Learning Sequence Analytics for Support in Learning Tasks

Manuel Valle Torre

Centre for Education and Learning, Delft University of Technology. Van Mourik Broekmanweg 6, Delft, The Netherlands

Abstract

Learning Analytics is currently undergoing a self-reflection process, with calls to bring the *learning* back into Learning Analytics (LA). At the same time, there is increasing interest in using temporal and sequential analysis to describe and understand learner behaviour, but current work is mostly exploratory, with promising potential for educational interventions. Obtaining sequential patterns from learner data can provide insights into processes that can be missed by descriptive analytics, where the order and timing of actions make a difference. A current scenario is in practical exercises in Data Science Education where teachers cannot access the learners' development process and are therefore unable to properly assess them or provide detailed feedback. Furthermore, there are phases of data science that may not generate data records, such as setting an objective, or intermediate evaluation and reflection. The introduction of Large Language Models (ChatGPT, Github Copilot) as coding assistants has the potential to obtain such data in the form of conversation. By using learning sequence analytics on the interactions between LLM and learners, integrated with the data traces in their programming and analytics environments, we can identify and understand behavioural patterns that can be used to assess, monitor and support the learning process in Data Science Education.

Keywords

Learning Analytics, Learning Sequences, Sequence Analysis, Adaptive Support, Data Science Education, Large-Language Models

1. Introduction

Learning analytics is working on measuring and optimising learning, supported by educational theory. One line of research focuses on observing learning as a temporal process, while students work on a task [1]. This is especially motivated by the prevalence of systems where it is impossible for teachers to monitor all students since it happens in systems and at a large scale, where assessment is often reduced to aggregated values [2]. By analysing sequences of learners' data, patterns can be identified and used to discover behaviours in the learning process that would be missed in a summative assessment. For the purpose of this work, a *learning sequence* is defined as an ordered list of actions with context-specific features, where each *unit* is a learning action, systematically mapped from one or more data traces [3]. However, the ideas and techniques for sequence analysis come from many research fields related to Educational Technologies, so there are no standard definitions. Likewise, there are no best practices of which analysis methods are applied for which purposes or how they relate to different educational settings [4]. Using the definition of learning sequence above, we conducted a literature review to develop an overview of the tasks, analysis methods and interventions in educational research with learning sequences.

For the next stage of this doctoral project, learning sequence analysis will be implemented in tasks of Data Science Education (DSE). Integrating many domains, DSE requires an important amount of practical work, and there are plenty of tools that allow interactive exercises and projects for this practice. Still, the assessment is often reduced to simple automated tests or the final output of a project [5]. Therefore, the process, strategies and challenges in the process, such as intermediate evaluations or debugging, cannot be identified and addressed by the teachers. Additionally, the tools and systems where students work may not be able to track vital parts of the process, such as objective definition, problem decomposition, method selection or reflection [6]. The introduction of chatbots using Large-language Models (LLMs) in education can provide a new data source to help in tracking processes that would otherwise be on search engines and notebooks. With students using the LLMs as data science assistants, we can collect the records of their interactions, especially on the concepts and procedures for which they need clarification [7]. Combining data traces and chatbot interactions would provide a better picture of the students' data science process, and the use of learning sequence analytics would allow us to identify relevant strategies and challenges in a variety of data science tasks. Furthermore, the chatbot provides a natural interface for timely intervention, where the student can be guided within the learning task.

For this project, a learning environment will be designed to evaluate these processes and reveal what happens when students interact with data science practice

Proceedings of the Doctoral Consortium of the 18th European Conference on Technology Enhanced Learning, 4th September 2023, Aveiro, Portugal

m.valletorre@tudelft.nl (M. Valle Torre)
0000-0002-0456-8360 (M. Valle Torre)

tasks. The environment will be based on interactive notebooks, with a chatbot interface, recording both the programming and conversational interactions. First, a lab experiment will be conducted as a pilot to observe the learner's actions in a data science environment, advancing our understanding of behaviours that are often missed in final assessments. This understanding will be used for research using the learning environment in tasks of data analytics and data science courses to assess potential interventions, and their impact on learning outcomes. In the following, I describe the results of the literature review and explain current and future work.

