Mining sequential patterns with high usage variation

Yingbin Zhang
University of Illinois at Urbana-Champaign
yingbin2@illinois.edu

Luc Paquette
University of Illinois at Urbana-Champaign
lpaq@illinois.edu

ABSTRACT
Sequential pattern mining is a useful tool in understanding learning processes, but identifying the most relevant patterns can be a challenge. Typical sequential pattern mining algorithms and interestingness metrics mainly focus on finding behavior patterns common across all students. However, educational researchers also care about individual differences. This study proposes a method for finding sequential patterns which usage have high variation across students. This method borrows techniques from the field of lag sequential analyses and meta-analyses. It uses the log odd ratio to model the individuals' usage of a sequential pattern and the heterogeneity test to examine the usage variation. We applied this method to analyzing student action logs in a virtual experimental environment and present preliminary results illustrating how the identification of sequential patterns with high usage variation provides interesting information about students' learning behavior. The proposed approach adds a way for understanding individual differences in learning processes.

Keywords
Sequential pattern mining, learning behavior differences, log odds ratio, lag sequential analysis, heterogeneity test

1. INTRODUCTION
Sequential pattern mining (SPM) aims to find the temporal associations between events [1]. For example, whether students read relevant material after answering a question incorrectly. Such sequential behaviors are named sequential patterns. SPM has shown its potentials in helping researchers understand learning behavior [2, 3].

However, there are challenges when applying SPM in education. One important challenge is that SPM algorithms may generate excessive sequential patterns, most of which are uninteresting or irrelevant to the research purpose [2]. This increases the difficulty of making meaningful interpretations and producing actionable pedagogical insights. To address this challenge, researchers select sequential patterns using interestingness metrics, such as the support value e.g., [4, 5]. The support value of a sequential pattern is the proportion of students that shown this pattern. As such, patterns with high support values will reflect similarities in the learners' behavior.

Educational researchers also care about differences among students [6]. The understanding of individual differences in learning is essential for providing learners with adaptive scaffolding. To address this need, this study proposes a method borrowing from lag sequential analyses and meta-analyses that uses log odd ratio and the heterogeneity test to select sequential patterns based on their variation in usage across learners.

2. Methodology
Let $E = \{e_1, e_2, \dots, e_p\}$ be a set of p unique events that may occur within a specific learning environment, such as answering a question and asking for a hint. Let $S_m = \{i_1, i_2, \dots, i_n\}$ be a sequence of N temporally ordered items with each i_j being a subset of E. A sequence is a student's learning process data, such as action logs in an intelligent tutoring system. Each i_j usually contains one event because students rarely initiate two different actions simultaneously. Let $e_x \rightarrow e_y$ be a sequential pattern where e_y occurs after e_x (e_x and e_y may be the same event). Let $\overline{e_x}$ denote an event other than e_x. If there are $i_k = e_x, i_l = e_y$, and $k < l$ in S_m, S_m contains $e_x \rightarrow e_y$ [8].

2.1 Using log odds ratio to model sequential pattern usage
If we fix the gap between $e_x \rightarrow e_y$ to a constant c, we may use methods from the field of lag-sequential analyses to quantify students' usage on $e_x \rightarrow e_y$ [8]. Fixing the gap to c means that we only consider $i_k = e_x, i_l = e_y$, and $l - k = c$ as an occurrence of $e_x \rightarrow e_y$. For example, $c = 1$ means that we only count the case where e_y directly follows e_x. Lag-sequential analysis utilizes statistics from contingency table analyses to quantify the usage of $e_x \rightarrow e_y$, such as the odds ratio and the log odds ratio [8].

