Human-Centered Data Science for Educational Technology Improvement using Crowd Workers

Steven Moore[1], JohnStamper[2]
Carnegie Mellon University, HCII
StevenJamesMoore@gmail.com[1], jstamper@cs.cmu.edu[2]

Soniya Gadgil
Eberly Center for Teaching Excellence and Educational Innovation
soniyag@andrew.cmu.edu

ABSTRACT: Learning curves (LC) provide a concise way to visualize student learning over time. Analysis of these curves can identify which knowledge components (KC) might be misaligned or at the very least where a problem in the system exists. While beneficial to system and course improvement, this analysis is time consuming and can be taxing when hundreds of KCs are present. Utilizing crowd workers, LCs can be mapped to categories and rank ordered, indicating which need improvement the most. Leveraging the categorization and rankings from these workers, a finer grained grouping can be achieved that indicates which LCs need attention first and foremost. This creates a more efficient analysis, helps to maintain the iterative cycle of system and course improvement, and provides another step towards leveraging crowdsourcing for educational improvement.

Keywords: Crowdsourcing, Visual Analytics, Data Analytics, E-learning

1 INTRODUCTION

The proliferation of data on students interacting with online learning environments has enabled new opportunities for understanding student performance in recent years (Baker & Inventado, 2014). It enables the construction of models on how students progress through the learning process and assists in identifying the gaps in their knowledge. Building on these student models for the purpose of tracking student learning over time has been a key area of focus in the educational technology community for many years as well (Murray, 2003). Cognitive Tutors, such as those from Carnegie Learning, utilize student models and are adaptive to student knowledge by tracking the mastery of skills or knowledge components (KCs) (Fanscali et al., 2013). The models that map KCs are generally created with the help of subject matter experts and cognitive scientists. Unfortunately, these knowledge component models (KCMs) do not always correctly model skills, which can impede student learning. When a KCM for a cognitive tutor is incorrectly modeled, it can cause incorrect problem selection and waste valuable student time on skills they have already mastered.

Learning analytics can address this problem and presents an opportunity for continuous improvement of the models using data driven techniques (Stamper & Koedinger, 2011). At present, DataShop (Stamper et. al, 2010) has user interface affordances that utilize a new framework for learning curve (LC) categorization to assist in identifying areas of improvements in the student models of the educational technology. The analysis of these LCs to provide insights into student

models has been around for many years (Anderson, Conrad, & Corbett, 1989). In addition to using these curves to improve student models, the algorithmic use of fitting learning curves has been used to improve upon cognitive models used in intelligent tutoring systems (Cen et al., 2006). While this categorization can assist in identifying which KCs might be misaligned or incorrect in the KCM, the process is still time consuming.

The use of crowd workers is common with educational technology, but often in a way that leverages the workers or users specific content knowledge (Anderson, 2011; Weld et al., 2012). Recently, crowdsourcing has become increasingly popular for content development in the educational domain (Porcello & Hsi, 2013; Paulin & Haythornthwaite, 2016). We propose a workflow that takes a slightly different approach, utilizing crowd workers in a way that does not necessarily leverage their domain expertise or have them develop content in anyway, while still benefiting from their input. This proposed workflow will leverage crowd workers to help with a time consuming and often tedious part of LC analysis that is necessary for course and educational system improvements. We look to utilize crowd workers in order to both better categorize and to provide a priority-ordered ranking of the learning curves for a given dataset, so that the largest improvements can be made in the quickest time frame.

For this proposed workflow, crowd workers from Amazon's Mechanical Turk, known as turkers, will be recruited to review a set of learning curves. They will select which category each LC fits under and assign it a unique rank order, based on how much it needs to be improved. This ranking of the LCs will be made available to the workflow user, providing a priority view of which LCs to focus their limited time on. It will extend the categorization currently offered by DataShop, utilizing the LC images from the learning curve visualization component in LearnSphere.

