of typical groups of MOOC participants [14, 21]:

- "Ghosts" – participants who enrolled in a course, but never accessed any course content, i.e. never actually participated in the course.
- "Observers" – participants who enrolled in a course, accessed course content (video, lecture notes), but ignore any tests or tasks.
- "Non-completers" – participants who use MOOC content as auxiliary in their studies or work. They have no intention to complete the course, so the majority of such students drops out.
- "Passive participants" – these participants access course content, watch the video lectures, take tests, communicate with other students and lecturers online, but ignore difficult tests or bigger projects.
- "Active Participants" – participants with a high motivation level working on any type of course content, participating in projects, actively communicating with other students and lecturers.

To identify these groups of participants, the analysis requires not only the grades, but also the information on their access to course content. The lecturers of openedu.ru have no direct access to this information. Nevertheless, we can assume that the first two groups consist of 70–80 % of the participants from Table 2 who enrolled in the course, but never performed any of the tasks. Previously, we described two clusters of the participants who were active in the first weeks, but dropped out at different stages of the course, thus corresponding to the "Non-completers" group. We can assume that the cluster of the participants who had stable performance throughout the course corresponds to the "Active Participant" group. The "Passive participants" apparently consists of the remaining cluster.

As for the prediction on the results of the participant's studies, in most cases the researchers use a different approach and initial data. For example, the predication mechanism uses information on the results of this particular student in other courses [10]. Or along with the grades the researchers engaged additional information, such as household income, the participant's sex, etc. [12], which is usually impossible to access for MOOCs. The papers also present statistics for the participants accessing the content (links referrals, time and mode of watching video, video paused, etc.) [13].

However, we should note that the approach to predicting whether a participant finishes the course or drops out based only on grades for the accomplished tasks presented in this paper showed that it can still be of interest. Lecturers can carry out this kind of analysis in a timely manner. At the same time, it is only applicable to the courses with strict deadlines for test paper admittance during the course. If the deadlines are absent or the main deadline is the date of the course ending, this approach is irrelevant.

Conclusion

This paper analyses grade reports of the participants of "Data management" course at an open education website openedu.ru. Grade reports are the kind of data lecturers use when running courses at the website. This MOOC system does not offer any tools to analyze students' progress in the courses yet.

As a result of clustering we identified 4 characteristic groups of participants. Moreover, the same clustering pattern persisted in all 5 sessions of the course analyzed.

We also demonstrated the grade reports obtained in the first half of the course are sufficient enough to predict whether students drop out or complete the course with high accuracy. To improve the prediction accuracy, we attempted to train the models using balanced samples. However, this approach did not result in any significant improvement of classification accuracy.

The classifiers show different accuracy for various groups of students. This allows to assume that using different algorithms for various groups of participants can benefit the prediction accuracy.

The lecturers can use the results to keep the participants in the course. For example, certain students may require some measures taken beforehand to provide incentives for them to remain in the course and complete the studies. These measures can include new content offers or task notifications sent to them via e-mail.

This kind of analysis can also help to indentify the reasons the students drop out of the course which can be taken into account to correct its structure. For instance, the lecturer can change difficult tasks and recommend additional content to certain groups of students. This can increase the number of participants who complete the course.

Thus, as a result of the research we obtained new knowledge on the course participants useful for the lecturers. They can take it into account while planning next MOOC sessions. The lecturers of openedu.ru and similar systems can apply this approach to conduct the analysis of grade reports for their courses. We can assume that, provided the courses have structures close to the one described in this paper, the results of identifying characteristic groups of participants may be similar. We can also expect the approach to predict whether a student drops out or completes the course used in the paper to be applicable for other courses as well.

REFERENCES

1. **Gelman B., Revelle M., Domeniconi C., Johri A., Veeramachaneni K.** Acting the same differently: A cross-course comparison of user behavior in MOOCs. *Proceedings of the 9th International Conference on Educational Data Mining*, EDM 2016, Pp. 376–381. Available: http://www.educationaldatamining.org/EDM2016/proceedings/paper_136.pdf (Accessed: 01.11.2019).

2. **Nesterov S.A., Smolina E.M.** Analysis of the results of distance learning in the format of the massive open online course. *Proceedings of XXII International Conference on Systems Analysis in Engineering and Control*, St. Petersburg: Politeh-Press, 2018, Vol. 2, Pp. 379–383. (rus)

3. **Protasova I.V., Tolstobrov A.P., Korzhik I.A.** Metodika analiza i povysheniya kachestva testov v sisteme elektronnogo obucheniya MOODLE [Method of test quality analysis and improvement in MOODLE e-learning system]. *Vestnik Voronezhskogo gosudarstvennogo universiteta. Seriya: Sistemnyy analiz i informatsionnyye tekhnologii*, 2014, No. 3, Pp. 61–72. (rus)

