Goguadze, G., Sosnovsky, S., Isotani, S. & McLaren, B.M. (2011). Evaluating a bayesian student model of decimal misconceptions. In the *Proceedings of the 4th International Conference on Educational Data Mining*.

Evaluating a Bayesian Student Model of Decimal Misconceptions

G. GOGUADZE
Saarland University, Germany,
S. SOSNOVSKY
DFKI GmbH, Germany,
S. ISOTANI
Carnegie Mellon University, U.S.,
AND
B. M. MCLAREN
Carnegie Mellon University, U.S.

Among other applications of educational data mining, evaluation of student models is essential for an adaptive educational system. This paper describes the evaluation of a Bayesian model of student misconceptions in the domain of decimals. The Bayesian model supports a remote adaptation service for an Intelligent Tutoring System within a project focused on adaptively presenting erroneous examples to students. We have evaluated the accuracy of the student model by comparing its predictions to the outcomes of students' logged interactions from a study with 255 school children. Students' logs were used for retrospective training of the Bayesian network parameters. The accuracy of the student model was evaluated from three different perspectives: its ability to predict the outcome of an individual student's answer, the correctness of the answer, and the presence of a particular misconception. The results show that the model's predictions reach a high level of precision, especially in predicting the presence of student misconceptions.

Key Words and Phrases: Student model evaluation, Bayesian networks, Bayesian student modeling

1. INTRODUCTION

The quality of an adaptive educational system (AES) critically depends on the quality of its student modeling. The system might implement a sound adaptation strategy and provide students with well-designed learning content, but if its estimations of students' knowledge are incorrect, the adaptive interventions it produces are unlikely to be effective. In recent years, significant efforts have been expended to develop a methodology for layered evaluation of AES that allows examining student modeling components in isolation (BRUSILOVSKY, ET AL., 2004). Various approaches have been used for measuring the goodness of a particular student modeling mechanism (SOSNOVSKY, & BRUSILOVSKY, 2005), guiding the improvement of a student model (SM) (MARTIN ET AL., 2011) or selecting the best SM configuration among several alternatives (YUDELSON, ET AL., 2008). All of these evaluations have been based on rigorous analyses of students' logs generated by the systems.

In this paper, we describe the development of a Bayesian network (BN) SM and a data mining study aimed at validating its quality. The model represents students' misconceptions in the domain of decimals. It was designed within the framework of the

Authors' addresses: G. Goguadze, Saarland University, Saarbrücken, Germany, E-mail:george@activemath.org; S. Sosnovsky, DFKI, Saarbrücken, Germany, E-mail: sosnovsky@gmail.com; S. Isotani, B. M. McLaren, Human-Computer Interaction Institute, Carnegie Mellon University, Pittsburgh, PA, U.S., E-mail: sisotani@gmail.com, bmclaren@cs.cmu.edu

AdaptErrEx project[1], which focuses on presenting and adapting erroneous examples (step-by-step solutions to decimal math problems in which at least one step is incorrect) to remediate students' misconceptions.

The evaluation of the model was done based on the data logs of 255 middle school students working with test problems in the domain of decimals. Data from 70% of the students was used for training model parameters. The remaining 30% of the data was used as the test set to compute three metrics, estimating how well the model predicts:

1. the exact answer to the next problem tackled by a student;
2. the correctness of the next answer provided by a student; and
3. the presence of a misconception the student has.

In order to compute these metrics, we compared the predictions of the individual SMs with the students' results on a delayed posttest. Although the values achieved for all three metrics could potentially be improved, they by far exceed the baseline of a random prediction. These results support our belief that the model is capable of accurate adaptation and encourage us to continue investigating ways to improve it.

2. MODELING STUDENTS' MISCONCEPTIONS IN ADAPTERREX

BNs are well-established tools for representing and reasoning about uncertainty in student models (CONATI, ET AL., 2002; MILLÁN, ET AL., 2010). Perhaps the closest example to the BN-based SM developed for AdaptErrEx is the SM of the DCT tutor that helped students learn decimal comparisons (STACEY ET AL., 2003). In the DCT's model, the misconceptions were represented as two probabilistic nodes identifying basic judgments used by a student for comparing decimals (e.g. "longer decimals are larger") and possible misconceived reasons for such judgments (e.g. "because longer integers are larger"). The causal relation between the two nodes was modeled with conditional probabilities defining the chance a student would come up with a basic judgment if she had a particular finer-grained misconception. The evidence nodes representing learning tasks were conditionally dependent on both misconception nodes.

