Microgenetic Designs for Educational Data Mining Research

Taylor Martin[1]
Ani Aghababyan[3]
Utah State University
Instructional Technology and Learning Sciences
2830 Old Main Hill
Logan, UT 84322-2830
taylor.martin@usu.edu

Nicole Forsgren Velasquez[2]
Jason Maughan[4]
Utah State University
Management of Information Systems
3515 Old Main Hill
Logan, UT 84322-3515
nicolefv@gmail.com

Philip Janisiewicz[5]
University of Texas at Austin
Curriculum and Instruction
1 University Station D5700
Austin, TX 78712
pjanisiewicz@gmail.com

ABSTRACT
Educational Data Mining (EDM) methods can expand the reach of microgenetic research. This paper presents an example of our pilot work using microgenetic analysis in the context of fraction game data, where we characterize student activity based on clustered sequences of actions. We cluster sequences by the similarity between them, calculated using optimal matching techniques.

1. INTRODUCTION

Microgenetic research investigates processes of learning ([8]; [11]), rather than simply focusing on products of learning. Three main elements distinguish microgenetic research designs: 1) studies occur when the topic is likely to learned, 2) observations of learning behavior are dense, and 3) analysis is conducted on an instance by instance basis [6]. To date, the grain size for these studies has been fairly large, and the number of time points has been relatively small. These elements can be greatly improved using EDM methods. A few researchers have begun expanding microgenetic methodology using EDM methods (e.g., [2]; [4]; [5]) but this work is still in early stages. In addition, many EDM researchers use methods and conduct analyses that could be productive for microgenetic research (even if they do not place their work within the microgenetic paradigm). Some of these approaches include process and sequence mining, some uses of hidden markov models, and dynamic bayesian networks.

2. REFRACTION

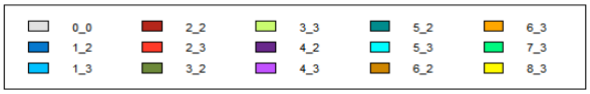
Third grade students (approximately 8–9 years old) played Refraction (http://play.centerforgamescience.org/refraction/site/), an online game based on fraction learning through splitting. In the game level used for this study, students create laser beams of 1/6 and 1/9 using a combination of 1/2 and 1/3 splitters. Students played the level twice: once at the start of gameplay (the prelevel) and again after playing the series of game levels (the postlevel). As students could stop play at any time, we had uneven numbers who completed the prelevel (N = 3,258) and the postlevel (N = 1,127).

3. ANALYSIS

Our unit of analysis is a "board state," or the configuration of the mathematical pieces of the game after a student makes a change. The two attributes of a board state we included in our analysis were initial splitter used (1/2 splitter or 1/3 splitter) and node depth. Solving the level requires starting with a 1/3 splitter, so the initial splitter variable indicates the quality of the board state. Node depth is the number of nodes, or levels of splitting, there are on a board state.

We employed the Needleman-Wunsch algorithm for optimal matching in R using the package TraMineR version 1.8-8, ([7]; [10]). This algorithm computes the "cost" of transforming one sequence into another based on insertions, deletions or substitutions. We set all costs equally at 1 as our events are all of the same type. To account for the discrepancies in our sequence lengths (prelevel range 1-82; postlevel range 1- 140), we used Abbott's normalization approach to standardize optimal matching distances ([1]; [7]).

We then used the distance matrix generated with optimal matching in a hierarchical cluster analysis [9] using Ward's, single linkage, and weighted average methods. We evaluated the number of clusters using dendrograms, and referenced group membership to ensure no clusters were too small. Finally, we inspected a visualization of each cluster solution for interpretation. The solution using Ward's analysis with seven clusters performed best (See Figures 1-6).



The first number in the pair is the node depth. The second number is for the 1/2 or 1/3 initial splitter.

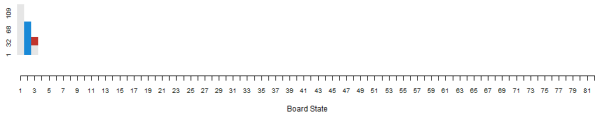


Figure 1. Minimal.

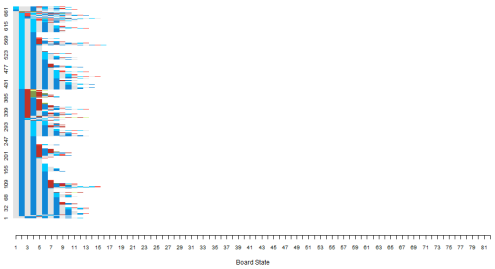


Figure 2. Halves.

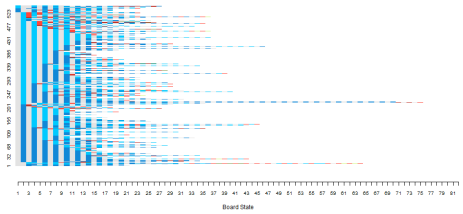


Figure 3. Exploring Halves.

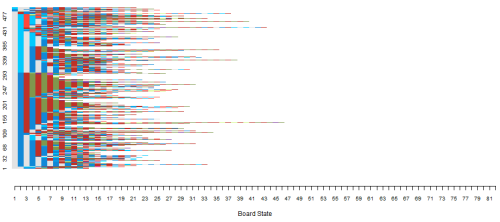


Figure 4. Exploring.