4. **Tolstobrov A.P., Protasova I.V., Korzhik I.A.** Sistema analiza statistiki testirovaniya kak sredstvo samootsenki prepodavatelem elektronnogo obrazovatelnogo resursa [The system of analysis of testing statistics as a means of self-assessment by a teacher of an electronic educational resource]. *Sovremennyye informatsionnyye tekhnologii i IT-obrazovaniye* [*Modern Information Technologies and IT-Education*], 2013, No. 9, Pp. 133–141. (rus)

5. **Nesterov S.A.** Analiz statistiki vypolneniya testovykh zadaniy v srede distantsionnogo obucheniya MOODLE [Analysis of quiz statistics in LMS MOODLE]. *Modern Information Technologies and IT-Education*, 2016, Vol. 12, No. 4, Pp. 62–67. (rus)

6. **Romero C., Ventura S.** Educational Data Mining: A review of the state of the art. *Journal IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews Archive*, 2010, Vol. 40, Issue 6, Pp. 601–618.

7. **Belonozhko P.P., Karpenko A.P., Khramov D.A.** Analiz obrazovatelnykh dannykh: napravleniya i perspektivy primeneniya [Analysis of educational data: directions and prospects of application]. *Journal Naukovedenie*, 2017, Vol. 9, No. 4. Available: http://naukovedenie.ru/PDF/15TVN417.pdf (Accessed: 01.11.2019). (rus)

8. **Algarni A.** Data Mining in education. *International Journal of Advanced Computer Science and Applications*, 2016. Vol. 7(6), Pp. 456–461. DOI: 10.14569/IJACSA.2016.070659

9. **Nesterov S.A., Smolina E.M.** Methods of Data Mining in analysis of the results of distance learning. *Proceedings of XXIII International Conference on Systems Analysis in Engineering and Control*, June 10–11, 2019. St. Petersburg: Politeh-Press, 2019, Vol. 3, Pp. 407–412. (rus)

10. **Sweeney M., Lester J., Rangwala H., Johri A.** Next-term student performance prediction: A recommender systems approach. *JEDM*, 2016, Vol. 8, Issue 1, Pp. 22–51.

11. **Villanueva A., Moreno L.G., Salinas M.J.** Data Mining techniques applied in educational environments: Literature review. *Digital Education Review*, 2018, No. 33, Pp. 235–266.

12. **Salal Y.K., Abdullaev S.M.** Using of Data Mining techniques to predict of student's performance in Industrial Institute of Al-Diwaniyah, Iraq. *Bulletin of the South Ural State University. Ser. Computer Technologies, Automatic Control & Radioelectronics*, 2019, No. 19, Pp. 121–130. DOI: 10.14529/ctcr190111

13. **Yang D., Kraut R., Rose C.** Exploring the effect of student confusion in massive open online courses. *JEDM*, 2016, Vol. 8, Issue 1, Pp. 52–83.

14. **Tabaa Y., Medouri A.** LASyM: A learning analytics system for MOOCs. *International Journal of Advanced Computer Science and Applications*, 2013, Vol. 4, Issue 5, Pp. 113–119. DOI: 10.14569/IJACSA.2013.040516

15. **Andreyeva N.V., Nesterov S.A.** Data management: online course. Available: *https://openedu.ru/course/spbstu/DATAM/* (Accessed: 01.11.2019). (rus)

16. **Bruce A., Bruce P.** *Practical statistics for Data Scientists*. O'Reilly Media, 2017.

17. **Lantz B.** *Machine learning with R*. Packt Publishing, 2015.

18. **Wickham H., Grolemund G.** *R for Data Science: Import, tidy, transform, visualize and modeling data*. St. Petersburg: Alfa-kniga Publ., 2018. 592 p. (rus)

19. **Barsegyan A.A., Kupriyanov M.S., Stepanenko V.V., Kholod I.I.** *Metody i modeli analiza dannykh: OLAP i Data Mining* [*Methods and models of data analysis: OLAP and Data Mining*]. St. Petersburg: BHV–Petersburg Publ., 2004. 336 p. (rus)

20. **Gras J.** *Data Science. Nauka o dannykh s nulya* [*Data Science from Scratch*]. St. Petersburg: BHV–Petersburg Publ, 2017. 336 p. (rus)

21. **Hill P.** Emerging student patterns in MOOCs: A (Revised) graphical view. Available: *https://eliterate.us/emerging-student-patterns-in-moocs-a-revised-graphical-view/* (Accessed: 01.11.2019).

Received 10.11.2019.

СПИСОК ЛИТЕРАТУРЫ

1. **Gelman B., Revelle M., Domeniconi C., Johri A., Veeramachaneni K.** Acting the same differently: A cross-course comparison of user behavior in MOOCs // Proc. of the 9th Internat. Conf. on Educational Data Mining. 2016. Pp. 376–381 // URL: http://www.educationaldatamining.org/EDM2016/proceedings/paper_136.pdf (Дата обращения: 01.11.2019).