A different approach to BN-based domain and student modeling, focused on skills rather than misconceptions, is described in (COLLINS, ET AL., 1996). The domain model here is a hierarchy of skills, where the probability of mastering a super-skill is conditionally dependent on mastery of the sub-skills. The bottom-level skills are probabilistically connected with the evidence nodes representing the test questions.

In AdaptErrEx we have followed an approach that is a combination of the two prior approaches. Based on the results of an extensive literature review of students' learning of decimals, we identified the most frequently occurring decimal misconceptions and organized them into a taxonomy based on their differences and similarities (ISOTANI, MCLAREN, & ALTMAN, 2010). The resultant taxonomy connects commonly observed misconceptions to the possible higher-level reasons for the misconceptions (e.g., the misconception "longer decimals are larger" is connected to the reason "student treats decimals like integers") providing means for diagnosing students' learning difficulties.

To account for dependencies between misconceptions, a BN was built, where each misconception is represented by a probabilistic node with two possible alternatives (present/absent). The taxonomic relations between the nodes are accompanied by tables of conditional probabilities representing the influence of the probability of presence of a parent misconception on the probabilities of presence of its child misconceptions.

The evidence nodes in the network represent problems. They can be connected to one or more misconceptions. The evidence nodes contain several alternatives, where each alternative corresponds to a possible answer the student might give to the problem. Every evidence node alternative is probabilistically connected to the corresponding

[1] See http://www.cs.cmu.edu/~bmclaren/projects/AdaptErrEx

misconception node alternatives. This means that presence/absence of a misconception influences the likelihood of a student giving a certain answer to the problem.

Overall, the developed network contains twelve misconception nodes, where seven nodes represent the most typical decimal misconceptions and five nodes serve as higher-level reasons for their occurrence. The misconception nodes are connected to 126 evidence nodes representing possible answers to decimal problems. The problems are divided into three isomorphic problem sets (set A, set B and set C), each set containing 42 problems. In order to ensure that the results are not driven by differences in the problems, the roles of problem sets A, B and C were counterbalanced across student groups. Each set was used either for a pretest, an immediate posttest, or a delayed posttest. In total, there are six possible combinations of the sets (ABC, ACB, BAC, BCA, CAB and CBA) depending on the role each set plays. Consequently, students were randomly assigned to one of the six groups, facing one of the six sequences of tests.

3. EVALUATING THE ACCURACY OF MODEL'S PREDICTIONS

This section summarizes our approach to evaluating the AdaptErrEx BN's capability to predict the effective state of student's learning. The approach consists of three steps:

- training the domain model based on the pretest data from 70% of the students;
- learning student models using the logs of the remaining 30% of the students;
- evaluating the accuracy of the model's predictions using 3 different metrics.

3.1. Training the initial domain model

Parameter estimation is a well-known challenge in the field of BNs. In our case, these parameters include prior probabilities for misconception nodes and conditional probability tables for links between the nodes. To complete this task, we supplied initial estimations of network parameters and then refined them with the training algorithm.

For the training set we randomly selected 70% of the students participating in the study. Based on the pretest logs of these students, from taking one of the three tests (A, B, or C), the prior probabilities for misconception nodes and the conditional probabilities for evidence nodes of all three problem sets A, B and C are computed. In this way, the resulting BN represents the initial state of knowledge of decimals (more specifically, the predicted state of misconceptions) for a typical student from the target population. The prior probabilities of misconception nodes quantify how likely such an average student is to have a particular misconception. The conditional probabilities encode the strength of a causal relation among misconceptions and between the misconceptions and the problems.

