Optimizing Partial Credit Algorithms to Predict Student Performance

Korinn Ostrow, Christopher Donnelly, Neil Heffernan
Worcester Polytechnic Institute
100 Institute Road
Worcester, MA 01609
{ksostrow, cdonnelly, nth}@wpi.edu

ABSTRACT

As adaptive tutoring systems grow increasingly popular for the completion of classwork and homework, it is crucial to assess the manner in which students are scored within these platforms. The majority of systems, including ASSISTments, return the binary correctness of a student's first attempt at solving each problem. Yet for many teachers, partial credit is a valuable practice when common wrong answers, especially in the presence of effort, deserve acknowledgement. We present a grid search to analyze 441 partial credit models within ASSISTments in an attempt to optimize per unit penalization weights for hints and attempts. For each model, algorithmically determined partial credit scores are used to bin problem performance, using partial credit to predict binary correctness on the next question. An optimal range for penalization is discussed and limitations are considered.

Keywords

Partial Credit, Student Modeling, Next Question Correctness, Adaptive Tutoring Systems, Maximum Likelihood, Grid Search

1. INTRODUCTION

Adaptive tutoring systems provide rich feedback and an interactive learning environment in which students can excel, while teachers maintain data-driven classrooms by using the systems as powerful assessment tools. Simultaneously, these platforms have opened the door for researchers conducting minimally invasive educational research at scale while offering new opportunities for student modeling. Still, they are commonly restricted to measuring performance through binary correctness on each problem. Arguably the most popular form of student modeling within computerized learning environments, Knowledge Tracing, is rooted in the binary correctness of each opportunity or problem a student experiences within a given skill [1]. Knowledge Tracing (KT) drives the mastery-learning component of renowned tutoring systems including the Cognitive Tutor series, allowing for real time predictions of student knowledge, skill mastery, or next problem correctness [4]. Similar modeling methods consider variables that extend beyond correctness but rarely escape the binary nature of the construct, including Item Response Theory [2] and Performance Factors Analysis [9]. By restricting input to a binary metric across questions, these modeling techniques fail to consider a continuous metric that is commonplace for many teachers: partial credit.

Partial credit scoring used within adaptive tutoring systems could provide more individualized prediction and thus establish models with better fit. It is likely that binary correctness has remained the default for learning models due to the inherent difficulty of defining a universal algorithm to generalize partial credit scoring across platforms. Some of the onus may also fall on users' familiarity with current system protocol; students tend to avoid using system feedback regardless of the benefits it may provide because requesting feedback results in score penalization. However, the primary goal of these platforms is generally to promote student learning rather than simply acting as an assessment tool, and thus, binary correctness is flawed.

The present study considers data from ASSISTments, an online adaptive tutoring system that provides assistance and assessment to over 50,000 users around the world as a free service of Worcester Polytechnic Institute. Researchers have previously used ASSISTments data to modify student-modeling techniques in a variety of ways including student level individualization [7], item level individualization [8], and the sequence of student response attempts [3]. Previous work has also shown that naïve algorithms and maximum likelihood tabling methods that consider hints and attempts to predict next problem correctness can be successful in establishing partial credit models meant to supplement KT [10; 11]. More recently, algorithmically derived partial credit scoring resulted in stand-alone tabled models using data from only the most recent question and yet showing goodness of fit measures on par with KT at lower processing costs [6]. However, we hypothesize that some conceptualizations of partial credit may lead to better predictive models than others. Rather than subjectively defining tables or algorithms, a data driven approach should be considered. Thus, considering student performance within the ASSISTments platform, the current study employs a grid search on per unit penalizations of hints and attempts to ask:

1. Based on penalties for hints and attempts dealt per unit, is it possible to algorithmically define partial credit scoring that optimizes the prediction of next problem correctness?
2. Does the optimal model of partial credit differ across different granularities of dataset analysis?

Establishing an optimal partial credit metric within ASSISTments would allow teachers using the tool to more accurately assess student knowledge and learning, while allowing students to alter their approach to system usage by taking advantage of adaptive feedback. The optimization of partial credit scoring would also enhance student modeling techniques and offer a new approach to answering complex questions within the domain of educational data mining.

2. DATA

The ASSISTments dataset used for the present study is comprised solely of assignments known as Skill Builders. This type of assignment requires students to correctly answer three consecutive questions to complete the problem set. Questions are randomly pulled from a large pool of skill content and are typically presented with tutoring feedback, most commonly in the form of hints. The dataset has been de-identified and is available at [5] for further investigation.