Let the frequency of pairs of consecutive events where the first event is e_x and the second event is e_y $n(e_x \rightarrow e_y) = a_m$, Let the frequency of pairs of consecutive events where the first event is e_x but the second event is not e_y $n(e_x \rightarrow \overline{e_y}) = b_m$. Let the frequency of pairs of consecutive events where the first event is not e_x but the second event is e_y $n(\overline{e_x} \rightarrow e_y) = c_m$. Let the frequency of pairs of consecutive events where the first event is not e_x and the second event is not e_y $n(\overline{e_x} \rightarrow \overline{e_y}) = d_m$. The odds ratio of $e_x \rightarrow e_y$ in S_m can be calculated as $\frac{a_m d_m}{b_m c_m}$, while the log odds ratio is $log\frac{a_m d_m}{b_m c_m}$. However, there is measurable bias in this expression when the sample is small. A slightly modified version is often used to reduce bias [9]:

$$Y_m(e_x \rightarrow e_y) = log\frac{(a_m + \frac{1}{2})(d_m + \frac{1}{2})}{(b_m + \frac{1}{2})(c_m + \frac{1}{2})}. \quad (2)$$

The log odds ratio of $e_x \rightarrow e_y$ represents the relative likelihood that e_y occurs after e_x during a student's learning, considering the

probability that e_y occurs after an event other than e_x. If a sequential pattern contains more than two events, researchers may segment a sequential pattern into two sub-patterns and represent the sequential pattern as one sub-patterns follows another. For example, $e_x \rightarrow e_y \rightarrow e_z \rightarrow e_w$ may be represented as $e_{x \rightarrow y} \rightarrow e_{z \rightarrow w}$. This preprocessing has been used in computing the confidence value of sequential patterns longer than two events [10]. Then, the above procedure can be used to calculate the log odds ratio.

The variance of $Y_m(e_x \rightarrow e_y)$ is:

$$V_m(e_x \rightarrow e_y) = \frac{1}{a_m + \frac{1}{2}} + \frac{1}{b_m + \frac{1}{2}} + \frac{1}{c_m + \frac{1}{2}} + \frac{1}{d_m + \frac{1}{2}}. \quad (3)$$

$V_m(e_x \rightarrow e_y)$ characterizes the imprecision of the log odds ratio and decreases as the length of S_m increases. The log odds ratio based on a long sequence is more precise than that based on a short sequence [9].

2.2 Ranking sequential patterns by variation across users

We can examine whether the log odds ratio varies across participants via the heterogeneity test used in meta-analyses [11]. One commonly used heterogeneity test is the Q test [12]. In meta-analyses, Q is the weighted sum of the squared deviations of each study's effect estimate from the weighted mean of all studies' effect estimates. The weighting for each study is the inverse of the variance of the study's effect estimate. Thus, in terms of the variation of the log odds ratio of $e_x \rightarrow e_y$, Q can be calculated using the formula:

$$Q(e_x \rightarrow e_y) = \sum \frac{(Y_m - \overline{Y})^2}{V_m}, \quad (4)$$

where $\overline{Y}$ is the weighted mean of log odds ratios, i.e.,

$$\overline{Y} = \frac{\sum \frac{Y_m}{V_m}}{\sum \frac{1}{V_m}}. \quad (5)$$

Q follows a chi-square distribution with $k - 1$ degrees of freedom, where k is the number of sequences or participants. Thus, if $Q(e_x \rightarrow e_y)$ is higher than the critical value for a given significance level (e.g., 0.05), we may conclude that the usage of $e_x \rightarrow e_y$ has statistically significant variation across participants. Moreover, for the same dataset, the number of participants is constant, and thus, the Qs of all sequential patterns follow the same chi-square distribution and are comparable. However, it is difficult to interpret Q because its magnitude is influenced by the number of participants. The I^2 index overcomes this issue [13].

$$I^2 = \begin{cases} \frac{Q - (k-1)}{Q} * 100\%, if\ Q > (k-1) \\ 0, else \end{cases}. \quad (6)$$

$I^2(e_x \rightarrow e_y)$ can be interpreted as the proportion of variation in the log odds ratio of $e_x \rightarrow e_y$ due to true between-participants variance. Ranking sequential patterns by Q and I^2 produces the same results because k is fixed for the same dataset.

