The real world significance of performance prediction

Zachary A. Pardos
Department of Computer Science
Worcester Polytechnic Institute
100 Institute Rd.
Worcester, MA 01609, USA
zpardos@wpi.edu

Qing Yang Wang
Department of Computer Science
Worcester Polytechnic Institute
100 Institute Rd.
Worcester, MA 01609, USA
wangqy@wpi.edu

Shubhendu Trivedi
Department of Computer Science
Worcester Polytechnic Institute
100 Institute Rd.
Worcester, MA 01609, USA
shubhendu_trivedi@ieee.org

ABSTRACT

In recent years, the educational data mining and user modeling communities have been aggressively introducing models for predicting student performance on external measures such as standardized tests as well as within-tutor performance. While these models have brought statistically reliable improvement to performance prediction, the real world significance of the differences in errors has been largely unexplored. In this paper we take a deeper look at what reported errors actually mean in the context of high stakes test score prediction as well as student mastery prediction. We report how differences in common error and accuracy metrics on prediction tasks translate to impact on students and depict how standard validation methods can lead to overestimated accuracies in these prediction tasks. Two years of student tutor use and corresponding student state test scores are used for the analysis of test prediction while a simulation study is conducted to investigate the correspondence between performance prediction error and latent knowledge prediction.

Keywords

Evaluation, Cross-validation, Interpretation of error, Simulation

1. INTRODUCTION

An open question among EDM researchers and policy makers with an interest in EDM techniques is what impact the techniques reported on will have on students and what performance to expect under real world model training constraints. The majority of analytical papers presented in the literature using educational datasets use n-fold cross-validation. This has become an expected standard and a justifiable one which offers clear statistical reliability benefits over a single test and train hold out strategy. However, as an applied field it is important to take a step back from the manipulation of datasets and consider what factors may impact the expected performance in a real world deployment of a method. Often the culprit of inflated cross-validated accuracy is the disregard for time constraints in temporal data. Because this type of data is predominant in the field due to the temporal nature of studying learning, it is especially important to keep violations of time in mind in the evaluation and reporting of our models.

Data leakage [1] is the more general term for using information during training or prediction that should not legitimately be available. This kind of leakage of data from the future has been prevalent in many data mining competitions including the 2010 KDD Cup on educational data [2]. In that competition, for example, a student's answers from Unit 2 could be used to predict her responses to questions of a related skill in Unit 1. While the models which were designed to predict that type of test set may very well also push the state of the art in real world prediction scenarios, the prediction accuracies reported in that competition do not reflect real world performance expectation. Furthermore, the relative rankings of algorithms in the competition may vary when future information is not available. We investigate the effect of leakage on the task of predicting end of year state test scores in section 2.

Removing leakage from evaluation adds confidence in replicating the reported error in real world deployment, however; of equal significance to deployment considerations is the real world meaning of the error and its implications for students. Recent work on ensemble learning with educational data [3] chronicles the various models introduced in the past years which track student knowledge. A common practice among these papers has been to compare the error of a newly introduced method to that of a longer established method. Generally, the merit of the new method is compelling if it demonstrates a statistically reliable improvement in error over the established method. With larger educational datasets becoming widely available, such as the 20M row 2010 KDD Cup dataset[1] [2] or the 1M row ASSISTments Platform dataset[2], finding statistical differences in models can be achieved even with prediction error differences among models only discernible at the third or fourth decimal. This raises the question of whether or not statistical tests are a useful yard stick when large datasets are being analyzed and more importantly it raises the question of what errors and various magnitudes of differences in errors actually mean in terms of their impact on students. The most practical application of these models, and a reason for their high relevancy in the literature, is to predict when a student has attained mastery of a particular skill. Improved accuracy of these knowledge tracing models is appealing because it presumes that the prediction of mastery will also be more accurate and thus reduce the amount of over and under practice on skills, a time saving benefit that teachers greatly appreciate. In section 3 we run a simulation study to investigate the meaning of errors in terms of knowledge assessment. In the simulation study both student knowledge and response data is generated from a standard model of learning. We evaluate the generated response data with several models to evaluate the correspondence between performance prediction metrics and accurately inferring when mastery was attained.