Ultimately this workflow looks to be a first step in getting more towards the human-in-the-loop aspect for LearnSphere and leverage crowd workers for work in the learning sciences. We want to leverage the human judgement ability and classification to build upon the existing classification of LCs by DataShop and to make the analysis portion more efficient. This will ultimately assist in the continuous iterative improvement cycle needed in many educational systems.

2 WORKFLOW METHOD

2.1 Data Inputs

The input into this workflow comes from the learning curve visualization component. This LC component outputs a series of Portable Network Graphic (png) images that correspond to the LCs for each present knowledge component in the initial dataset. These images make up the file output of the LC component, which is the primary input into our proposed workflow. The current output size and file names for the images are appropriate for the workflow's needs. While the image file sizes are small, in order to keep the bandwidth and latency low, we suggest compressing the resulting file in a ZIP file to be used with our workflow. Our workflow can then unzip the images and use them for the model, however as it stands making use of the currently output file from the LC component is also functional.

These LC images are already anonymized regarding their content area, as the images are all titled "KC" followed by an incremental number, as shown in Figure 1. The png files are also similarly named, which assists with confidentiality as well as mapping the image to the corresponding KC that is used in the LC visualization component and any prior analysis ones. Additionally, the images have

the toggleable option to include their DataShop curve categorization next to the curve's title. While we suggest leaving this off, the proposed workflow could include it depending on what the user ultimately wanted to achieve or how they expect it to bias the crowd workers, if at all.

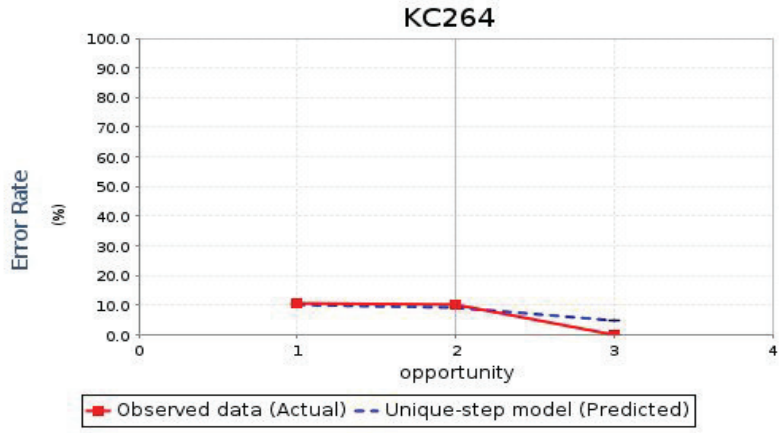

Figure 1: Example learning curve, with the assigned KC title

A second and optional input for the workflow is the XML output from the LC visualization component. This data can be used to provide additional information in the table columns that this workflow outputs. This accompanying data, along with the classification and rankings of the LCs, can provide a high-level view of how well the KCs are mapped via a concise tabular view. It is also trivial to map this XML data back to the corresponding LC, as the KC name and number joins the two. As a first pass and minimally viable workflow, the file name for the images contains enough information to construct the appropriate output for this model without this secondary input.

2.2 Workflow Model

Once the image files and optional corresponding XML data have been input into this model, the images need to be grouped for their presentation to the crowd workers. The workflow will have a configurable input detailing the size of these groups, which corresponds to how many LCs a turker will be reviewing. By default, we suggest a value of ten, as it requires a low amount of time and lends itself to having a commonly quantifiable ranking scale, in this case ranking from 1-10. The second configurable option offered is the grouping of LCs by categories, as labeled in the learning curve visualization component. With this option, enabled by default, the component will select LCs from the same category to present to the crowd workers when possible. For instance, when enabled ten LCs from the too little data category may be selected. If there are not enough for a given category, the component will fill in the rest with LCs from a different category, so that the assigned grouping count is always met. As a first pass for this workflow, each LC will only be reviewed by a single turker, unless LCs are needed to fill in the gaps for groups that do not meet the group size parameter.

