Psicothema

Psicothema 2018, Vol. 30, No. 3, 322-329
doi: 10.7334/psicothema2018.116

ISSN 0214 - 9915 CODEN PSOTEG

www.psicothema.com

Discovering learning processes using Inductive Miner: A case study with Learning Management Systems (LMSs)

Alejandro Bogarín[1], Rebeca Cerezo[2] and Cristóbal Romero[1]
[1] Universidad de Córdoba and [2] Universidad de Oviedo

Abstract

Background: Process mining with educational data has made use of various algorithms for model discovery, principally Alpha Miner, Heuristic Miner, and Evolutionary Tree Miner. In this study we propose the implementation of a new algorithm for educational data called Inductive Miner. **Method:** We used data from the interactions of 101 university students in a course given over one semester on the Moodle 2.0 platform. Data was extracted from the platform's event logs; following preprocessing, the mining was carried out on 21,629 events to discover what models the various algorithms produced and to compare their fitness, precision, simplicity and generalization. **Results:** The Inductive Miner algorithm produced the best results in the tests on this dataset, especially for fitness, which is the most important criterion in terms of model discovery. In addition, when we weighted the various metrics according to their importance, Inductive Miner continued to produce the best results. **Conclusions:** Inductive Miner is a new algorithm which, in addition to producing better results than other algorithms using our dataset, also provides valid models which can be interpreted in educational terms.

Keywords: Educational Data Mining (EDM), Educational Process Mining (EPM), model discovery algorithms, Inductive Miner, Learning Management Systems (LMSs).

Resumen

Descubriendo procesos de aprendizaje aplicando Inductive Miner: un estudio de caso en Learning Management Systems (LMSs). **Antecedentes:** en la minería de procesos con datos educativos se utilizan diferentes algoritmos para descubrir modelos, sobremanera el Alpha Miner, el Heuristic Miner y el Evolutionary Tree Miner. En este trabajo proponemos la implementación de un nuevo algoritmo en datos educativos, el denominado Inductive Miner. **Método:** hemos utilizado datos de interacción de 101 estudiantes universitarios en una asignatura de grado desarrollada en la plataforma Moodle 2.0. Una vez prepocesados se ha realizado la minería de procesos sobre 21.629 eventos para descubrir los modelos que generan los diferentes algoritmos y comparar sus medidas de ajuste, precisión, simplicidad y generalización. **Resultados:** en las pruebas realizadas en nuestro conjunto de datos el algoritmo Inductive Miner es el que obtiene mejores resultados, especialmente para el valor de ajuste, criterio de mayor relevancia en lo que respecta al descubrimiento de modelos. Además, cuando ponderamos con pesos las diferentes métricas seguimos obteniendo la mejor medida general con el Inductive Miner. **Conclusiones:** la implementación de Inductive Miner en datos educativos es una nueva aplicación que, además de obtener mejores resultados que otros algoritmos con nuestro conjunto de datos, proporciona modelos válidos e interpretables en términos educativos.

Palabras clave: minería de datos educativos, minería de procesos educativos, algoritmos de descubrimiento, *Inductive Miner*, sistemas de gestión del aprendizaje.

The increase in internet access in recent years has allowed a huge number of students to experience higher education learning in Computer Based Learning Environments (CBLEs) (Broadbent, & Poon, 2015), generally through widely used Learning Management Systems (LMSs). These systems are ubiquitous in higher education, with 99% of US colleges and universities currently reporting that they have an LMS in place (Dahlstrom, Brooks, & Bichsel, 2014). These LMSs have provided very useful, fairly easy to access information for institutions and stakeholders by collecting data on all student activities at different levels of granularity, from enrollment in a particular program to student performance (Trcka, & Pechenizkiy, 2009). Various educational agents closer to the teaching-learning process have recently started to explore the adoption of these techniques to gain insight into online learners' study processes (Papamitsiou, & Economides, 2014) through Educational Data Mining (EDM), being the first decade of the twenty first century the kick-off of EDM (Peña-Ayala, 2014).

Educational Data Mining (EDM) techniques have been applied extensively to find interesting patterns from large volumes of educational data (Dutt, Ismail, & Herawan, 2017; Romero, & Ventura, 2007). However, EDM techniques are not generally aimed at discovering, analyzing or visualizing the complete educational process, they do not focus on the process but on the result. To allow analysis in which the process plays the central

Received: February 2, 2018 • Accepted: May 9, 2018
Corresponding author: Rebeca Cerezo
Departamento de Psicología
Universidad de Oviedo
33003 Oviedo (Spain)
e-mail: cerezorebeca@uniovi.es

role there is a new line of data-mining research called Educational Process Mining (EPM) (Romero, & Ventura, 2013). Nowadays, the use of Process Mining (PM) in the educational domain is in its early stages and has given rise to EPM research, which is one of the current promising techniques in the EDM firmament (Reimann, Markauskaite, & Bannert, 2014). Although both EDM and EPM start from data, there are some significant differences between them.

EDM can be generally understood as the application of Data Mining (DM) to the specific type of dataset that comes from learning environments in order to address educational questions (Romero, & Ventura, 2010; Weijters, Van Der Aalst, & De Medeiros, 2006). It focuses on the analysis of large data sets in the service of educational science that focuses on modeling and improving learning processes through the use of that data. EPM bridges the gap between EDM and educational science, as it combines data analysis with modeling, and insighting in educational processes. PM is process-centric (Pechenizkiy, Trcka, Vasilyeva, Van Aals &, De Bra, 2009) thereby making unknown (or only partially known) processes explicit. PM, in contrast to DM, is interested in end-to-end processes rather than local patterns (Van der Aalst, 2016).