Best practices for calculating statistically reliable difference between predictions is an open question, however, a frequent approach is to calculate a paired t-test of prediction squared errors [2] or a Wilcoxn signed rank sum test on per student Area Under the Curve (AUC) also referred to as A' [3].

[1] http://pslcdatashop.web.cmu.edu/KDDCup/downloads.jsp

[2] http://teacherwiki.assistment.org/wiki/Assistments_2009-2010_Full_Dataset

2. PREDICTION OF STATE TEST SCORES

In this section we evaluate the effect of leakage in predicting state test scores and also provide an analysis of the impact of error differences on student test score prediction.

2.1 Dataset

We used two datasets [4] collected from the 2004-2005 and 2005-2006 school years usage of the ASSISTments Platform among 8th grade math students in four Massachusetts high schools. The datasets had 627 and 760 students respectively. Both datasets were organized with one row per student and six features summarizing each student's usage in the system for that year. The per student features were: overall percent correct, number of problems answered, percent correct on scaffold questions, average time spent per problem, average attempts per problem, and average number of hints requested per problem. The seventh feature is the student's end of year math state test score, which is the target being predicted. The state test is the Massachusetts Comprehensive Assessment System (MCAS). The minimum raw score for the test is 0 and the maximum score is 54. The raw score is scaled to a score between 200 and 280. The scaled score contains four ranges that correspond to the following proficiency categories shown in Table 1.

Score Range	Category
200-218	Failing
220-238	Needs Improvement
240-258	Proficient
260-280	Advanced

Table I. Proficiency categories for the MCAS test

While the scaled score ranges always map to the same categories, the raw score mapping to scaled score changes yearly and is only computed after all tests are received and evaluated by the state. This presents an additional challenge for category prediction; however, scaling is just one of many sources of change between the two years' data. Changes in the tutor as well as changes in student instruction outside the tutor also contribute variance and are a part of why a two year train/test procedure might be more difficult to predict than a one year cross-validated.

The MCAS test is a high stakes test because of the significance of scoring in the Failing or Advanced categories. Failing category students cannot graduate high school, regardless of their class grade, while students who score in Advanced receive an automatic state college scholarship. For this reason, interested parties want to know a prediction of the student's category, not just raw score.

2.2 Methodology

Two prediction algorithms were used and two hold out strategies. Multiple algorithms were chosen not for the sake of comparison but rather to see if the relative performance of the algorithms changes between hold out strategies. An algorithm that does not fit the cross-validated set very well may capture the appropriate level of generality to be better in the train/test scenario. The two prediction algorithms chosen were linear regression, used in prior work with this dataset [4] and Random Forests [6], a highly affective algorithm from the machine learning community. We also include a K-means clustering technique that claimed to improve prediction accuracy of algorithms on this same dataset [5]. This K-means enhancement is an ensemble technique [3] and we include it to see if it underperforms in the train/test hold out.

Two hold out strategies were used, one to demonstrate a typical 5-fold cross-validation hold out which contains future information leakage and the other strategy uses the previous year's data to train the algorithm and uses the next year's data to test on. For the cross-validation, the 2005-2006 data was used. The second hold out strategy represent a realistic scenario where only historic data is able to be used to train a model whereas the cross-validated hold out allows information about test outcomes and scaling which the algorithms should not legitimately have access to.

The actual distribution of students in the different categories is presented as well as the predicted distribution according to the various algorithms paired with the two hold out strategies. The error metrics used are Mean Absolute Difference (MAD) and Root Mean Squared Error (RMSE). The formula for RMSE is:

$\sqrt{\frac{\sum_{i=1}^{n}(predicted_i - actual_i)^2}{n}}$ where n is the number of students.

2.3 Results

The results presented in Tables II and III are the algorithm errors in predicting the raw test scores. Scaled score false positive and negative evaluation is reported in Tables VI and V.

