Combining clustering and sequential pattern mining to detect behavioral differences in log data: conceptualization and case study

Juan Antonio Martínez-Carrascal [1,2*] [0000-0002-7696-6050], Elena Valderrama [2] [0000-0001-7673-2310] and Teresa Sancho-Vinuesa[1] [0000-0002-0642-2912]

[1] Universitat Oberta de Catalunya, Rambla del Poblenou, 156, 08018 Barcelona, Spain
[2] Universitat Autonoma de Barcelona, Engineering School, Campus UAB, 08193 Bellaterra, Spain
jmartinezcarra@uoc.edu

Abstract. Many on-campus universities are shifting their methodologies towards blended learning models. In these models, students cover some content online –normally associated with traditional lectures and quiz practice–, and attend also on-campus activities. Online activity is recorded in Learning Management Systems (LMS) logs, where interactions of the students with content are recorded.

In this article, we propose a method to detect differences in behavior considering only data recorded in the LMS log. We begin by clustering students based on activity log information. This process is carried out on a periodical basis. Clustering results are translated into meaningful states and then sequenced. The generated sequence is mined through sequential pattern mining (SPM).

Besides method description, we apply our method to a specific case-study to prove its validity. In particular, we analyze differences between passing and failing students in a blended-learning course. We prove that the method can generate meaningful sequences which – once analyzed – show relevant behavioral differences between students who pass and those who fail. In particular, failing students show more disengagement patterns than those who pass, while working attitudes - in particular, continuous working - are more common among the passing group.

Keywords: behavioral analysis, sequential pattern mining, process mining, student modelling, learning analytics.

1 Introduction

Learning management systems (LMSs) are nowadays a common piece in any university. Some of them were in fact born thanks to these systems. On the other side, some traditional on-campus universities were initially reluctant to develop them on their campuses. The perceived added value of the campus life – including campus classes – was considered a core value which did not admit changes.

This initial view is nowadays no more than a reminder of the days where the Internet was stilly maturing. Today, it would be difficult to find any university which has not adopted some kind of LMS [1]. LMSs have contributed to nuclear changes in the way instructors teach and students learn [2]. Beyond its relevant participation in the learning process, it has provided researchers with a valuable item: data.

Data coming from LMSs can be analyzed alone, combined with pre-existing student data, or with data coming from other systems. A whole set of data-mining techniques has been developed to solve different problems that range from performance prediction to analysis of social engagement [3]–[5]. These techniques can be analyzed from a computational perspective – which in the educational field leads to Educational Data Mining (EDM) [6] – or looking for its application in the learning process – giving place to Learning Analytics (LA) [7], [8]-.

Among these techniques, educational process mining (EPM) has gained popularity in recent years [9], [10]. EPM derives from generic process mining (PM) [11] but applies specifically to educational data. Inside this subset, sequential pattern mining (SPM) [12] tries to find interesting subsequences in a dataset. In the learning scenario, potential applications of this technique covers a broad range of topics. While our interest focuses on the detection of behavioral differences, other applications include better curriculum design or performance prediction [13]–[15].

In this paper we describe a method based on SPM to detect behavioral differences between passing and failing students in a blended-learning course based on log data. The method is tested on a first-year Engineering course offered at a public university. The subject under analysis was designed with blended-learning approach and lasts for 12 weeks, including three main milestones which correspond to intermediate tests performed in weeks 4 and 8 and a final test in week 12. The edition we analyze has a total of 337 students enrolled – 199 of them passing the subject -. A detailed description of the course structure can be found in [16], where results show that both groups effectively behave differently.

Our goal is to evaluate if we can detect these differences through a new method. The method we describe begins by clustering log activity on a per-week basis. Output of this clustering process is then interpreted to determine the weekly state of the student. We iterate this process in consecutive periods to create a sequence of states. Data is then split in two groups – associated to passing and failing students – and SPM is used to detect behavioral differences. In particular, we raise the following research question (RQ):

- RQ: Can the method exposed collect and show behavioral differences among both group of students?