2. **Нестеров С.А., Смолина Е.М.** Анализ результатов дистанционного обучения в формате массового открытого онлайн-курса // Системный анализ в проектировании и управлении: сб. науч. тр. XXII Междунар. науч.-практ. конф. СПб.: Изд-во Политехн. ун-та, 2018. Ч. 2. С. 379–383.

3. **Протасова И.В., Толстобров А.П., Коржик И.А.** Методика анализа и повышения качества тестов в системе электронного обучения MOODLE // Вестник Воронежского государственного университета. Сер.: Системный анализ и информационные технологии. 2014. № 3. С. 61–72.

4. **Толстобров А.П., Протасова И.В., Коржик И.А.** Система анализа статистики тестирования как средство самооценки преподавателем электронного образовательного ресурса // Современные информационные технологии и ИТ-образование. 2013. № 9. С. 133–141.

5. **Нестеров С.А.** Анализ статистики выполнения тестовых заданий в среде дистанционного обучения MOODLE // Современные информационные технологии и ИТ-образование. 2016. Т. 12. № 4. С. 62–67.

6. **Romero C., Ventura S.** Educational Data Mining: A review of the state of the art // J. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews Archive. 2010. Vol. 40. Issue 6. Pp. 601–618.

7. **Белоножко П.П., Карпенко А.П., Храмов Д.А.** Анализ образовательных данных: направления и перспективы применения // Науковедение. 2017. Т. 9. № 4 // URL: http://naukovedenie.ru/PDF/15TVN417.pdf (Дата обращения: 01.11.2019).

8. **Algarni A.** Data Mining in education // Internat. J. of Advanced Computer Science and Applications. 2016. Vol. 7(6). Pp. 456–461. DOI: 10.14569/IJACSA.2016.070659

9. **Нестеров С.А., Смолина Е.М.** Методы интеллектуального анализа данных в задачах оценки результатов дистанционного обучения // Системный анализ в проектировании и управлении: сб. науч. тр. XXIII Междунар. науч.-практ. конф. СПб.: Политех-Пресс, 2019. Ч. 3. С. 407–412.

10. **Sweeney M., Lester J., Rangwala H., Johri A.** Next-term student performance prediction: A recommender systems approach // JEDM. 2016. Vol. 8. Issue 1. Pp. 22–51.

11. **Villanueva A., Moreno L.G., Salinas M.J.** Data Mining techniques applied in educational environments: Literature review // Digital Education Review. 2018. No. 33. Pp. 235–266.

12. **Salal Y.K., Abdullaev S.M.** Using of Data Mining techniques to predict of student's performance in Industrial Institute of Al-Diwaniyah, Iraq // Bulletin of the South Ural State University. Ser. Computer Technologies, Automatic Control & Radioelectronics. 2019. No. 19. Pp. 121–130. DOI: 10.14529/ctcr190111

13. **Yang D., Kraut R., Rose C.** Exploring the effect of student confusion in massive open online courses // JEDM. 2016. Vol. 8. Issue 1. Pp. 52–83.

14. **Tabaa Y., Medouri A.** LASyM: A learning analytics system for MOOCs // Internat. J. of Advanced Computer Science and Applications. 2013. Vol. 4. Issue 5. Pp. 113–119. DOI: 10.14569/IJACSA.2013.040516

15. **Андреева Н.В., Нестеров С.А.** Управление данными: онлайн-курс // URL: https://openedu.ru/course/spbstu/DATAM/ (Дата обращения: 01.11.2019).

16. **Bruce A., Bruce P.** Practical statistics for Data Scientists. O'Reilly Media, 2017.

17. **Lantz B.** Machine learning with R. Packt Publishing, 2015.

18. **Уикем Х., Гроулмунд Г.** Язык R в задачах науки о данных: импорт, подготовка, обработка, визуализация и моделирование данных. Пер. с англ. СПб.: Изд-во «Альфа-книга», 2018. 592 с.

19. **Барсегян А.А., Куприянов М.С., Степаненко В.В., Холод И.И.** Методы и модели анализа данных: OLAP и Data Mining. СПб.: БХВ-Петербург, 2004. 336 с.

20. **Грас Дж.** Data Science. Наука о данных с нуля. СПб.: БХВ-Петербург, 2017. 336 с.

21. **Hill P.** Emerging student patterns in MOOCs: A (Revised) graphical view // URL: https://eliterate.us/emerging-student-patterns-in-moocs-a-revised-graphical-view/ (Дата обращения: 01.11.2019).

Статья поступила в редакцию 10.11.2019.

THE AUTHORS / СВЕДЕНИЯ ОБ АВТОРАХ

Nesterov Sergei A.
Нестеров Сергей Александрович
E-mail: nesterov@saiu.ftk.spbstu.ru

Smolina Elena M.
Смолина Елена Михайловна
E-mail: smolensk9595@mail.ru