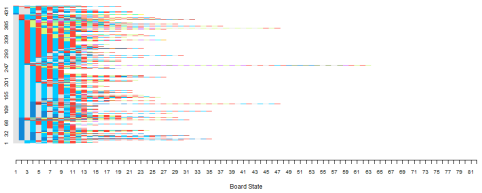


Figure 5. Exploring Thirds.

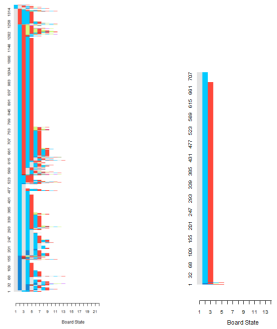


Figure 6. Thirds (left) and Efficient (right).

4. RESULTS

a. *Minimal* (N prelevel = 129; N postlevel = 4): very short sequences, very low node depths, and all 1/2 initial splitters.

b. *Halves* (N prelevel = 651; N postlevel = 26): medium sequences, shift from low node depth to higher, shift from mostly 1/2 initial splitter to some 1/3 initial splitters, show "reset" pattern, or clearing laser and starting over.

c. *Exploring Halves* (N prelevel = 536; N postlevel = 14): very similar to the Halves cluster, except longer sequences.

d. *Exploring* (N prelevel = 341; N postlevel = 165): greater exploring: many long sequences, try higher node depths, use both the 1/2 and 1/3 initial splitters.

e. *Exploring Thirds* (N prelevel = 359; N postlevel = 90): very similar to the Exploring cluster, except mostly 1/3 initial splitters.

f. *Thirds* (N prelevel = 859; N postlevel = 490): relatively short sequences, mostly node depth of 1 or 2, mostly 1/3 splitter.

g. *Efficient* (N prelevel = 383; N postlevel = 387): very short sequences, nearly all sequences identical: the start state, a state with node depth of 1 and 1/3 initial splitter, and a state with node depth of 2 and 1/3 initial splitter.

Most students, regardless of cluster membership on the prelevel, were in the Thirds or Efficient clusters on the postlevel (see Table 1). The marginal homogeneity nonparametric test for related samples of ordinal data [3] showed that this change was significant ($p < .001$).

Table 1. *Change in Cluster Membership From Pre- to Postlevel: Percentage of Students*

				Postlevel			
Prelevel	Minimal	Halves	Exploring Halves	Exploring	Exploring Thirds	Thirds	Efficient
Minimal	0%	3%	0%	23%	3%	40%	31%
Halves	0%	6%	3%	18%	9%	45%	18%
Exploring Halves	1%	3%	1%	21%	13%	42%	20%
Exploring	0%	1%	2%	12%	10%	50%	26%
Exploring Thirds	0%	0%	2%	21%	10%	43%	24%
Thirds	1%	1%	1%	12%	7%	44%	35%
Efficient	1%	2%	0%	6%	4%	40%	47%

Students in the Thirds and Efficient clusters were more likely to succeed on the both the pre- and postlevels than those in the other clusters; prelevel $\chi^2(1,6) = 1353.39$; $p < .001$; postlevel $\chi^2(1,6) = 605.32$; $p < .001$.

5. CONCLUSIONS

While this case demonstrates the utility of this approach as a microgenetic method in EDM, our next steps will be to extend this method to examine change over days or weeks of a learning event, to further test the utility of this method.

6. REFERENCES

[1] A. Abbott and A. Hrycak. Measuring resemblance in sequence data: An optimal matching analysis of musicians' careers. American journal of sociology, 96(1), 1990.

[2] R.S.J.d. Baker, A. Hershkovitz, L.M. Rossi, A.B. Goldstein, S.M. Gowda. Predicting Robust Learning With the Visual Form of the Moment-by-Moment Learning Curve. Journal of the Learning Sciences, 22 (4), 639-666, 2013.

[3] W. Barlow. Modeling of categorical agreement. In P. Armitage & T. Colton (Eds.), The encyclopedia of biostatistics (pp. 541-545). New York, NY: Wiley, 1998.

[4] M. Berland, T. Martin, T. Benton, C. Petrick Smith, & D. Davis. Using Learning Analytics to Understand the Learning Pathways of Novice Programmers. Journal of the Learning Sciences, 22(4), 564–599, 2013.

[5] P. Blikstein, M. Worsley, C. Piech, A. Gibbons, M. Sahami, & S. Cooper. Programming Pluralism: Using Learning Analytics to Detect Patterns in Novices' Learning of Computer Programming. International Journal of the Learning Sciences Special Issue on Learning Analytics, 2013.

[6] L. K. Fazio and R. S. Siegler. Microgenetic learning analysis: A distinction without a difference. Human Development, 56(1): 52-58, 2013.

[7] A. Gabadinho, G. Ritschard, M. Studer, and N. S. Müller. Mining sequence data in r with the traminer package: A users guide for version 1.2. Geneva: University of Geneva, 2009.

[8] D. Kuhn. Metacognitive development. Current directions in psychological science, 9(5): 178-181, 2000.

[9] M. Lorr. Cluster analysis for social scientists. Jossey-Bass San Francisco, 1983.

[10] S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of molecular biology, 48(3): 443-453, 1970.

[11] R. S. Siegler and K. Crowley. The microgenetic method: A direct means for studying cognitive. American Psychologist, 46(6), 606, 1991.