Data-driven Proficiency Profiling - Proof of Concept

Behrooz Mostafavi
Department of Computer Science
North Carolina State University
Raleigh, NC 27695
bzmmostaf@ncsu.edu

Tiffany Barnes
Department of Computer Science
North Carolina State University
Raleigh, NC 27695
tmbarnes@ncsu.edu

ABSTRACT

Data-driven methods have previously been used in intelligent tutoring systems to improve student learning outcomes and predict student learning methods. We have been incorporating data-driven methods for feedback and problem selection into Deep Thought, a logic tutor where students practice constructing deductive logic proofs. In this latest study we have implemented our data-driven proficiency profiler (DDPP) into Deep Thought as a proof of concept. The DDPP determines student proficiency without expert involvement by comparing relevant student rule scores to previous students who behaved similarly in the tutor and successfully completed it. The results show that the DDPP did improve in performance with additional data and proved to be an effective proof of concept.

Keywords

Data-driven, Tutoring system, Student classification, Clustering

1. INTRODUCTION

Data-driven methods have been used in intelligent tutoring systems to accurately predict student behavior and improve learning outcomes. In contrast to individualized tutoring systems based on experts developing complex and context specific models of behavior, data-driven systems reduce the need for expert involvement in designing the system, and can potentially adapt to new users without refinement of a behavioral model. We have been incrementally augmenting the Deep Thought logic tutor with data-driven methods for formative feedback and problem selection to improve student learning and reduce tutor dropout. Our long term goal is to create an intelligent tutor for logic proof construction that is fully data-driven and can adapt to students learning logic with varying curricular requirements without the need for further expert input. To this end, the next step in our work is to replace the expert-authored assessment parameters built into our problem selection system

with a data-driven proficiency calculation that approximates the original system's performance. Deep Thought utilizes a data-driven mastery learning system (DDML) consisting of 6 strictly ordered levels of proof problems. Each level is split into a higher proficiency track with a lower number of complex problems, and a lower proficiency track with a greater number of simpler problems. The first level of problems are the same for all students, and are used to estimate their initial proficiency. Proficiency is calculated using the knowledge tracing of all rule-application actions taken in the tutor. These action scores are compared to the average score thresholds of corresponding problems solved by past exemplars – students who have successfully completed the entire tutor and have therefore demonstrated sufficient proficiency in the subject matter. The difference between each score and its threshold score is weighted based on expert-decided rule priorities; actions that involve rules that are of greater importance to solving proofs in a given level are weighted higher than actions that involve rules that are not as important to the solutions in that level, based on expert determination. This system was shown to increase student completion and reduce tutor dropout over unordered and hint-based versions of Deep Thought [5].

Since the DDML system uses expert-decided priorities for each of the rule application actions when calculating a student's proficiency, any new problems or levels added to the system would require expert involvement to determine which rules were prioritized in each new or altered level. To this end, we developed our data-driven proficiency profiler (DDPP), a data-driven method of determining student proficiency that can replace the current expert-decided rule priorities in Deep Thought. The DDPP compares student rule scores to the rule threshold scores of clustered exemplars, weighting each rule score based on primary component importance (determined by previous student data). This method was previously compared to proficiency calculations using the minimum rule scores and average rule scores of exemplars, also weighted by primary component importance, to see how these methods compare to each other and to the expert authored system in an exploratory study[6]. We found that additional data was required to improve the DDPP's proficiency calculations.

For this latest study we have augmented the DDPP with additional data from the previous semester's use of Deep Thought and fully implemented the DDPP into the tutor as a proof of concept. We hypothesized that the DDPP will compare comparably to the DDML with expert decided rule priorities. The results show that while there was some im-

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions.acm.org.

LAK '16, April 25-29, 2016, Edinburgh, United Kingdom

© 2016 ACM. ISBN 978-1-4503-4190-5/16/04...\$15.00

DOI: <http://dx.doi.org/10.1145/2883851.2883935>

provement with students who used the DDPP in dropout and path prediction accuracy, more data is required to improve the system. However, the results are very promising as a proof of concept. We expect that the DDPP will improve in path prediction accuracy with additional data collected from students using the tutor in successive versions.