3. Example

This section applied the proposed method to a dataset of student action logs collected from a virtual experiment environment called LabBuddy [14].

3.1 Data

3.1.1 Participants

The data were collected from a graduate-level enzymology course at a university in the Netherlands. Participants were 76 graduate students in this course. The average age was 22.91 years old (SD = 1.80). Around 64.47% of the students were female.

3.1.2 LabBuddy

The course helped students prepare for the laboratory classes using LabBuddy. LabBuddy in this course contained a self-directed learning task, which included six research questions offered by a virtual tutor, Professor Kabel. Students start with proposing hypotheses for each question and make an experimental design via a flow chart to test the hypotheses (Figure 1). Each block in the flow chart represents a chemical method and contains details about the method. Each block also contains some closed questions that students must answer correctly before implementing the method and getting the raw data. Students do some calculations based on the raw data to get the results. The details, raw data, and calculations of a method are located in different subblocks of a block. If students are struggling with a closed question, they may request hints or the correct answer. Once students obtain the results, they may consult Professor Kabel to interpret them and either accept or reject their hypotheses. Students used LabBuddy for an average of 7.5 h distributed over three days. Their action logs were used for analysis.

3.2 Analyses

We preprocessed the action logs by removing redundant successive repeated actions (e.g., multiple selections of the same block) and contextualizing some actions (e.g., is the submitted answer to a closed question correct?). The preprocessing resulted in 19 unique events. The average number of events in a student's action log was 995 (SD = 363). Then, we implement our methods via the following procedure:

1. Apply the cSPADE algorithm to find frequent sequential patterns with support no less than 0.5. We used this algorithm because it allows us to fix the gap between events in a sequential pattern, a prerequisite for calculating the log odds ratios. The gap was fixed to 1 in the analysis. For simplicity, we only focused on sequential patterns containing two events. This step generated 81 frequent sequential patterns.
2. For each student, compute the log odds ratio, variance, and the number of occurrences of each frequent sequential pattern.
3. For each frequent sequential pattern, conduct the Q test and calculate the I^2 index, the average log odds ratio, and the average occurrence. As the Q test was run 81 times, we used the Benjamini-Yekutieli correction to control the false discovery rate [15].

Note we only apply our method to frequent sequential patterns because the variation of a sequential pattern across participants would be low if few participants used a pattern (i.e., it was infrequent).

