Exploring the Role of Small Differences in Predictive Accuracy using Simulated Data

Juraj Nižnan, Jan Papoušek, and Radek Pelánek

Faculty of Informatics, Masaryk University Brno
{niznan,jan.papousek,xpelanek}@mail.muni.cz

Abstract. Research in student modeling often leads to only small improvements in predictive accuracy of models. The importance of such improvements is often hard to assess and has been a frequent subject of discussions in student modeling community. In this work we use simulated students to study the role of small differences in predictive accuracy. We study the impact of such differences on behavior of adaptive educational systems and relation to interpretation of model parameters. We also point out a feedback loop between student models and data used for their evaluation and show how this feedback loop may mask important differences between models.

1 Introduction

In student modeling we mostly evaluate models based on the quality of their predictions of student answers as expressed by some performance metric. Results of evaluation often lead to small differences in predictive accuracy, which leads some researchers to question the importance of model improvements and meaningfulness of such results [1]. Aim of this paper is to explore the impact and meaning of small differences in predictive accuracy with the use simulated data. For our discussion and experiments in this work we use a single performance metric – Root Mean Square Error (RMSE), which is a common choice (for rationale and overview of other possible metrics see [15]). The studied questions and overall approach are not specific to this metric.

Simulated students provide a good way to study methodological issues in student modeling. When we work with real data, we can use only proxy methods (e.g., metrics like RMSE) to evaluate quality of models. With simulated data we know the "ground truth" so we can study the link between metrics and the true quality of models. This enables us to obtain interesting insight which may be useful for interpretation of results over real data and for devising experiments. Similar issues are studied and explored using simulation in the field of recommender systems [7, 17].

We use a simple setting for simulated experiments, which is based on an abstraction of a real system for learning geography [12]. We simulate an adaptive question answering system, where we assume items with normally distributed difficulties, students with normally distributed skills, and probability of correct

answer given by a logistic function of the difference between skill and difficulty (variant of a Rasch model). We use this setting to study several interrelated question.

1.1 Impact on Student Practice

What is the impact of prediction accuracy (as measured by RMSE) on the behavior of an adaptive educational system and students' learning experience?

Impact of small differences in predictive performance on student under-practice and over-practice (7-20%) has been demonstrated using real student data [18], but insight from a single study is limited. The relation of RMSE to practical system behavior has been analyzed also in the field of recommender systems [2] (using offline analysis of real data). This issue has been studied before using simulated data is several studies [5, 6, 10, 13]. All of these studies use very similar setting – they use Bayesian Knowledge Tracing (BKT) or its extensions and their focus is on mastery learning and student under-practice and over-practice. They differ only in specific aspects, e.g., focus on setting thresholds for mastery learning [5] or relation of moment of learning to performance metrics [13]. In our previous work [16] have performed similar kind of simulated experiments (analysis of under-practice and over-practice) both with BKT and with student models using logistic function and continuous skill.

In this work we complement these studies by performing simulated experiments in slightly different setting. Instead of using BKT and mastery learning, we use (variants of) the Rasch model and adaptive question answering setting. We study different models and the relation between their prediction accuracy and the set of items used by the system.

1.2 Prediction Accuracy and Model Parameters

Can RMSE be used to identify good model parameters? What is the relation of RMSE to the quality of model parameters?

In student modeling we often want to use interpretable models since we are interested not only in predictions of future answers, but also in reconstructing properties of students and educational domains. Such outputs can be used to improve educational systems as was done for example by Koedinger at al. [9]. When model evaluation shows that model A achieves better prediction accuracy (RMSE) then model B, results are often interpreted as evidence that model A better reflects "reality". Is RMSE a suitable way to find robust parameters? What differences in metric value are meaningful, i.e., when we can be reasonably sure that the better model really models reality in better way? Is statistical significance of differences enough? In case of real data it is hard to answer these question since we have no direct way to evaluate the relation of a model to reality. However, we can study these questions with simulated data, where we have access to the ground truth parameters. Specifically, in our experiments we study the relation of metric values with the accuracy of reconstructing the mapping between items and knowledge components.

1.3 Feedback between Data Collection and Evaluation

Can the feedback loop between student models and adaptive choice of items influence evaluation of student models?

We also propose novel use of simulated students to study a feedback loop between student models and data collection. The data that are used for model evaluation are often collected by a system which uses some student model for adaptive choice of items. The same model is often used for data collection and during model evaluation. Such evaluation may be biased – it can happen that the used model does not collect data that would show its deficiencies. Note that the presence of this feedback loop is an important difference compared to other forecasting domains. For example in weather forecasting models do not directly influence the system and cannot distort collected data. In student modeling they can.