Algorithm	RMSE	MAD
Linreg + K-means	9.193	7.240
Linreg	9.262	7.358
RF + K-means	9.399	7.463
RF	9.420	7.540

Table II. MCAS prediction error using 5-fold cross-validation on the 2005-2006 data

Linear regression with bagged K-means resulted in the most accurate prediction of test scores according to RMSE and MAD error metrics. The more complex prediction technique of random forests did not fare as well with regular linear regression beating random forests with bagged K-means and random forests alone. The RMSE difference between the best and worse algorithm was 0.227, or 2.4% worse than the best score.

We now compare to the other hold out strategy where the same 2005-2006 test scores are being predicted except using data from 2004-2005 to train. Prediction results of this second hold out strategy are shown in Table III.

Algorithm	RMSE	MAD
Linreg + K-means	9.748	7.957
Linreg	9.817	8.044
RF + K-means	9.941	8.204
RF	10.106	8.337

Table III. MCAS prediction error of 2005-2006 test scores using 2004-2005 data.

Table III shows that the relative rankings of the algorithms have not changed using this hold out strategy but the overall errors have increased. The RMSE difference between the best score using cross-validation versus using the previous year's data is 0.555, or 6% worse than the better score. This difference is more than twice the difference between the best and worse algorithms in Table II. What does this level of difference mean to actual student score prediction? To investigate this we look deeper at the predicted score category compared to the actual category of the

two best algorithms using each hold out strategy. For the train/test hold out strategy we used the '04-'05 scaling to transform the '05-'06 raw predictions to categorical predictions.

	real	pred.	false pos.	false neg.	sensitivity
Adv.	0.083	0.016	0.003 (2)	0.841 (53)	0.159
Prof.	0.176	0.140	0.099 (62)	0.672 (90)	0.328
Need.	0.364	0.530	0.444 (215)	0.321 (89)	0.679
Fail.	0.377	0.315	0.171 (81)	0.445 (128)	0.554

Table IV. Statistics for Linear regression + bagged K-means prediction of the cross-validated 2005-2006 data

Table IV shows the real percentage of students that fall in to the four proficiency categories as well as the predicted percentages according to the prediction algorithm. False positives, false negatives and sensitivity are also shown. Sensitivity is the probability that students who belong to that category will be properly placed into that category. We can observe that not many students scored in the advanced category and that the majority of the distribution (74.1%) lies in the Needs Improvement and Failing categories. Two students were placed in advanced that did not belong there and 53 students were not placed in that category that belonged there. For failing, 81 students were placed there that did not belong there and 128 were not placed there who belonged there.

	real	pred.	false pos.	false neg.	sensitivity
Adv.	0.083	0.000	0.000 (0)	1.000 (63)	0.000
Prof.	0.176	0.152	0.1132 (71)	0.664 (89)	0.336
Need.	0.364	0.654	0.654 (286)	0.235 (65)	0.765
Fail.	0.377	0.193	0.089 (42)	0.634 (182)	0.366

Table V. Statistics for Linear regression + bagged K-means prediction of 2005-2006 data using 2004-2005 data to train

In Table V we can see a different distribution that places the bulk of the classification into the Needs Improvement category. Looking again at the most important categories, the advanced had no false positives but had 63 (100%) students placed outside of advanced that should have been in advanced. For failing, 42 students were improperly classified into that category while 182 were improperly left out of the category.

The cross-validated hold out misclassifies more students into Failing while the non cross-validated hold out fails to correctly classify more students as Failing. Both hold out strategies fail to classify all or most of the advance students as advanced.

This analysis demonstrates the areas of improvement for this test score prediction task, particularly in correctly identifying Advanced students. The hold out analysis also shows that while the previous year training strategy resulted in 6% worse error, it is still performing reasonably well compared to the cross-validated result in important category classification areas according to the statistical analysis. This more in-depth analysis gives us confidence that deployment of this prediction method in a real world setting would result in raw test score predictions of within 12% of actual (8/54). Misclassification of Advanced and Failing students is an aspect that needs improvement on the algorithm end, perhaps with ensemble techniques or the addition of more features engineered from the logged data.