2. Background

Learning Sequences, in the context of this work, refer to the sequences of actions that learners execute during a learning task, such as solving a programming problem or creating a concept map from reading materials [8, 9]. Such learning tasks are situations designed for the students to actively engage with the learning materials, which require them to integrate domain knowledge with general critical thinking and analytical skills [10]. In these tasks, learners have a large number of operators or actions available in the problem space, making the combinations of actions that lead to a solution difficult to observe. As a consequence, the assessment and monitoring of learners are particularly difficult for teachers [11]. This is increasingly challenging with the current scale provided by educational technologies and platforms. Learning analytics, however, can use the data produced by the systems in modern online education to scale monitoring and assessment capabilities [12]. Furthermore, by analysing the sequence of learning actions, an educational system can use all the information available about a learner throughout their current process, provide timely support and enable a formative assessment approach for teachers [13]. A growing field of education where many processes are lost in a computer system, but where learners would benefit from such timely support, is Data Science [6].

2.1. Data Science Education

Data has revolutionized various scientific disciplines, paving the way for unprecedented data collection and analysis capabilities [14]. This data-driven transformation extends from engineering to computational and natural science studies, where data analytics techniques can support processes such as scientific hypothesis evaluation, system modelling and simulation, and decision-making [15]. As the complexity and diversity of data continue to grow, the need for data skills becomes more pronounced in the future workplace. The educational landscape is evolving to meet these demands, equipping students with the necessary data analytics skills that will become a necessity in their professional lives.

Data science education (DSE) is the process of teaching and learning data science skills and concepts, such as programming, statistics, machine learning, data visualization, and data ethics. DSE can be delivered through various formats, such as online courses, university programs, bootcamps, and competitions; with a variety of programming languages and tools [5]. Some of the main challenges for DSE include the lack of standard topics, practical activities, and methods, which makes it difficult to the scope of a course, and the need to adapt to a wide variety of students' backgrounds. Furthermore, the assessment of activities is reduced to a final product, or automated grading systems on clearly defined tasks, as they depend on pre-written test codes [6].

2.2. Generative AI in Education

Generative AI and Large Language Models are relatively new in the mainstream, specifically due to the availability of ChatGPT, launched in November of 2022, and now used by more than 100 million people [16]. Without information from OpenAI, it is difficult to know how many students use ChatGPT, or if they use it for any study-related activities. However, two surveys from March [17] and September [18] 2023, state that more than 40% of the respondents have experience with ChatGPT and around 20% of university students have reported using it for their homework.

While there are several ways educators and institutions are dealing with AI in the classroom, from banning it to embracing it as a mandatory part of the curriculum, the popularity of LLMs has brought important change in education [19]. In consequence, there has been an explosion of research on Generative AI and education, focusing on large-scale risk and opportunity analyses [20, 19], on the ability of LLMs to assist teachers [21, 22] or solve existing homework and exams [23, 24]. The last ones mention the risks and potentials for students, for instance, that it can facilitate cheating or explain difficult concepts in many different ways.

The focus of this work lies in the overlap of these topics: to use learning sequence analysis to understand and support practical work in data science education, including activities that traditionally happen outside of the learning environment.

2.3. Research Questions

The primary research question of this PhD dissertation is *how can learning analytics be used to support learners when solving complex problems in technology-enhanced learning environments?* To address this, I seek to answer

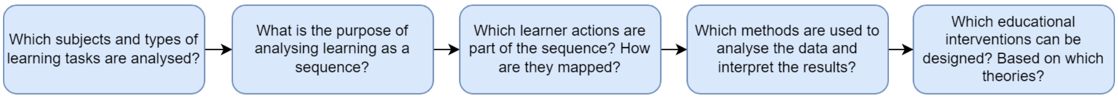
Which subjects and types of learning tasks are analysed?
What is the purpose of analysing learning as a sequence?
Which learner actions are part of the sequence? How are they mapped?
Which methods are used to analyse the data and interpret the results?
Which educational interventions can be designed? Based on which theories?

Figure 1: Learning Sequence Analysis - Framework

3 main questions: For my work, I've chosen to focus on sequential analytics, making the first question **how are learning sequences analyzed to describe and support learners' actions in computer systems?**

- How are data records transformed into sequence units for analysis?
- Which methods are used to analyze learning sequences? For which purposes?
- Which interventions are designed and implemented using the insights from learning sequence analysis?

Furthermore, I will use Data Science education as the complex problem to use for evaluation, an exciting and relevant domain, which together with current developments in Artificial Intelligence technology leads to the second question: **how do learners work on Data Science problems in a technology-enhanced environment?**

- Which are the main actions and phases of a data science task?
- How to describe it as a sequence to determine behaviour patterns and support the learning process?

Finally, for the activities of data science where no programming is involved, integrating a chatbot as a tutor/assistant and analysing the interactions might provide the information we need. This leads to the third question, **how are learners using LLMs when dealing with a data science problem?**

- How can we identify the students' prompting or questioning tactics?
- How can we leverage the interactions between learners and LLMs to understand the complete data science learning process?

