Using Machine Learning to Understand Students' Learning Patterns in Simulations

Wonkyung Jang, Josh Francisco, Nethra Ranganathan, Kathleen Marley McCarroll and Kihyun Ryoo
wkjang@live.unc.edu, francisco97j@unc.edu, netranga@live.unc.edu, kathymcc@live.unc.edu, khryoo@email.unc.edu
The University of North Carolina at Chapel Hill

Abstract: This study explores how unsupervised machine learning (ML) techniques can identify productive learning patterns as students conduct virtual experiments using a simulation. The log data from 52 pairs of eighth-grade students were analyzed using two ML techniques, Finite Mixture Model (FMM) and Sequential Pattern Mining (SPM). The results show four levels of learning patterns (i.e., Low Activity, Random Interaction, High Analysis, Tasked Exploration), each of which have unique, sequential actions. This study shows the potential value of unsupervised ML for understanding which types of interactions with simulations could facilitate students' understanding of complex scientific phenomena.

Keywords: machine learning, simulation, log data, learning patterns

Introduction

The Next Generation Science Standards (NGSS; NGSS Lead States, 2013) require students to understand complex scientific phenomena through science practices, such as carrying out investigations, collecting and analyzing data, and engaging in argument from evidence. Computer simulations can create such learning opportunities by allowing students to design virtual experiments by manipulating variables and displaying the outcomes in multiple forms, including molecular animations and dynamic graphs (e.g., Plass et al., 2012). During their experiments, students can also collect and analyze different types of data to identify relations among the variables and deepen their understanding of complex scientific systems (Stieff, 2011).

However, research has shown that students often struggle with how to conduct scientific investigations using simulations, which can lead to an incomplete understanding of scientific phenomena (Scalise et al., 2011). For instance, students may run too few or too many trials when testing hypotheses (Gobert et al., 2012). Some students also randomly change variables, rather than carefully designing their experiments (McElhaney & Linn, 2011). Even after successful trials, some students may not know how to interpret multiple sources of evidence to make sense of the underlying mechanism (Kanari & Millar, 2004).

Research has shown that productive learning patterns can be identified from log data using unsupervised machine learning (ML) techniques that can uncover hidden patterns in raw or unlabeled data (Amershi & Conati, 2009; Khalid & Prieto-Alhambra, 2019; Qiao & Jiao, 2018). For example, Shih, Koedinger, and Scheines (2010) used unsupervised Hidden Markov Models to discover student learning tactics while incorporating student-level outcome data. Bernardini and Conati (2010) used Class Association Rule Mining to automatically identify common interaction behaviors.

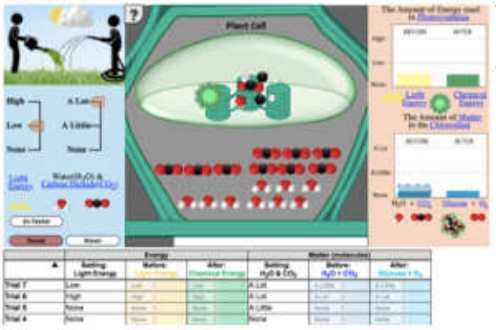
In particular, recent studies have shown the value of Finite Mixture Model (FMM) and Sequential Pattern Mining (SPM) for finding patterns within complex datasets (e.g., interactions in simulations collected during science inquiry instruction) (e.g., Baker & Inventado, 2014; McLachlan & Peel, 2000). For example, FMM can handle high-dimensional low sample size (HDLSS) datasets (i.e., # of observation is much smaller than # of variables) and model multivariate data from populations suspected to include separate subpopulations (Melnykov & Maitra, 2010). SPM can discover important learning subsequences that appear in the same relative order but not necessarily contiguous (e.g., <a(bc)dc> is a subsequence of <a(abc)(ac)d(cf)>) (Zaki, 2001).

