Analyzing student inquiry data using process discovery and sequence classification

Bruno Emond
National Research Council Canada
bruno.emond@nrc.gc.ca

Scott Buffett
National Research Council Canada
scott.buffett@nrc.gc.ca

ABSTRACT

This paper reports on results of applying process discovery mining and sequence classification mining techniques to a data set of semi-structured learning activities. The main research objective is to advance educational data mining to model and support self-regulated learning in heterogeneous environments of learning content, activities, and social networks. As an example of our current research efforts, we applied temporal data mining analysis techniques to a PSLC DataShop data set [17, 18, 19, 20]. First, we show that process mining techniques allow for discovery of learning processes from student behaviours. Second, sequential pattern mining is used to classify students according to skill. Our results show that considering sequences of activities as opposed to single events improved classification by up to 230%.

1. INTRODUCTION

The Learning Performance Support Systems program (LPSS) at the National Research Council Canada aims at delivering a personal learning environment (LPSS.me), software algorithms, and prototypes to enable Canada's training and development sector to offer learning solutions to industry partners that will address their immediate and long-term skills challenges. The main elements of the personal learning environment include a common platform architecture, a personal learning assistant, a personal cloud, learning resources repository network, personal learning records, and analytics to discover and assess competencies. The program is at an early stage of development.

One of the main thrusts within this research program seeks to advance and apply educational data mining to model and support self-regulated learning in heterogeneous environments of learning content, activities, and social networks. Our initial position points towards a complementary use of latent knowledge estimation and performance prediction methods [3], and temporal data mining methods. A main research trend in educational data mining consists of analyzing students' performance within intelligent tutoring systems, focusing on the correctness of previous questions or the number of hints and attempts students needed in order to predict their future performance [6]. Predictive mathematical models resulting from this analysis characterize, through parameter values, some information contained in the sequence of actions leading to student performances, but do not represent explicitly those sequences. Over the years there has been a growing interest to examine explicitly learning sequences as a complementary approach. Process and sequence mining have been applied for the analysis of content sequencing and curriculum sequencing [5, 15], group behaviour sequences in collaborative software development tasks [16], problem solving behaviours over a shared tabletop [14], as well as self-regulated learning and meta-cognition [7].

The remainder of this paper consists of a short presentation of temporal data mining, followed by process mining and sequence mining analyses of a semi-structured inquiry learning activity data set [17, 18, 19] obtained from the Pittsburgh Centre for Science and Learning DataShop [8]. We show that process mining techniques allow for the discovery of learning processes, and that sequential pattern mining can used to identify the level of skill exhibited by each student.

2. TEMPORAL DATA MINING

Temporal data mining refers to the extraction of information and knowledge from potentially large collections of temporal or sequential data [12]. According to Laxman and Sastry [9], sequential data refers to any type of data where data points are explicitly ordered, either by time stamps or some other sequencing mechanism. This includes data such as moves in a chess game or commands entered by a computer user, but also other forms of data that are not explicitly time-stamped but are still otherwise ordered, such as text or protein sequences.

Temporal data is often divided into two categories: sequences that consist of continuous, real-valued data points taken at regular intervals, which are referred to as *time series data*, and sequences that may be represented by compositions of nominal symbols from a particular alphabet, which are referred to as *temporal sequences* [2]. As the field of time series analysis has a long history with many established techniques, the more recent field of temporal data mining instead focuses on information extraction from temporal sequences.

Given a set of temporal sequences, the general tasks of tem-

poral data mining consist of 1) prediction, 2) classification, 3) clustering, 4) search and retrieval, and 5) pattern discovery. These tasks can be accomplished using a number of established techniques in the area. A few of the more prevalent techniques include: A) *Sequential pattern mining*: The goal of sequential pattern mining [1] is to identify highly frequent sequences that appear within a database of ordered items or events; B) *Sequence classification*: Sequence classification [11] attempts to assign a candidate sequence to one of possibly several classes of existing sequences, typically according to either similarity or common features such as frequent sub-sequences; C) *Episode mining*: Frequent episodes [13] are sets of partially ordered events that are found to occur close together frequently and consistent with the specified partial order; and D) *Process mining*: Process mining refers to the extraction of process-related information from event logs [21]. Process mining algorithms are used to build a model of the business process by representing the different ways cases in the process can be executed. However, there are some key differences between business processes and learn flows [4].

