Multimodal Learning Analytics for the Qualitative Researcher

Author Name, Institution, Email

Abstract: The area of learning analytics is often viewed as a tool for supporting quantitative analysis. Based on previous research, this association between quantitative analysis and learning analytics does seem to be the trend. However, certain researchers have proposed the use of multimodal learning analytic techniques as a viable and valuable contribution to more qualitative research methodologies. This paper examines that idea by trying to use the output from an algorithm that learns discriminating features, as the starting point for video observations. Ultimately, the analysis suggests that there is utility in leaning on machine learning to help identify important patterns in the data, provided that those patterns are contextualized and studied using the original video data. Additionally, the work makes clear the need for better tools for conducting these types of multimodal analyses.

Introduction

The past decade has seen the emergence and expansion of research at the intersection of data mining and education research. In particularly, the fields of educational data mining, and learning analytics have gained increased traction as novel ways for studying learning (Baker & Yacef, 2009; Martin & Sherin, 2013). While educational data mining and learning analytics have traditionally focused on data derived from cognitive tutors, learning management systems and other computer mediated experiences, multimodal learning analytics (Blikstein & Worsley, 2016), a growing subfield of learning analytics, has been proposed to study more collaborative, human-to-human interactions. Moreover, researchers within this sub-field have positioned multimodal learning analytics as a set of techniques that have relevance for qualitative researchers. The goal of this paper is to examine this claim. In particular, the analysis will use techniques from multimodal learning analytics to learn an algorithm that can distinguish between students who recorded positive learning gains, from those who recorded negative learning gains. Key elements of the algorithm are then used to extract segments of video that correspond to behaviors that positively and negatively correlate with learning gains. Finally, these video segments are used to qualitatively draw inferences about how students learned from the experience.

Prior literature

Multimodal learning analytics uses non-traditional sensory data (gestures, gaze, speech, emotions, digital pen traces, etc.) to study and model student learning. As with any sub-field involving "analytics," there is a presumed reliance on computational techniques, as well as an assumed level of automaticity. Despite this assumed automaticity, multimodal learning analytics is heavily influenced by human intuitions with many researchers adopting mixed-method approaches (Prieto & Rodriguez-Triana, 2017). From the design and implementation of the specific algorithms, to the labelling of training data, human intelligence plays is essential. However, the mere inclusion of human input at some point during the analytic process does not truly speak to the disposition of qualitative research. Instead, we seek is a high level of intimacy with the data.

Worsley & Blikstein (2014) provide an interpretation of research that combines learning analytics and qualitative analysis. Within this study, the researchers hand-coded student actions into one of five possible states (build, test, adjust, undo or plan). They then used machine learning to identify patterns within each user's processes, and studied how those processes differed by the participant's level of expertise.

Prieto & Sharma (2017) and Liu & Stamper (2017) take an approach that is motivated by reducing the amount of human coding required. Prieto & Sharma (2017) leverages four measures for cognitive load in the context of mobile eye tracking in classrooms. They use those measures to identify regions of the video that have high agreement, and, in this way reduce the amount of data that they analyze. Liu & Stamper (2017) take a similar approach in identifying regions of video that have been tagged within a cognitive tutor. Their tool then allows them to automatically extract video segments that correspond to any number of important events.

The aforementioned papers offer potential paradigms for multimodal learning analytics to integrate with qualitative research practices. This paper presents another paradigm, that, in a sense, sits at the intersection of these two approaches. More specifically, the approach will rely on machine learning to identify characteristics or behaviors that seem to correlate with learning, and then generate a subset of videos that a human can study to better understand the semantics of the behaviors.

Methods

The analysis presented in this paper is based on a subset of the Engineering Design with Everyday Materials Multi-modal Dataset (Worsley, 2017). This dataset includes 54 students from a community college in the United States and features nearly 27 hours data. Participants worked in pairs to complete two engineering design tasks. The first task asked students to use one sheet of printer paper to construct a structure that could support one or more engineering textbooks at least three inches above a table. Students had six minutes to complete this task. In the second task, students had 10 minutes to build a structure that could support a mass of 0.5 lb. as high off the table as possible using limited household materials. This task involved similar engineering principles as the first task, but with greater variability in the materials, and with the added goal of height optimization. While scores for how well each structure performed are included in the dataset, they are not used within this analysis. Instead, this analysis focuses on learning gains. The learning gains are based on the student's identification of principles or mechanisms that confer stability to three example structures (a bridge, a ladder and an igloo). The recorded learning gains are the change in the proportion of a student's responses that refer to a structure's configuration or geometry (i.e. triangles, wide base, and symmetry). Referring to a structure's geometry contrasted with references to a structure's material (i.e. metal, wood). Students who increased the proportion of configuration/geometry-based principles on the post-test, relative to their pre-test, were classified as having learned. Explaining why some students seem to recognize the relative importance of geometry while others did not is the primary focus of this analysis. Along with the learning gains information, the dataset also includes multi-modal data. The focal data for this study is derived from the Xbox Kinect Sensor which generated skeletal tracking, head pose and audio data. For the sake of simplicity, this analysis focuses head pose, or, more generally, eye gaze, as derived from frontal images of each student. Head pose was calculated using OpenFace (Baltrušaitis, Robinson, & Morency, 2016). The section that follows briefly describes how the data was used.