To answer the question, we describe the method in detail and apply it to the aforementioned case study. Once done, we analyze if different behavioral patterns can be found in the sequences present in the failing and passing group. Results will reveal that behavior is different for the passing and failing group, and specific patterns which anticipate potential failure arise. This fact, combined with the simplicity to gather and process data, opens interesting applications such as predictor or alarm indicator.

2 Theoretical framework

Research on learning analytics [7], [8] focuses on different aspects of learning, being one of them the analysis of the learning process itself [9], [17]. In this context, studies that apply process analysis tool to learning environments have emerged with rising interest in recent years, as different compilations show [9], [10].

The process approach was common in other disciplines [18], in particular in business and industry. In the learning scenario, specific areas of interest include curriculum mining, computer-based assessment or LMS log analysis. Expected results include the detection of learning difficulties, learning flows or sequential patterns[9].

While processes can be designed based on theoretical behavior, complex or unstructured processes need a process discovery stage [18]. The learning process totally fits into this categorization. Process discovery constitutes a discipline on its own, where one of the approaches is to model processes based on log records, that are translated into real models, including the detection of variants in the same process[19]–[21]. The field is promising, but initial models usually conform *spaghetti-like* graphs, which require simplification [22]. In addition, direct interpretation is not straightforward.

For these reasons, some studies focus on analysis of sub-processes, or look for meaningful sequence of actions. The basics for these techniques were introduced in [23] and constitute the core of SPM. In particular, SPM deals with ”sequences of events, items, or tokens occurring in an ordered metric space appear often in data and the requirement to detect and analyze frequent subsequences is a common problem.“ [24].

SPM applications are common in different knowledge areas ranging from medicine to business processes [12]. Applications are broad, depending on the topic under analysis. For instance in the medical field, they range from prescription [25] to detection of specific patterns, such as gene mutation [26].

In the learning field, we can find different studies and applications. [27] looks for better understanding of the learning process through the use of a specific algorithm. [13], [28] show applications in recommender systems. Other uses include impact of behavior during practical sessions on final performance [29] or discovering of navigational patterns [14]. Applications to behavioral analysis can also performed through SPM [30].

Practical implementation of SPM can be done in two different ways: either a-priori or using pattern growth. A-priori methods are based on [23], and rely on the hypothesis that if a sequence is not frequent, super sequences based on it can neither be frequent. Pattern-growth methods are based on [31]. Any of these methods can provide, for each

of the detected sequences, its support, which indicates the percentage of items where the sequence is present.

In order to feed these SPM algorithms with data coming from log files, preprocessing is needed, as a sequence of states is required as input. [32] remarks that one of the challenges when translating logs into process is the granularity of log data. In order to translate this log data into states, an aggregation is needed, as working with low-granularity data can be useless [14].

An interesting approach to this aggregation is the use of clustering based on log data, which can be found in studies such as [33], [34]. In particular, in [34] clustering is used to detect groups, while SPM is used to detect sequences in that group. Following this research line, we plan to create the sequences based on the clustering process itself. In other words, we do not plan to cluster students, but to cluster states on a periodical basis that will finally constitute a sequence.

To feed clustering algorithms, commonly interesting variables include the number of lectures viewed, quizzes assessed or time between consecutive sessions. These parameters show relevant for student success, with relevance depending on the case under study [35]. Some of these parameters require a deep knowledge of the course structure. While it is relatively simple to know how many times a user logins per week, knowing how many lectures or quizzes a user performs requires prior classification of LMS items – associating a category to each individual item–.

Regarding the algorithm itself, different techniques can be used. [36] performs a compilation of potentially interesting algorithms which are common in e-Learning problems.

3 Methodology

Linking these ideas, we build a sequence of states for each user and time interval. The states are created through clustering based on data extracted from log files. These states are sequenced and a search is performed. Analysis will look for relevant patterns that can show behavioral differences.

We decided to analyze activity on a per-week basis for two reasons. First, due to the dynamics of the blended course. Instructors anticipate contents the students should cover before attending class, and this is normally done on a weekly basis. Second reason influencing this decision takes into account student habits. While some students are full-time students and can possibly cover contents during the week, some others may only be able to cover them during the weekend. The weekly analysis accommodates this situation, showing whether students are effectively engaged during the different weeks of the course.