3. TEMPORAL EDM ANALYSIS

To demonstrate the potential of temporal data mining in the analysis of educational data, we conducted a study utilizing process mining and sequential pattern mining to discover learning processes and to identify the level of student skill using a data set [17, 18, 19] taken from the Pittsburgh Science of Learning Center DataShop [8]. This data set contains data on 148 middle school students performing activities logged while working within a micro-world, where students engage in "scientific inquiry" to study liquid phase change. Here, the students form hypotheses and conduct experiments as they investigate whether container size, heat level, substance amount, and cover status affected the boiling/freezing point of water, or the time it took to freeze/boil. All students' fine-grained actions were attributed a time stamp and recorded by the system. These actions included: interactions with the inquiry support widgets, interactions with the simulation including changing simulation variable values and running/pausing/resetting the simulation, and transitioning between inquiry tasks [18].

Given that we are mostly interested in the discovery of self-regulated learning, the fact that students had a moderate degree of freedom to choose their own procedures for conducting experiments, less than in purely exploratory learning environments though [19], was an interesting data set for studying sequences of student behaviours and how they correlate with student success.

3.1 Process Mining and Discovery

Process mining offers a set of techniques and tools to discover sequential patterns represented as workflows. The analysis in this section was performed using the *Inductive visual miner* [10]. We were interested to discover, from the log of students inquiry activities, similar process models to the one depicted in Figure 1. For this discovery analysis, we limited ourselves to the whole data set, and we did not try to distinguish between groups of students. The purpose was to explore and compare the actual processes that students followed to the expected process from the author of the learning environment given in Figure 1, rather than suggest alternative learning processes. The log file contained 29679 events for 147 students. The overall distribution of inquiry activities indicated that 58.1% were spent in analysis, 19.1% in experiment, 18.4% in hypothesis formation, and 4.4% in observation.

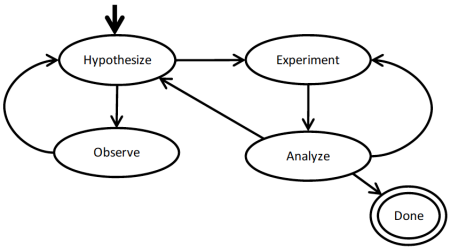

Figure 1: Intended learning paths during scientific inquiry.

As indicated in Figure 1, the intended learning process contains many possible loops while students progress in their scientific inquiry. Figure 2 and Figure 3 show respectively discovered process models from the transactions log using 100% of the events and sequences, and the top 70% most frequent events and sequences. From the visual comparison of the process model for 100% of the data (Figure 2), and the intended process of Figure 1, it is clear that there is a lot of variability in students transitioning between inquiry steps, given that the model is mostly disjunctive, with sequences resulting from loops. However, after leaving out the 30% most infrequent events and event sequences from the data, we discover a process model, Figure 3, that has some resemblance to the intended inquiry process, representing explicitly the sequence of hypothesize to experiment or analyze. Notice that the observation inquiry step is not part of the model because of the low frequency of its related events, which indicates a difference with the intended learning process, or more accurately, a tendency by the students to avoid the observation stage.

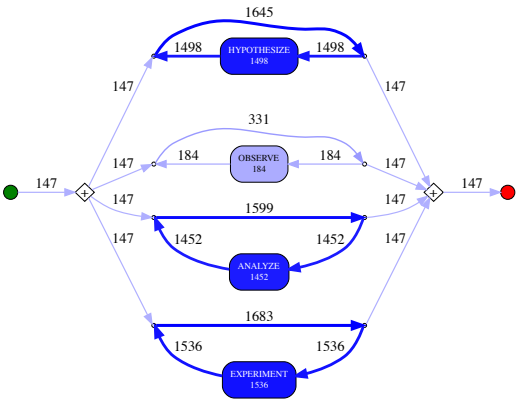

Figure 2: Process model using 100% of events and sequences (from top to bottom: hypothesize, observe, analyse, experiment).

