Educational Data Mining: An Advance for Intelligent Systems in Education

Ryan S. Baker
Teachers College, Columbia University

Computer-based technologies have transformed the way we live, work, socialize, play, and learn. Today, the use of data collected through these technologies is supporting a second-round of transformation in all of these areas. Over the last decades, the methods of data mining and analytics have transformed field after field. Scientific fields such as physics, biology, and climate science have leveraged these methods to manage and make discoveries in previously unimaginably large datasets. The first journal devoted to data mining and analytics methods in biology, *Computers in Biology and Medicine*, began publication as long ago as the 1970s. In the mid-1990s and into the first decade of the new millennium, data mining and analytics began to occupy a prominent place in business practices. Now, it seems like one can hardly go to an airport without seeing an advertisement from a consultancy, promising to increase profits through analytics, or an advertisement for a master's degree in analytics or data science.
In the Education field, data mining and analytics also have huge transformational potential: discovering how people learn, predicting learning, and understanding real learning behavior. By achieving these goals, educational data mining can be used to design better and smarter learning technology and to better inform learners and educators.

Initial Obstacles and Efforts

The development of data mining and analytics in the field of Education was fairly late, compared to other fields. The first articles and workshops on data mining in Education began only a decade ago, the first conference series (the International Conference on Educational Data Mining) began in 2008, and the first journal (the *Journal of Educational Data Mining)* began publication in 2009. A second group of researchers coalesced, forming the Conference on Learning Analytics and Knowledge in 2011. A third group created ACM Learning @ Scale in 2014. Educational data mining has even seen its own massive open online course (MOOC), Big Data and Education, taught by the author of this article on Coursera, now available as a massive online open textbook (MOOT) on YouTube and the Columbia University webpage. A second MOOC, taught by George Siemens and colleagues through EdX, has been announced for Fall 2014. Attention to educational data mining by the broader AI community increased with the Knowledge Discovery and Data Mining (KDD) Cup on educational data, held in 2010.
There has been considerable progress, however, in the last few years. Models have been articulated for a number of constructs, starting with models of learning, and advancing in sophistication to models that can infer *when* during the learning process a student is more likely to have learned (helping us understand the factors leading to those moments of learning). An example of a moment-by-moment learning graph is shown in the related sidebar.

[insert sidebar 1]

Beyond learning, researchers have been able to use data mining methods to model a range of constructs, including affect (emotion in context), engagement, meta-cognition, and collaboration. These models, also called *automated detectors*, are able to detect whether these constructs are present in real time, based on trace data from learners. These models have in turn been used to study the contexts when these constructs emerge and what their impacts are—for example, what student decisions lead to being more likely to be retained in an academic program,[1] how does user interface design impact student engagement,[2] and does the same surface behavior have different impacts depending on how much the student knows about the content domain?[3] More unsupervised methods have been used to discover previously unknown patterns in student learning, and these patterns have been studied to determine which patterns have an important educational impact.[4]
One of the reasons for the later emergence of data mining in Education, as compared to other fields, was that the availability of large datasets in usable formats emerged later in Education than other fields. School records were often stored in paper form in file cabinets up into the current millennium, and data from online learning systems was often in difficult-to-use formats, didn't store essential semantic information, or wasn't stored at all. Even today, a surprising amount of educational data is stored in the form of the screen locations of clicks and mouse movements,

or perhaps with reference to arbitrarily named objects, making meaningful data analysis difficult.
Another probable reason for this later emergence has to do with the history of the field of education itself. So far, relatively loosely operationalized educational theories and common sense have driven educational design decisions. Although empirical research has been a component of education for more than a century, empirical research has mostly involved a very small, humanly observable scale, or very large randomized controlled trials that, despite their scope, are designed to answer only a single question: Which of two approaches works better?

Significant Advances and Remaining Challenges

One major advance that shot the field of educational data mining forward was the advent of the Pittsburgh Science of Learning Center DataShop,[5] for almost a decade the largest open repository of data on the interactions between students and educational software (see the related sidebar). While many projects since then have chosen to use their own formats rather than adopt the PSLC DataShop format and infrastructure, the PSLC DataShop influenced practice both generally and specifically. The sheer existence of the PSLC DataShop, and its use in analyses, created awareness of the usefulness of educational data. But its specific format also created awareness of the need to store educational interaction data semantically, storing not just whether a student was correct or incorrect, but the context of the behavior and the skills involved in the task they were undertaking. Projects such as the Assessment Data Aggregator for Game Environments (ADAGE) and the Generalized Intelligent Framework for Tutoring (GIFT),[6,7] which attempt to create standards for logging educational software data (respectively, game data and military training data) have been influenced considerably by the DataShop.