2. RELATED WORK

The Cognitive Algebra Tutor[9] is an algebra tutor that models student behavior based on the cognitive theory ACT-R and student data gathered from several previous studies. This almost entirely data-driven tutor improved student performance, and the authors noted that it would be improved the more data was collected. However, this system still took a long time and a great deal of expert involvement to design and improve. More recently Lee and Brunskill[4] examined the benefits and drawbacks to basing model parameters on existing data from individual students in comparison to data from an entire population, specifically as it pertained to the number of practice opportunities a student would require to master a skill. The authors estimated that using individualized parameters would reduce the number of practice opportunities a student would need to master a skill. Gonzalez et al.[2] demonstrated a data-driven model which automatically generated a cognitive and learning model based on previous student data in order to discover what skills students learn at any given time, and when they use skills they have learned. The resulting model predicted student behavior without the aid of previous domain knowledge and performed comparably to a published model.

Cluster-based classification in particular has several advantages when applied to data-driven tutoring, in that it can reveal unexpected learning behaviors which may not be incorporated into the expert-decided classification processes. Experts classify using their perception of the average students' performance[8] [10], which may be different from the actual participant group. Cluster-based classification methods, however, are able to classify and update classifications based on actual student behaviors. For example, Kizilec et al. [3] clustered MOOC learners into different engagement trajectories, and revealed several trajectories that are not acknowledged by MOOC designers. Despotovic-Zrakic et al. [1] adapted different course-levels, learning materials, and content in Moodle, an e-learning platform, for students in different clusters. Results showed that students with adapted course design had better learning gains, and a more positive attitude towards the course. The majority of previous work clustered students solely on their overall performance statistics. In contrast, our method clusters students based on their application of specific knowledge components throughout the tutor.

3. THE PROFICIENCY PROFILER

The Data-driven Proficiency Profiler (DDPP) is a system which calculates student proficiency at the end of each level in Deep Thought based on how a given student performs in comparison to exemplars who employed similar problem solving strategies (see Fig. 1), with rule scores weighted as determined through principal component analysis. Based on how similar exemplars were assigned in subsequent levels, the DDPP can determine the best proficiency level for a new student. In contrast to the DDML system previously

employed, this proficiency calculation and rule weighting is entirely data-driven, with no expert involvement.

We determined similar problem solving strategies among the exemplars by clustering the exemplars' rule scores (*KCs*) based on hierarchical clustering. Expert weighting was replaced by principal component analysis (PCA) of the frequency of the rules used for each exemplar for each level, accounting for 95% variance of the results. This percentage was determined through incremental testing at 5% intervals. From the results generated, 25% is the percentage that maximized accuracy. For each rule, its PCA importance value is the new weight for that rule score. Unlike expert authored weights, these rule score weights are based on each rule's importance as determined by the data.

When a new student uses the tutor, the student's rule scores are calculated throughout the level. At the end of each level, the DDPP examines each student's individual rule score and assigns it to a cluster for that rule. The DDPP then finds which clusters the scores for the most important rules fall into for that level (based on the same PCA based weighting), and then classifies that student into a *type* based on the set of clusters the student matches (see Fig. 1, right). Finally the system assigns the student to a proficiency track based on data from the matching type of exemplars, and how those exemplars were placed in the next level. In the event that a new student doesn't match an existing type in the exemplar data, their proficiency is calculated using the average scores, as in the original DDML system.

In a previous exploratory study, The DDPP method was compared alongside proficiency calculations using the minimum rule scores and average rule scores of exemplars, also weighted by primary component importance, to see how these methods compare to each other and to the expert authored system. Testing was performed on data collected from two courses using Deep Thought with the DDML system. Our results showed that the DDPP performed about as well as the average method overall, and did not approximate the accuracy of the original system. However our data set was very limited, and we hypothesized that the system would improve with additional data. The previous study was also exploratory in nature, and did not involve testing the DDPP in a system used by students[6].