Using these two unsupervised ML techniques can determine distinct learning categories in log data and provide further insight into subsequential learning patterns within each category. Given the limited research on the use of FMM and SPM in the field of education, this study explores how these two techniques can identify productive interaction patterns that can enhance eighth-grade students' science learning within a simulation.

Methods

As part of a larger NSF project, 52 pairs of eighth-grade students from a low-income middle school completed a web-based inquiry project on photosynthesis for three to four days. During the project, pairs used a simulation to explore what happens to energy, matter, and plant growth during photosynthesis by conducting virtual

experiments (see Figure 1). Before and after using the simulation, pairs answered identical prediction and reflection questions about the target concepts. The simulation provided scaffolded prompts to engage them in making a hypothesis, designing an experiment, collecting data, and analyzing data to draw a conclusion. All pairs' actions with the simulation (e.g., a sequence of hyperlink and button clicks, students' responses to a data table) were logged with time-stamps. During the study, at least one researcher was present in each classroom to observe student interactions with the simulation and videotape selected pairs.



Categories of Actions	Description and Examples
Experimentation	Actions related to conducting experiments and collecting data (e.g., manipulating variables, resetting a simulation, running a simulation, changing the speed settings, entering data)
Data Analysis	Actions related to drawing conclusions based on the evidence collected (e.g., answering the analysis question)
Information Seeking	Actions related to obtaining additional information (e.g., clicking vocabulary, clicking instruction button)

Figure 1. Simulation interface and student actions.

Analysis methods

Student responses to the prediction and reflection questions were scored using a seven-scale rubric that rewarded elaborated connections between multiple key concepts. Student actions from the log data were analyzed using the following three steps. First, we used exploratory factor analysis to consolidate the large number of actions into a smaller set of latent factors. As a result, 48 actions were grouped into three categories (see Table 1). Second, we used FMM with the Expectation-Maximization (EM) algorithm to approximate complex probability densities which present multimodality, skewness, and heavy tails of log data. Third, we explored how student sequential interaction patterns vary across the identified levels using SPM.

Results

Using FMM with the EM algorithm, the results show four levels of learning patterns based on the frequency of actions and learning gains (see Table 1). The Low Activity group is characterized by the lowest amounts of actions in all three categories (experimentation, analysis, and information seeking), as well as the lowest learning gains from the prediction to the reflection questions. The results suggest that this group may have been distracted or may have misunderstood how to use the simulation.

Table 1: Description of learning pattern levels

Level	Group	Pair Proportion	Actions			Learning Gains
			Experimentation	Data Analysis	Information Seeking	
1	Low Activity	67% (N=35)	Low	Moderate	Low	Low
2	Random Interaction	15% (N=8)	High	Moderate	High	Low
3	High Analysis	13% (N=7)	Moderate	High	Low	Moderate
4	Tasked Exploration	3% (N=2)	Low	Moderate	High	High

In contrast, the Random Interaction group is characterized by its high number of experimental runs and low learning gains (see Figure 2). Our preliminary findings from video and classroom observations reveal that this group appeared to make random actions with the simulation, such as running the simulation multiple times without changing the required variables, rather than carefully designing their experiments toward the learning goals of the activity. Consistent with the findings from previous research (e.g., McElhaney & Linn, 2011), such actions led to low improvement in their understanding of the target concepts.

The High Analysis group is characterized by its high frequency of analysis actions, moderate experimentation runs, and moderate learning gains. The results indicate that as students repeatedly analyze data

collected, they might have more opportunities to reflect on the key scientific ideas and revise their initial understanding of the concepts. Such processes could have helped them develop a better understanding of the abstract concepts of energy and matter in photosynthesis.