[insert sidebar 2]

Another significant challenge for educational data mining for online learning is the specific features of educational data. While many types of data have sequential aspects, the distribution of educational data over time has unique characteristics; for example, a skill may be encountered 80 times during a school year, but separated over time and in the context of quite different activities. At the same time, the student's affective state over the last minute might also be a strong determinant of current behavior. Additionally, the hierarchy in educational data is important, with data aggregating meaningfully into problems, lessons, activities, curricula, classrooms, school cohorts, regions, and so on. Recent research, for instance, suggests that urbanicity (whether a student lives in a metropolitan area or a rural area) can be an important factor in how students respond to educational technology.[8] Developing methods and practices for handling these issues appropriately has been a significant ongoing challenge for researchers in this area.
Another important recent development for educational data mining is the growing realization that not all key information is stored in one data stream. Many recent projects have leveraged outcome data of various types, from college enrollment data and standardized exam data, to human coding of student engagement integrated with data streams. For collecting this type of human coding, the author's group has created an Android app for coding educationally meaningful constructs in real time and synchronizing them with log data, an app now used by dozens of researchers from Massachusetts to Utah to the Philippines and India.[9] Leveraging a combination of human judgment and data mining methods, within a supervised learning paradigm, has facilitated the development of automated detectors of student meta-cognition, disengagement, and emotion.
Another important development is an improvement in model quality, driven by continuing improvements in methodology, especially methods for feature engineering. For example, sensor-free models of student affect published in the last two years typically achieve almost double the model quality seen 3–5 years ago, as Figure 1 shows. Increasingly, affect detectors are also more extensively validated—for new students, new populations, and new content. However, while validation standards are improving, this achievement is incomplete. Though some projects are steadily improving model quality and the degree of validation, thoroughly validated detectors still remain a minority of the models published at conferences and journals. Similarly, corporations that make strong claims about the power of their analytics methods in education are still not being asked by the market to provide evidence of the accuracy of their models.

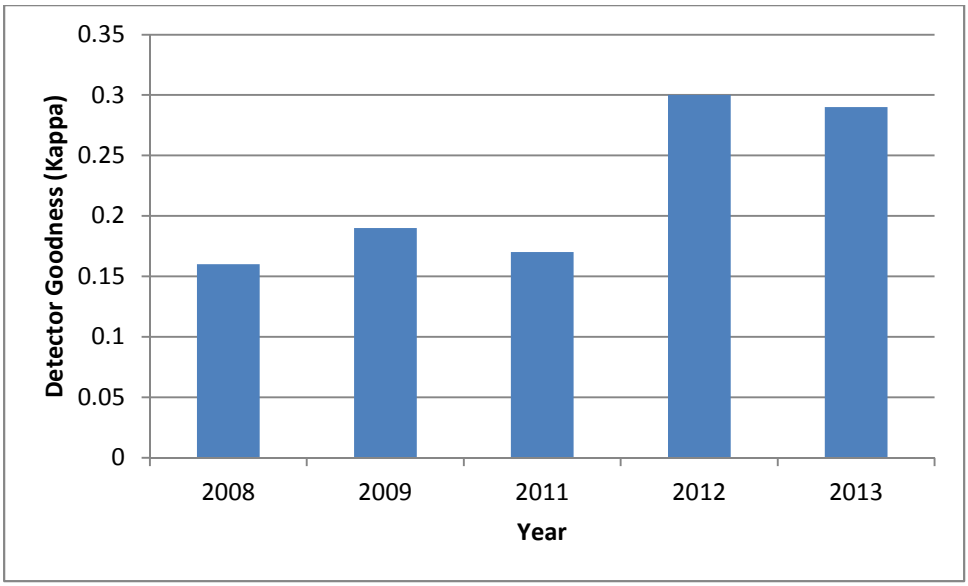

Figure 1. Performance of a sensor-free affect detector model, year-by-year. Model performance is shown here as Cohen's Kappa (how much better than chance the model is). Human agreement on student affect varies by method, but quantitative field observation methods generally achieve Cohen's Kappa of around 0.6; that is, current methods for sensor-free affect detection are about half as good as human beings at recognizing affect.

One of the challenges in this type of model development, as well as in other areas of education research, is limited access to training data. For example, while a system might be used by hundreds of thousands of students, it could be impractical to obtain training labels of affect (often collected by expert field observers) or data on long-term outcomes, for more than a few hundred or few thousand students. As such, collecting representative data becomes essential, as does paying attention to validation methods that inform developers about whether a model is representative, such as conducting population-level cross-validation on a diverse sample of students. By taking these steps, it becomes possible to increase confidence that the resulting models will be relevant for the full population of students where the model's predictions would be useful.