Clustering results for the different weeks is then be analyzed and labelled according to meaningful patterns. Once done, a sequence is created for each student. This sequence will be meaningful, as labels will have been assigned to each of the states.

Finally, and in order to validate the method with a practical case study, we will apply it to a real case. We separate students into two groups. This segmentation will be done according to academic result – pass or fail – For each of these groups, we will perform

SPM. After getting the results, comparison will be made between groups, in order to validate if the method properly detects differences in support for specific patterns between groups.

3.1 Clustering: input data and algorithms

As outlined in the introduction, we will focus in online activity. The data we analyze corresponds to that gathered in the LMS. We do not include data from other sources, neither take into account any evaluative marks obtained by the students. We consider the global amount of online activity performed, but also classify interactions according to the kind of content covered.

Online activity will be classified based on the categorization of the tasks students are instructed to follow. On each class, students are suggested to cover specific contents before to prepare forthcoming sessions. These contents fit into one of these categories:

- Lectures: which correspond to encapsulated videos provided prior to face-to-face sessions.
- Problem sets: where the student can test to what extent she has acquired knowledge properly
- Evaluative quizzes: that correspond to quizzes that have impact on final grade
- Specific non-assessed contents: which correspond to contents related to the subject, and which are covered in classes, but that are not assessed in any evaluation, and do not have impact on the final grade
- Suggested readings.

Instructors were asked to detect and inform relevant content for each week of the course. Students are specifically instructed to cover these contents before attending specific face-to-face sessions. A total of 150 items were considered. Table 1 shows the type and amount of activities considered.

Table 1. Number of items in the course for each content type

Kind of activity	Number of items
Lectures	54
Problem sets	31
Evaluative quizzes	7
Specific non-assessed contents	12
Suggested readings	46

Data for each of these five kind of content are kept along with the number of login sessions the user performs in the period under consideration. For each week and student, we summarize the number of items in each category – for instance, a student can watch 5 lectures, review 3 problem sets and do not perform any other kind of activity –, performing 4 login sessions. This information will constitute the input to the clustering algorithm. To be able to extract this information, instructors must classify contents in advance in order to properly account each access. It is interesting to note

that once this is done, the approach is computationally simple, and data to feed the clustering algorithm is readily available.

Our idea when running this experiment was to obtain a clear view of different group behaviors for each week of the course. For instance, we expected to detect a cluster containing students who show high activity, or a cluster clearly focused on evaluative assessments. Clustering in the different stages should be consistent to allow comparison among weeks – same label should indicate same behavior –. In this way, we could check temporal evolution. For instance, a sequence could indicate that a student begins in the 'high activity' group and then changes to the 'assessment oriented' during the following week.

Regarding the clustering technique, we opted to select k-Means as clustering algorithm for being commonly used [36]. In this technique, clusters are created based on the distance to a centroid, which constitutes the center of this cluster. Interpretation of centroid data will allow to assign meaningful labels to the clusters obtained.

To implement this approach, we needed to consider the number of clusters in advance. Literature indicates 4-5 clusters is a common number for this kind of environments[33], [37]. Our tests will be done considering k=4 as a potentially interesting number of clusters, as our scenario can be considered similar to [33] in terms of course methodology – flipped –, course duration and LMS data as main data source.

The translation form cluster labels to meaningful naming will be done based on centroid information analysis. Centroids will be kept to allow proper interpretation and appropriate labelling of the cluster. We must keep in mind that the sequence we look for would be useless having non-meaningful states such as 'cluster *n*'. Analyzing centroids will allow us to interpret the meaning of the cluster, and label a particular group with meaningful attributes such as 'high activity' or 'assessment oriented'.

Besides individual week cluster labeling, centroids will also be used to provide cluster coherence among different weeks. In other words, the same detected behavior should map to the same label, even among different weeks. For clarity purposes, we will try to keep the same number of clusters – and interpretation if possible - for all weeks. This analysis will be done manually, as human intervention is needed to properly interpret cluster results, and to provide coherence among weeks.