3. Preliminary Results and Proposed Approach

This article, part of the ECTEL23 Doctoral Consortium was written at the end of the second year of my PhD, out of 4, as it is common in the Netherlands. The current state and next steps are described below.

3.1. Literature Review

To answer the first questions in Figure 1, a systematic literature review on learning sequence analytics has been executed and submitted for review. The selected records were published between 2010 and 2023, from a search performed in SCOPUS and Web of Science, including 74 works that were full-research articles, in English, where sequence analysis methods were used on data produced during a learning task. The analysis considered 5 main sections: the types of tasks to be studied, the purpose of using sequences for the main analysis, the learning actions that build the sequences, the analysis methods implemented, and the educational interventions designed. The literature review is currently under review and the main takeaways are the following: The transformation from raw data to **sequence units** is essential and will define the analysis and interpretability of the results. Domain-specific knowledge is required to identify the sequence units as learning actions, it facilitates the interpretation of patterns but limits generalizeability. Using theory to define the sequence units, such as self-regulated learning actions, can be used to generalize the results, but it is not always possible. Sequence **analysis methods** can be grouped by their scope, and the purpose of the sequence, shown in Figure 2. For example, the scope of the analysis can be to identify strategies in the whole sequence, common when using Process Mining or Probability Models. On the other hand, methods such as Sequential Pattern Mining can be used to find context-specific tactics, which can be as short as 2-3 learning actions. Support methods, such as machine learning classifiers, are often used to determine learner characteristics from the obtained sequential patterns. Most reviewed articles mention potential educational interventions with the obtained insights, but only a few implement such interventions, with 2 of the articles evaluating them.

3.2. Experimental Study

To evaluate the trends in Section 3.1, the next step of my project is to analyse a learning task using learning sequence analysis, with an educational intervention grounded both in the obtained insights and in learning theory. The idea is to use Data Science tasks to understand the tactics and strategies that users implement when working on them. These tasks require both analytical skills and domain knowledge, as well as a certain

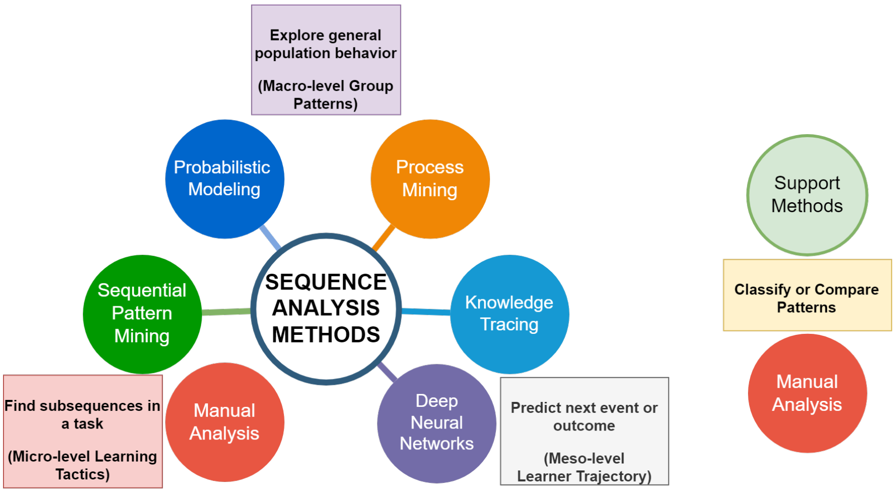
Explore general population behavior
(Macro-level Group Patterns)
Probabilistic Modeling
Process Mining
SEQUENCE ANALYSIS METHODS
Sequential Pattern Mining
Knowledge Tracing
Find subsequences in a task
(Micro-level Learning Tactics)
Manual Analysis
Deep Neural Networks
Predict next event or outcome
(Meso-level Learner Trajectory)
Support Methods
Classify or Compare Patterns
Manual Analysis

Figure 2: Sequence Analysis Methods, Purpose and Scope

proficiency in programming and statistics. As such, they can be conceptualized as creative problem solving, with intermediate evaluations and reflections, which can be difficult to observe and assess by teachers [5]. This makes them a perfect use case for learning sequence analysis.

3.3. Tooling and Method

A customised Jupyter Notebook system is currently in development, integrating a chat interface and detailed logging capabilities. The system will track the development and execution of Python code used for all data processing and visualization, as well as the outputs and errors from the interpreter. The chatbot interface is able to use different LLMs in the background, currently running on OpenAI's API, but capable of working with a local model as well. A logger on the chatbot interface will be used to store the learner-agent interactions as a sequence of periodic text edits, executed requests and agent responses.