The goal of EPM is to extract knowledge from event logs recorded by an educational system (LMSs, MOOCs, etc.) and EPM algorithms discover process models of student behavior. There are a great number of PM algorithms for discovering underlying processes from event logs, and they have been used in a wide range of application domains. Most of the work has concentrated on supporting company processes in business contexts (Van Der Aalst, 2011) and although there is a large body of previous research in applying EPM, the algorithms that have been used to report quality metrics to address educational issues are limited to Alpha Miner, Heuristic Miner and Evolutionary Tree Miner (Bogarín, Cerezo, & Romero, 2018):

Alpha Miner (AM): This was the first discovery algorithm and served as the base for the development of later, improved algorithms (Van der Aalst, 2016). Its main limitation is that it doesn't use frequencies, and so does not guarantee soundness, and is only suitable for event logs without noise, quite infrequently fact in learning data.

Heuristic Miner (HM): It has three significant improvements over the Alpha Algorithm. First, it takes frequencies and significance into account, so it can filter out noisy or infrequent behavior, which makes it less sensitive to noise and incomplete logs (Bogarin et al., 2014). Second, it can detect short loops. Third, it allows single activities to be skipped. It does not, however, guarantee sound educational process models.

Evolutionary Tree Miner (ETM): this is a genetic algorithm that optimizes the educational process model based on user-defined quality metrics. In addition, it works with process trees so unsound models will not be considered. By using a genetic algorithm for process discovery, it gains flexibility to change the weighting of different fitness factors, so process discovery can be guided based on the weighted average of predefined quality factors depending on the importance of each factor for the user (Buijs, Van Dongen, & van Der Aalst, 2012).

These algorithms have provided new ways of discovering, monitoring, and improving processes in different educational contexts such as computer-supported collaborative learning, curriculum mining, computer-based assessment, software repositories, professional training, 3D Educational Virtual Worlds, Structured Inquiry Cycle in informal adult learning, and of course, in MOOCs, LMSs and Hypermedia Learning Environments (Bogarin et al., 2018).

Regarding to the LMSs field, Trcka, Pechenizkiy, & Van der Aalst (2010) showed the potential of PM for extracting knowledge from student exam traces in LMSs. In Bogarin, Romero, Cerezo, & Sánchez-Santillán (2014) the authors used data clustering in order to produce more accurate PM models of student behavior. In a similar environment, Reiman et al. in 2014 proposed the use of PM with learning traces based on theoretical principles of Self-Regulated Learning (SRL). Using those principles, Bannert, Reimann, & Sonnenberg (2014) detected differences in frequencies of SRL events using PM techniques. In other research Mukala, Buijs, & Van Der Aalst (2015) used PM techniques in order to trace and analyze successful and unsuccessful student learning patterns based on MOOC data. In later research they also made use of alignment-based conformance checking to analyze students' learning patterns (Mukala, Buijs, Leemans, & Van der Aalst, 2015). Along similar lines, Emond and Buffett (2015) applied process discovery mining and sequence classification mining techniques to model and support SRL in heterogeneous learning environments. Finally, Vidal, Vázquez-Barreiros, Lama, & Mucientes (2016) used logs from a CBLE to extract the learning flow structure using PM, and to obtain the underlying rules that control students' adaptive learning by means of decision tree learning.

Based on current literature, the process discovery algorithm known as *Inductive Miner (IM)* has not been applied to educational datasets until now (Bogarín et al., 2018). In this paper, we propose the use of this algorithm for improving models previously obtained by EPM with other discovery algorithms. Different PM algorithms have been proposed, however no existing algorithm returns good quality metrics in all cases, while IM is being extensively used in business with very promising results (Leemans, Fahland, & van der Aalst, 2013). IM means an improvement over the Alpha and Heuristics miners that makes it easier to explore an event log; it is able to cope with infrequent behavior and large event logs, while ensuring soundness (Leemans, Fahland, & van der Aalst, 2014). It is also expected to produce more sound learning process models. Our objective is to compare the performance of this algorithm with previously used PM algorithms, and the ultimate goal is to be able to produce better process models about student behavior when using CBLEs. Below, we address the study method but describing before the preprocessing data process. Following the results we discuss EPM and its educational value.

Method

Participants

We used data from 101 undergraduate students (mean age=20.23; SD=1.01; female=83%) studying for a degree in psychology at a university in the North of Spain, who completed an online course using the corporate LMS Moodle 2.0.

Instruments

The log file provided by the LMS was the data collection instrument in this study. The data provided by Moodle contains all of each student's events recorded during their interactions with the

LMS, summarized in six attributes (see Table 1). It was necessary to preprocess and filter the Moodle log file; this is essential when we use real event logs and the data is often noisy (Romero, Ventura, & García, 2008).