With the LC images formed into their given groups, conforming to the two configurable parameters, the next step is to make the assignment, known as a HIT, for Mechanical Turk. Amazon offers a variety of free APIs that can be used to programmatically generate an assignment on the platform using different common programing languages. These APIs will be leveraged, along with a provided HTML template file, to embed the LC images so that the turkers can review them. The first part of the HIT will explain the task at hand, which involves turkers reviewing a series of graphs and

ranking them in terms of which need the most improvement. In this case, needing improvement corresponds to which LCs do not demonstrate learning or a good fit for the given KC. To provide these workers with a frame of reference and background information on these concepts, a brief yet informative exert will be used to explain LCs and their corresponding five categories. An example of such text can be found below (Stamper et al., 2010).

> "A learning curve visualizes changes in student performance over time. The line graph displays opportunities across the x-axis, and a measure of student performance along the y-axis. A good learning curve reveals improvement in student performance as opportunity count (i.e., practice with a given knowledge component) increases."

After the turkers read the HIT instructions and the exert regarding LCs and their categories, using concise language pulled from DataShop, they will be presented with five learning curves. Each of these LCs will distinctively fall into one of the five aforementioned categories, such as the LC for too little data having a single point or the LC for low and flat having five points that all remain in the 10-15% range. To establish a baseline for accuracy, the turkers will first be asked to categorize each of these five curves. In addition to categorizing them, they will be asked to rank the LCs in a unique order of 1-5, where 1 indicates the present LC that needs the least improvement (such as a good one) and 5 indicating an LC that needs the most improvement (such as still high). These LCs are to be ranked in comparison to one another, so all five will be proximally located near one another in the interface. This is done in order to gauge their accuracy of the presented information and interpretation of the LCs. If they incorrectly categorize or fail to rank an LC in an order that is far off, their results will not be included in the output.

Following the accurate completion of this baseline portion, the turkers will be instructed to perform the same task for a set of the grouped LCs that were input from the learning curve visualization component. An example with the default configuration enabled might present ten LCs from the still high category, all located near on another on the same page, and ask the turker to again select which category each curve would fall into and how they would uniquely rank order each curve in terms of needing improvement. Note in Figure 2, showing an example of how an LC might be presented, the values 4 and 9 are greyed out since each ranking can only be used once per grouping. Once all presented LCs are ranked and their perceived category is selected, the turker can submit their HIT for completion.

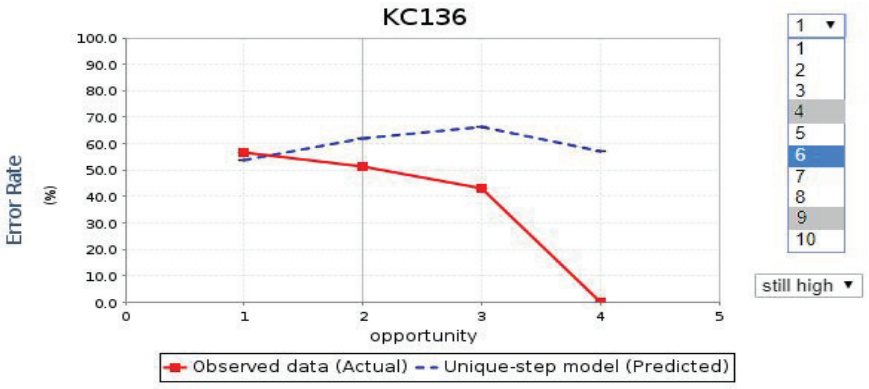

Figure 2: Prototype of how the learning curves can be ranked and categorized by a crowd worker

Currently the price point suggested for this proposed workflow task is 0.70 USD, based on similar image review tasks present on the platform while still offering a living wage amount. Additionally, this task is expected to take no more than five minutes to complete and even faster if they repeat the task for a new set of LCs.

2.3 Workflow Outputs

The primary output will be a text file to display, similarly to the results view for imported data from DataShop or another source. This will display correctly in an HTML friendly format and have the option to be downloaded, so that it can be imported into another environment, such as R Studio or Excel, for further analysis. This output text file will contain the tab separated data in an organized view with each row corresponding to an input learning curve. XML data from the learning curve visualization component, consisting of the DataShop assigned category, number of curve points, and KC name will be present for the columns of the table. It will be trivial to map this output data back to other data frames consisting of more detailed KC and LC information, since these rows can be matched by the common LC name found in both. Additionally, two more columns providing the crowd worker assigned LC category and ranking order will be present in the file. These two columns are the core analysis addition, their usefulness is detailed in the following discussion section.