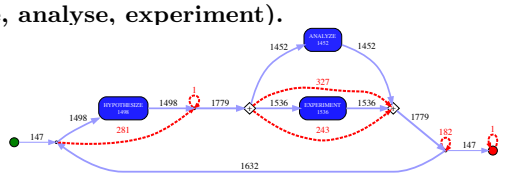

Figure 3: Process model using 70% of most frequent events and sequences (from left to right: hypothesize, analyse (top), experiment (bottom)).

Another element of interest was the sequence of problems students address during their inquiry. The overall distribution of student activities within those problems were relatively balanced with 30.7% in "container size", 24.9% in "amount of substance", 23.0% in "level of heat", and 21.4% in "cover status". Figure 4 shows a process model including 100% of events and event sequences. The process model clearly indicates a bias towards starting from the container size problem, followed by equivalent choices from the three other problems. This is likely a consequence of the the container size being the default value at the start of the inquiry session, which is a restriction on the student self-regulated learning processes.

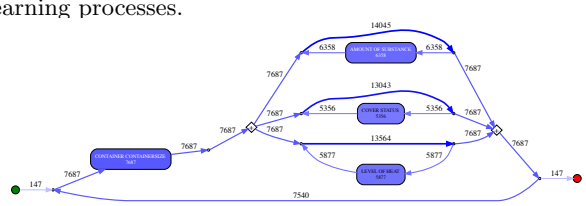

Figure 4: Process model of problems sequence using 100% of events and event sequences (from left to right: container size, amount of substance (top), cover status (middle), level of heat (bottom)).

Interestingly though, one would expect that the inquiry steps would be grouped (follow each other closely) within each problem. An inspection of a process model for an event classifier including the combination of both inquiry steps (hypothesize, observe, experiment, analyze) and problems (container size, amount of substance,level of heat, cover status) with 100% of events and sequences reveals only three groups of steps and not four as one would expect. In Figure 5, 1) the leftmost group is focused on inquiry steps applied to container size, and amount of substance, 2) the middle group to level of heat, amount of substance, and cover status, and 3) the rightmost group to cover status. This distribution of steps indicates that the four problems were not explored completely independently by the students, which manifest a strategy to explore concurrently the effect of different factors. However, this strategy might be different when comparing students with good and poor results and should be explored in a subsequent analysis.

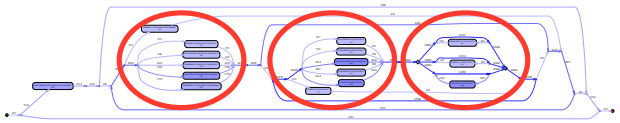

Figure 5: Three groups of problems and inquiry steps combination sequences.

3.2 Sequence classification

The second phase of our study was to explore the potential of sequential pattern mining in the identification of the level of skill exhibited by each student. Since sequences of student activity in the data set were not explicitly labelled as "skilled", "unskilled", etc., we considered two other metrics to measure skill exhibited: 1) number of times the student got an answer wrong, and 2) total time taken to complete the experiments. We used leave-one-out cross validation, applying our sequence classification learning algorithms on the training set and attempting to classify each test student as having either the high/low number of incorrect answers, or high/low time to complete, depending on the test.

Figure 6 shows the results of classifying students as "high number of incorrect steps". Success of the classifiers are measured by likelihood ratio (LR), which indicates how much more likely a positive example will be classified as positive than a negative example. The left-hand chart shows the success in classifying whether a student is in the bottom 50% in terms of number of incorrect answers, for varying maximum sequence size. Thus, a maximum sequence size of 1 represents the case where sequential relations are not considered, and only the presence/absence of certain actions are used for the classification. Observe that the LR is close to 1 in this case, meaning that we are no more likely to classify a positive case as positive or negative. The LR then increases steeply by 230% to 2.3 as sequences of size 2 are considered, before levelling off at about 1.75 for size 3 and greater. The right-hand chart then demonstrates how the classifier improves as we use sequences (max size 4) to classify students into the categories of worst 50%, 40%, 30%, 20% and 10%. Figure 7 depicts the results similarly for classifying students as "long time to complete". While not as dramatic, the positive effect of utilizing sequential information is demonstrated here as well.