We converted the students' names into IDs (Identifiers) to maintain their anonymity. Then, we deleted duplicate records, and instructor, system administrator and test user records. We used only four attributes (Time, Full Name, Action and Information) which was sufficient for our research purposes; the name of the course (the same for all records) and the IP address were not relevant. Then, we filtered some irrelevant actions in our log file. So, from the original 42 actions that Moodle stored by default, we selected the 16 actions that were relevant to the learning process and academic performance for this course (Cerezo, Sánchez-Santillán, Paule-Ruiz, & Núñez, 2016). In addition, we used high level coding (Fayyad, Piatetsky-Shapiro, & Smyth, 1996) with five action labels (Planning, Learning, Executing, Review and Forum Peer Learning) in order to produce more easily understandable models (see Table 2) in accordance with assumptions of SRL theory. Following that, we transformed the original Excel log file into the XES (eXtensible Event Stream) file which is required to implement process mining using the ProM framework.

Subsequently, we will consider the student as the "case" and the union between action and high level codification attributes as the "event classes" in only one attribute, for example: URL (Uniform Resource Locator) view-LEARNING, quiz view-PLANNING, and so on. In this way, each row in our preprocessed event logs is an event class (action and high level codification attributes), that is carried out by a case (student) on a specific date (timestamp). The traceability for each case will be the different event classes carried out by a student.

Additionally, we also used the students' final marks. This is a file containing each student's ID and final mark (a numerical value on a 10-point scale). We transformed this continuous value into a categorical value using traditional Spanish academic grading: from 0 to 4.9 is a fail and from 5 to 10 is a pass. Using their performance, we were able to group the students and label them Pass or Fail. Clustering by marks during preprocessing is useful for comparing the performance of different algorithms and for assessing the practical application and theoretical value of the resultant models. In this way we can divide each log file into three different files: All (containing events for all students on the course), Pass (containing only events of students who pass the course) and Fail (containing only events of students who fail the course).

Finally, we produced sub-files by unit in order to analyze student behavior more thoroughly. The course was made up of different units that can be thought of as lessons with different content but similar processes. For this reason we preprocessed the information attribute in each record in order to ascertain which unit it belonged to. Once the preprocessing was done the file was ready for EPM with ProM software (Romero, Cerezo, Bogarín, & Sánchez-Santillán, 2016). Table 3 shows the final number of cases and number of events in each unit after preprocessing.

Table 1
Attributes of a Moodle event log file

Attribute	Description
Course	The name of the course
IP Address	The IP of the device used to access Moodle
Time	The date they accessed Moodle
Full Name	The name of the student
Action	The action that student performed
Information	More information about the action

Table 2
Codification of the attribute actions

Low level Moodle Action	High Level Codification
assign submit	EXECUTING
assign view	PLANNING
forum add discussion	FORUM PEER LEARNING
forum add post	FORUM PEER LEARNING
forum update post	FORUM PEER LEARNING
forum view discussion	FORUM PEER LEARNING
forum view forum	FORUM PEER LEARNING
page view	LEARNING
quiz attempt	EXECUTING
quiz close attempt	EXECUTING
quiz continue attempt	EXECUTING
quiz review	REVIEW
quiz view	PLANNING
quiz view summary	PLANNING
resource view	LEARNING
URL view	LEARNING

Procedure

The experiment took the form of an assignment in the curriculum of a compulsory 3rd year subject completed entirely outside teaching hours. The course was made up of different units that were delivered to the students on a weekly basis during one semester. Students were asked to participate in an eTraining program about SRL and study strategies related to the subject topic (Cerezo, Núñez, Rosario, Valle, Rodríguez, & Bernardo, 2010). The instructor strongly suggested that students approached the assignments for each unit in the following order: understand the theoretical content, put them in practice through the corresponding task, share their experience about the week's topic in the forum; a learning path supported by SRL theory (Nuñez

Table 3
Number of cases and events per unit at the datasets

Units	Number of cases	Number of events
Unit 1	101	1782
Unit 2	101	2103
Unit 3	100	2192
Unit 4	101	2946
Unit 5	100	2514
Unit 6	101	1612
Unit 7	95	2067
Unit 8	87	1931
Unit 9	86	1699
Unit 10	87	1163
Unit 11	84	1620

et al., 2011). However, the students were free to follow their own learning path and the only compulsory assignments for each unit were to complete the weekly practical task and to post at least one comment in each unit forum.

Data analysis

Data analysis had three steps: log file preprocessing (previously described in the *Instruments* section), process discovery, and algorithm evaluation and interpretation (Figure 1).

In order to compare the discovered PM models we executed the most commonly used educational process discovery algorithms provided by the ProM framework (Van der Aalst, 2016): AM algorithm, HM Algorithm, ETM, and finally, the object of this study, *Inductive Miner*. To that end we compared some evaluation measures of the models obtained based on four quality forces (see Figure 2) that measure how well an educational process model describes the observed data:

- *Fitness* quantifies the extent to which the discovered model can accurately reproduce the cases recorded in the log.
- *Precision* shows the proportion of the behavior represented by the model which is not seen in the event log.
- *Generalization* assesses the extent to which the model will be able to reproduce future behavior of the process and can be seen as a measure of confidence in the precision.
- *Simplicity* captures the complexity of a process model in terms of readability.

All indexes are important for process discovery. However, it only makes sense to consider precision, generalization and simplicity if fitness is acceptable (Buijs et al., 2012; Van der Aalst, 2016). Existing process discovery algorithms typically consider, at most, two out of the four main quality dimensions because these four quality forces pull in different directions and whenever one is optimized, quality is usually lost in other measures. In light of this, we used a new *overall* measure proposed by Buijs et al., in 2012, to balance these four measures together, allocating them different weights (see Figure 2) (*Fitness*: weight 10; *Precision*: weight 5; *Generalization*: weight 1; *Simplicity*: weight 1).