The dataset used in the present study is a compilation of Skill Builders from the 2012-2013 school year, containing data for 866,862 solved problems. Recorded data includes students' performance on the problem (i.e., binary correctness, hint count, attempt count), variables that identify the problem itself (i.e., problem type, unique problem identification number) and information pertaining to the assignment housing the problem (i.e., unique identifiers for assignments, skill type, teachers, and schools). The dataset was representative of 120 unique skills and 24,912 unique problems, solved by 20,206 students.

On average, students made 1.53 attempts per problem (SD = 15.08). The minimum number of attempts was 0 (i.e., a student who opened the problem and then left the tutor), while the maximum number of attempts was a daunting 12,246 (i.e., a student who hit 'Enter' repeatedly for a prolonged period of time, likely out of frustration or boredom). Students made a total of 1,324,226 attempts across all problems. The majority of problems (74.9%) had just one logged attempt per student (typically correct answers), while 15.1% of problems carried only two logged attempts.

Hint usage among all students averaged 0.61 hints per problem (SD = 1.29). The minimum number of hints used was 0 (i.e., no feedback requested), while the maximum number of hints used was 10. Interestingly, the maximum number of hints available for any particular problem was 7. Thus, a handful of students who logged more than 7 hints were accessing the tutor in multiple browser windows (i.e., cheating). On average there were 3.22 hints available per problem (SD = 0.89). The majority of problems contained 3 hints (44.6%), 4 hints (28.9%), or 2 hints (18.2%). Although there were 2,768,299 hints available across all problems, students only used 529,394 hints, or approximately 19% of available feedback. Bottom out hints, or those providing the problem's solution, were only used on 146,742 (16.9%) of problems.

Additional analyses were performed on the 261,787 problems that students answered incorrectly out of the original 866,862 problems solved. Within this subset of data, students made an average of 2.75 attempts per problem (SD = 27.40). Students also used an average of 2.02 hints (SD = 1.63). This subset of problems had 860,131 total hints available, of which students used 528,644 hints (61.5%).

Hint usage would likely increase if partial credit scoring was implemented within the ASSISTments platform. In many classrooms, binary first attempt scoring has created an environment in which students are afraid to use hints although they would benefit from feedback, as they know they will receive no credit. Further, the dataset suggests that once students are marked wrong, they are more likely to jump through all available hints and seek out the answer (56% of incorrect first attempts led to bottom out hinting). This reflects another substantial downfall in the system's current protocol: once the risk has passed, so has the drive to learn. The implementation of partial credit scoring has the potential to alleviate this misuse.

3. METHODS

The present study presents an extensive grid search of potential per hint and per attempt penalizations. The full dataset was used to define partial credit scores algorithmically based on per unit penalizations ranging from 0 to 1 in increments of 0.05 for both hints and attempts. Thus, for each solved problem in the dataset, 441 partial credit scores were established based on each possible combination of per unit penalization. For example, in a model in which each attempt earned a penalization of 0.05, and each hint earned a penalization of 0.1, a student who made three attempts and used one hint would receive a penalty of 0.25 ((3x0.05) + (1x0.1)), effectively scoring 0.75 on that problem. This process was used to score each problem in the dataset for each possible penalty combination, with a floored per problem score of 0 (students could not receive negative scores). This method was similar to that presented by Wang & Heffernan in the Assistance Model [10] which established a tabling method to calculate probabilities of next problem correctness based on combinations of hints and attempts that resulted in twelve possible bins or parameters.

For each of the 441 partial credit models, a maximum likelihood tabling method was employed using five fold cross validation. Within each model, a modulo operation was used on each student's unique identification number to assign students to one of five folds. Note that this method resulted in folds that all represented approximately 20% of students in the dataset. Maximum likelihood probabilities for next problem correctness were then calculated for each partial credit score within each model. Table 1 presents an average of test fold probabilities for the model in which each attempt and each hint are penalized 0.1. For instance, a student using two attempts (2 x 0.1) and one hint (1 x 0.1) would be penalized 0.3, thus falling into the score bin of 0.7 (PC Score). Following through with this example, based on 11,174 problems solved that fit this scoring structure, the average of known binary performance on the following problem was 0.599. This value becomes the prediction for next problem correctness for students scoring 0.7 on the current problem.