3. INFERRING STUDENT KNOWLEDGE

In this section we conduct a simulation study to observe the correspondence between performance prediction error and knowledge inference error. In particular, at which opportunity does the model infer knowledge has been attained compared with the opportunity at which the simulated student attained knowledge in the synthesized data. This correspondence is compared with the prediction error of each model. The significance of performance prediction is looked at from a different angle than in the previous section. Instead of measuring the effect of leakage on prediction, we look at how performance prediction corresponds to a different objective, that of inferring student knowledge. This type of inference of knowledge is used in the Cognitive Tutors [8]. Reported performance prediction improvements often come with the presumption that knowledge inference accuracy is also improved.

3.1 Dataset

For this dataset we synthesized data for 500 simulated students answering 50 questions each of the same skill. The simulation generated 50 responses per student in addition to 50 knowledge states per student. Student responses are either 0, representing an incorrect answer, or 1, representing a correct answer. Student knowledge states are also 0 or 1 corresponding to the skill being known or not known.

3.2 Methodology

The standard Bayesian Knowledge Tracing [7] model was used to simulate data. This is a Hidden Markov Model of learning where a student is either in the learned or unlearned state and evidence of their past response history can be used as evidence to infer the probability of their current knowledge state as well as the probability of a correct answer on the next problem opportunity. The model has four parameters: prior, learn, guess and slip and these parameters were fixed to values of 0.30, 0.09, 0.14 and, 0.09 respectively for the generation of the data.

A 5-fold student level cross-validation was run using six different knowledge tracing models to attempt to recover the parameters and predict simulated student response and also infer the probability of knowledge at each opportunity. The six models included: 1) the ground truth model (GT) using the real generating parameters 2) a model that let Expectation Maximization (EM) iterate until convergence 3-6) these models kept three parameters at their ground truth values and increased the fourth by 0.20. For example, model "gt_guess" has the guess parameter set to 0.34 instead of 0.14 while all other parameters remain at ground truth level. These models were included so we may observe the sensitivity of the various parameters on performance and knowledge prediction. RMSE was again used to evaluate results as this has been a popular metric to evaluate within-tutor prediction and was the metric used to score results in the 2010 KDD Cup challenge [2]. AUC was used in place of MAD as AUC has also been popular in the user modeling literature to score prediction accuracy. AUC can only be used with binary prediction classes and so it was not applicable to the MCAS scoring. AUC is an accuracy metric with a 0.50 score being no better than chance and a score of 1 being a perfect prediction. Statistics comparing the correspondence between the time that simulated student knew the skill and the time that the inferred probability of knowledge was 0.95 or above were also calculated. The threshold of 0.95 is common in Cognitive tutors [8] for determining that a student has mastered a skill and allowing them to move on in the curriculum.

3.3 Results

model	RMSE	AUC
EM	0.4273	0.7260
GT	0.4296	0.7191
gt_prior	0.4307	0.7154
gt_guess	0.4367	0.7998
gt_slip	0.4373	0.7241
gt_learn	0.4773	0.6480

Table VI. Cross-validated simulation performance prediction results for the eight models.

Table VI shows that the best model according to RMSE was the EM model. The gt_guess model was best according to AUC. It is somewhat surprising that the ground truth model, although close in RMSE, was not the best. The EM model converged to within 0.01 of GT parameter values, so the slightly improved accuracy may be due to chance that this particular simulated population skewed towards the EM converged parameters. AUC is a rank order estimation of accuracy and thus, so long as predictions correlate with responses, the predictions can be poor and still attain high AUC. Nevertheless, this result is surprising. The worse model, according to both AUC and RMSE, was the model which increased the learn rate parameter by 0.20. This suggests that learn rate is a sensitive parameter to prediction error and a potentially worthwhile area to focus on for student prediction improvement.

	Median un/over predicted	Mean absolute difference	students over practiced	students under practiced
gt_learn	1	2.37	409	56
GT	2	2.53	469	21
gt_prior	2	2.53	469	21
gt_slip	2	2.64	473	17
EM	2	2.68	475	18
gt_guess	4	4.81	494	0

Table VII. Under and over practice amounts on average caused by model inference in students' knowledge using a mastery threshold of 0.95 probability.