So far this feedback has not been thoroughly studied in student modeling. Some issues related to this feedback have been discussed in previous work on learning curves [6, 11, 8]. When a tutoring system uses mastery learning, students with high skill drop out earlier from the system (and thus from the collected data), thus a straightforward interpretation of aggregated learning curves may be misleading. In this work we report experiment with simulated data which illustrate possible impact of this feedback loop on model evaluation.

2 Methodology

For our experiments we use a simulation of a simplified version of an adaptive question answering systems, inspired by our widely used application for learning geography [12]. Fig. 1 presents the overall setting of our experiments. System asks students about items, answers are dichotomous (correct/incorrect), each student answers each item at most once. System tries to present items of suitable difficulty. In evaluation we study both the prediction accuracy of models and also sets of used items. This setting is closely related to item response theory and computerized adaptive testing, specifically to simulated experiments with Elo-type algorithm reported by Doebler et al. [3].

Simulated Students and Items We consider a set of simulated students and simulated items. To generate student answers we use logistic function (basically the Rasch model, respectively one parameter model from item response theory): $P(correct|\theta_s, d_i) = 1/(1 + e^{-(\theta_s - d_i)})$, where θ_s is the skill of a student s and d_i is difficulty of an item i.

To make the simulated scenarios more interesting we also consider multiple knowledge components. Items are divided into disjoint knowledge components and students have different skill for each knowledge component. Student skills and item difficulties are sampled from a normal distribution. Skills for individual knowledge components are independent from one another.

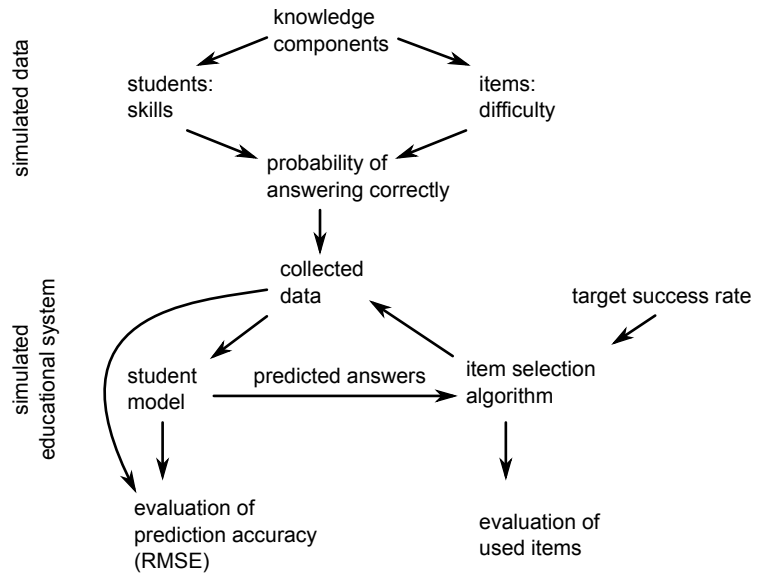

Fig. 1. Setting of our experiments

Item Selection Algorithm The item selection algorithm has as a parameter a target success rate t. It repeatedly presents items to a (simulated) student, in each step it selects an item which has the best score with respect to the distance of the predicted probability of correct answer p and the target rate t (illustrated by gray dashed line in Fig. 3). If there are multiple items with the same score, the algorithm randomly selects one of them.

Student Models Predictions used by the item selection algorithm are provided by a student model. For comparison we consider several simple student models:

- Optimal model – Predicts the exact probability that is used to generate the answer (i.e., a "cheating" model that has access to the ground truth student skill and item difficulty).
- Optimal with noise – Optimal model with added (Gaussian) noise to the difference $\theta_s - d_i$ (before we apply logistic function).
- Constant model – For all students and items it provides the same prediction (i.e., with this model the item selection algorithm selects items randomly).
- Naive model – Predicts the average accuracy for each item.
- Elo model – The Elo rating system [4, 14] with single skill. The used model corresponds to the version of the system as described in [12] (with slightly modified uncertainty function).
- Elo concepts – The Elo system with multiple skills with correct mapping of items to knowledge components.

– Elo wrong concepts – The Elo system with multiple skills with wrong mapping of items to knowledge components. The wrong mapping is the same as the correct one, but 50 (randomly choosen) items are classified incorrectly.