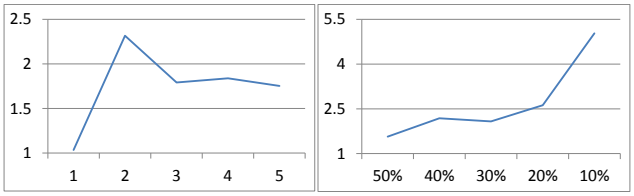

Figure 6: LR for classifying as "high number of incorrect steps".

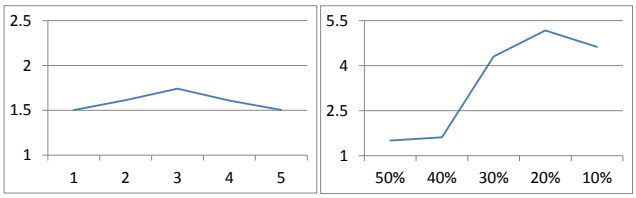

Figure 7: LR for classifying as "long time to complete".

4. CONCLUSION

One of the main thrusts within the Learning Performance Support Systems program (LPSS) at the National Research Council Canada seeks to advance and apply educational data mining to model and support self-regulated learning in heterogeneous environments of learning content, activities, and social networks. The program is at an early stage of development and our initial position points towards a complementary use of latent knowledge estimation and performance prediction methods [3], and sequence mining methods. In order to support the validity of our argument that sequential data analytics holds great potential for the analysis of student knowledge and skill acquisition, we demonstrated the application of discovery process mining and sequence mining in classifying students according to success using a data set of semi-structured learning activities [17, 18, 19] taken from the Pittsburgh Science of Learning Center DataShop [8].

Using process mining tools we were able to discover in-

quiry learning patterns in relationships with inquiry learning steps, learning problems, and a combination of those. Our analysis showed some differences between the semi-structured process intended by the developers of the learning environment and the actual processes followed by the students. We also showed that process mining techniques allow for the discovery of learning processes, and that considering sequences of events as features we can improve classification by up to 230% over considering single, non-sequential events. Given the learning process patterns discovered in the initial analysis of the students inquiry activity log, the next process mining discovery analysis will be to compare the inquiry processes of students having low and high correct outcomes.

5. ACKNOWLEDGEMENT

We would like to thank the Pittsburgh Science of Learning Center for providing the data supporting this analysis. We used the 'Science Sim State Change January 2010' data set accessed via the PSLC DataShop [8]. We thank Ken Koedinger from Carnegie Mellon for his help in choosing this data set. This work is part of the National Research Council Canada program Learning and Performance Support Systems (LPSS), which addresses training, development and performance support in all industry sectors, including education, oil and gas, policing, military and medical devices.

6. REFERENCES

[1] R. Agrawal and R. Srikant. Mining sequential patterns. In *Proc. of the 11th Int'l Conference on Data Engineering*, pages 3–14. IEEE, 1995.

[2] C. M Antunes and A. L. Oliveira. Temporal data mining: An overview. In *KDD workshop on temporal data mining*, pages 1–13, 2001.

[3] R. S. Baker and A. T. Corbett. Assessment of robust learning with educational data mining. *Research and Practice in Assessment*, 9:38–50, 2014.

[4] R. Bergenthum, J. Desel, A. Harrer, and S. Mauser. Modeling and mining of learnflows. In K. Jensen, S. Donatelli, and J. Kleijn, editors, *LNCS Transactions on Petri Nets and Other Models of Concurrency*. Springer, Springer-Verlag : Berlin Heidelberg, 2012.

[5] M. C. Desmarais and R. S. Baker. A review of recent advances in learner and skill modeling in intelligent learning environments. *User Modelling and User-adapted Interaction*, 22:9–38, 2012.