Analytics summary

The initial dataset of 54 students was reduced to ten students due to missing head pose data for several students. The head pose data is recorded as values for pitch, roll and yaw. These values correspond to looking up and down, left and right and tilting one's head. Values for yaw (looking left or right) were transformed such that they corresponded to looking away from one's partner or looking towards one's partner. This transformation aims to eliminate the positional dependence of seating arrangement, as students were seated one next to the other. Values for pitch were left unchanged, but were included within the analysis because they could proxy for identifying when a participant is looking more towards the materials, or elsewhere. Values for roll (tilting one's head left or right) were omitted because they did not have a clear significance in this analysis.

The next step was to automatically cluster the data. Before clustering**,** however, data underwent column-based normalization. Each value was divided by the standard deviation of the column to minimize bias. K-means (N=5) was used to cluster the data from both of the tasks. In this way, it is assumed that the postural enactments are sufficiently similar across the two tasks as to warrant a combined clustering step. After the clustering step, every time step, for each participant was associated with one of five possible clusters.

The clustered data was used to compute the proportion of time participants spent using each cluster across each task. This analysis focuses on cluster frequency from the second task.

The proportion of time in each cluster was used to train a decision tree classifier that learns cluster frequency values that correlate with positive and negative learning gains. Note: This process did not follow the usual machine learning convention of cross-validation or doing a training-testing split. The reason for this is an explicit interest in identifying the differences between those with differential learning gains in this dataset.

The nodes of the decision tree were used to identify which clusters would be worthwhile to examine via human coding. Specifically, a custom script was used to identify contiguous segments of video that included the cluster of interest. Extracted segments were a minimum of three seconds long. A random selection of no more than five video segments per participant were extracted for human observation.

Finally, a team of researchers performed open coding on the videos to determine how the videos from different clusters were semantically different from one another. Conducting this process involved repeated observation of the videos among the team of five researchers.

Results

A key, and somewhat surprising, result that emerged from looking at the output of the cluster frequency aggregation step, was that the cluster frequencies for several of the clusters perfectly correlated with learning. Table 1 indicates that participants with a c3 (cluster 3) proportion of less than 0.2 all recorded negative learning gains. Similarly, participants with larger c3 proportions recorded higher learning gains. On the other hand, lower c1 (cluster 1) scores seem to correlate with positive learning gains.

Table 1: Participant cluster frequencies and learning gains

USER	LEARNING	C0	C1	C2	C3	C4
10_1	-0.636	0.302	0.132	0.266	0.139	0.161
10_2	-0.500	0.320	0.192	0.261	0.129	0.098
9_1	-0.017	0.367	0.180	0.152	0.168	0.133
9_2	0.208	0.323	0.148	0.167	0.248	0.113
14_1	0.000	0.245	0.118	0.143	0.332	0.162
14_2	0.193	0.280	0.107	0.112	0.398	0.103
27_1	0.381	0.393	0.070	0.137	0.232	0.168
27_2	0.444	0.437	0.057	0.098	0.247	0.162
23_1	0.178	0.306	0.012	0.082	0.424	0.176
23_2	0.533	0.512	0.012	0.048	0.298	0.131

With the apparent observation that cluster 1 and cluster 3 frequency appear to correlate with student learning on this activity, the next step was to study the values that characterize those clusters and delve into the videos. Table 2 highlights the pitch and yaw values for each cluster.

Table 2: Cluster centroid values (positive pitch corresponds to up, positive yaw is towards their partner)

CLUSTER	0	1	2	3	4
PITCH	-0.72517	0.475809	-1.24574	0.391357	-2.00303
YAW	1.630276	0.331196	-0.09873	2.031234	1.652582

While the numbers suggest some difference between the clusters, the numbers are not of much benefit in determining how the different head poses would result in, or correlate with, differential learning gains on the activity. Hence, video coding was used to gain more detailed insights.