For a pilot experiment, participants will be asked to work on a data science problem using a Large Language Model (LLM) as an assistant. This pilot will be performed in a controlled lab setting, to evaluate the system and generate a general overview of data science actions and LLM interactions. Phases, subphases and actions of the data science process will be identified using examples from existing Jupyter notebook repositories [25]. LLM interactions will be coded using problem decomposition, from a prompt engineering perspective [26]. The integrated sequence will be analysed using Learning Sequence Analytics on two scopes: micro-level pattern mining [27] to identify programming or statistics tactics, such as specific data cleaning techniques or statistical tests, and one of the methods on the macro-level [28] to determine general data science strategies.

For the second study, the system will be used as optional practice for different phases of data science, part of a real course. Finally, the presence of certain tactics or the use of problem-solving strategies identified can be used to classify learners, for instance, by their proficiency in programming, statistics, analytics, or any of the disciplines within data science. In future work, we plan to implement a middle layer in the interface between the LLM chatbot and the learner to provide support within the assistant, with interventions based on sequential patterns to optimise the learning support.

4. Contribution to TEL

The first contribution of this work is the literature review on learning sequence analytics, including the lessons learned on learning tasks, design, methods, and educational interventions in the context of sequence analysis. The second will be a better understanding of the processes that students execute when working on data science tasks. This would also serve as an evaluation of the findings from the literature review on learning sequence analytics. The third is an exploration of the interactions between learners and LLMs, what they ask and how they use them, bringing us closer to understanding the potential of LLMs in education.

Acknowledgments

This work is funded by the Leiden-Delft-Erasmus Centre for Education and Learning.

References

[1] I. Molenaar, A. F. Wise, Temporal aspects of learning analytics - grounding analyses in concepts of time, in: C. Lang, G. Siemens, A. F. Wise, D. Gašević, A. Merceron (Eds.), The Handbook of Learning Analytics, 2 ed., SoLAR, Vancouver, Canada, 2022, pp. 66–76.

[2] K. D. Wang, J. M. Cock, T. Käser, E. Bumbacher, A systematic review of empirical studies using log data from open-ended learning environments to measure science and engineering practices, British Journal of Educational Technology 54 (2023) 192–221. doi:10.1111/bjet.13289.

[3] P. H. Winne, Construct and consequential validity for learning analytics based on trace data, Computers in Human Behavior 112 (2020) 106457. doi:10.1016/j.chb.2020.106457.

[4] S. Knight, A. Friend Wise, B. Chen, Time for Change: Why Learning Analytics Needs Temporal Analysis, Journal of Learning Analytics 4 (2017). doi:10.18608/jla.2017.43.2.

[5] T. Donoghue, B. Voytek, S. E. Ellis, Teaching Creative and Practical Data Science at Scale, Journal of Statistics and Data Science Education 29 (2021) S27–S39. doi:10.1080/10691898.2020.1860725.

[6] A. Schwab-McCoy, C. M. Baker, R. E. Gasper, Data Science in 2020: Computing, Curricula, and Challenges for the Next 10 Years, Journal of Statistics and Data Science Education 29 (2021) S40–S50. doi:10.1080/10691898.2020.1851159.

[7] X. Tu, J. Zou, W. J. Su, L. Zhang, What Should Data Science Education Do with Large Language Models?, 2023. arXiv:2307.02792.

[8] T. J. Nokes, C. D. Schunn, M. T. H. Chi, Problem Solving and Human Expertise, in: P. Peterson, E. Baker, B. McGaw (Eds.), International Encyclopedia of Education (Third Edition), Elsevier, Oxford, 2010, pp. 265–272. doi:10.1016/B978-0-08-044894-7.00486-3.

[9] K. D. Wang, S. Salehi, M. Arseneault, K. Nair, C. Wieman, Automating the Assessment of Problem-solving Practices Using Log Data and Data Mining Techniques, in: Proceedings of the Eighth ACM Conference on Learning @ Scale, ACM, Virtual Event Germany, 2021, pp. 69–76. doi:10.1145/3430895.3460127.

[10] A. Proske, S. Narciss, H. Koerndle, The Exercise Format Editor: A multimedia tool for the design of multiple learning tasks, 2004, pp. 149–164.

[11] M. Eagle, D. Hicks, B. Peddycord, T. Barnes, Exploring networks of problem-solving interactions, in: Proceedings of the Fifth International Conference on Learning Analytics And Knowledge, LAK '15, Association for Computing Machinery, New York, NY, USA, 2015, pp. 21–30. doi:10.1145/2723576.2723630.