Data We generated 5,000 students and 200 items. Items are divided into 2 knowledge components, each user has 2 skills corresponding to the knowledge components and each item has a difficulty. Both skills and difficulties were sampled from standard normal distribution (the data collected from the geography application suggests that these parameters are approximately normally distributed). The number of items in a practice session is set to 50 unless otherwise noted.

3 Experiments

We report three types of experiments, which correspond to the three types of questions mentioned in the introduction.

3.1 Impact on Student Practice

Our first set of experiments studies differences in the behavior of the simulated system for different models. For the evaluation of model impact we compare the sets of items selected by the item selection algorithm. We make the assumption that the algorithm for item selection using the optimal model generates also the optimal practice for students. For each user we simulate practice of 50 items (each item is practiced at most once by each student). To compare the set of practiced items between those generated by the optimal model and other models we look at the size of the intersection. We assume that bigger intersection with the set of practiced items using the optimal model indicates better practice. Since the intersection is computed per user, we take the mean.

This is, of course, only a simplified measure of item quality. It is possible that an alternative model selects completely different set of items (i.e., the intersection with the optimal set is empty) and yet the items are very similar and their pedagogical contribution is nearly the same. However, for the current work this is not probable since we are choosing 50 items from a pool of only 200 items. For future work it would be interesting to try to formalize and study the "utility" of items.

Noise Experiment The optimal model with noise allows us to easily manipulate differences in predictive accuracy and study their impact on system behavior. Experiment reported in the left side of Fig. 2 shows both the predictive accuracy (measured by RMSE) and the impact on system behavior (measured by the size of the intersection with the optimal practiced set as described above) depending on the size of noise (we use Gaussian noise with a specified standard deviation). The impact of noise on RMSE is approximately quadratic and has a slow rise –

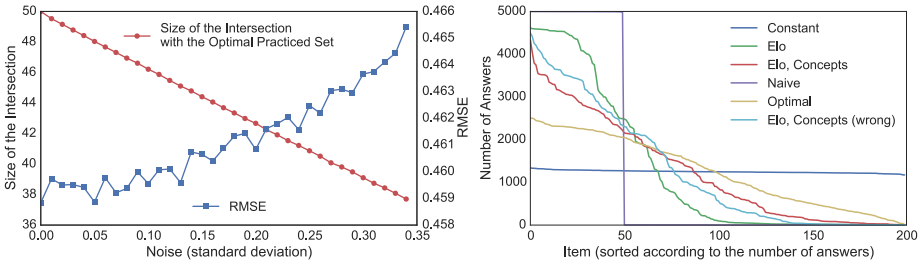

Fig. 2. Size of the intersection with the optimal practiced set of items and RMSE depending on Guassian noise in optimal model (left side). Distribution of answers over the items based on the given model (right side).

this is a direct consequence of the quadratic nature of the metric. The impact on used items is, however, approximately linear and rather steep. The most interesting part is for noise values in the interval $[0, 0.1]$. In this interval the rise in RMSE values is very small and unstable, but the impact on used items is already high.

Model Comparison Right side of the Fig. 2 shows the distribution of the number of answers per item for different models. The used models have similar predictive accuracy (specific values depend on what data we use for their evaluation, as discussed below in Section 3.3), yet the used model can dramatically change the form of the collected data.

When we use the optimal model, the collected data set covers almost fairly most items from the item pool. In the case of worse models the use of items is skewed (some items are used much more frequently than others). Obvious exception is the constant model for which the practice is completely random. The size of the intersection with the optimal practiced set for these models is – Constant: 12.5; Elo: 24.2; Elo, Concepts: 30.4; Elo, Concepts (wrong): 28.5; Naive: 12.0. Fig. 3 presents a distribution of answers according to the true probability of their correctness (given by the optimal model). Again there is a huge difference among the given models, especially between simple models and those based on Elo.

3.2 Prediction Accuracy and Model Parameters

Metrics of prediction accuracy (e.g., RMSE) are often used for model selection. Model that achieves lower RMSE is assumed to have better parameters (or more generally better "correspondence to reality"). Parameters of a selected model are often interpreted or taken into account in improvement of educational systems. We checked validity of this approach using experiments with knowledge components.

