A Data Driven Approach to the Discovery of Better Cognitive Models

Kenneth R. Koedinger [1] and John C. Stamper[2]
[1]koediger@cmu.edu, [2]john@stamper.org
[2]Human Computer Interaction Institute, Carnegie Mellon University

Abstract. Cognitive models composed of knowledge components are an integral part of intelligent tutors and drive many of the instructional decisions that these systems make. Most of these models are designed by educators and subject experts. Today vast amounts of data, collected from many intelligent tutors, allow us to analyze and improve the current cognitive models through educational data mining. In this research, we show how we identified, in the tutor data, potential improvements to existing cognitive models and then evaluated those improvements using statistical analysis and cross validation.

1 Introduction and Experiment

Educational data mining provides a great opportunity to discover better cognitive models. A correct cognitive model is one that is consistent with student behavior. It predicts task difficulty and transfer between instruction and test. Multiple algorithms have been developed for automated discovery of the attributes or factors that make up a cognitive model. This research provides the basis for an infrastructure for automatically applying such algorithms to data sets and discovering better cognitive models. We show how data analysis tools such as those in the PSLC DataShop [2] can be used to identify areas for improvement, and then discuss how to quantitatively evaluate the new models.

We tested a number of Knowledge Component (KC) models on the Geometry Area ’96 data set (from DataShop), which indicate how well 59 students performed (i.e., how often they were correct without tutor help) on 139 unique geometry task items, that is, steps in multi-step problems presented by the Geometry Cognitive Tutor. Two measures of quality, the Bayesian Information Criteria (BIC) and the root mean squared error (RMSE) on the held-out test data in a 3-fold cross validation were calculated for each KC model. The KC models represent different ways of sorting the 139 items into groups that measure student acquisition of the same KC, like all items requiring application of the circle area formula. In all cases, predictions of student performance are based on Additive Factors Model (AFM). It is important to point out that both BIC and cross-validation adjust for over-fitting, which can result from unnecessary addition of model parameters. Thus, the differences between the models are likely to be of practical significance. A new model was discovered using the visualization and analysis tools provided by the PSLC DataShop. This model is better (lower BIC and lower RMSE) than the “original” production rule cognitive model in Geometry Cognitive Tutor, which was created by cognitive scientists and domain experts [3]. It is also better than any of the models discovered by our existing automated approach to KC model discovery called Learning Factors Analysis [1]. The DataShop tools show that most of the KCs have appropriate learning curves, like circle-area and trapezoid-area, where the error rate starts high and then goes down. The curve for compose-by-addition curve, however, is flat.

This kind of curve suggests that this KC may not be correctly defined – some task items it summarizes may require different knowledge than others. The compose-by-addition KC characterizes the step in complex geometry area problems where the student must find an irregular area, like the area of a sidewalk around a pool by subtracting the area of the pool from the whole area. Closer inspection of such problems reveals that some of them provide "scaffolding" by indicating that the student should find the component areas first (whole area and pool area) before finding the target irregular shape (the sidewalk) and other problems do not provide this hint. The new model distinguishes steps in such problems where such decomposition planning is required and ones where it is not. Using data to discover and confirm the existence of such "hidden" planning knowledge is an interesting cognitive science achievement – such a hidden or cognitive component is not directly apparent in student behavior, in contrast with the original compose-by-addition KC, which is adding or subtracting two numbers. It is also important for instructional design. We have used this insight to redesign this unit of the cognitive tutor to better help students acquire this difficult and important problem decomposition skill (such non-trivial problems are frequently seen on standardized tests). A 'close-the-loop' in vivo experiment is being run this spring to test whether these designs yield improved robust learning as compared to the existing tutor.

2 Conclusion and Future Work

This work demonstrates the use of tools in the DataShop to discover a better cognitive model, even in a domain (Geometry) where there has been considerable attention and prior cognitive analysis. The approach described here to discover cognitive models has a heavy component of human expertise. Using data to optimize cognitive models and improve instructional systems is a tremendous opportunity for EDM. The achievement will be greater to the extent that the discovered models involve deep or integrative KCs not directly apparent in surface task structure, like the problem decomposition skill we identified in Geometry. In addition, future work should compare the statistical model structure of competing discovery algorithms to shed new light on the nature or extent of regularities or laws of learning, like the power or exponential shape of learning curves and whether or not there are systematic individual differences in student learning rates.

References

[1] Cen, H., Koedinger, K. R., & Junker, B. (2006). Learning Factors Analysis: A general method for cognitive model evaluation and improvement. In M. Ikeda, K. D. Ashley, T.-W. Chan (Eds.) *Proceedings of the 8th International Conference on Intelligent Tutoring Systems*, 164-175. Berlin: Springer-Verlag.

[2] Koedinger, K.R., Baker, R.S.J.d., Cunningham, K., Skogsholm, A., Leber, B., Stamper, J. (2009) A Data Repository for the EDM community: The PSLC DataShop. In Romero, C., Ventura, S., Pechenizkiy, M., Baker, R.S.J.d. (Eds.) Handbook of Educational Data Mining. Boca Raton, FL: CRC Press.

[3] Koedinger, K. R. & Aleven, V. (2007). Exploring the assistance dilemma in experiments with Cognitive Tutors. Educational Psychology Review, 19 (3): 239-264.