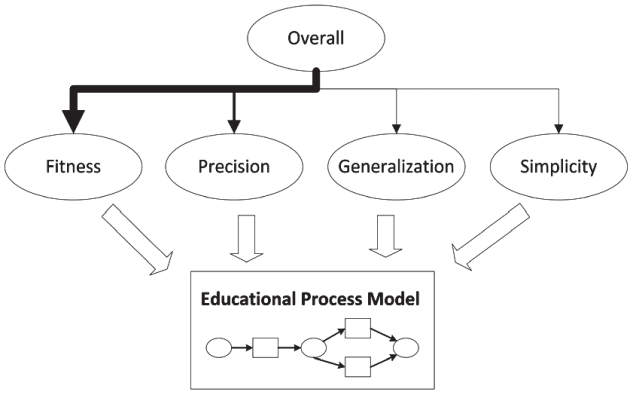

Figure 2. Model quality metrics

Results

Table 4 shows the results of the four algorithms in the *overall* evaluation metric. The IM algorithm scores the highest values in every unit, followed by ETM, then HM and AM depending on the sub-file.

In the *overall* metric, the IM algorithm scored highest, and the same is true if we consider each quality metric separately. Table 5 shows the performance of every algorithm in sub-file 4, where the students showed the most interaction with the LMS resulting in a

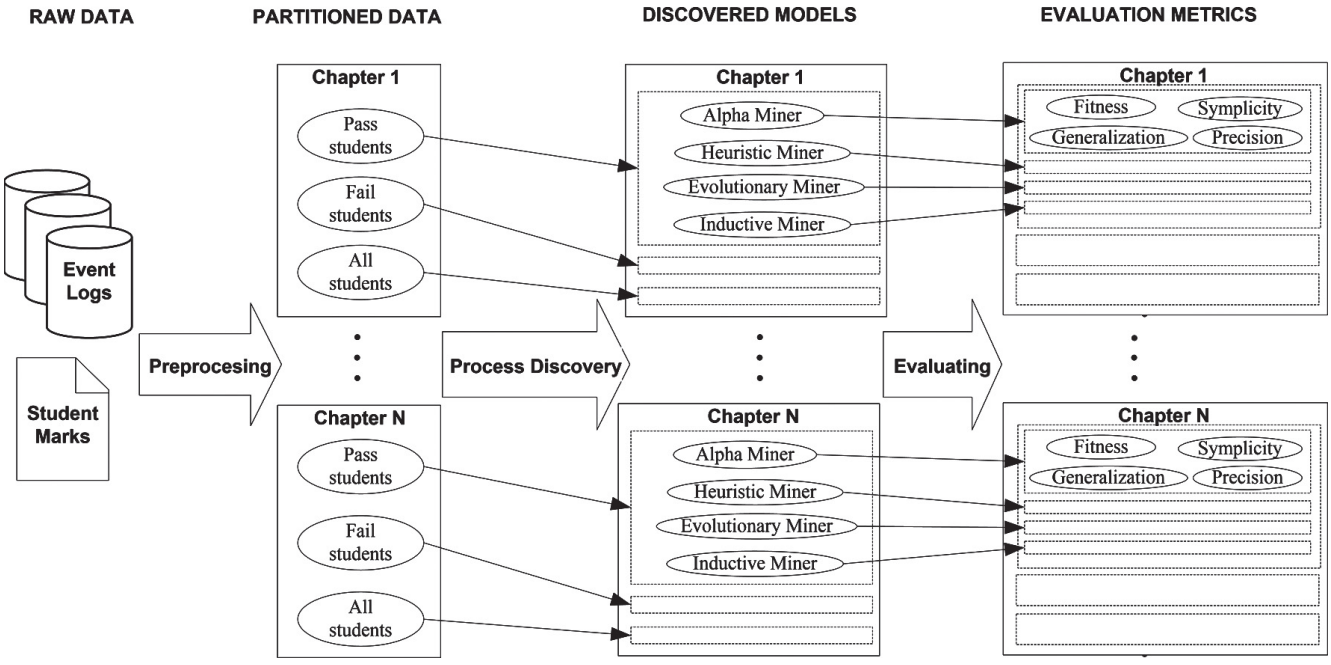

Figure 1. Procedure followed for carrying out EPM

higher number of cases and events, and subsequently complexity of modeling extraction. The IM algorithm scored the highest values in *fitness*, the metric which is fundamental for considering the other quality indexes, and also in *generalization*. It also scored the highest in *simplicity*, *along with* ETM, which indicates that the obtained models are easy to interpret and not *spaghetti-like models*. Nevertheless, it did not achieve the best score for *precision*. The highest scoring models in this metric were ETM and Heuristic Miner, although ETM scored better on *generalization*. The same table also shows the effect of clustering the data, with the quality indexes improving when clustering before, as expected.