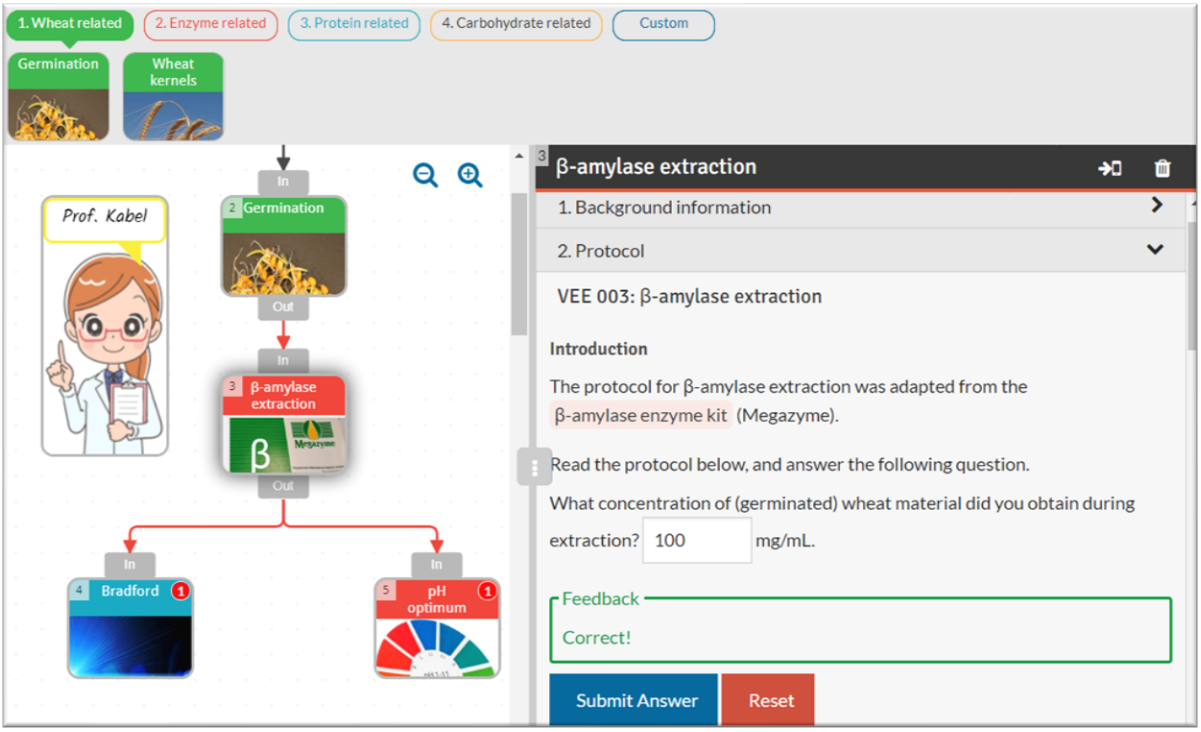

Figure 1. The LabBuddy learning environment.

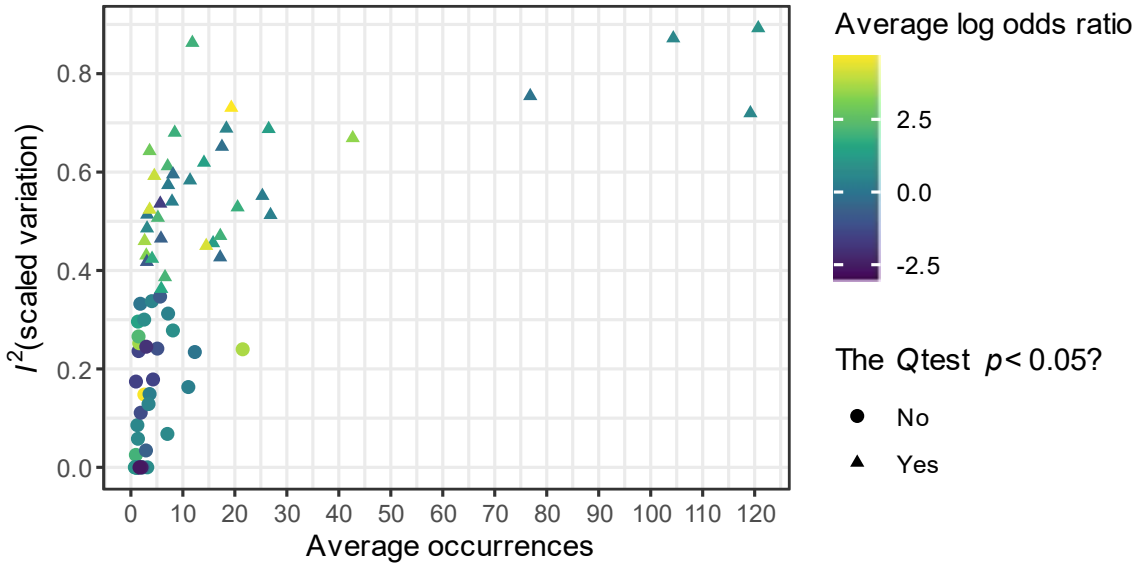

Figure 2. The I^2 indexes, average log odds ratios, and average occurrences of sequential patterns.