3.2 Establishing and mining sequence of actions

In order to properly model student behavior through the course, we map the information obtained through the clustering process into a sequence of situations. For instance, assuming the clustering process leaves three groups, labeled as 'Low activity' (1), 'Quiz oriented'(2) and 'Low login' (3), a sequence such as (1,2,3,3,3,3) would mean the user begins by performing low (1), she then has a quiz-oriented week (2), and after that four weeks with low login activity (associated to the 3333 in the sequence.

This sequence will be treated as a sequence of states that will be mined with sequence mining tools. Data will be split between passing and failing students, in order to detect differences in support for the most relevant patterns. As noted in the theoretical framework, different techniques exist [12]. We selected generalized sequential pattern

(GSP) algorithm for being commonly used [12], due to the existence of proven implementations, and considering performance is not a constraint (n=337 students). Our focus will be set on the interpretation of results, and not on the algorithm itself. We will keep sequence and associated support for each of the groups under analysis. An open-source implementation will be used [38].

4 Results

4.1 Clustering

The clustering process was carried out with k-means. x-means was previously used to explore the potential number of interesting clusters and confirmed k=4 was a proper number, which could accommodate clustering results for the different weeks.

The centroid analysis provided also interesting results. While almost all studies suggest there is a low activity group and a high activity group, some other behaviors exists. For instance, gamers who try to game the system and perform high number of quizzes but do not follow lessons in such a way. Table 2 shows centroid data for the first two weeks:

Table 2. Sample of centroid data (first two weeks)

Week	Cluster ID	Lectures	Quizzes	Evaluative	Non assessed	Suggested	Login sessions
1	cl0	18,93	9,64	8,93	24,86	0,86	5,75
1	cl1	2,51	1,27	1,50	1,03	0,07	1,78
1	cl2	15,12	6,24	6,79	12,30	0,54	5,06
1	cl3	9,31	4,06	5,07	4,50	0,27	4,31
2	cl0	3,32	1,53	2,06	1,30	0,17	2,52
2	cl1	18,21	6,94	9,14	11,30	0,87	6,54
2	cl2	35,82	9,82	10,45	22,82	2,55	7,18
2	cl3	11,28	4,24	5,53	4,43	0,42	5,29

As Table 2 shows, for our case study the group with higher values for login sessions per week shows also higher activity in the different categories. This finding suggests that the clustering is really showing the amount of work performed by the student. For instance, the low on-line activity group – which for Table 2 would be cl1 for week 1 or cl0 for week 2 - is always present, showing low performance in all items (lectures, quizzes, …). The other three clusters are graded according to their amount of work. For this reason, we identified the clusters as low (L), medium (M), high (H) and extreme (E) activity.

4.2 Segmentation

The results of the clustering process were compiled into sequences for each of the students. As stated, we segmented the dataset into two groups according to final academic result. This segmentation is performed in order to detect differences in patterns between the passing and the failing group.

Once segmentation is performed, GSP algorithm is run on each of the resulting sequence dataset.

4.3 Relevant patterns for the failing group

GSP algorithm run on the failing group dataset looking for sequences with a minimum support of 0.6. Among the resulting sequences, Table 3 shows those with greater support. As it could be expected, long periods of low-activity are present among those students who finally fail the subject. It is also noticeable that most sequences include one or more low activity periods (L).

Table 3. Sequences that are present in higher percentage among the failing group (Top 10)

Sequence	% failing students showing sequence
L	0.95
L - L	0.88
M	0.84
L - L - L	0.80
M - L	0.74
M - M	0.73
L - L - L - L	0.72
L - L - L - L - L	0.69
M - L - L	0.68
H	0.68

4.4 Relevant patterns for the passing group

We carried out the same process for the passing group. Again, we used 0.6 as minimum support. Most common sequences and their support are shown in Table 4:

Table 4. Sequences that are present in higher percentage among the passing group (Top 10)

Sequence	% passing students showing sequence
M	0.97
M - M	0.92
H	0.87

M - M - M	0.84
L	0.84
M - H	0.83
M - L	0.81
H - H	0.77
H - M	0.76
M - M - M - M	0.74

4.5 Comparison of patterns

Despite Tables 3 and 4 already show noticeable differences, we perform a specific search to determine the support for sequences in the failing group inside the passing group. In this case, support for a specific sequence can be below 0.6, as it can be common only in the failing group. We also sort the table according to this difference in support. Results are shown in Table 5:

Table 5. Support for Top-10 failing sequences among the passing group

Sequence	% failing students showing sequence	% passing students showing sequence	Difference
L - L - L - L - L	0.69	0.3	0.38
L - L - L - L	0.72	0.42	0.29
L - L - L	0.8	0.56	0.24
L - L	0.88	0.68	0.2
L	0.95	0.84	0.11
M - L - L	0.68	0.62	0.05
M - L	0.74	0.81	-0.06
M	0.84	0.97	-0.13
M - M	0.73	0.92	-0.18
H	0.68	0.87	-0.19

5 Discussion

Results in Tables 3,4 and 5 allow us to answer the RQ raised in the introduction. The application of the described method to our case study has proven valid to detect differences in behavior between the two groups under study: students who pass and those who fail behave differently. That means that behavior is kept in the state sequence. We deepen into the process itself and its results for this particular case.

The process described uses SPM to mine sequences generated through clustering. Clustering is the initial stage, and in our case, produced pure activity groups. Students

who show higher volume of activity show it on all kind of items. In particular, and for instance, students with higher number of login sessions show also higher lecture activity and higher quiz completions.

This fact can be compared with other clustering analysis present in the literature. A similar scenario – university course, first year engineering, computer science topic and flipped design – can be found in [33]. In this case, a clustering process is also performed aimed to detect student strategies. As a key different to our study, assessment data is included into the clustering process. The study detects four initial clusters, two oriented to assessments – formative or summative – and two related to content – one more oriented to video lecture and one to reading materials-.

[37] provides also a study of two courses focused on activity. Four clusters are also identified, being two of them clearly identified as highly active and low active. This study includes only activity, gathered also from a LMS platform. Clusters showing activity show also higher activity for the values considered – in this case, resource view, forum view and forum participation – with the exception of one single group showing least forum activity.

In the MOOC environment, this kind of studies is also present to analyze engagement in courses [39], [40]. We believe this scenario shows relevant differences in behavior to our case. This reason explains different pattern detection, such as samplers or returners. We believe this behavior is common in MOOC courses, but not so much in regular university courses.

Regarding behavioral sequences, Table 3 shows more common behavior for failing students. 95% of them show low activity in at least one of the weeks, and almost 90% in two consecutive weeks. A week with low activity is present in top-5 sequences, and in 7 out of 8 of those sequences with a minimum of two items. For the passing group, the most common situation is to follow medium or high engagement combinations.

Differences become more evident if we have a look at Table 5. While almost 70% of failing students show 5 consecutive weeks of low engagement with content, only 30% of passing students show this behavior. At the same time, it is also noticeable that sequences showing two or more low access weeks show the higher differences with passing students.

In fact, and according to Table 4, sequences which include at least one week of medium or high activity are more common in passing students. That indicates that passing students perform higher volume of online activity. From a pedagogical point of view, the interpretation of results in Table 5 shows that while one disengagement week makes no major difference, failure probability increases as the number of disengaged weeks does. In other words, the continuous detection of low online activity can indicate the student is more likely to fail.

Besides specific interpretation of this case study, we believe the approach provides an interesting insight to log analysis and its transformation into a process model. Log processing is simple and no specific restrictions have been imposed to algorithms for clustering or SPM. We have also indicated a potential selection of specific tools, such as k-means clustering and GSP.

We consider two parameters make the approach particularly attractive. First, the system is easy to implement. Only data obtained from log is needed. No sociological,

preexisting or data coming from other record systems is needed. Classification does not require prior categorization as the work is done according to individual behavior in relation to the group.

Second, the process takes into account not only static values, but a dynamic picture of the student. A low engagement week may not be relevant, but it can become a problem if two consecutive weeks – or more – are accumulated. In short, the model is capturing a relevant part of the learning process. And this learning process is not a static picture. Analysis can only be done when the process is seen in perspective. In this sense, we consider the method depicted can provide a new and interesting insight to many problems related to research in learning analytics.