Using the maximum likelihood probabilities for next problem correctness within each test fold as predicted values, residuals were then calculated by subtracting predictions directly from actual next problem binary correctness (i.e., 1 – 0.725 = 0.275; 0 – 0.571 = -0.571). This approach was used rather than selecting an arbitrary cutoff point to classify a prediction as correct or incorrect in the binary sense (i.e., values greater than or equal to 0.6 serve as predictions of correctness) because it reduced the potential for researcher bias.

Table 1. Probabilities averaged across test folds for the model in which the penalization per hint and per attempt is 0.1

PC Score	n	Max. Likelihood NPC
0	149,504	0.467
0.1	422	0.571
0.2	685	0.581
0.3	1,055	0.578
0.4	1,784	0.574
0.5	3,442	0.583
0.6	6,623	0.585
0.7	11,174	0.599
0.8	18,679	0.662
0.9	49,972	0.725
1.0	476,523	0.802

4. RESULTS

For each model, residuals were used to calculate RMSE, R^2 & AUC at three levels of granularity: problem level, student level, and skill level. Heat maps are only presented here for RMSE, as the other metrics established almost identical maps. Metrics representing greater model fit are depicted using the purple end of the spectrum, while those representing poorer fit are represented using the red end of the spectrum. Further, a series of ANOVAs were conducted to compare each set of models within the same penalization level for attempts and hints. For example, the 21 models in which attempt penalty was set to 0.2 were compared to all other sets of attempt penalty models to investigate significant differences across penalties. This method was used rather than comparing each model with all other models using paired samples t-tests, as the resulting 194,481 analyses (441^2) would greatly inflate the rate of Type I error without unrealistic corrections.

Initial analysis was performed at the problem level; residuals were calculated for each problem that contained next problem correctness metrics and goodness of fit measures were averaged across the dataset. Each metric followed a similar structure in which low attempt penalties appear to result in better fitting models, while hint penalty does not appear to be significant. Thus, partial credit scoring algorithms using lower penalties for attempts were better at predicting next problem performance, as depicted in Figure 1. The ANOVA results depicted in Table 2 suggest that differences in attempt penalty models were significant. Thus, the set of models with per attempt penalties of 0.1 differed significantly from the set of models with per attempt penalties of 0.8. Differences among hint penalty models were not reliably significant. Figure 1 also suggests that the current binary scoring protocol used by ASSISTments results in predictive models that are inadequate. First attempt binary correctness is the equivalent of the model in which per attempt and per hint penalty are both set to 1, or the upper right corner of each heatmap). This model resulted in consistently poor fit metrics, suggesting that modeling techniques such as KT should employ continuous or binned partial credit values as input as they enhance next problem prediction ability. It has not yet been investigated how this alteration would change the prediction of other variables commonly predicted through KT, such as latent student knowledge or skill mastery.

Student level analysis was undertaken using a subset of the original data file. At this granularity, goodness of fit metrics were calculated for each student and averaged across students to obtain final metrics for each of the 441 models. As the ASSISTments system measures completion of a Skill Builder as three consecutive correct answers, a number of high performing students had limited opportunity counts within skills. For students with too few data points, it was not possible to calculate R^2 and AUC. Therefore, student level analysis incorporated 7,429 students from the original dataset, or 651,849 problem logs. Answering our second research question, it appears as though the region of optimal partial credit values observed at the problem level remains consistent at the student level, as shown in Figure 2. ANOVA results depicted in Table 2 show reliably significant differences across attempt penalty models but not across hint penalty models.

Skill level analysis was also undertaken using a subset of the original data file. One skill did not have enough data based on a low number of users and high mastery within those users, and was

Table 2. ANOVA results for groups of attempt and hint penalty models at each level of analysis

			Attempt Penalty			Hint Penalty		
Level	*Min*	*Max*	*F*	*p*	R^2	*F*	*p*	R^2
Problem								
RMSE	.430	.435	302.70	.000	.935	0.95	.519	.043
AUC	.626	.655	295.46	.000	.934	1.14	.304	.052
R^2	.070	.091	304.34	.000	.935	0.95	.525	.043
Student								
RMSE	.424	.429	222.49	.000	.914	1.34	.149	.060
AUC	.578	.593	208.19	.000	.908	1.42	.106	.063
R^2	.096	.110	374.52	.000	.947	0.80	.715	.037
Skill								
RMSE	.423	.429	517.85	.000	.961	0.55	.944	.026
AUC	.624	.647	250.17	.000	.923	0.72	.805	.033
R^2	.073	.090	510.96	.000	.961	0.49	.971	.023

Note. For all models, df = (20, 420).