We take several models with different (random) mappings of items to knowledge components and evaluate their predictive accuracy. We also measure the

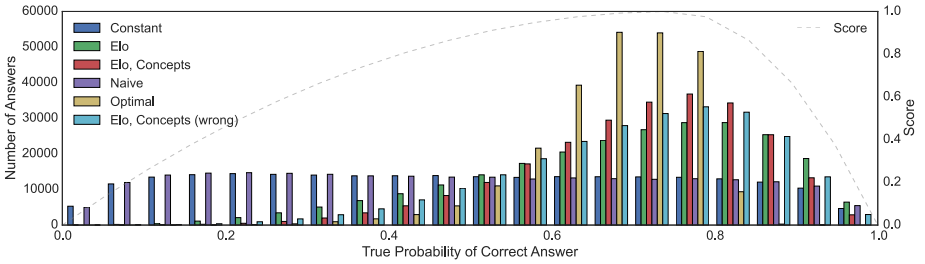

Fig. 3. Distribution of answers according to the true probability of correct answer. The gray dashed line stands for the score function used by the algorithm for item selection.

quality of the used mappings – since we use simulated data, we know the ground truth mapping and thus can directly measure the quality of each mapping. Quality is expressed as the portion of items for which the mapping agrees with the ground truth mapping. The names of the knowledge components are irrelevant in this setting. Therefore, we compute quality for each one-to-one mapping from the names of the components in the model to the names of the components in the ground truth. We select the highest quality as the quality of the model's item-to-component mapping. To focus only on quality of knowledge components, we simplify other aspects of evaluation, specifically each student answers all items and their order is selected randomly.

These experiments do not show any specific surprising result, so we provide only general summary. Experiments show that RMSE values correlate well with the quality of mappings. In case of small RMSE differences there may be "swaps", i.e., a model with slightly higher RMSE reflects reality slightly better. But such results occur only with insufficiently large data and are unstable. Whenever the differences in RMSE are statistically significant (as determined by t-test over different test sets), even very small differences in RMSE correspond to improvement in the quality of the used mappings. These results thus confirm that it is valid (at least in the studied setting) to argue that a model A better corresponds to reality than a model B based on the fact that the model A achieves better RMSE than the model B (as long as the difference is statistically significant). It may be useful to perform this kind of analysis for different settings and different performance metrics.

3.3 Feedback between Data Collection and Evaluation

To study feedback between the used student model and collected data (as is described in subsection 1.3) we performed the following experiment: We choose one student model and use it as an input for adaptive choice of items. At the same time we let all other models do predictions as well and log answers together with all predictions.

Fig. 4 shows the resulting RMSE for each model in individual runs (data collected using specific model). The figure shows several interesting results. When

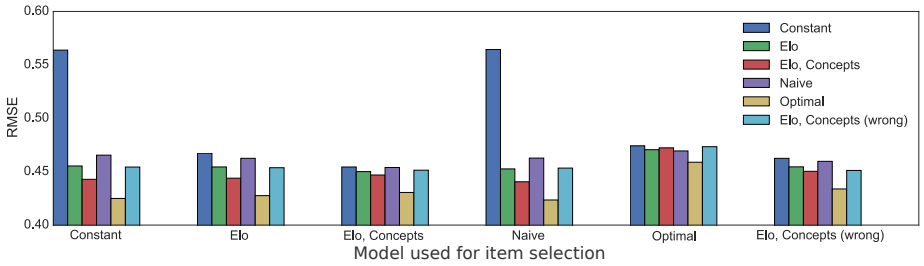

Fig. 4. RMSE comparison over data collected using different models.

the data are collected using the optimal model, the RMSE values are largest and closest together; even the ordering of models is different from other cases. In this case even the constant model provides comparable performance to other models – but it would be very wrong to conclude that "predictive accuracy of models is so similar that the choice of model does not matter". As the above presented analysis shows, different models lead to very different choice of items and consequently to different student experience. The reason for small differences in RMSE is not similarity between models, but characteristics of data ("good choice of suitable items"), which make predictions difficult and even a naive predictor comparatively good.

Another observation concerns comparison between the "Elo concepts" and "Elo concepts (wrong)" models. When data are collected by the "Elo concepts (wrong)" model, these two models achieve nearly the same performance, i.e., models seem to be of the same quality. But the other cases show that the "Elo concepts" model is better (and in fact it is by construction a better student model).

4 Conclusions

We have used simulated data to show that even small differences in predictive accuracy of student models (as measured by RMSE) may have important impact on behavior of adaptive educational systems and for interpretation of results of evaluation. Experiments with simulated data, of course, cannot demonstrate the practical impact of such small differences. We also do not claim that small differences in predictive accuracy are always important. However, experiments with simulated data are definitely useful, because they clearly illustrate mechanisms that could play role in interpretation of results of experiments with real student data. Simulated data also provide setting for formulation of hypotheses that could be later evaluated in experiments with real educational systems.