In studying the videos, the first step was to consider if the types of actions (e.g., building, planning, undoing, etc.) that students were taking in the different clusters were substantively different. However, no clear differences emerged across the different groups along this dimension. Instead, the key difference had everything to do with where the participants were looking, or, more importantly, where they were working. Looking at the figures (Figure 1) we see one potential difference between the two head poses. Namely, in cluster 1, the student's attention is focused on building higher off the table than in cluster 3. We saw this difference across a number of participants in the sample. The other key difference that we observed was that in cluster 1, more of the frames involved the student looking forward, while in cluster 3, the students spent more time looking toward their partner.

Figure 1. Images a,b,c,d (from left to right) provide examples of cluster 1 (a and c) and cluster 3 (b and d).

Discussion

In summary, the analysis above relied on machine learning and multimodal data to pinpoint behavioral differences between participants with differential learning gains during a set of engineering design activities. Identification of those behavioral differences emerged based on unsupervised clustering of head pose data. When we examined the numeric data for the salient differences, it was unclear as to what was occurring. One could certainly postulate plausible arguments about learner engagement based on those numbers, but this would ultimately have been unsatisfying. Instead, a closer look at the video data made it apparent that differences existed in the height at which students were building and/or designing, as well as in the extent of engagement with their collaborator. This may point to differences in how much students were focusing on the base of the structure, something that is important to conferring stability to a design, as well as non-content related issues of collaboration quality. Hence, the differences in pose, while represented in the form of two clusters, corresponded to a larger set of behaviors that necessitated thoroughly engaging the video.

The above analysis is not intended to provide the only explanation for the observed differential learning gains. Moreover, the qualitative analysis of the videos was in no way exhaustive. Instead, this work aims to provide an example of how to make use of multimodal learning analytic techniques for supporting qualitative analysis. The algorithm took care of the discrimination between those who learned and those who did not, while the task of making sense of those differences rested on the researcher.

An additional point of discussion is the noted challenges in doing this work. Even after collecting and synchronizing the data, both of which relied on custom developed computer programs, there was the task of extracting the salient features, and subsequently identifying a subset of videos to analyze via human inference. These steps involved custom Python scripts and a number of machine learning libraries. Adopting this strategy more broadly will likely require robust, interpretable, and easy to use analysis and visualization tools (Liu & Stamper, 2017; Rodríguez-Triana, Prieto, Holzer, & Gillet, 2017).

Conclusion

Data mining and computational techniques are gaining increasing prevalence within the education research community. However, these strategies and tools are primarily being utilized by researchers with more of a quantitative orientation. It is the goal of this paper to demonstrate one way for leveraging multimodal learning analytics in a more qualitative fashion. While there remains a need for more robust and easy to use tools, there is a profound opportunity for researchers to deeply examine their data through a different set of lenses. Furthermore, as more qualitative researchers look to these tools to support and facilitate their analysis, technology designers will be able to better tailor the tools to the needs of qualitative researchers.

References

Baker, R. S. J. D., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *JEDM-Journal of Educational Data Mining*, *1*(1), 3–17.

Baltrušaitis, T., Robinson, P., & Morency, L.-P. (2016). OpenFace: an open source facial behavior analysis toolkit. In *IEEE Winter Conference on Applications of Computer Vision*.

Blikstein, P., & Worsley, M. (2016). Multimodal Learning Analytics and Education Data Mining : using computational technologies to measure complex learning tasks, *3*(2), 220–238.

Liu, R., & Stamper, J. (2017). Multimodal Data Collection and Analysis of Collaborative Learning through an Intelligent Tutoring System. *LAK Workshops '17*, 1–6.

Martin, T., & Sherin, B. (2013). Learning analytics and computational techniques for detecting and evaluating patterns in learning: An introduction to the special issue. *Journal of the Learning Sciences*, *22*, 511–520.

Prieto, L., & Rodriguez-Triana, M. J. (2017). Towards Novel Researcher Tooling Based on Multimodal Analytics. In *Cross Multimodal Learning Analytics Workshop*.

Prieto, L., & Sharma, K. (2017). Scaling up Mobile Eye-Tracking Classroom Studies: Two Approaches. *Euroopean Conference on Technology Enhanced Learning, Eye Tracking Enchanced Learning Workshop Proceedings*.

Rodríguez-Triana, M. J., Prieto, L. P., Holzer, A., & Gillet, D. (2017). The multimodal study of blended learning using mixed sources: Dataset and challenges of the SpeakUp case. *CEUR Workshop Proceedings*, *1828*, 68–73. https://doi.org/10.1145/1235

Worsley, M. (2017). Engineering Design with Everyday Materials Multi- modal Dataset. *CEUR Workshop Proceedings*. Retrieved from http://ceur-ws.org/Vol-1828/paper-16.pdf

Worsley, M., & Blikstein, P. (2014). Analyzing Engineering Design through the Lens of Computation. *Journal of Learning Analytics*, *1*(2), 151–186.