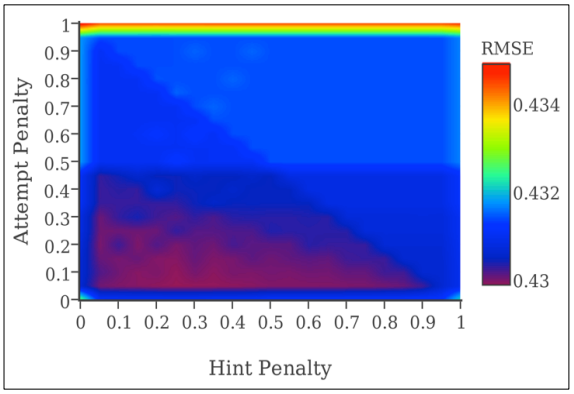

Figure 1. Problem Level RMSE

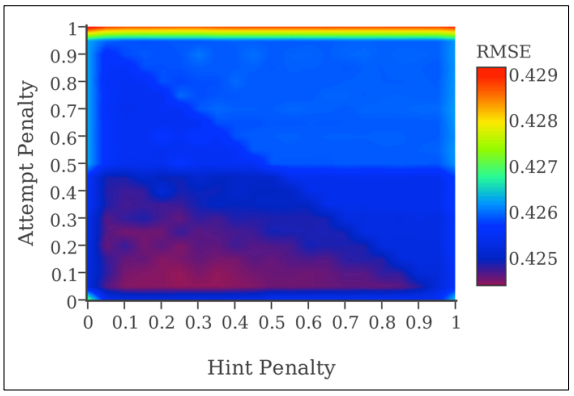

Figure 2. Student Level RMSE

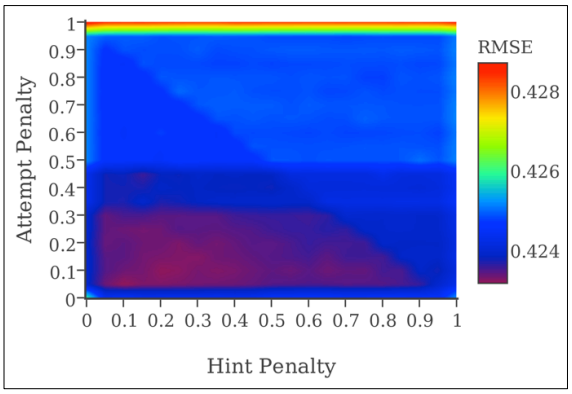

Figure 3. Skill Level RMSE

excluded from skill level analysis, resulting in a file with 119 skills. At this granularity, goodness of fit metrics were calculated for each skill and averaged across all skills to obtain final metrics for each of the 441 models. Results are depicted in Figure 3. The heat map shows that the region of optimal penalization has grown more concise, showing optimal fit among models with low per hint and per attempt penalties (< 0.3). ANOVA results depicted in Table 2 again suggest reliably significant differences in all metrics across attempt penalty models but not across hint penalty models.

Post-hoc analyses were conducted on ANOVA results using multiple comparisons to examine significant differences between attempt penalty and hint penalty model groups when considering problem level AUC. Using a Bonferroni correction to reduce Type I error, this process resulted in a series of significance estimates for penalty group comparisons (i.e., all models where attempt penalty is 0.1 compared to all models where attempt penalty is 0.3 results in a non-significant difference, $p = 0.88$). Results suggested that models close in penalty were less likely to differ significantly than models with greater difference in penalty. For instance, models with an attempt penalty of 0.1 were significantly different than those with an attempt penalty of 0.4, but were not significantly different than those with an attempt penalty of 0.2. This information can be used to help optimize partial credit penalizations, as it may be more motivating and productive for students to receive smaller penalizations. Such information could also allow systems like ASSISTments to define a range of possible penalizations that could then be refined by the teacher, providing all users with a greater sense of control.

5. DISCUSSION & CONTRIBUTION

The initial findings of a grid search on partial credit penalization through per unit hint and attempt docking suggest that the implementation of partial credit within adaptive tutoring systems can be established using a data driven approach that will ultimately produce stronger predictive models of student performance while enhancing the way adaptive tutoring systems are used by students and teachers.