Table 4
Comparison of algorithms based on the *overall* metric

Units	AM	HM	ETM	IM
Unit 1	0.676	0.666	0.793	**0.797**
Unit 2	0.666	0.618	0.752	**0.781**
Unit 3	0.583	0.493	0.675	**0.712**
Unit 4	0.452	0.597	0.715	**0.747**
Unit 5	0.582	0.533	0.649	**0.659**
Unit 6	0.577	0.621	0.742	**0.793**
Unit 7	0.612	0.664	0.724	**0.773**
Unit 8	0.724	0.732	0.750	**0.796**
Unit 9	0.516	0.510	0.744	**0.784**
Unit 10	0.685	0.700	0.827	**0.856**
Unit 11	0.553	0.563	0.735	**0.778**

Along with quality metrics, Figures 3 and 4 show two of the resultant visualization of the models for sub-file 4. In order to understand and interpret the IM-generated models it is necessary to understand what each element means: the boxes are the activities carried out by the students, the number in the box is the frequency, the arrows indicate the direction of the process, and the number above the arrows is the frequency of the transition between these two actions. Each model begins with an initial node and ends with a final node.

Looking at the learning path followed by students in the fail cluster (see Figure 3), the first activity they do is the *quiz attempt*, followed by the *quiz view summary* and the *quiz view*. In other words, they start doing activities related to the quiz, which are one of the two compulsory course assignments. In the middle part of the model they do forum-related activities such as *forum view forum* and *forum add post*, which are the other compulsory activities. Following that there are parallel actives *-quiz review, quiz continue attempt-*, finishing with *page view* and *URL view* which would have been the logical starting point for the learning path suggested by the instructor.

Students in the pass cluster (Figure 4) started their study process by visiting the *forum view discussion*, after which the model splits into different possible routes. One route continues

Table 5
Comparison of algorithms based on *fitness, precision, generalization, simplicity*, and *overall* in sub-file 4

Algorithm	Cluster	*Fitness*	*Precision*	*Generalization*	*Simplicity*	*Overall*
Alpha Miner	Fail	0.765	0.197	0.422	0.636	0.570
Heuristic Miner	Fail	0.491	0.521	0.487	0.653	0.509
ET Miner	Fail	0.684	0.709	0.873	0.913	0.716
Inductive Miner	Fail	0.96	0.322	0.957	0.882	**0.768**
Alpha Miner	Pass	0.863	0.164	0.464	0.666	0.622
Heuristic Miner	Pass	0.526	0.707	0.603	0.732	0.596
ET Miner	Pass	0.712	0.691	0.841	0.901	0.725
Inductive Miner	Pass	0.959	0.315	0.962	0.882	**0.765**
Alpha Miner	All	0.581	0.198	0.414	0.466	0.452
Heuristic Miner	All	0.472	0.868	0.483	0.611	0.597
ET Miner	All	0.693	0.715	0.719	0.923	0.715
Inductive Miner	All	0.87	0.443	0.867	0.909	**0.747**

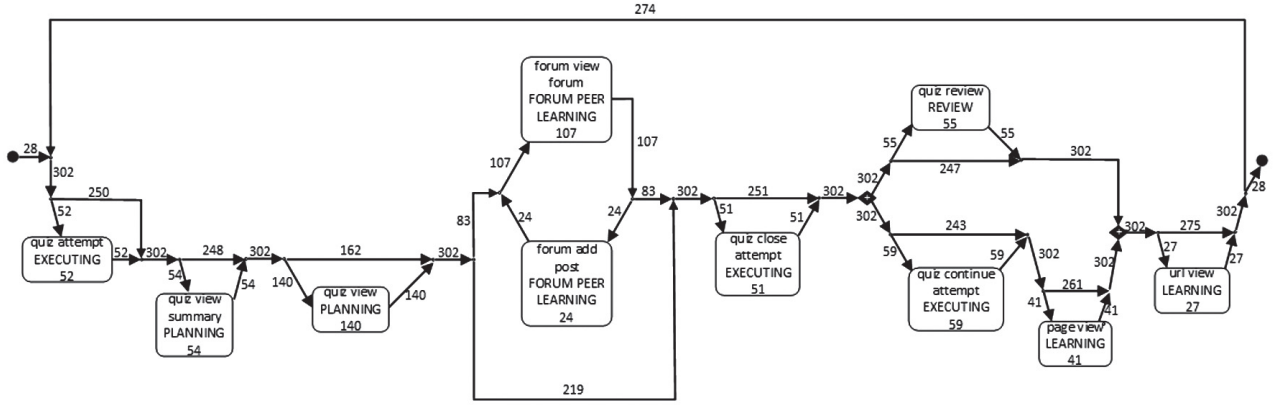

Figure 3. Visualization of failing students' learning path in sub-file 4

via the *URL view* action, the second route involves continuing the study process with forum-related activities *-forum view forum, forum add post* and *forum update post-*. There is also a third route, in which student do actions related to the quizzes *quiz attempt, quiz view summary* and *quiz continue attempt*. The general model finishes with the *quiz close attempt* and *quiz review* actions.

Discussion

This research focuses on the analysis of learning processes based on EPM. We applied PM techniques to educational data in order to discover learning processes, compare algorithm performance, and extract educational implications to guide future work. However, PM algorithms cannot be directly applied to educational problems, preprocessing is necessary first and only then can the mining methods be applied to the problems (Dutt et al., 2017). Therefore, we also described the preprocessing required before discovery could take place.