6 Conclusions

The process described generates a sequence of states based on behavioral clustering. This sequence is then analyzed in order to detect differences between two groups of students (passing and failing). Results show that behavior is effectively different and that this difference is contained in the sequences analyzed.

While we have focused on the method itself, we envision two groups of potential applications of this process. First, the use as a potential failure indicator. Second, as a detector of points of disengagement during the course, which could lead to curriculum redesign.

In order to implement potential applications in any of these groups, a previous extension of the study would be advisable. This extension can be done to successive editions of the same course, or to other courses, opening interesting research lines.

In the first case, results could be potentially extended to analyze forthcoming editions of the same course. While the issue of portability has not been addressed, we believe the study could open a different approach in prediction processes. The method could be carried out on a per-week basis as described and raise alarms when sequences indicating failure are detected.

Regarding portability to other courses, it would be interesting to compare results among courses, and even deep into the pedagogical implications of course type and methodology in results. For instance, results could help to detect differences not only in terms of passing and failing groups, but can detect differences among on-campus or on-line courses, methodologies - i.e. blended, flipped, MOOC- or even topic – STEM vs social -. This extensions could allow deeper comparison of results with some references analyzed in this study (for instance [37], [39], [41] for the MOOC case).

Finally, and while our interest has remained on pure non-grading activity data, the method could also be extended to other scenarios, and include other aspects in the clustering process, such as sociological data or even impact of specific learning activities (i.e. quizzes or evaluative assessments). Authors are open to collaborate in these open scenarios, in particular looking for practical applications and contributions to better learning designs.

References

[1] W. Lasanthika and W. Tennakoon, "Assessing the Adoption of Learning Management Systems in Higher Education," *GATR Glob. J. Bus. Soc. Sci. Rev.*, vol. 7, no. 3, pp. 204–209, Sep. 2019, doi: 10.35609/gjbssr.2019.7.3(5).

[2] H. Coates, R. James, and G. Baldwin, "A critical examination of the effects of learning management systems on university teaching and learning," *Tert. Educ. Manag.*, vol. 11, no. 1, pp. 19–36, Mar. 2005, doi: 10.1007/s11233-004-3567-9.

[3] F. Castro, A. Vellido, À. Nebot, and F. Mugica, "Applying data mining techniques to e-learning problems," *Stud. Comput. Intell.*, 2007, doi: 10.1007/978-3-540-71974-8_8.

[4] C. Romero, S. Ventura, P. G. Espejo, and C. Hervás, "Data mining algorithms to classify students," in *Educational Data Mining 2008 - 1st International Conference on Educational Data Mining, Proceedings*, 2008.

[5] C. Vieira, P. Parsons, and V. Byrd, "Visual learning analytics of educational data: A systematic literature review and research agenda," *Comput. Educ.*, vol. 122, 2018, doi: 10.1016/j.compedu.2018.03.018.

[6] R. S. J. D. Baker and K. Yacef, "The State of Educational Data Mining in 2009 : A Review and Future Visions," *J. Educ. Data Min.*, 2009, doi: http://doi.ieeecomputersociety.org/10.1109/ASE.2003.1240314.

[7] G. Siemens, "Learning analytics: Envisioning a research discipline and a domain of practice," in *ACM International Conference Proceeding Series*, 2012, doi: 10.1145/2330601.2330605.

[8] R. Ferguson, "Learning analytics: drivers, developments and challenges," *Int. J. Technol. Enhanc. Learn.*, vol. 4, no. 5, pp. 304–317, 2012, doi: 10.1504/IJTEL.2012.051816.

[9] A. Bogarín, R. Cerezo, and C. Romero, "A survey on educational process mining," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 8, no. 1, p. e1230, Jan. 2018, doi: 10.1002/widm.1230.

[10] M. A. Ghazal, O. Ibrahim, and M. A. Salama, "Educational process mining: A systematic literature review," in *Proceedings - 2017 European Conference on Electrical Engineering and Computer Science, EECS 2017*, Jul. 2018, pp. 198–203, doi: 10.1109/EECS.2017.45.

