labelling help, and explanatory captions, I did not know what I was looking for or at in the three different student designs. As a result, I was unable to make comparisons from one solution to the other.

« 9 » I see in Figure 1 there is a reflection space for students to make comments on their own work using multiagent modelling. I wish we had learned more about these. In constructionist projects, all participants should be acknowledged as actors in research and learning. They are all part of the same narrative. In an exploratory project like this one, student voices are essential data and must be recorded and reported on. Why bother to include a space for student comments and then not include their observations in the final results?

Narrative wrap-up

« 10 » I said at the outset of my commentary that the article by Hjorth and Wilensky is important. But perhaps I should have said *potentially* important: potentially, because although important things are hinted at, they are never quite revealed. The technical tool certainly warrants attention, but without being privy to some of the hows and whys mentioned previously, I found it difficult to evaluate the nature of concept changes being explored. Here we have no narrative account of the methodology either from the researchers' point of view or from the students'. Without this complementary perspective, we are getting an incomplete picture of the experiment. For a very good introduction to the benefits of narrative approaches, see Brian Schiff (2017). The best methodology for constructionists should include both qualitative and quantitative techniques in the same narrative.

References

Behrens J. (1997) Principles and procedures of exploratory data analysis. Psychological Methods 2(2): 131–160.

Benzécri J.-P. (1982) Histoire et préhistoire de l'analyse des données, Dunod, Paris.

Schiff B. (2017) A new narrative for psychology. Oxford University Press, New York.

Özdemir G. & Clark D. (2007) An overview of conceptual change theories. Eurasia Journal of Mathematics, Science & Technology Education 3(4): 351–361.

James E. Clayson is Professor Emeritus at the American University of Paris (AUP), where he has taught applied mathematics and visual thinking for thirty years. He specializes in building computational environments where liberal arts undergraduates can explore the power of building personal models that link the visual, the qualitative and the quantitative. Jim was educated at MIT, the University of Chicago and the School of Oriental and African Studies (University of London).

Received: 23 May 2019
Accepted: 13 June 2019

Machine Learning and the Perils of Prolific Pattern Finding

Bruce Sherin
Northwestern University, USA
bsherin/at/northwestern.edu

> **Abstract** • Horth and Wilensky have taken an important first step in introducing a new method for capturing changes in student thinking, one that draws from the field of machine learning. However, I argue that there is much work to be done by educational researchers, as we seek to understand how best to apply methods from machine learning, and to appropriately interpret the results they produce.

Introduction

« 1 » Arthur Hjorth and Uri Wilensky's target article describes work that makes multiple contributions to educational research. These contributions include, for example, the development of a novel simulation environment and an associated social studies curriculum, both of which are well grounded in prior research. As the authors indicate (§1), the design of curricula for social studies is an area that is understudied. However, the authors clearly intend for the main contribution of the target article to be the presentation of a novel analysis method. Thus, it is there that I focus this commentary.

« 2 » The authors' analysis focuses on student responses to one short written prompt that students were given before and after the curricular activities (§28). In response to this prompt, each student wrote a short response. These responses were then coded by the authors in terms of a set of "causal nodes." Finally, the authors used Association Rule Mining (ARM) to identify a set of rules that connect the causal nodes. It is this last piece – the use of ARM – that is the most novel aspect of this approach.

« 3 » There is much to be admired in this novel approach. In essence, what we are seeing is the use of a technique from machine learning to create a new type of *quantitative* method, one that is intended for use with the sort of data that is often produced in studies of student learning. This new method is akin to traditional quantitative methods in the loose sense that multi-step mathematical algorithms are used to reduce complex data to a relatively small number of values. However, it is unlike most traditional quantitative methods in an important respect: methods from machine learning are intended for data exploration and pattern discovery, rather than hypothesis testing. In this way, these new methods may surprisingly be better suited to the aims of researchers who have, in the past, found themselves more drawn to qualitative methods than traditional quantitative methods (Sherin, Kersting & Berland 2018).

« 4 » However, because the use of this sort of data in our field is new, there is little in the way of conventions, established ground rules, or rules of thumb to guide the authors' analysis, and our own reading of their work. We do not yet know, for example, how much data is sufficient for an analysis, what counts as a "striking" pattern, and what values of measures are small, and which are large. Without these existing conventions and rules of thumb, the authors have had no choice but to invent them; when faced with a choice, they had to make their best guess.

« 5 » Given the newness of the approach, this is exactly what the authors should be doing. However, it is worth taking the time to reflect on the authors' decisions, and to start laying the groundwork for shared conventions and rules of thumb. This will also help us to decide what we can conclude from the particular results presented in the target article.

Identification of rules

« 6 » I begin with the process whereby the authors identified "interesting" rules. In this regard, the key passages are in §48 and §49. Here we are told that the authors initially identified 433 rules. They then narrowed the set to those 47 rules that appeared in at least seven of the 41 total responses in both the pretest and posttest results. Next, they used the *lift* measure to further narrow the set of interesting rules to just five. In particular, they identified rules with a change of at least .15 in lift between the pretest and posttest responses. To an untrained reader, this sequence of steps might seem radically ad hoc.