We proposed the application of the IM algorithm as a new way to discover learning processes in LMSs; an extensive literature review suggested that this is the first study to apply IM to educational data (Bogarin et al., 2018). Based on our results, we can draw three important conclusions. Firstly, the IM algorithm produces the best fitness. This is significant because none of the other quality indexes should be considered in isolation; it only makes sense to consider them all together if the fitness is acceptable (Buijs et al., 2012; van Dongen, 2007). Secondly, the results show that the balance of quality forces (*overall*) are better in IM than in the other contemporary PM algorithms. And thirdly, both metrics, taken together or individually, are even better when we apply clustering to improve subsequent mining, as previously seen with educational (Bogarin et al., 2014; Bogarin et al., 2018) and business data (Bose, & van der Aalst, 2009). It seems that, applying the IM algorithm to discovering learning models opens a new field in the research, development, and understanding of PM applied to educational issues.

Process discovery is one of the most challenging process mining tasks; starting from a simple log, a process model is constructed capturing the behavior seen in the log (Van der Aalst, 2011). However, apart from quality compliance, the resultant model needs to be able to reproduce the behavior seen in the log file in an understandable educational process, giving EPM a practical meaning rather than just ideas and theories.

With that in mind, we selected and interpreted two of the most challenging discovered models. If, in addition to the raw actions in the failing cluster, we look at the high level coding, the models lead us to conclude that the students who failed did not follow the learning path suggested by the instructor and promoted by SRL theories leading to quality learning results. Based on the assumptions of SRL (Zimmerman, 1990), starting executing before planning or learning leads to low quality learning or failure, as seen in this study.

If we look at the high-level coding of those students who passed, we can see that although they did not follow the instructors' suggestions exactly, they did follow the logic of a successful learning process. These results are in line with those from Lust, Elen, and Clarebout (2013a, 2013b), who found that only a minority of students regulated their behavior in line with course requirements. Passing students started with actions indicating comprehension and learning of the materials, three different routes can then be seen: two task oriented groups, one socially focused giving a leading role to collaborative learning in the forums, and another more individually focused; and finally, a non-task or learning oriented group. These different learning profiles are in accordance with data previously obtained by Cerezo et al, 2016 also using LMS interaction data. All three routes concluded with the executing and reviewing actions suggested by the instructor and SRL rationale, leading to successful achievement in varying degrees.

The PM models also allow us to examine which specific actions the students performed. It is interesting to see the actions related to forum-supported collaborative learning. Students in

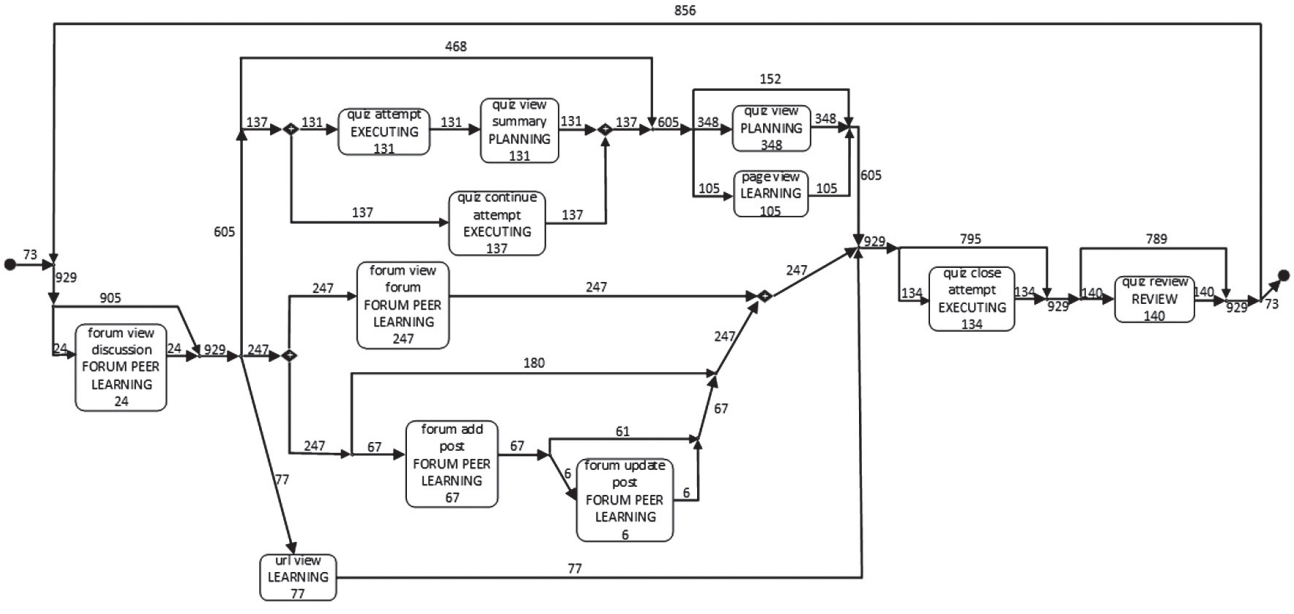

Figure 4. Visualization of passing students' learning path in sub-file 4

the pass cluster performed actions such as *forum update post* and *forum view discussion*, which do not appear in the model from the fail cluster. This is very valuable since forum behavior has been previously related to student achievement in LMSs (Romero, López, Luna & Ventura, 2013).