Another trend that still needs to be developed is the actual use of these types of detectors in intelligent educational systems of various sorts. There are far more published examples of detectors than there are of detectors being used to drive intervention, though some positive examples do exist. In perhaps the most notable example, the Purdue Signals Project (now Ellucian) provided instructors with reports of whether students were at risk of dropping or failing a course, and scaffolded instructors in how to intervene, leading to better outcomes for learners.[1] Examples also exist in fully automated systems. Carnegie Learning has been leveraging automated models of student knowledge—now fit using data mining methods—to tailor the quantity of instruction students receive on different topics, leading to positive learning outcomes.[10] More complex models have also been used in automated systems, but only with relatively small-scale studies in classrooms and laboratory settings.[11-13] These systems have found benefits from automated agents that adapt to disengagement and affect, but these approaches have not yet been deployed at scale.

Overall, educational data mining methods have been successful at modeling a range of phenomena relevant to student learning in online intelligent systems. Models are achieving better accuracy every year, and are being validated to be more generalizable over time. They've been used in a small number of projects to improve student outcomes, and have been successful. The key goals for this community in the next decades include better standards for validation—bringing better practices to the whole field—and more deployment into intelligent systems. To achieve that goal, more partnerships will be needed between educational data mining researchers and the broader intelligent systems community. This partnership is desirable and we can expect it will support cross-fertilization for both communities.

Acknowledgments

I would like to thank Judy Kay and Kalina Yacef for their helpful comments and suggestions.

References

1. K.E. Arnold and M.D. Pistilli, "Course Signals at Purdue: Using Learning Analytics to Increase Student Success," *Proc. 2nd Int. Conf. Learning Analytics & Knowledge*, 2012, pp. 267–270.
2. R.S.J.d. Baker et al., "Educational Software Features that Encourage and Discourage 'Gaming the System'," *Proc. 14th Int'l Conf. Artificial Intelligence in Education*, 2009, pp. 475–482.
3. V. Aleven et al., "Toward Meta-Cognitive Tutoring: A Model of Help Seeking with a Cognitive Tutor," *Int'l J. Artificial Intelligence and Education*, vol. 16, no. 2, 2006, pp. 101–128.
4. A.J. Bowers, "Analyzing the Longitudinal K-12 Grading Histories of Entire Cohorts of Students: Grades, Data Driven Decision Making, Dropping Out and Hierarchical Cluster Analysis," *Practical Assessment Research and Evaluation*, vol. 15, no. 7, 2010, pp. 1–18.
5. K.R. Koedinger et al., "A Data Repository for the EDM Community: The PSLC DataShop," *Handbook of Educational Data Mining*, CRC Press, 2010, pp. 43–56.
6. V.E. Owen and R. Halverson, "ADAGE: Assessment Data Aggregator for Game Environments," *Proc. Games, Learning, and Society Conf*., ETC Press, 2013.
7. R.A. Sottilare et al., *The Generalized Intelligent Framework for Tutoring (GIFT)*, US Army Research Laboratory, 2012; https://gifttutoring.org/projects/gift/wiki/Overview.
8. J. Ocumpaugh et al., "Population Validity for Educational Data Mining Models: A Case Study in Affect Detection," *British J. Educational Technology*, vol. 45, no. 3, 2014, pp. 487–501.
9. J Ocumpaugh, R.S.J.d. Baker, and M.M.T. Rodrigo, *Baker-Rodrigo Observation Method Protocol (BROMP) 1.0. Training Manual version 1.0*, tech. report, EdLab. Manila, Philippines: Ateneo Laboratory for the Learning Sciences, 2012.
10. K.R. Koedinger and A.T. Corbett, "Cognitive Tutors: Technology Bringing Learning Science to the Classroom," *The Cambridge Handbook of the Learning Sciences*, Cambridge Univ. Press, 2006, pp. 61–78.
11. I. Arroyo et al., "Repairing Disengagement with Non-Invasive Interventions," *Proc. 13th Int'l Conf. Artificial Intelligence in Education*, 2007, pp. 195–202.
12. I. Arroyo et al., "The Impact of Animated Pedagogical Agents on Girls' and Boys' Emotions, Attitudes, Behaviors and Learning," *Proc. 2011 IEEE 11th Int'l Conf. Advanced Learning Technologies*, 2011, pp. 506–510.
13. S. D'Mello et al., "A Time for Emoting: When Affect-Sensitivity Is and Isn't Effective at Promoting Deep Learning," *Proc. 10th Int'l Conf. Intelligent Tutoring Systems*, 2010, pp. 245–254.