3 DISCUSSION

One of the key goals of this workflow is to build upon the three-step process of LearnSphere: import, analysis, and visualization. An aspect that commonly gets neglected, but is essential in connecting this iterative process, is refinement. Many educational systems and courses often take initial efforts to construct appropriate content, but they unfortunately fail to iterate on these efforts after evaluation. While issues like a lack of continued funding might affect this lack of effort, the time such efforts take is a large barrier. This workflow looks to mitigate that by using crowd workers to further categorize and rank the LC visualizations so the ones needing the biggest improvement can easily come to light. The idea is the largest improvement and impact can be made back into a course, by addressing these most troublesome and ill-fitting KCs as identified by the crowd workers in their LC review.

While DataShop currently categorizes curves, it can benefit from having a knowledgeable human assist in the categorization process. For larger datasets, there might be hundreds of curves which fall into a given category. This automatic grouping becomes less useful when the user is unable to easily assess which curves might need the most attention, especially from such a large collection that would be difficult to display all at once. Having crowd workers take these categories and rank the LCs in them in order of which appear to need the most attention provides a better way to efficiently select which KCs to work on. Even with fine tuned parameters, the categories assigned by DataShop sometimes do not accurately group or portray KCs that need attention. For instance, the two LCs in Figure 3 are categorized the same, yet it is clear the bottom LC is representative of a KC that would need attention by comparison. It also allows the comparison of human categorized LCs to the categories assigned by another EDM workflow, in this case DataShop.

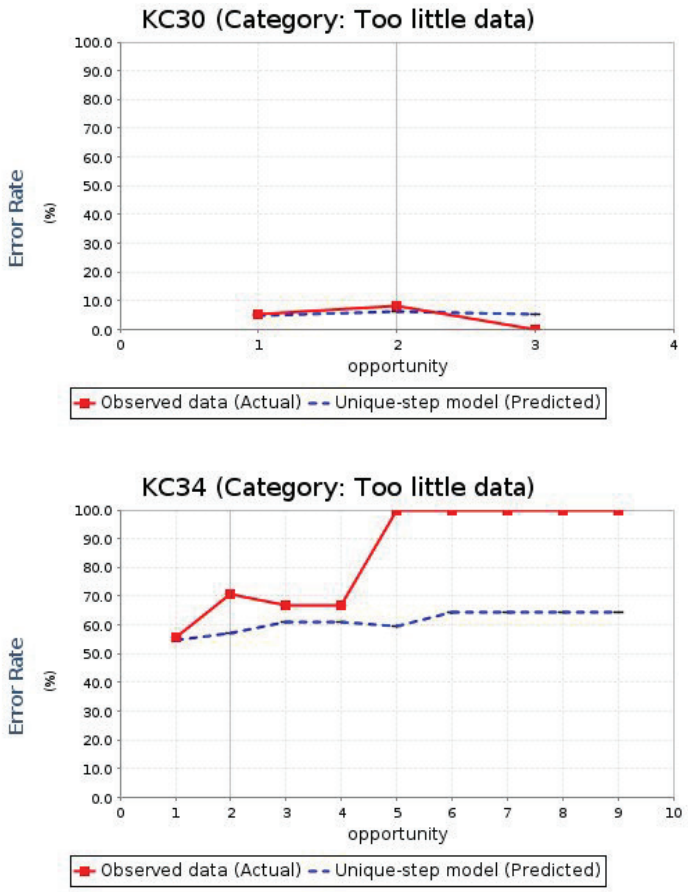

Figure 3: Two learning curves placed in the same category, with the bottom curve demonstrating lower learning than the learning curve at the top

Integrating humans in the loop helps to not only more accurately identify which KCs need improvement, but helps to maintain this iterative process of using the data these systems output to improve them. This is a key aspect of several professions, such as learning engineers and instructional designers, one that is often difficult to maintain and time-consuming. This offers a cheaper and faster alternative, all while removing a more tedious aspect of the process, so that these professionals can leverage their expertise where it counts.