It should be also noted that IM models are able to discover meaningful learning processes similar to those previously obtained using alternative algorithms (Mukala, Buijs, & Van Der Aalst, 2015; Mukala, Buijs, et al., 2015). However, in this case the instructor can visualize and interpret the behavior model of students' learning paths thanks to their simplicity. Process discovery algorithms often result in spaghetti-like process models (Van der Aalst, 2011), which are very hard to read. However, IM strongly focuses on simplicity and generally results in simple models (Buijs et el., 2012).

In conclusion, learning model discovery with IM could be a promising resource for preventing learning failure in LMSs. With insight into at-risk students' distance-learning progress, we can strategically design preventive interventions based on Adaptive Hypermedia Learning Environments (Brusilovsky, & Millán, 2007) or early detection and remedial actions through real time modeling. PM is not restricted to the past, but also relevant to the present (recommendation and real-time conformance checking), and the future (prediction) (Van der Aalst, Schonenberg, & Song, 2011). In a wider sense, the scope in academic contexts is also extensive, from allow universities to invest in those resources which are shown to be most useful for preventing school drop-out (Areces, Rodríguez Muñiz, Suárez Álvarez, de la Roca, & Cueli, 2016) to the contribution of social networks to learning (Sanmamed, Carril, & Alvarez de Sotomayor, 2017).

Finally, in order to generalize the good performance of IM with educational data, it would be interesting to test the algorithm in different CBLEs, such as alternative LMSs or the emerging MOOCs. Modeling learning process in MOOCs would be a very challenging prospect in terms of simplicity and readability.

Acknowledgements

Authors gratefully acknowledge the financial subsidy provided by Spanish Ministry of Science and Technology TIN2017-83445-P and EDU2014-57571-P. We have also received funds from the European Union and the Principality of Asturias, through its Science, Technology and Innovation Plan (grant GRUPIN14-053).

References

Areces, D., Rodríguez Muñiz, L. J., Suárez Álvarez, J., de la Roca, Y., & Cueli, M. (2016). Information sources used by high school students in the college degree choice. *Psicothema*, *28*(3), 253-259. doi: 10.7334/psicothema2016.76

Bannert, M., Reimann, P., & Sonnenberg, C. (2014). Process mining techniques for analysing patterns and strategies in students' self-regulated learning. In: *Metacognition and learning*, *9*(2), 161-185. doi:10.1007/s11409-013-9107-6

Bogarín, A., Romero, C., Cerezo, R., & Sánchez-Santillán, M. (2014). Clustering for improving educational process mining. In M. Pistilli, J. Willis, & D. Koch (Eds.), *Proceedings of the Fourth International Conference on Learning Analytics And Knowledge* (pp. 170-181). Indianapolis, USA: ACM. doi:10.1145/2567574.2567604

Bogarín, A., Cerezo, R., & Romero, C. (2018). A survey on educational process mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *8*(1). doi:10.1002/widm.1230

Bose, R. J. C., & van der Aalst, W. M. (2009, September). Trace clustering based on conserved patterns: Towards achieving better process models. In U. Dayal, J. Eder, J. Koehler & H. Reijers (Eds.), *Proceedings of the International Conference on Business Process Management* (pp. 170-181). Berlin, Heidelberg: Springer.

Broadbent, J., & Poon, W. L. (2015). Self-regulated learning strategies & academic achievement in online higher education learning environments: A systematic review. *The Internet and Higher Education*, *27*, 1-13. doi:10.1016/j.iheduc.2015.04.007

Buijs, J. C., Van Dongen, B. F., & van Der Aalst, W. M. (2012). On the role of fitness, precision, generalization and simplicity in process discovery. In R. Meersman, H. Panetto, T. Dillon, S. Rinderle-Ma, P, Dadam, X. Zhou, S. Pearson, A. Ferscha, S. Bergamaschi, & I. F. Cruz, *Proceedings of the OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"* (pp. 305-322). Berlin: Springer. doi:10.1007/978-3-642-33606-5_19

Brusilovsky, P., & Millán, E. (2007). User models for adaptive hypermedia and adaptive educational systems. In P. Brusilovski, A. Kobsa & W. Nejdl (Eds.), *The adaptive web* (pp. 3-53). Berlin: Springer.

Cerezo, R., Sánchez-Santillán, M., Paule-Ruiz, M. P., & Núñez, J. C. (2016). Students' LMS interaction patterns and their relationship with achievement: A case study in higher education. *Computers & Education*, *96*, 42-54. doi:10.1016/j.compedu.2016.02.006

Cerezo, R., Núñez, J. C., Rosario, P., Valle, A., Rodríguez, S., & Bernardo, A. (2010). New Media for the promotion of self-regulated learning in higher education. Psicothema, *22*(2), 306-315.

Dahlstrom, E., Brooks, D. C., & Bichsel, J. (2014). *The current ecosystem of learning management systems in higher education: Student, faculty, and IT perspectives* (Research report) Retrieved from http://www. educause. edu/ecar. 2014 EDUCAUSE. CC by-nc-nd

Dutt, A., Ismail, M. A., & Herawan, T. (2017). A systematic review on educational data mining. *IEEE Access*, *5*, 15991-16005. doi:10.1109/ACCESS.2017.2654247

Emond, B., & Buffett, S. (2015, June). *Analyzing Student Inquiry Data Using Process Discovery and Sequence Classification*. Paper presented at the International Educational Data Mining Society, Madrid, Spain.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, *39*(11), 27-34. doi: 10.1145/240455.240464

Leemans, S. J., Fahland, D., & van der Aalst, W. M. (2013, August). *Discovering block-structured process models from event logs containing infrequent behaviour*. Paper presented at the International Conference on Business Process Management, Beijing, China.