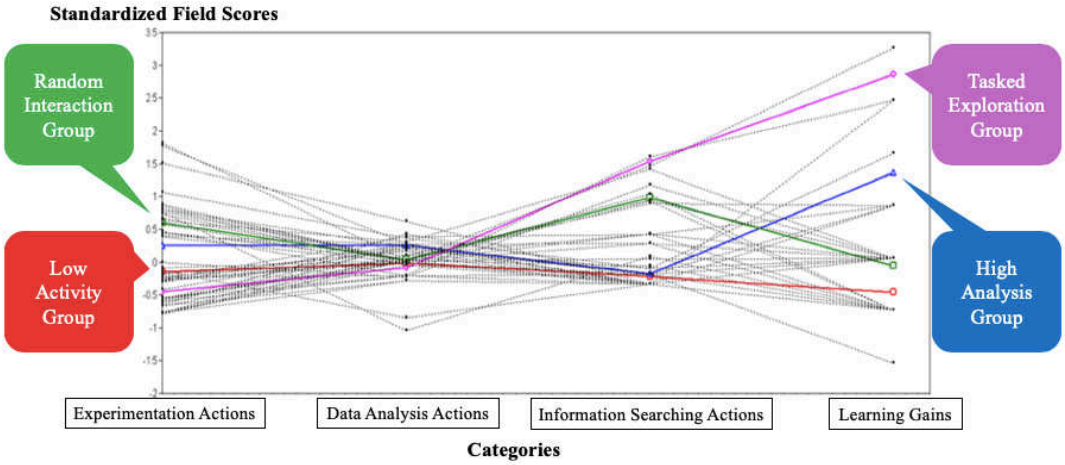


Figure 2. Groups of action frequencies for three categories related to learning gains.

The Tasked Exploration group is characterized by its low experiment runs, high frequency of information seeking, and high learning gains. Although they show less frequent experimentation actions than the High Analysis group, they appear to engage in purposeful investigations by manipulating the key variables required to answer a targeted hypothesis. It is possible that as they obtain additional information about scientific terms (e.g., light energy) and how to navigate the simulation, they developed more effective strategies to successfully complete their investigations.

Furthermore, the SPM analysis reveals unique sequential action patterns within each group (see Table 2). While the Low Activity and the High Analysis group demonstrate patterns of consistent experiment runs, the Random Interaction group only shows patterns of basic necessary functions to complete the task, such as reading the instructions and submitting the analysis response (e.g., IH, Sa, and S). Compared to other groups, the Low Activity and High Analysis groups appear to be more inclined towards utilizing the default variable settings for their experiment runs. However, the SPM analysis shows that the High Analysis group continues running experiments, while the Low Activity group immediately attempts to reach analysis.

Table 2: Sequential pattern mining in groups

Action (Detailed)			Group	Sequence	Support*
IH	Beginning: Instruction and Hypothesis		Low activity group	IH → 1S	0.916
R(XX)	Experiment (Energy Setting, Matter Setting) (e.g., R(LL) – (Low Sunlight, Low # Reactants)			IH → 1S → S	0.833
R()a	After experiment, ready to analyze			IH → R(LL)a → 1S	
R()f	Run experiment fast		Random interaction group	IH → 1S → S	1
R()m	Pause and continue animation during experiment run			IH → Sa → S	
Re	Reset experiment			IH → 1S → Sa → S	
1S	First submission before analysis		High analysis group	IH → S	1
Sa	Saving analysis			IH → 1S → S	
S	Submitting analysis			R(LL) → S	0.875
				IH → 1S → Sa → S	
				IH → Sa → S	
				IH → B → S	

**Note: The support's lower confidence bound is 0.8.*

Conclusions and implications

This study shows the value of FMM and SPM to identify both productive and unproductive interaction patterns when eighth-grade students conduct experiments using a simulation. Although these unsupervised ML

techniques are widely used in the fields of statistics and computer science, they have not been used to explore middle school students' learning patterns within simulation environments. The findings of this study show that FMM can categorize students into different learning groups based on the frequency of actions in the simulation, as well as the improvement in their understanding of scientific phenomena. When supported by SPM, unique subsequential patterns of each group can even be detected. Such findings can inform how to design tailored scaffolding for engaging students in effective interactions using simulations.