4. STUDY

The DDPP was augmented with additional exemplars from the pool of students who used Deep Thought in Spring 2015. Exemplars from the Spring 2015 class were added to the existing dataset of exemplars. This dataset was then clustered based on the same hierarchical clustering methods for the DDPP described in [6]. The DDPP was then implemented into Deep Thought. This version of Deep Thought was used as a mandatory homework assignment in an undergraduate computer science course in Fall 2015. Students were required to solve 1-3 problems per level depending on proficiency, another 1-2 problems were randomly presented as worked examples. As students used Deep Thought, their actions and progress were recorded. The study results encompass three test groups; the control group, DDML(new) ($n = 65$), where students used Deep Thought with the DDML used to calculate proficiency between levels; the control comparison group, DDML(old) ($n = 47$), data from students who used Deep Thought with the DDML in identical conditions to the DDML(new) group the previous year; and the DDPP group

Figure 1: The Data-Driven Proficiency Profiler. (Left) At each level interval, exemplar KC scores are clustered, and exemplars are assigned a cluster for each KC Score. (Center) KCs that make up 25% of importance in the current level are used to assign exemplars to types. (Right) New student scores are assigned to clusters, and compared to existing types to determine next level path.

($n = 27$), students who used the version of Deep Thought where the DDPP is used for between level assessment rather than DDML, and all other conditions are identical. Students in all three groups were taught by the same instructor, using the same curriculum. We examine data from the DDML(old) group in order to analyze the consistency in the DDML's performance between years.

5. RESULTS

There were no significant results between the DDML control group (DDML-Control) and the DDPP group. However there were some interesting trends in tutor dropout, rule application scores, and path prediction accuracy that indicate the DDPP's ability to improve and perform well with additional data.

Path Prediction Accuracy

Table 1 shows the path prediction accuracy of the DDPP and the DDML-Control group in comparison to the path prediction accuracy of the DDPP and the DDML at the time of the exploratory study (DDPP(E) and DDML(E) respectively)[6]. As previously mentioned in [6], path prediction accuracy is based on the possibility that a student may not finish a level in the same proficiency path that the system originally assigned them to due to skipping a certain number of times within a level. If the student did not finish the level on the same proficiency path, it is an indication that the proficiency calculation system may have initially assigned the student to the wrong proficiency path. Therefore we can calculate the accuracy of the original system by determining how often students who completed the entire tutor changed proficiency paths throughout. Given $S_{sameTrack}$ as the number of students who finished a level on the same proficiency track, and S_{total} as the total number of students who completed the level, the path prediction accuracy for each level (*LevelAccuracy*) is calculated as follows:

$$LevelAccuracy = \frac{S_{sameTrack}}{S_{total}} \quad (1)$$

There was improvement in the path prediction accuracy in the DDPP between the exploratory study and this current study by 23-30% in each level. Additionally, while the

DDPP in the exploratory study ranged from 52-74% accuracy, the DDPP in the current study was never less than 80% accurate. In Levels 3 and 5, the DDML was slightly less accurate than the DDPP. This confirms our earlier hypothesis that additional data would improve the DDPP's performance and enable it to calculate student proficiency much more accurately.

Table 1: Path prediction accuracy of the DDPP and DDML-Control in the new study, and the accuracy of the DDPP and DDML-Control in the exploratory study (E) at the end of each level.

	DDPP	DDML	DDPP(E)	DDML(E)
Lvl 1	88.9%	93.8%	58.0%	94.0%
Lvl 2	92.6%	93.8%	66.0%	96.0%
Lvl 3	81.5%	75.4%	52.0%	88.0%
Lvl 4	81.5%	89.2%	58.0%	86.0%
Lvl 5	100%	89.0%	74.0%	92.0%

Table 2 shows the number of types identified in each level, the number of those types that were matched, and the percentage of students that were successfully assigned to a type. These metrics are an indication of the DDPP's ability to distinguish varying problem solving methods at each level, as well as its ability to match new students to a type. Overall the percentage of students assigned to a type was very high, between 82% and 100%. The number of types found increased steadily over the course of the tutor, from 14 types in Level 1 to 44 types in Level 5, indicating an increasing number of possible problem solving types as the tutor problems are more complex. The number of those types that were matched ranged from 9-15.

Student Dropout

Students are more likely to perform better in-tutor when given problems in their zone of proximal development through scaffolding of major concepts [7]. Proper scaffolding increases learning, and therefore increases the likelihood of further progression through the tutor and exposure to higher-level concepts. Figure 2 shows the rate and level of student dropout, shown via the percentage of students assigned the tutor that were still using it at any given level. The DDPP

Table 2: The number of types identified per level from historical data (DDML), the number of types matched with students using the DDPP system, and the percentage of students who were matched an existing type in the DDPP system.