3.2. Learning specific student models

After the initial training/calibration, the BN was ready to learn the specific models of individual students. In order to do this, we fed the activity logs of the remaining 30% of the students to the network. Only answers to the pretest and immediate posttest were used on this step. This evidence back-propagated to the relevant misconception nodes and updated all posterior probabilities, thus individualizing the networks. The resulting collection of BNs contained individual misconception models for every student in the test set. Each resulting individual SM took into account both the collective traits of the target population and the history of idiosyncratic behavior of the corresponding student.

3.3. Estimating accuracy of the student model's predictions

The BNs obtained in Step 2 can be used to make individual predictions about students. Based on such predictions, an AES could control the individual learning experiences of its students. We identified three types of predictions and verified their average accuracy by comparing the models of the students from the test set with their results on the delayed posttest. The three prediction types were: predicting the actual student answer, predicting

the correctness of the student answer, and predicting the presence of a student misconception. The notion of accuracy in these three cases was defined as follows:

I. A prediction of the actual student answer is accurate if the alternative chosen by a student for a posttest problem had the highest probability in this student's BN trained in Step 2. The corresponding metric is computed as a percentage of accurate predictions.

II. A prediction of the correctness of the student's answer is accurate if:

- the student answers correctly to a delayed posttest problem and the probability of the correct alternative for this problem's node is maximum in the BN trained for this student in Step 2;
- or the student answers incorrectly to a delayed posttest problem and the probability of the correct alternative for this problem's node is less than the sum of probabilities of incorrect alternatives in the BN trained in Step 2.

The corresponding metric is computed as a percentage of accurate predictions.

III. A prediction of the presence of a misconception is defined as follows. Based on the state of a misconception node, the student is believed to have a corresponding misconception if its probability is greater than 0.5. This prediction is considered accurate if during the delayed posttest the student has shown more evidence of having the misconception than not having it (and vice-versa). The evidence is quantified as an average rate of misconception occurrence in the students' answers in the delayed posttest. The average rate of misconception occurrence is computed in the following way:

Let P(M) be the probability of the presence of a misconception M in a Bayesian model, Npos(M) – the number of student's answers that provide evidence for the misconception M, Nneg(M) – the number of correct answers to the problems that can diagnose M, and N(M) – the total number of problems that address this misconception. Then, the model prediction is said to be accurate if and only if:

$$\left[(P(M) \geq 0.5) \& \left(\frac{Npos(M)}{N(M)} \geq 0.5\right)\right] \vee \left[(P(M) < 0.5) \& \left(\frac{Nneg(M)}{N(M)} \geq 0.5\right)\right]$$

4. EXPERIMENT SETTINGS AND EVALUATION RESULTS

The data for the evaluation came from an empirical study conducted in a middle school classroom in Pittsburgh, PA (U.S.A.) during the fall of 2010. Overall, 255 students from 6th-8th grades participated in the study. The study had several subsections conducted over multiple sessions, including a pretest, treatment problems, an immediate posttest, and (one week later) a delayed posttest. The 126 test problems were split into 3 isomorphic problem sets (A, B, and C) and the roles of these problem sets being pretest, posttest or delayed posttest were counterbalanced across student groups. The learning materials came from the domain of decimals. The MathTutor web-based system (ALEVEN, ET AL., 2009) was used to deliver the materials to the participants. Students took 4 to 5 sessions to complete all of the materials.

MathTutor logs all students' interactions, as well as diagnostic information in the PSLC DataShop (KOEDINGER ET AL., 2010). In total, we analyzed 31,049 student interaction events, which resulted from each of the 255 students solving up to 126 problems. The accuracy values were calculated for the test set (i.e., the 77 students; data from these students was not used in the training phase). Using the metrics defined in Section 2.3, we evaluated the accuracy of the predictions of our SM. As a result of the calculation, the average accuracy of predicting the actual answer of the students in the delayed posttest was 60%, whereas the average accuracy of predicting the answer correctness was 69%. The average accuracy of predicting misconceptions was 87% (σ=0.148). Similar studies on evaluating the accuracy of predictions of a BN student model of the DCT tutor (NICOLSON ET AL., 2001), achieve comparable accuracy for predicting misconceptions (80-90%). However, our model estimates students'

misconceptions with higher granularity (a node per misconception compared to one node for all misconceptions).