[6] H. Duong, L. Zhu, Y. Wang, and N. T. Heffernan. A prediction model that uses the sequence of attempts and hints to better predict knowledge: "better to attempt the problem first, rather than ask for a hint". In *Proceedings of the 6th International Conference on Educational Data Mining, Memphis, Tennessee, USA, July 6-9, 2013* [6], pages 316–317.

[7] J. S. Kinnebrew, K.M. Loretz, and G. Biswas. A contextualized, differential sequence mining method to derive students' learning behavior patterns. *Journal of Educational Data Mining*, 5:190–219, 2009.

[8] K.R. Koedinger, R.S.J.d. Baker, K. Cunningham, A. Skogsholm, B. Leber, and J. Stamper. *A Data Repository for the EDM community: The PSLC DataShop. In Romero, C., Ventura, S., Pechenizkiy, M., Baker, R.S.J.d. (Eds.) Handbook of Educational Data Mining*. CRC Press, Boca Raton, FL, 2010.

[9] S. Laxman and P. Sastry. A survey of temporal data mining. *Sadhana*, 31(2):173–198, 2006.

[10] S. J. J. Leemans, D. Fahland, and W. M. P. van der Aalst. Process and deviation exploration with inductive visual miner. In *In Twelve International Conf. on Business Process Management, Accepted Demonstration 46*, Eindhoven, Netherlands, 2014.

[11] N. Lesh, M. J. Zaki, and M. Ogihara. Mining features for sequence classification. In *Proc. of the fifth ACM SIGKDD international conf. on Knowledge discovery and data mining*, pages 342–346. ACM, 1999.

[12] N. Mamoulis. Temporal data mining. In Ling Liu and M Tamer Özsu, editors, *Encyclopedia of Database Systems*. Springer, New York, NY, 2009.

[13] H. Mannila, H. Toivonen, and A. I. Verkamo. Discovering frequent episodes in sequences extended abstract. In *Proceedings the first Conference on Knowledge Discovery and Data Mining*, pages 210–215, 1995.

[14] R. Martinez, K. Yacef, J. Kay, A. Al-Qaraghuli, and A. Kharrufa. Analysing frequent sequential patterns of collaborative learning activity around an interactive tabletop. In *Proc. of the Fourth Int'l Conference on Educational Data Mining*, Eindhoven, Netherlands, 2011.

[15] M. Pechenizkiy, N. Trcka, P. De Bra, and P Toledo. Currim: Curriculum mining. In *Proceedings of the 4th International Conference on Educational Data Mining*, pages 216–217, 2012.

[16] D. Perera, J. Kay, I. Koprinska, K. Yasef, and O. Zaiane. Clustering and sequential pattern mining to support team learning. *IEEE Transactions on Knowledge and Data Engineering*, 21:759–772, 2009.

[17] M. Sao Pedro, R. Baker, and J. Gobert. Improving construct validity yields better models of systematic inquiry, even with less information. In J. Masthoff, B. Mobasher, M. Desmarais, and R. Nkambou, editors, *Proceedings of the 20th Conference on User Modeling, Adaptation, and Personalization*, pages 249–260, Montreal, Canada, 2012.

[18] M. Sao Pedro, R. Baker, and J. Gobert. What different kinds of stratification can reveal about the generalizability of data-mined skill assessment models. In J. Masthoff, B. Mobasher, M. Desmarais, and R. Nkambou, editors, *Proceedings of the 3rd Conference on Learning Analytics and Knowledge*, Leuven, Belgium, 2013.

[19] M. Sao Pedro, R. Baker, J. Gobert, O. Montalvo, and A. Nakama. Leveraging machine-learned detectors of systematic inquiry behavior to estimate and predict transfer of inquiry skill. *User Modeling and User-Adapted Interaction*, 23:1–39, 2013.

[20] M.A. Sao Pedro. *Real-time Assessment, Prediction, and Scaffolding of Middle School Students' Data Collection Skills within Physical Science Simulations*, 2013.

[21] W.M.P. van der Aalst. *Process Mining: Discovery, Conformance and Enhancement of Business Processes*. Springer-Verlag, Berlin, 2011.