Level	1	2	3	4	5
Types Found (DDML)	14	17	19	22	44
Types Matched (DDPP)	10	9	11	9	15
% Students Matched (DDPP)	100	96.3	82.4	94.1	89.2

Figure 2: The percentage of students at the end of each level of the tutor.

resulted in the lowest dropout rate in comparison to the previous and current iteration of the DDML. Additionally, all of the students stayed in the tutor until Level 6, where roughly 7% of the students stopped. In the current version of the DDML the students started dropping out around Level 3, while in the older DDML study the students started dropping out as early as Level 2. As a result, a far higher percentage of the students using the DDPP were exposed to the advanced concepts introduced in the problems in later levels of Deep Thought than any set of students using the DDML. A higher percentage of the students using the DDPP also completed the entire tutor in comparison to the students using the DDML in earlier semesters or the current semester. 93% of the students using the DDPP completed the entire tutor, while only 91% of the students using the DDML in this current study and 90% of the students using the DDML in previous semesters completed the entire tutor.

Rule Application

Figure 3 shows the median percentage of correct rule applications at each level of the tutor. A correct rule application is when a student uses a logical rule in the correct way in the proof. We track the change in percentage in correct rule applications in order to track whether students are applying rules more accurately as they progress through the tutor. The DDPP, new DDML, and old DDML all showed an increase in the correct percentage of rule applications between

Figure 3: The median percentage of correct rule application in each level of the tutor. Each level's calculation is independent of the other levels.

Levels 1 and 2. After Level 2 the percentage decreases, and then levels out, for students in this or the previous study using the DDML. For students using the DDPP on the other hand, there was a steady increase in the percentage of correct rule applications as they progress through the tutor. This steady increase in understanding is supported by the post-test results. The median Post-Test Scores for DDML-Control is 75%, for DDPP is 87.5%.

6. DISCUSSION

While there were no statistically significant results between the DDPP and the DDML-Control group, there were several metrics that showed a trend toward improvement. The previous exploratory study [6] showed some potential for the DDPP based on path prediction accuracy, but it was far less accurate than the DDML. The DDPP path prediction accuracy at the time of the exploratory study ranged from 52% to 74%. Since the DDPP was augmented with additional student data however, the path prediction accuracy increased to a range of 81.5% to 100%. This indicates that the DDPP was able to place students on a proper proficiency path much more often once there was additional data.

The number of types found, matched, and the percentage of students that were matched are additional indications of the DDPP's improved performance. In the exploratory study, there were only at most 26 types found, and the number of those types that were matched to students in the test set ranged from 0 to 10. In this latest study, there are a couple notable changes. The number of types found at each level increased overall, indicating that the DDPP was able to find a greater variety of problem solving strategies. This is especially noticeable at Level 5, where 44 types were found by the DDPP as opposed to 26 in the exploratory study. The other notable change is that more of these types are matched with students. The number of types matched ranged from 9 to 15, indicating that the DDPP was able to find types that correspond to the way multiple students solve problems and was able to detect those types in new students. Proportionally, there are still very few types matched to a student, particularly in the later levels. Level 4 only matched 9 out

of 22 possible types, and Level 5 only matched 15 out of 44 types. However, the percentage of students that were matched was very high throughout all levels of the tutor. This would indicate that a majority of the students are being assigned to a proportionally small number of types. The percentage of students assigned to a type fluctuates however, which could be due to a lack of data that would include still more problem solving types at each level.

Considering the rate of tutor dropout and the rule application accuracy throughout the levelsthe DDPP guided the students through the tutor in a way that was overall beneficial. None of the students using the DDPP dropped out of the tutor before Level 6, so more of those students were able to be exposed to the advanced concepts presented in the later level problems. This dropout occurs much later than in the older or newer DDML groups. This indicates that the DDPP was able to properly assign students to the proficiency path that contained problems close to their proximal development in terms of difficulty, reducing the frustration that would cause students to stop their progress in the tutor prematurely.