[12] J. Seifried, S. Brandt, K. Kögler, A. Rausch, The computer-based assessment of domain-specific problem-solving competence—A three-step scoring procedure, Cogent Education 7 (2020) 1719571. doi:10.1080/2331186X.2020.1719571.

[13] J. Saint, D. Gašević, W. Matcha, N. A. Uzir, A. Pardo, Combining analytic methods to unlock sequential and temporal patterns of self-regulated learning, in: Proceedings of the Tenth International Conference on Learning Analytics & Knowledge, ACM, Frankfurt Germany, 2020, pp. 402–411. doi:10.1145/3375462.3375487.

[14] J. Gibson, T. Mourad, The growing importance of data literacy in life science education, American Journal of Botany 105 (2018). doi:10.1002/ajb2.1195.

[15] C. J. Lynch, R. Gore, A. J. Collins, T. S. Cotter, G. Grigoryan, J. F. Leathrum, Increased Need for Data Analytics Education in Support of Verification and Validation, in: 2021 Winter Simulation Conference (WSC), 2021, pp. 1–12. doi:10.1109/WSC52266.2021.9715485.

[16] F. Duarte, Number of ChatGPT Users (2023), https://explodingtopics.com/blog/chatgpt-users, 2023.

[17] M. T. Nietzel, More Than Half Of College Students Believe Using ChatGPT To Complete Assignments Is Cheating, https://www.forbes.com/sites/michaeltnietzel/2023/03/20/more-than-half-of-college-students-believe-using-chatgpt-to-complete-assignments-is-cheating/, 2023.

[18] A. Prothero, How Students Use AI vs. How Teachers Think They Use It, in Charts, Education Week (2023).

[19] M. M. Rahman, Y. Watanobe, ChatGPT for Education and Research: Opportunities, Threats, and Strategies, Applied Sciences 13 (2023) 5783. doi:10.3390/app13095783.

[20] M. Farrokhnia, S. K. Banihashem, O. Noroozi, A. Wals, A SWOT analysis of ChatGPT: Implications for educational practice and research, Innovations in Education and Teaching International 0 (2023) 1–15. doi:10.1080/14703297.2023.2195846.

[21] W. C. H. Hong, The impact of ChatGPT on foreign language teaching and learning: Opportunities in

education and research, Journal of Educational Technology and Innovation 5 (2023).

[22] J. Jeon, S. Lee, Large language models in education: A focus on the complementary relationship between human teachers and ChatGPT, Education and Information Technologies (2023). doi:10.1007/s10639-023-11834-1.

[23] K. Malinka, M. Peresíni, A. Firc, O. Hujnák, F. Janus, On the Educational Impact of ChatGPT: Is Artificial Intelligence Ready to Obtain a University Degree?, in: Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education V. 1, ITiCSE 2023, Association for Computing Machinery, New York, NY, USA, 2023, pp. 47–53. doi:10.1145/3587102.3588827.

[24] J. Qadir, Engineering Education in the Era of ChatGPT: Promise and Pitfalls of Generative AI for Education, in: 2023 IEEE Global Engineering Education Conference (EDUCON), 2023, pp. 1–9. doi:10.1109/EDUCON54358.2023.10125121.

[25] L. Quaranta, F. Calefato, F. Lanubile, KGTorrent: A Dataset of Python Jupyter Notebooks from Kaggle, in: 2021 IEEE/ACM 18th International Conference on Mining Software Repositories (MSR), 2021, pp. 550–554. doi:10.1109/MSR52588.2021.00072.

[26] M. Baer, K. T. Dirks, J. A. Nickerson, Microfoundations of strategic problem formulation: Strategic Problem Formulation, Strategic Management Journal 34 (2013) 197–214. doi:10.1002/smj.2004.

[27] Y. Wang, T. Li, C. Geng, Y. Wang, Evaluating Student Learning Effect Based on Process Mining, in: H. Florez, M. Leon, J. M. Diaz-Nafria, S. Belli (Eds.), Applied Informatics, Communications in Computer and Information Science, Springer International Publishing, Cham, 2019, pp. 59–72. doi:10.1007/978-3-030-32475-9_5.

[28] K. Akhuseyinoglu, P. Brusilovsky, Data-Driven Modeling of Learners' Individual Differences for Predicting Engagement and Success in Online Learning, in: Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization, UMAP '21, Association for Computing Machinery, New York, NY, USA, 2021, pp. 201–212. doi:10.1145/3450613.3456834.