Our first research question was answered with a resounding "Yes," certain algorithmically derived combinations of partial credit penalization are better than others when used to predict next problem performance. Optimal partial credit models were visible in heat maps spanning three levels of data granularity and remained relatively consistent across granularities, thus answering our second research question. ANOVAs revealed that differences in attempt penalty models were consistently significant across dataset granularities, while differences in hint penalty models were not reliable. This finding is likely due to the fact that hint usage is lower and less distributed than attempt count across problems in the dataset, and it is possible that this finding would diminish in a system that more readily promoted the use of tutoring feedback without penalization, or a system already employing partial credit scoring.

The partial credit models that we define here as optimal, based on their ability to predict next problem performance, were models with per hint and per attempt penalties of 0.3 or less. Additional analyses revealed that at the problem level, there should be no reliable difference in predictive ability of a model penalizing 0.3 per attempt from a model penalizing 0.1 per attempt, with variable hint penalization. This finding suggests that less penalization is just as effective, offering an opportunity to consider student motivation and affect when defining a partial credit algorithm. This grid search also revealed that partial credit metrics outperform binary metrics when predicting next problem performance, as previously shown in [6]. Thus, it is possible to improve prediction of student performance within adaptive tutoring systems simply by implementing partial credit scoring. It should also be noted that a leading limitation of the approach presented here is that we have only been predicting next problem correctness, rather than latent variables such as skill mastery or student knowledge. It is possible that optimizing partial credit would also provide benefits for the prediction of latent effects, but further research is necessary in this domain.

6. ACKNOWLEDGMENTS

We acknowledge funding from NSF (1316736, 1252297, 1109483, 1031398, 0742503, 1440753), ONR's "STEM Grand Challenges," and IES (R305A120125, R305C100024). Thanks to S.O. & L.P.B.O.

7. REFERENCES

[1] Corbett, A.T. & Anderson, J.R. (1995). Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge. *User Modeling and User-Adapted Interaction*, 4: 253-278.

[2] Drasgow, F. & Hulin, C.L. (1990). Item response theory. In M.D. Dunnette & L.M. Hough (Eds.), Handbook of Industrial and Organizational Psychology, Vol. 1, pp 577-636. Palo Alto, CA: Consulting Psychologists Press.

[3] Duong, H.D., Zhu, L., Wang, Y., & Heffernan, N.T. (2013). A Prediction Model Uses the Sequence of Attempts and Hints to Better Predict Knowledge: Better to Attempt the Problem First, Rather Than Ask for a Hint. In S. D'Mello, R. Calvo, & A. Olney (Eds.) Proceedings of the 6th Int Conf on EDM. pp. 316-317.

[4] Koedinger, K.R. & Corbett, A.T. (2006). Cognitive tutors: Technology bringing learning science to the classroom. In K. Sawyer (Ed.), *The Cambridge handbook of the learning sciences* (pp. 61-78). NY: Cambridge University Press.

[5] Ostrow, K. (2014). Optimizing Partial Credit Data. Accessed 12/8/14. https://tiny.cc/OptimizingPartialCredit

[6] Ostrow, K., Donnelly, C., Adjei, S. & Heffernan, N. (2015). Improving Student Modeling Through Partial Credit and Problem Difficulty. In Russell, Woolf & Kiczales (Eds.), Proceedings of the 2nd ACM Conf on L@S. pp. 11-20.

[7] Pardos, Z.A. & Heffernan, N.T. (2010). Modeling Individualization in a Bayesian Networks Implementation of Knowledge Tracing. In Proceedings of the 18th Int Conf on UMAP. pp. 255-266.

[8] Pardos, Z.A., & Heffernan, N.T. (2011). KT-IDEM: Introducing Item Difficulty to the Knowledge Tracing Model. In Joseph A. Konstan et al. (Eds.): UMAP 2011, LNCS 6787, pp. 243-254.

[9] Pavlik, P.I., Cen, H., Koedinger, K.R. (2009). Performance Factors Analysis - A New Alternative to Knowledge Tracing. In: Proceedings of the 14th Int Conf on AIED, pp. 531-538.

[10] Wang, Y. & Heffernan, N.T. (2011). The "Assistance" Model: Leveraging How Many Hints and Attempts a Student Needs. The 24th International FLAIRS Conference.

[11] Wang, Y. & Heffernan, N. (2013). Extending Knowledge Tracing to Allow Partial Credit: Using Continuous versus Binary Nodes. In K. Yacef et al. (Eds.) AIED 2013, LNAI 7926, pp 181-188.