The rule application accuracy shown in Figure 3 indicates that students using the DDPP increased their understanding of logic rules. All the groups show an increase in rule application accuracy from Level 1 to Level 2. However, whereas the students in the old and new DDML group maintained a steady level of rule application accuracy throughout the rest of the tutor, the students in the DDPP increased the percentage of accurate rule applications throughout the tutor, to the point where those students had a higher median percentage of correct rule applications than either of the DDML groups. This confirms that the DDPP led the students through the tutor in a way that enabled them to learn concepts and improve on how they apply them.

7. CONCLUSIONS & FUTURE WORK

We have conducted a proof of concept study in order to determine the performance of our data-driven proficiency profiler (DDPP) in comparison to a data-driven mastery learning system based on average score thresholds and expert-determined rule priorities. We had previously conducted an exploratory study on the DDPP and determined that the DDPP's performance would improve with additional data. This was the first semester where we could test the effectiveness of the DDPP with additional data and fully incorporated into Deep Thought.

In terms of the DDPP's performance, the results were not statistically significant between the DDPP group and either of the DDML groups. However, the trends in student dropout and rule application accuracy indicate that the DDPP performs at least as well as the DDML in guiding students through Deep Thought so they reach later levels while gaining greater understanding of the knowledge components. The improvements in path prediction accuracy, types matched, and percentage of students matched to a type between the DDPP at the time of the exploratory study and the DDPP incorporated into Deep Thought confirms that the DDPP does improve with additional data. Overall, the results demonstrate that the DDPP could feasibly replace the current proficiency calculation based on the average rule scores and expert-determined rule priorities, resulting in Deep Thought as a tutor being completely data-driven in its behavior.

Future work with the DDPP would involve incorporating additional Deep Thought tutor data from this semester, which we hypothesize would improve the DDPP's performance in a more significant way. Data from the Fall 2015 students who completed Deep Thought will be clustered and the results will be incorporated into the next iteration of the DDPP used in Deep Thought. This next iteration would also be tested on a larger pool of students, as the number of students using the DDPP was still limited in size.

8. ACKNOWLEDGEMENTS

This material is based on work supported by the National Science Foundation under Grants 1432156 and 0845997.

9. REFERENCES

- [1] M. Despotovic-Zrakic, A. Markovic, Z. Bogdanovic, D. Barac, and S. Krco. Providing adaptivity in moodle lms courses. *Educational Technology Society*, 15(1):326–338, 2012.
- [2] J. Gonzalez-Brenes and J. Mostow. What and when do students learn? Fully data-driven joint estimation of cognitive and student models. In *International Conference on Educational Data Mining (EDM 2013)*, pages 236–239, 2013.
- [3] R. Kizilcec, C. Piech, and E. Schneider. Deconstructing disengagement: analyzing learner subpopulations in massive open online courses. In *International Conference on Learning Analytics and Knowledge (LAK 2013)*, pages 170–179, 2013.
- [4] J. Lee and E. Brunskill. The impact on individualizing student models on necessary practice opportunities. In *International Conference on Educational Data Mining (EDM 2012)*, pages 118–125, 2012.
- [5] B. Mostafavi, M. Eagle, and T. Barnes. Towards Data-driven Mastery Learning. In *International Conference on Learning, Analytics, and Knowledge (LAK 2015)*, 2015.
- [6] B. Mostafavi, Z. Liu, and T. Barnes. Data-Driven Problem Profiling. In *International Conference on Educational Data Mining (EDM 2015)*, pages 335 – 341, 2015.
- [7] T. Murray and I. Arroyo. Toward Measuring and Maintaining the Zone of Proximal Development in Adaptive Instructional Systems. In *International Conference on Intelligent Tutoring Systems (EDM 2002)*, pages 289 – 294, 2002.
- [8] E. Perez, L. Santos, M. Perez, J. de Castro Fernandez, and R. Martin. Automatic classification of question difficulty level: Teachers' estimation vs. students' perception. In *IEEE Frontiers in Education Conference*, pages 1–5, 2012.
- [9] S. Ritter, J. Anderson, K. Koedinger, and A. Corbett. Cognitive tutor: Applied research in mathematics education. *Psychonomic Bulletin/Review*, 14(2):249–255, 2007.
- [10] G. van de Watering and J. van der Rijt. Teachers and students perceptions of assessments: A review and a study into the ability and accuracy of estimating the difficulty levels of assessment items. *Educational Research Review*, 1(2):133–147, 2006.