Given the exploratory nature of this study, there are several limitations to be addressed in future research. First, given that our study analyzed log data from only 52 pairs and FMM is sensitive to parameters of selected models (Melnykov & Maitra, 2010), future research should involve a larger number of students to obtain stable clustering results. Second, the simulation used in our study provided carefully designed scaffolding to help pairs engage in science practices while exploring the concepts of energy and matter in photosynthesis. Students may have different interaction patterns when using more open-ended simulations with limited scaffolding or when exploring different scientific phenomena (e.g., physics). Future research should investigate different types of simulations with varying levels of scaffolding, variables, interactivity, and scientific concepts.

References

Amershi, S., & Conati, C. (2009). Combining unsupervised and supervised classification to build user models for exploratory learning environments. *Journal of Educational Data Mining*, *1*(1), 18-71.

Baker, R. S., & Inventado, P. S. (2014). Educational data mining and learning analytics. In J. A. Larusson, & B. White (Eds.), *Learning Analytics: From Research to Practice* (pp. 61-75). New York, NY: Springer.

Bernardini, A., & Conati, C. (2010). Discovering and recognizing student interaction patterns in exploratory learning environments. In V. Aleven, J. Kay, & J. Mostow (Eds.), *Proceedings of the 10th International Conference on Intelligent Tutoring Systems* (pp. 125-134). Berlin, Heidelberg: Springer.

Gobert, J. D., Sao Pedro, M. A., Baker, R. S., Toto, E., & Montalvo, O. (2012). Leveraging educational data mining for real-time performance assessment of scientific inquiry skills within microworlds. *Journal of Research in Science Teaching, 41*(7), 748-769.

Kanari, Z., & Millar, R. (2004). Reasoning from data: How students collect and interpret data in science investigations. *Journal of Research in Science Teaching, 41*(7), 748-769.

Khalid, S., & Prieto-Alhambra, D. (2019). Machine learning for feature selection and cluster analysis in drug utilization research. *Current Epidemiology Reports, 6*(3), 364-372.

McElhaney, K. W., & Linn, M. C. (2011). Investigations of a complex, realistic task: Intentional, unsystematic, and exhaustive experimenters. *Journal of Research in Science Teaching*, *48*(7), 745-770.

McLachlan, G. J., & Peel, D. (2000). *Finite Mixture Models.* Hoboken, NJ: John Wiley & Sons.

Melnykov, V., & Maitra, R. (2010). Finite mixture models and model-based clustering. *Statistics Surveys, 4*, 80-116.

NGSS Lead States. (2013). *Next Generation Science Standards: For States, By States*. Washington, DC: The National Academies Press.

Plass, J. L., Milne, C., Homer, B. D., Schwartz, R. N., Hayward, E. O., Jordan, T., ... & Barrientos, J. (2012). Investigating the effectiveness of computer simulations for chemistry learning. *Journal of Research in Science Teaching, 49*(3), 394-419.

Qiao, X., & Jiao, H. (2018). Data mining techniques in analyzing process data: a didactic. *Frontiers in Psychology, 9*, 2231.

Scalise, K., Timms, M., Moorjani, A., Clark, L., Holtermann, K., & Irvin, P. S. (2011). Student learning in science simulations: Design features that promote learning gains. *Journal of Research in Science Teaching, 48*(9), 1050-1078.

Shih, B., Koedinger, K. R., & Scheines, R. (2010). Unsupervised discovery of student learning tactics. In: R. Baker, A. Merceron, P. Pavlik Jr. (Eds.), *Proceedings of the 3rd International Conference on Educational Data Mining* (pp. 201-210). Pittsburgh, PA: Educational Data Mining.

Stieff, M. (2011). Improving representational competence using molecular simulations embedded in inquiry activities. *Journal of Research in Science Teaching*, *48*(10), 1137-1158.

Zaki, M. J. (2001). SPADE: An efficient algorithm for mining frequent sequences. *Machine Learning, 42*(1), 31-60.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. 1552114.