A final goal of this workflow is to see where we can leverage crowdsourcing for work in the learning sciences that is not directly related to content creation or curation. The concept of a LC may sound filled with jargon at first, but it can be boiled down to basic line graph interpretation. Workers can still contribute meaningful input to the process, despite not having an explicit background in a domain related to learning sciences. Other aspects of educational data analysis can benefit from breaking down the task in a similar way, so crowd workers can contribute without needing such expertise. This lack of expertise might also provide a unique lens to look at the problem, categorization, etc. in a way that provides beneficial insights into improvements. This workflow's code can also be leveraged for components at different parts of the workflow, not just following the visualization portion like this component functions. Other instances, especially regarding data preprocessing, may leverage from the review of crowd workers before moving onto the next component.

REFERENCES

Anderson, J. R., Conrad, F. G., & Corbett, A. T. (1989). Skill acquisition and the LISP tutor. Cognitive Science, 13(4), 467-505.

Anderson, M. (2011). Crowdsourcing higher education: A design proposal for distributed learning. MERLOT Journal of Online Learning and Teaching, 7(4), 576-590.Automated Feedback Generation for Introductory Programming Assignments.

Balakrishnan, R. (2006, March). *Why aren't we using 3D user interfaces, and will we ever?* Paper presented at the IEEE Symposium on 3D User Interfaces. http://dx.doi.org/10.1109/vr.2006.148

Baker, R. S., & Inventado, P. S. (2014). Educational data mining and learning analytics. In Learning analytics (pp. 61-75). Springer New York.

Cen, H. et al. 2006. Learning Factors Analysis – A General Method for Cognitive Model Evaluation and Improvement. ITS '06 (2006), 164–175.

Fancsali, S. E., Ritter, S., Stamper, J., & Nixon, T. (2013). Toward "hyperpersonalized" Cognitive Tutors. In AIED 2013 Workshops Proceedings Volume (Vol. 7, pp. 71-79).

Gagné, M., Forest, J., Vansteenkiste, M., Crevier-Braud, L., van den Broeck, A., Aspeli, A. K., . . . Westbye, C. (2015). The Multidimensional Work Motivation Scale: Validation evidence in seven languages and nine countries. *European Journal of Work and Organizational Psychology, 24*(2), 178-196. http://dx.doi.org/10.1080/1359432x.2013.877892

Murray, T. (2003). An Overview of Intelligent Tutoring System Authoring Tools: Updated analysis of the state of the art. In Authoring tools for advanced technology learning environments (pp. 491-544). Springer, Dordrecht.

Paulin, D., & Haythornthwaite, C. (2016). Crowdsourcing the curriculum: Redefining e-learning practices through peer-generated approaches. The Information Society, 32(2), 130-142.

Porcello, D., & Hsi, S. (2013). Crowdsourcing and curating online education resources. Science, 341(6143), 240-241.

Stamper, J., Koedinger, K.R. (2011) Human-machine Student Model Discovery and Improvement Using DataShop. In Kay, J., Bull, S. and Biswas, G. eds. Proceeding of the 15th International Conference on Artificial Intelligence in Education (AIED2011). pp. 353-360. Berlin Germany:Springer.

Stamper, J., Koedinger, K., d Baker, R. S., Skogsholm, A., Leber, B., Rankin, J., & Demi, S. (2010). PSLC DataShop: A data analysis service for the learning science community. In International Conference on Intelligent Tutoring Systems (pp. 455-455). Springer, Berlin, Heidelberg.

Weld, D. S., Adar, E., Chilton, L., Hoffmann, R., Horvitz, E., Koch, M., ... & Mausam, M. (2012, July). Personalized online education—a crowdsourcing challenge. In Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence (pp. 1-31).