« 7 » Now, the literature is replete with guidance for identifying interesting association rules (e.g., Bayardo & Agrawal 1999; Lavrač, Flach & Zupan 1999; Omiecinski 2003). However, in virtually all cases, these techniques use multiple measures, and require the ad hoc introduction of cut-off values. So, in that regard, the authors of the target article are in good company.

« 8 » However, the research reported in the target article is unlike most published uses of association rules in that the amount of data is quite small. So, perhaps more than prior work, we need to be perspicacious in our acceptance of any patterns uncovered; we need to be careful that we are seeing meaningful patterns, and not just noise. Indeed, thinking about the 47 rules that survived the first cut, we would expect some inherent variability in lift from pre- to posttest, and thus some random variability in the deltas that are observed.

« 9 » Let us look more closely at some of the values of lift that are identified. For the five interesting rules, lift on the posttest responses ranges from 1.17 to 1.37. It is a little hard to know what to make of these numbers. Are these relatively large numbers? Suppose a rule were supported seven times on the posttest responses, but the rule would have been predicted to occur six times, if the causal nodes had been independent. This gives a lift of 1.17. Viewed in this light, the observed values of lift on the posttest do not seem very large, given the small scale of the data corpus. Similarly, the changes in lift from pre to posttest do not seem overly dramatic.

« 10 » Thus, it is hard to know whether the rules uncovered by the ARM analysis represent meaningful patterns in the data, as opposed to a statistical fluctuation. Furthermore, this is not a problem that is easily solved by introducing some sort of hypothesis testing (Lallich, Teytaud & Prudhomme 2007).

« 11 » One way to deal with these difficulties is to inspect the patterns that we uncover, and to look to see if those patterns add up to a coherent story, and one that makes sense given the context, as well as human inspection of the raw data. This sort of inspection is, indeed, evident in the target article, such as in the passages in §50 through §52. However, all of the rules involved just nine causal nodes, all of which were closely related to the explanation task given to the students. Thus, it seems reasonable to worry that we might still be able to tell a reasonable story about any rules involving this set of causal nodes, in any combination.

« 12 » This is perhaps the central danger in applying these methods from machine learning: namely, the methods are prolific in uncovering patterns, but they are perhaps too prolific. It is a problem faced by most attempts to apply machine learning to the social sciences, including my own work (Sherin 2013). Thus, my first and primary question is: How can we avoid the perils of prolific pattern-finding? (Q1)

Interpretation of the rules

« 13 » Assuming that we accept that the identified rules reflect meaningful and important patterns in the data, we are then faced with the task of understanding what these rules mean. The authors do attempt to assist the reader in interpreting the rules, for example in §48. But this is an inherently subtle business, and it is worth emphasizing that the use of ARM in the target article is of a very different sort than typical uses of ARM in machine learning.

« 14 » The authors tell us they are interested in *causal* reasoning. Furthermore, ARM produces rules of the form {a, b}→{c, d}, where, in this case, a, b, c, and d are causal nodes. However, I do not think there is any sense in which the arrow in the rule can be interpreted causally. The rules reflect cooccurrence and predictiveness; they do not necessarily tell us anything about the structure of the explanation that is given. But then how *should* we interpret the arrow? Does it tell us anything about the meaning of the rule?

« 15 » Note, further, that at the time of the pretest, some of the 5 interesting rules have a value of lift that is less than one. This means that, at the time of the pretest, the causal nodes on the right side of the rule are less likely to occur than they would be if their occurrence were simply independent of the nodes on the left. Does this indicate some sort of negative rule? These observations, together, lead to another question: How should we interpret the rules? In particular, what is the relationship between identified rules and the students' causal explanations? (Q2)

Conclusion

« 16 » The authors have taken an important first step in introducing a new method for capturing changes in student thinking, one that is drawn from the field of machine learning. But, as I hope to have illustrated in this commentary, there is much work to be done within the field of educational research as we seek to apply these new methods, and to appropriately interpret the results they produce.

References

Bayardo R. J. & Agrawal R. (1999) Mining the most interesting rules. In: Fayyad U., Chaudhuri S. & Madigan D. (eds.) Proceedings of the Fifth ACM SIGKDD international conference on Knowledge Discovery and Data Mining (KDD '99). ACM, New York: 145–154.

Lallich S., Teytaud O. & Prudhomme E. (2007) Association rule interestingness: Measure and statistical validation. In: Guillet F. J. & Hamilton H. J. (eds.) Quality measures in data mining. Springer, Berlin: 251–275.

Lavrač N., Flach P. & Zupan B. (1999) Rule evaluation measures: A unifying view. In: Džeroski S. & Flach P. (eds.) Inductive logic programming. Springer, Berlin: 174–185.