***Ryan S. Baker** is an associate professor in the Department of Human Development at Teachers College, Columbia University. Contact him at Baker2@exchange.tc.columbia.edu.*

[sidebar 1]

Moment-by-Moment Learning Model

Figure A shows an example of the moment-by-moment learning graph,[1] from a student using online learning software for genetics. This graph is derived from the Moment-by-Moment Learning Model, which is in turn derived from Bayesian Knowledge Tracing, a classic student modeling algorithm. This graph shows one student's learning over time on a specific skill; this skill was needed in 12 problems that the student encountered while using the learning software. The graph represents the probability that the student learned the relevant skill at each problem—in this case, high on the first four problems, lower on the fifth, and low afterwards. This graph tells us that the student had a period of steady learning that concluded after the fifth problem. However, the student continued to receive problems involving this skill even after reaching this point. As it turns out, patterns where learning is brief and concentrated seem to be associated with less retention and preparation for future learning than learning where the student repeatedly refines and consolidates their knowledge.[1]

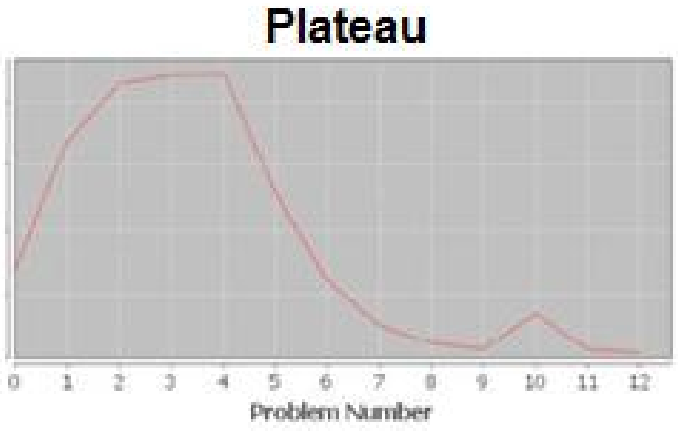

Figure A. One student's learning over time on a specific skill. This skill was needed in 12 problems that the student encountered while using the learning software. Learning was at a high plateau on the first four problems, lower on the fifth, and then low afterwards.

Reference

1. R.S.J.d. Baker et al., "Predicting Robust Learning with the Visual Form of the Moment-by-Moment Learning Curve," *J. Learning Sciences*, vol. 22, no. 4, 2013, pp. 639–666.

[sidebar 2]

Pittsburgh Science of Learning Center Datashop

The Pittsburgh Science of Learning Center (PSLC) DataShop has been the largest open repository of data on the interactions between students and educational software for almost a decade.[1] The PSLC DataShop was created as part of the Pittsburgh Science of Learning Center (now called LearnLab), which was founded in 2004 as part of the National Science Foundation's Science of Learning Center program that created multiple large-scale, long-term centers with the goal of advancing the frontier of the science of learning through integrative research. The PSLC DataShop created a database schema that was tailored for intelligent tutoring systems, but which has been utilized for other types of learning environments such as simulations, and which has influenced later databases of student interaction data. The PSLC DataShop played a key early role in educational data mining; DataShop data sets were used in 14 percent of the papers published in the first two years of the International Conference on Educational Data Mining conference (Baker & Yacef, 2009). It has continued to play an important role in this area of research since then, with DataShop data being used in more than 100 published papers.[2] It now contains data from almost 100 million interactions between a student and educational software, which occurred during more than 200,000 hours of interaction. It has supported research in a variety of areas, including research on the structure of knowledge domains,[3] the comparison of different student modeling approaches,[4] the study of differences in content across an entire year,[5] and the development of models of new constructs.[6]

References

1. K.R. Koedinger et al., "A Data Repository for the EDM Community: The PSLC DataShop," *Handbook of Educational Data Mining*, CRC Press, 2010, pp. 43–56.
2. K.R. Koedinger et al., "LearnLab's Data Shop: A Data Repository and Analytics Tool Set for Cognitive Science," *Topics in Cognitive Science*, vol. 5, no. 3, 2013, pp. 668–669.
3. K.R. Koedinger et al., "Using Data-Driven Discovery of Better Student Models to Improve Student Learning," *Proc. Int'l Conf. Artificial Intelligence in Education*, 2013, pp. 421–430.
4. Z.A. Pardos et al., "The Sum is Greater Than the Parts: Ensembling Models of Student Knowledge in Educational Software," *SIGKDD Explorations*, vol. 13, no. 2, 2011, pp. 37–44.
5. R.S.J.d. Baker et al., "Educational Software Features that Encourage and Discourage 'Gaming the System'," *Proc. 14th*

Int'l Conf. Artificial Intelligence in Education, 2009, pp. 475–482.

6. V. Aleven et al., "Toward Meta-Cognitive Tutoring: A Model of Help Seeking with a Cognitive Tutor," *Int'l J. Artificial Intelligence and Education*, vol. 16, no. 2, 2006, pp. 101–128.

6. R. Baker & K. Yacef, "The State of Educational Data Mining in 2009: A Review and Future Visions," *J. Educational Data Mining*, vol. 1, no. 1, 2009, pp. 3–17.