Leemans, S. J., Fahland, D., & van der Aalst, W. M. (2014). Process and Deviation Exploration with Inductive Visual Miner. *BPM (Demos)*, *1295*, 46.

Lust, G., Elen, J., & Clarebout, G. (2013a). Regulation of tool-use within a blended course: student differences and performance effects. *Computers & Education*, *60*(1), 385-395.

Lust, G., Elen, J., & Clarebout, G. (2013b, August). *Measuring students' strategy-use within a CMS supported course through students' tool-use patterns*. Paper presented at the 15th biennial conference EARLI 2013, Munich, Germany.

Mukala, P., Buijs, J. C. A. M., & Van Der Aalst, W. M. P. (2015). *Uncovering learning patterns in a MOOC through conformance alignments* (Research report). Retrieved from http://bpmcenter.org/wp-content/uploads/reports/2015/BPM-15-09.pdf

Mukala, P., Buijs, J. C., Leemans, M., & van der Aalst, W. M. (2015, December). *Learning Analytics on Coursera Event Data: A Process Mining Approach*. Paper presented at the SIMPDA, Viena, Austria.

Papamitsiou, Z., & Economides, A. A. (2014). Learning analytics and educational data mining in practice: A systematic literature review

of empirical evidence. *Journal of Educational Technology & Society, 17*(4), 49.

Pechenizkiy, M., Trcka, N., Vasilyeva, E., van Aalst, W., & De Bra, P. (2009, July). *Process mining online assessment data*. In *Educational Data Mining*. Paper presented at the International Conference on Educational Data Mining, Córdoba, Spain.

Peña-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works. *Expert systems with applications*, *41*(4), 1432-1462. doi:10.1016/j.eswa.2013.08.042

Reimann, P., Markauskaite, L., & Bannert, M. (2014). E-Research and learning theory: What do sequence and process mining methods contribute? *British Journal of Educational Technology*, *45*(3), 528-540. doi:10.1111/bjet.12146

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert systems with applications*, *33*(1), 135-146. doi:10.1016/j.eswa.2006.04.005

Romero, C., & Ventura, S. (2010). Educational data mining: a review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, 40*(6), 601-618. doi:10.1109/TSMCC.2010.2053532

Romero, C., & Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *3*(1), 12-27.

Romero, C., Ventura, S., & García, E. (2008). Data mining in course management systems: Moodle case study and tutorial. *Computers & Education*, *51*(1), 368-384. doi: 10.1016/j.compedu.2007.05.016

Romero, C., Cerezo, R., Bogarín, A., & Sánchez-Santillán, M. (2016). Educational process mining: a tutorial and case study using Moodle data sets. In *Data Mining and Learning Analytics: Applications in Educational Research* (pp. 1-28). Wiley & Blackwell. doi:10.1002/9781118998205.ch1

Romero, C., López, M. I., Luna, J. M., & Ventura, S. (2013). Predicting students' final performance from participation in on-line discussion forums. *Computers & Education, 68,* 458-472.

Sanmamed, M. G., Carril, P. C. M., & Álvarez De Sotomayor, I. D. (2017). Factors which motivate the use of social networks by students. *Psicothema, 29*(2), 204-210. doi: 10.7334/psicothema2016.127

Trcka, N., & Pechenizkiy, M. (2009). From local patterns to global models: Towards domain driven educational process mining. In *Proceedings of the Ninth International Conference on Intelligent Systems Design and Applications* (pp. 1114-1119). New Jersey: The Institute of Electrical and Electronics Engineers. doi:10.1109/ISDA.2009.159

Trcka, N., Pechenizkiy, M., & van der Aalst, W. (2010). Process mining from educational data. In C. Romero, S. Ventura, M. Pechenizkiy & R. Baker (Eds.), *Handbook of educational data mining* (pp. 123-142). Florida: Taylor & Francis.

van der Aalst, W. M. (2011). Process Discovery: An Introduction. In *Process Mining* (pp. 125-156). Berlin, Heidelberg: Springer. doi: 10.1007/978-3-642-19345-3_5

van der Aalst, W. M. (2016). *Process mining: data science in action*. Berlin, Heidelberg: Springer. doi:10.1007/978-3-662-49851-4

van der Aalst, W. M., Schonenberg, M. H., & Song, M. (2011). Time prediction based on process mining. *Information systems*, *36*(2), 450-475.

van Dongen, B. F. (2007). Process mining and verification. *Dissertation Abstracts International*, *68*(4).

Vidal, J. C., Vázquez-Barreiros, B., Lama, M., & Mucientes, M. (2016). Recompiling learning processes from event logs. *Knowledge-Based Systems, 100,* 160-174. doi:10.1016/j.knosys.2016.03.003

Weijters, A.J.M.M., van Der Aalst, W.M., & De Medeiros, A.A. (2006). Process mining with the heuristics miner-algorithm. *Technische Universiteit Eindhoven Technology Reports, 166*, 1-34.

Zimmerman, B. J. (1990). Self-regulated learning and academic achievement: An overview. *Educational psychologist, 25*(1), 3-17.