Table VII shows how each model performed at inferring when a student has mastered the skill. The median un/over practice column shows the median number of over or under practice opportunities. Average of absolutes column calculates the average absolute under/over prediction which takes the absolute value of the residual between inferred mastery opportunity and actual mastery opportunity. The lower this value, the better the model did at inferring exactly when a student learns and not letting them over or under practice the skill. As we can see by the "number over practiced" column, the vast number of simulated students are inferred to learn the skill after they have actually learned it. The worse over predictions was by the gt_guess model caused decreased confidence in knowledge when observing positive performance which further exacerbated the under prediction bias.

4. DISCUSSION

We have investigated the significance of performance prediction in the context of test score prediction and within-tutor knowledge inference. We have raised the issue of leakage in prediction evaluation and its role in cross-validation accuracy inflation. The result of leakage was a 6% increase in error from the best cross-validated model to the best model trained on the previous year's data. A 6% increase is reasonable for training on a separate cohort of students. An additional analysis of the results using a confusion matrix revealed a decrease in prediction of proficiencies at the extremes, and a tendency to predict more towards the average proficiency category with the previous year hold out.

Our simulation study revealed a clear bias towards knowledge under prediction among the knowledge tracing models. The inflated learning rate model, gt_learn, worked to offset some of this bias, reducing the median over prediction from 2 to 1 opportunity, which provided a better knowledge inference estimate but also resulted in the worst performance prediction score. This discord underscores the motivation behind studying the real impact of performance prediction on the intended objectives, although this magnitude of disparity warrants further investigation.

5. ACKNOWLEDGEMENTS

This research was supported by the National Science foundation via grant "Graduates in K-12 Education" (GK-12) Fellowship, award number DGE0742503. We thank Feng & Heffernan for the datasets and the Worcester Public Schools for their ongoing partnership and for making test scores available to us. We also acknowledge the many additional funders of ASSISTments Platform found here: http://www.webcitation.org/5ym157Yfr

6. REFERENCES

[1] S. Kaufman, S. Rosset, and C. Perlich. Leakage in data mining: formulation, detection, and avoidance. In Proceedings of the 2011 conference on Knowledge Discovery in Data Mining, San Diego. pp. 556-563.

[2] Pardos, Z.A., Heffernan, N. T.: Using HMMs and bagged decision trees to leverage rich features of user and skill from an intelligent tutoring system dataset. To appear in the Journal of Machine Learning Research W & CP, In Press

[3] Gowda, S., Baker, R.S.J.d., Pardos, Z., Heffernan, N. (2011) The Sum is Greater than the Parts: Ensembling Student Knowledge Models in ASSISTments. Proceedings of the KDD 2011 Workshop on KDD in Educational Data.

[4] Feng, M., Heffernan, N.T., & Koedinger, K.R. (2009). Addressing the assessment challenge in an Intelligent Tutoring System that tutors as it assesses. The Journal of User Modeling and User-Adapted Interaction. Vol 19: p243-266.

[5] Trivedi, S., Pardos, Z. & Heffernan, N. (2011) Clustering Students to Generate an Ensemble to Improve Standard Test Score Predictions In Biswas et al (Eds) Proceedings of the Artificial Intelligence in Education Conference 2011. Springer. LNAI 6738, Pages. 328–336.

[6] L. Breiman. Random forests. Machine Learning, 45(1):5-32, 2001.

[7] CORBETT, A. T., & ANDERSON, J. R. 1995. Knowledge tracing: Modeling the acquisition of procedural knowledge. User Modeling and User-Adapted Interaction, 4, 253-278.

[8] CORBETT, A. T., KOEDINGER, K., & HADLEY, W. S. 2001. Cognitive tutors: From the research classroom to all classrooms. In P. S. Goodman (Ed.), Technology enhanced learning: Opportunities for change. (pp. 235-263). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.