Omiecinski E. R. (2003) Alternative interest measures for mining associations in databases. IEEE Transactions on Knowledge and Data Engineering 15(1): 57–69.

Sherin B. L. (2013) A computational study of commonsense science: An exploration in the automated analysis of clinical interview data. Journal of the Learning Sciences 22(4): 600–635.

Sherin B. L., Kersting N. B. & Berland M. (2018) Learning analytics in support of qualitative analysis. In: Kay J. & Luckin R. (eds.) Rethinking learning in the digital age: Making the learning sciences count. Proceedings of the 13th International Conference of the Learning Sciences (ICLS 2018). Volume 3. International Society of the Learning Sciences, London: 464–471.

Bruce Sherin is a professor in the School of Education and Social Policy at Northwestern University. His work has focused on conceptual change in science – the process through which our everyday understanding of the natural world changes over time and with instruction. More recently, his work has focused on some methodological issues in the study of conceptual change. As part of this newer work, he applies techniques from natural language processing to interview protocols. He is also the designer and developer of Tactic Text, which is a web-based text-mining environment designed for qualitative data analysts.

Received: 26 May 2019
Accepted: 10 June 2019

The Role of Debugging in Knowledge Construction

José Armando Valente
State University of Campinas (UNICAMP), Brazil
jvalente/at/unicamp.br

> Abstract • Hjorth and Wilensky's target article describes two important tools for helping students debug their conceptual misconceptions: the NetLogo model, and results from Association Rule Mining. In this commentary, I focus on these tools' contributions to the debugging process, and the way they allow students to improve their conceptual knowledge.

« 1 » Digital technology can be an important resource for supporting knowledge-construction processes. Activities using these technologies help make the concepts and strategies learners use to solve problems explicit. In addition, based on the results students get, students can reflect on what they have done in terms of the concepts and strategies used, and, with this information, debug misconceptions they might have. Debugging, therefore, is an important step towards improving and constructing new knowledge. I found it particularly interesting that the present article, though it uses technologies in knowledge-constructing activities, does not mention the debugging process. Thus, I would like to discuss debugging in two contexts: with respect to the NetLogo model, and with respect to the ARM findings. Finally, I will address the teacher's role in debugging activities.

« 2 » The debugging concept was created within the computational area (Sussman 1975), and is a fundamental notion in the development of products using digital technologies. It is the process by which a programmer finds and corrects errors in a computer program. According to this concept, the error is seen as a bug, and debugging is the act of eliminating the bug. The use of debugging in education was developed by Seymour Papert (1980), who saw a learning opportunity in computer programming. Finding and fixing bugs is a unique context for learners to understand what they are doing and thinking.

« 3 » Although debugging was related to programming, this concept has been appropriated by other areas, such as the production process. The Japanese created the *kaizen* (改善) concept, referring to the constant improvement of ideas, working conditions, and operations performance (Imai 1986). These continuous improvements lead to the creation of new procedures for performing tasks and, as a result, become the source of knowledge generation. Thus, debugging can be seen as the motor for learning, which can be applied in any circumstance or domain.

« 4 » However, it is not enough to create opportunities for students to reflect upon what they have done. It is also unlikely that they will be able, by themselves, to come up with appropriate concepts to fix their problems. The role of the teacher is important, as well as how she approaches the learning situation.

Debugging related to the NetLogo model

« 5 » The use of the NetLogo model created the opportunity for the students to design interactive learning activities. In §25 the authors mention that the model lets students interactively articulate a casual explanation, test it, and potentially revise it. From §26 through §29, it is possible to understand that the articulation step refers to design and implementation phases; testing refers to improving and assessing the model's success, part of the data analysis phase; and revising is related to the second part of the data analysis phase.

« 6 » The article does not go into detail regarding the revision aspect of the interactive learning activity. In §56, the authors say that they "speculated" that collaborative simulation might have influenced the students' thinking. Thus, I wonder, is it possible to record the revisions students made during model simulation so that the analysis of these changes can contribute to an understanding of how students debug their misconceptions related to the question the target authors focus on in §28? (Q1) If we are trying to understand students' conceptual changes, the information provided by the revision process is important data to be analyzed and used for debugging purposes.

Debugging related to ARM findings

« 7 » The ARM findings can be seen as important information regarding patterns in changes during students' construction of causal nodes. This can be very helpful when analyzing students' conceptual change, both at the individual and at the classroom level. Yu Guo, Wanli Xing, and Hee-Sun Lee mentioned that ARM results allow for teachers to have important insights regarding students' knowledge. According to these authors "Such insights can be used to improve teaching and learning, and to inform the design of future instructional interventions" (Guo, Xing & Lee 2015: 268). Other authors (e.g., Berland, Baker & Blikstein 2014) have also mentioned the use of educational data mining for debugging activities.

« 8 » If we analyze the ARM results and the students' responses, it is possible to identify that there are still misconceptions regarding geographic/spatial location, commuting, and desires or possibilities related to income. Thus, the question is: How can teachers use the ARM results to help students debug their misconceptions? (Q2)