Simulated data also enable us to perform experiments that are not practical for realization with actual educational systems. For example in our experiment with the "feedback loop" we have used different student models as a basis for item selection. Our set of models includes even a very simple "constant model",

which leads to random selection of practiced item. In real setting we would be reluctant to apply such a model, as it is in contrary with the advertised intelligent behavior of our educational systems. However, experiments with this model in simulated setting provide interesting results – they clearly demonstrate that differences in predictive accuracy of models do not depend only on the intrinsic quality of used student models, but also on the way the data were collected.

Our analysis shows one particularly interesting aspect of student modeling. As we improve student models applied in educational systems, we should expect that evaluations of predictive accuracy performed over these data will show worse absolute values of performance metrics and smaller and smaller differences between models (even if models are significantly different), just because virtues of our models enable us to collect less predictable data.

References

1. Joseph E Beck and Xiaolu Xiong. Limits to accuracy: How well can we do at student modeling. In *Proc. of Educational Data Mining*, 2013.
2. Paolo Cremonesi, Yehuda Koren, and Roberto Turrin. Performance of recommender algorithms on top-n recommendation tasks. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 39–46. ACM, 2010.
3. Philipp Doebler, Mohsen Alavash, and Carsten Giessing. Adaptive experiments with a multivariate elo-type algorithm. *Behavior research methods*, pages 1–11, 2014.
4. Arpad E Elo. *The rating of chessplayers, past and present*, volume 3. Batsford London, 1978.
5. Stephen E Fancsali, Tristan Nixon, and Steven Ritter. Optimal and worst-case performance of mastery learning assessment with bayesian knowledge tracing. In *Proc. of Educational Data Mining*, 2013.
6. Stephen E Fancsali, Tristan Nixon, Annalies Vuong, and Steven Ritter. Simulated students, mastery learning, and improved learning curves for real-world cognitive tutors. In *AIED Workshops*. Citeseer, 2013.
7. Jonathan L Herlocker, Joseph A Konstan, Loren G Terveen, and John T Riedl. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)*, 22(1):5–53, 2004.
8. Tanja Käser, Kenneth R Koedinger, and Markus Gross. Different parameters-same prediction: An analysis of learning curves. In *Proceedings of 7th International Conference on Educational Data Mining. London, UK*, 2014.
9. Kenneth R Koedinger, John C Stamper, Elizabeth A McLaughlin, and Tristan Nixon. Using data-driven discovery of better student models to improve student learning. In *Artificial intelligence in education*, pages 421–430. Springer, 2013.
10. Jung In Lee and Emma Brunskill. The impact on individualizing student models on necessary practice opportunities. *International Educational Data Mining Society*, 2012.
11. R Charles Murray, Steven Ritter, Tristan Nixon, Ryan Schwiebert, Robert GM Hausmann, Brendon Towle, Stephen E Fancsali, and Annalies Vuong. Revealing the learning in learning curves. In *Artificial Intelligence in Education*, pages 473–482. Springer, 2013.

12. Jan Papoušek, Radek Pelánek, and Vít Stanislav. Adaptive practice of facts in domains with varied prior knowledge. In *Proc. of Educational Data Mining*, pages 6–13, 2014.
13. Zachary A Pardos and Michael V Yudelson. Towards moment of learning accuracy. In *AIED 2013 Workshops Proceedings Volume 4*, page 3, 2013.
14. Radek Pelánek. Application of time decay functions and Elo system in student modeling. In *Proc. of Educational Data Mining*, pages 21–27, 2014.
15. Radek Pelánek. Metrics for evaluation of student models. *Journal of Educational Data Mining*, 2015. To appear.
16. Radek Pelánek. Modeling student learning: Binary or continuous skill? In *Proc. of Educational Data Mining*, 2015.
17. Al Mamunur Rashid, Istvan Albert, Dan Cosley, Shyong K Lam, Sean M McNee, Joseph A Konstan, and John Riedl. Getting to know you: learning new user preferences in recommender systems. In *Proceedings of the 7th international conference on Intelligent user interfaces*, pages 127–134. ACM, 2002.
18. Michael V Yudelson and Kenneth R Koedinger. Estimating the benefits of student model improvements on a substantive scale. In *EDM 2013 Workshops Proceedings*, 2013.