[11] W. Van Der Aalst, "Process mining: Overview and opportunities," *ACM Transactions on Management Information Systems*, vol. 3, no. 2. pp. 1–17, Jul. 2012, doi: 10.1145/2229156.2229157.

[12] P. Fournier-Viger, J. Chun, W. Lin, R. U. Kiran, Y. S. Koh, and R. Thomas, "A Survey of Sequential Pattern Mining," 2017.

[13] M. Salehi *et al.*, "Personalized recommendation of learning material using sequential pattern mining and attribute based collaborative filtering," doi: 10.1007/s10639-012-9245-5.

[14] L. K. M. Poon, S. C. Kong, M. Y. W. Wong, and T. S. H. Yau, "Mining sequential patterns of students' access on learning management system," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2017, vol. 10387 LNCS, pp. 191–198, doi: 10.1007/978-3-319-61845-6_20.

[15] P. Fournier-Viger, J. Li, J. C. W. Lin, T. T. Chi, and R. Uday Kiran, "Mining cost-effective patterns in event logs," *Knowledge-Based Syst.*, vol. 191, Mar. 2020, doi: 10.1016/j.knosys.2019.105241.

[16] J. A. Martínez-Carrascal, D. Márquez Cebrián, T. Sancho-Vinuesa, and E. Valderrama, "Impact of early activity on flipped classroom performance prediction: A case study for a first-year Engineering course," *Comput. Appl. Eng. Educ.*, Mar. 2020, doi: 10.1002/cae.22229.

[17] P. Strecht, L. Cruz, C. Soares, J. Mendes-Moreira, and R. Abreu, "A Comparative Study of Classification and Regression Algorithms for Modelling Students' Academic Performance," 2015. Accessed: Apr. 11, 2019. [Online]. Available: http://www.educationaldatamining.org/EDM2015/proceedings/short392-395.pdf.

[18] W. Van Der Aalst, "Process mining: Overview and opportunities," *ACM Transactions on Management Information Systems*, vol. 3, no. 2. Jul. 2012, doi: 10.1145/2229156.2229157.

[19] N. Tax, N. Sidorova, R. Haakma, and W. M. P. van der Aalst, "Mining local process models," *J. Innov. Digit. Ecosyst.*, 2016, doi: 10.1016/j.jides.2016.11.001.

[20] A. Bolt, M. de Leoni, and W. M. P. van der Aalst, "Process variant comparison: Using event logs to detect differences in behavior and business rules," *Inf. Syst.*, vol. 74, pp. 53–66, May 2018, doi: 10.1016/j.is.2017.12.006.

[21] S. Rizvi, B. Rienties, J. Rogaten, and R. F. Kizilcec, "Investigating variation in learning processes in a FutureLearn MOOC," *J. Comput. High. Educ.*, vol. 32, no. 1, pp. 162–181, Apr. 2020, doi: 10.1007/s12528-019-09231-0.

[22] D. Fahland and W. M. P. van der Aalst, "Simplifying Mined Process Models: An Approach Based on Unfoldings," Springer, Berlin, Heidelberg, 2011, pp. 362–378.

[23] R. Agrawal and R. Srikant, "Mining sequential patterns," in *Proceedings - International Conference on Data Engineering*, 1995, pp. 3–14, doi: 10.1109/icde.1995.380415.

[24] C. H. Mooney and J. F. Roddick, "Sequential pattern mining - Approaches and algorithms," *ACM Computing Surveys*, vol. 45, no. 2. Feb. 2013, doi: 10.1145/2431211.2431218.

[25] A. P. Wright, A. T. Wright, A. B. McCoy, and D. F. Sittig, "The use of sequential pattern mining to predict next prescribed medications," *J. Biomed. Inform.*, vol. 53, pp. 73–80, Feb. 2015, doi: 10.1016/j.jbi.2014.09.003.

[26] A. S. Johnsha Ali, "Frequent Sequential Patterns (FSP) Algorithm for Finding Mutations in BRCA2 Gene," *Int. J. Recent Technol. Eng.*, no. 3, pp. 2277–3878, 2019, doi: 10.35940/ijrte.C6507.098319.