5. CONCLUSIONS AND FUTURE WORK

We have presented the design and evaluation of a Bayesian approach to modeling student misconceptions. Three different metrics have been compared for estimating how well the model predicts student's misconceptions. The evaluation results demonstrate high accuracy of models' predictions, yet leaving room for improvement.

Future work is planned in two main directions: improving the structure of the Bayesian model and enhancing the methods of evaluation of the model validity. We plan to experiment with different configurations of BNs, such as dynamic BNs, and the networks with soft evidence nodes. When adjusting the evaluation method we could experiment with additional parameters of the students such as gender, grade, or general math skills. Difficulty of the problems could be used here as well as an additional parameter in the computation of the accuracy metrics. For example, if the problem is very easy, the student is likely to solve it correctly even if the probability of having a misconception is high, and the other way round, difficult problems can be solved incorrectly even if the probabilities of misconceptions are low.

ACKNOWLEDGEMENTS

The U.S. Department of Education Institute of Education Sciences, grant# R305A090460, funded this work.

REFERENCES

ALEVEN, V., MCLAREN, B., & SEWALL, J., 2009. Scaling up programming by demonstration for intelligent tutoring systems development: An open-access website for middle school mathematics learning. *IEEE Transactions on Learning Technologies,* 2(2), 64-78.

BRUSILOVSKY, P., KARAGIANNIDIS, C., AND SAMPSON, D., 2004. Layered evaluation of adaptive learning systems. *Int'l Journal of Continuing Engineering Education and Lifelong Learning* 14 (4/5), 402 – 421.

CONATI, C., GERTNER, A. AND VANLEHN K., 2002. Using Bayesian networks to manage uncertainty in student modeling. *Journal of User Modeling and User-Adapted Interaction*, vol. 12(4), p. 371-417

COLLINS, J., GREER, & J., HUANG, S., 1996. Adaptive assessment using granularity hierarchies and Bayesian nets. In*: Lecture Notes in Computer Science*. Vol. 1086. pp. 569–577.

ISOTANI, S., MCLAREN, B., & ALTMAN, M., 2010. Towards intelligent tutoring with erroneous examples: A taxonomy of decimal misconceptions. Submitted as a poster paper to the *Tenth International Conference on Intelligent Tutoring Systems (ITS-2010)*. Pittsburgh, PA.

KOEDINGER, K., BAKER, R., CUNNINGHAM, K., SKOGSHOLM, A., LEBER, B., STAMPER, J., 2010. 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, pp 43-55, FL: CRC Press.

MARTIN, B., MITROVIC, A., KOEDINGER, K., & MATHAN, S., 2011. Evaluating and Improving Adaptive Educational Systems with Learning Curves, *J. of User Modeling and User Adapted Int.*, 21(3), Springer.

MILLÁN, E., LOBODA, T., & PÉREZ-DE-LA-CRUZ, J., 2010. Bayesian networks for student model engineering, *Computers & Education*, Vol. 55, Issue 4, 1663-1683.

NICOLSON, A., BONEH, T., WILKIN, T. , STACEY, K., SONENBERG, L., & STEINLE, V., 2001. A case study in knowledge discovery and elicitation in an intelligent tutoring application. In *Proceedings of the 17th Conference on Uncertainty in AI*, pp. 386-394, Seattle.

SOSNOVSKY, S., & BRUSILOVSKY, P., 2005. Layered Evaluation of Topic-Based Adaptation to Student Knowledge. In *Proceedings of 4th Workshop on the Evaluation of Adaptive Systems*, 47-56.

STACEY, K., SONENBERG, E., NICHOLSON, A., BONEH, T., & STEINLE, V., 2003. A teacher model exploiting cognitive conflict driven by a Bayesian network. In Peter Brusilovsky, Albert T. Corbett, Fiorella De Rosis (Eds), *User* Modeling *2003: Proceedings of the Ninth International Conference*. (pp. 352-362) New York: Springer-Verlag (ISBN 3540403817).

YUDELSON, M., MEDVEDEVA, O., AND CROWLEY, R., 2008. A multifactor approach to student model evaluation. User Modeling and User-Adapted Interaction, 18(4), 349-382.