3.3 Preliminary results

Figure 2 visualizes the relationships among the I^2 indexes, average log odds ratios, and average occurrences of the 81 sequential patterns. There were moderate positive relationships between the I^2 index and the average log odds ratio ($r = 0.57$, $p < 0.001$) as well as the average occurrences ($r = 0.36$, $p = 0.001$). Nevertheless, ranking sequential patterns by their variation between students results in a different set of selected patterns than ranking them by their similarities (average log odds ratios and occurrences) between students. Some sequential patterns had few average occurrences (e.g., less than 5) or negative average log odds ratios while still being used differentially by students (the adjusted p of the Q test < 0.05). Some sequential patterns had relatively high average occurrences (e.g., larger than 10) or average log odds ratios (e.g., larger than 1) but were used consistently across students (the adjusted p of the Q test > 0.05).

We investigated how I^2 might help us detect behavioral differences by looking more closely at two sequential patterns with distinct I^2: *Submitting an intermedia answer → Submitting an intermedia answer* and *Requesting a hint → Requesting a hint*. Both patterns had high values in average occurrences and log odds ratios (see Table 1). *Submitting an intermedia answer → Submitting an*

intermedia answer had a I^2 of 0.75 ($p < 0.001$), indicating that students had high variations in the usage on this pattern. Further analysis showed that, in 9.54% of all pairs of students (309/3,240), the log odd ratio was significantly different between the two students. This means that among 10 randomly sampled pairs, on average, there was one pair where the two students had significantly different probability of submitting two intermedia answers consecutively. In contrast, the usage on *Requesting a hint → Requesting a hint* was relatively consistent across students (I^2 = 0.24, p = 0.52). Analyses showed that, in only 1.6% of pairs, the log odd ratio was significantly different between two students.

Table 1. The metrics of two sequential patterns

Pattern	I^2	p for the Q test	Log odds ratio	Occurrences
RH.RH	0.24	0.52	3.76	21.51
SI.SI	0.75	0.00	4.81	19.32

Note. RH.RH: *Requesting a hint → Requesting a hint*. SS.WA: *Submitting an intermedia answer → Submitting an intermedia answer*.

4. Discussion

This study proposed a method for mining sequential patterns which usage has high variation across students. We applied the method to a dataset of student action logs in a virtual experimental environment. The preliminary results suggest that ranking sequential patterns by their variation across students results in a different selection of patterns than by their similarities across students. Moreover, the results demonstrated how the proposed method could capture individual differences in sequential behavior patterns. The approach adds a way for understanding individual differences in learning, which is critical in education.

The next step is to examine whether the sequential patterns with high variation are related to students' learning gains. Such investigation would contribute to our understanding of how differences in which sequential patterns may lead to differences in learning outcomes. The insights, in turn, would provide information about how the learning environment might scaffold the learners' interaction with the learning environment by prompting sequential behavior patterns beneficial to learning and discouraging patterns harmful to learning.

Our approach requires fixing the gap between events of a sequential patterns. This requirement limits flexibility. For example, researchers may regard *Submitting an intermedia answer → Requesting a hint → Submitting an intermedia answer* as an instance of *Submitting an intermedia answer → Submitting an intermedia answer*, but fixing the gap to 1 excludes this possibility. On the other hand, if fixing the gap to 2, *Submitting an intermedia answer* directly after *Submitting an intermedia answer* would not be regarded as an instance of *Submitting an intermedia answer → Submitting an intermedia answer*. The limitation is the same as the issue that the lag between the antecedent and consequent events must be fixed in a lag sequential analysis [8]. Addressing this issue is challenging but worthy of effort.

5. ACKNOWLEDGMENTS

We would like to thank the Laboratory of Food Chemistry, Wageningen University & Research for allowing us to use the dataset and access to LabBuddy.