[27] A. Bogarín, R. Cerezo, and C. Romero, "Discovering learning processes using inductive miner: A case study with learning management systems (LMSs)," *Psicothema*, vol. 30, no. 3, pp. 322–329, 2018, doi: 10.7334/psicothema2018.116.

[28] J. K. Tarus, Z. Niu, and A. Yousif, "A hybrid knowledge-based recommender system for e-learning based on ontology and sequential pattern mining," *Futur. Gener. Comput. Syst.*, vol. 72, pp. 37–48, Jul. 2017, doi: 10.1016/j.future.2017.02.049.

[29] R. Venant, K. Sharma, P. Vidal, P. Dillenbourg, and J. Broisin, "Using sequential pattern mining to explore learners' behaviors and evaluate their correlation with performance in inquiry-based learning," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2017, vol.

10474 LNCS, pp. 286–299, doi: 10.1007/978-3-319-66610-5_21.

[30] J. S. Kinnebrew, K. M. Loretz, and G. Biswas, "A Contextualized, Differential Sequence Mining Method to Derive Students' Learning Behavior Patterns," *JEDM | J. Educ. Data Min.*, vol. 5, no. 1, pp. 190–219, May 2013, doi: 10.5281/ZENODO.3554617.

[31] J. Han and J. Pei, "Mining frequent patterns by pattern-growth," *ACM SIGKDD Explor. Newsl.*, vol. 2, no. 2, pp. 14–20, Dec. 2000, doi: 10.1145/380995.381002.

[32] M. Zhou, Y. Xu, J. C. Nesbit, and P. H. Winne, "Sequential pattern analysis of learning logs: Methodology and applications," in *Handbook of Educational Data Mining*, 2010.

[33] J. Jovanović, D. Gašević, S. Dawson, A. Pardo, and N. Mirriahi, "Learning analytics to unveil learning strategies in a flipped classroom," *Internet High. Educ.*, vol. 33, pp. 74–85, Apr. 2017, doi: 10.1016/j.iheduc.2017.02.001.

[34] D. Perera, J. Kay, I. Koprinska, K. Yacef, and O. R. Zaane, "Clustering and sequential pattern mining of online collaborative learning data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 6, pp. 759–772, Jun. 2009, doi: 10.1109/TKDE.2008.138.

[35] R. Conijn, C. Snijders, A. Kleingeld, and U. Matzat, "Predicting student performance from LMS data: A comparison of 17 blended courses using moodle LMS," *IEEE Trans. Learn. Technol.*, 2017, doi: 10.1109/TLT.2016.2616312.

[36] A. Dutt, M. A. Ismail, and T. Herawan, "A Systematic Review on Educational Data Mining," *IEEE Access*, 2017, doi: 10.1109/ACCESS.2017.2654247.

[37] I. P. Ratnapala, R. G. Ragel, and S. Deegalla, "Students behavioural analysis in an online learning environment using data mining," in *2014 7th International Conference on Information and Automation for Sustainability: "Sharpening the Future with Sustainable Technology", ICIAfS 2014*, Mar. 2014, doi: 10.1109/ICIAFS.2014.7069609.

[38] P. Fournier-Viger *et al.*, "The SPMF open-source data mining library version 2," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2016, doi: 10.1007/978-3-319-46131-1_8.

[39] R. Ferguson and D. Clow, "Examining engagement: Analysing learner subpopulations in massive open online courses (MOOCs)," in *ACM International Conference Proceeding Series*, Mar. 2015, vol. 16-20-March-2015, pp. 51–58, doi: 10.1145/2723576.2723606.

[40] L. Shi and A. I. Cristea, "In-depth exploration of engagement patterns in MOOCs," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Nov. 2018, vol. 11234 LNCS, pp. 395–409, doi: 10.1007/978-3-030-02925-8_28.

[41] M. Khalil and M. Ebner, "Clustering patterns of engagement in Massive Open Online Courses (MOOCs): the use of learning analytics to reveal student categories," *J. Comput. High. Educ.*, vol. 29, no. 1, pp. 114–132, Apr. 2017, doi: 10.1007/s12528-016-9126-9.