6. REFERENCES

[1] Baker, R. 2010. Data mining for education. In B. McGaw, P. Peterson & E. Baker (Eds), *International Encyclopedia of Education* (pp. 112-118). Oxford, UK: Elsevier Ltd.

[2] Zhou, M., Xu, Y., Nesbit, J. C. & Winne, P. H. 2010. Sequential pattern analysis of learning logs: Methodology and applications. In *Handbook of Educational Data Mining* (pp. 107-121). Boca Raton: CRC Press.

[3] Moon, J. & Liu, Z. 2019. Rich representations for analyzing learning trajectories: Systematic review on sequential data analytics in game-based learning research. In A. Tlili & M. Chang (Eds), *Data analytics approaches in educational games and gamification systems* (pp. 27-53). Singapore: Springer.

[4] Jiang, Y., Paquette, L., Baker, R. S. & Clarke-Midura, J. 2015. Comparing novice and experienced students within virtual performance assessments. In *Proceedings of the 8th International Conference on Educational Data Mining* (Madrid, Spain, Jun. 2015). International Educational Data Mining Society, 136-143.

[5] Kang, J., Liu, M. & Qu, W. 2017. Using gameplay data to examine learning behavior patterns in a serious game. *Comput. Hum. Behav.*, 72, (2017), 757-770. http://doi.org/10.1016/j.chb.2016.09.062

[6] Malmberg, L., Lim, W. H., Tolvanen, A. & Nurmi, J. 2016. Within-students variability in learning experiences, and teachers' perceptions of students' task-focus. *Frontline Learning Research*, 4, 5 (2016), 62-82. http://doi.org/10.14786/flr.v4i5.227

[7] Agrawal, R. & Srikant, R. 1995. Mining sequential patterns. In *Proceedings of the eleventh international conference on data engineering* (Taipei, Taiwan, 1995). IEEE, 3-14.

[8] Bakeman, R. & Quera, V. *Sequential analysis and observational methods for the behavioral sciences.* Cambridge University Press, New York, NY, US, 2011.

[9] Dagne, G. A., Howe, G. W., Brown, C. H. & Muthén, B. O. 2002. Hierarchical modeling of sequential behavioral data: An empirical Bayesian approach. *Psychol. Methods*, 7, 2 (Jun. 2002), 262-280. http://doi.org/10.1037/1082-989X.7.2.262

[10] Fournier-Viger, P., Faghihi, U., Nkambou, R. & Nguifo, E. M. 2012. CMRules: Mining sequential rules common to several sequences. *Know.-Based Syst.*, 25, 1 (2012), 63-76. http://doi.org/10.1016/j.knosys.2011.07.005

[11] Huedo-Medina, T. B., Sánchez-Meca, J., Marín-Martínez, F. & Botella, J. 2006. Assessing heterogeneity in meta-analysis: Q statistic or I² index? *Psychol. Methods*, 11, 2 (2006), 193-206. http://doi.org/10.1037/1082-989X.11.2.193

[12] Cochran, W. G. 1954. The Combination of Estimates from Different Experiments. *Biometrics*, 10, 1 (1954), 101-129. http://doi.org/10.2307/3001666

[13] Higgins, J. P. T. & Thompson, S. G. 2002. Quantifying heterogeneity in a meta-analysis. *Stat. Med.*, 21, (2002), 1539-1558. http://doi.org/10.1002/sim.1186

[14] Van Der Kolk, K., Beldman, G., Hartog, R. & Gruppen, H. 2012. Students Using a Novel Web-Based Laboratory Class Support System: A Case Study in Food Chemistry Education. *J. Chem. Educ.*, 89, 1 (Jan. 2012), 103-108. http://doi.org/10.1021/ed1005294

[15] Benjamini, Y. & Yekutieli, D. 2001. The Control of the False Discovery Rate in Multiple Testing under Dependency. *The Annals of Statistics*, 29, 4 (Jan. 2001), 1165-1188.