Master Thesis Research Proposal

Lieuwe Rekker

January 29, 2013

1 Introduction

Many current intelligent tutor systems (ITS) use learner models that can give indications of to what extent a student has mastered a particular skill and even how fast students learn. The quality of these models is generally measured by how well these models predict 'one question (item) into the future' performance, which is also what algorithms are trying to maximize. Parameter values from these models are also inspected and stated to represent students' knowledge levels, how fast students learn, how difficult questions are etc. It would thus be wise to inspect other factors beyond the accuracy of the model to gain some idea of whether they actually convey some stable real world factor or are rather part of a more black-box like interior of a model that performs well on its prediction task. [Bec07] clearly brings up that this is indeed a problem: widely differing parameter settings can yield similar, optimal performance. This proposal proposes research to explore if parameters of learner models hold meaning in real life.

The approach of this research is from the perspective that models approach reality, but are necessarily not precise mirrors of it. Initially generated data will be used to see how well parameter values can be retrieved at all and to see what happens to parameters when there is a mismatch between model and 'reality' (i.e. training a model on data generated by a different model). In the case of real data, original parameters or models are unknown. Over multiple training runs on different data-sets, variation in parameter values can be inspected. High variation here indicates that (despite possible good prediction performance of the model as a whole) the parameter does not represent anything in reality.

An important factor for how well a model works is how much data is available. Therefor amount of data will be varied to inspect its role in parameter estimates. This also functions to remove lack of data as the leading cause for varying parameters. One reason for the mismatch between model and reality mentioned before is that not all data that is appropriate to how students learn is available and thus these cannot be incorporated in the model. Nevertheless most ITS collect more data than what is currently used for the learner models. One way in which this data is already being used is to estimate whether students are actually engaged with the ITS (see for example [Bak07]). The second factor that will be experimented with is to incorporate some form of engagement variable into the model to see to what extend the mis-estimation of parameters is caused by this. Additionally it is hoped that adding engagement to the model lowers the variation in parameters, leading to more stable models.

2 Related work

In [Bec07] Beck goes beyond investigating simply the accuracy of a model (knowledge tracing in this case) and also looks at the parameter values. In his research a big problem comes forward: multiple global maxima: there are multiple (very different) parameter settings that lead to similar, near-optimal accuracies. This shows that the trained model could have parameter values that although the performance of the model is good are otherwise meaningless.

Engagement has been a focus of research on ITSs as it is seen as important for learning and many talk of the necessity for the ITS to intervene when a student's engagement is low. Some research utilizes extra equipment to obtain data from which to estimate engagement. This equipment ranges from cameras and microphones to EEG equipment. Others opt to only use information that is already gathered by ITSs or that could easily be gathered by them (e.g. by using screen replays [])

as this research is probably not useful after all (except maybe for using the model as input) this paragraph will probably be removed In [Bak07] Baker only uses data already obtained by the ITS to model 'engagement' (behavior). The model is trained on engagement frequencies as observed during usage of the tutor and distinguishes four states: working on the tutor, on-task conversation (not working in the tutor, but for example discussing its contents with someone else), off-task behavior (in any way not paying attention to the tutor) and gaming the system (interacting with the tutor in a non-constructive way, e.g. asking for hints until the answer is revealed or random guessing).

Little research has been done on the interaction between engagement and models of student performance. In [Bec05] the authors notice a specific relation between relative answer time and probability that a question is answered correctly. Based on this observation, their model assumes that the faster a student answers a question (both relative to how fast the student generally is and how long the question should take), the more likely the student is to be disengaged and just having a guess at the question. The resulting model looks like an IRT model, but rather than skill uses time spent on the question relative to the length of the question and skill of the student to obtain a probability that the student was engaged. Engagement here was shown to correlate at .25 with gains on an external test.

In [JW06] a more elaborate model is made: an IRT-model modeling skill is combined with a Markov-chain modeling motivation. Both models play a role in performance on questions, while additional information (such as time spend on a question) also play a role in the motivation model. Motivation itself here is a latent variable. The prediction of performance is then either the result from the IRT-model (in the motivated case) or chance or 0 in the unmotivated choice (depending on whether students are random guessing or repeatedly asking for hints). Both these papers use a model where 'attention' is a latent variable to be discovered and neither model takes learning itself into account, which is an essential part of this proposal.

3 Research Question

The question I would like to answer in my masters thesis is: "Do learner models and their parameter values accurately mirror mechanics in reality?". This question will be approached by training models on data and then inspecting them. Two specific factors that can have an influence on how well the model resembles reality are examined: the amount of data and the exclusion of an influencing variable. Two different angles are taken on answering this question: one looking at models based on generated data and the other looking at models based on real-data.

The first factor taken into account is the amount of data. Preferably data is increased up to the point where the parameter estimates (or their variance) stabilizes, so that the best possible estimate can be examined. The second factor is the influence of leaving out a variable that is considered to have an influence on learning. For this variable engagement will be used. Please note that the exact nature of this variable here is not important. One might argue that this variable represents motivation, attention or emotional disposition or something else along those lines or many different factors combined rather than engagement. The bottom line though is that such factors seem likely to play a similar role in learning. An issue is that there is no engagement variable in the collected data, just models that derive this (or similar) variables from other collected data. In this research initially the ground truth values for 'attention' will be used. If parameter estimates improve over the simpler model, noise will be added to this variable to examine at what accuracy this additional information no longer has a positive effect. In the case of the real data experiment, additional experiments will be run using the engagement values from an attention model from the literature.

The first angle taken in this research is to train models based on generated data. This way the true learning rate is known and an upper bound can be established for how well this particular model can retrieve it. Additionally comparing how models do on data generated by other models might provide some insights into what happens to learning rate estimates when the 'true' model and the trained model differ. The second angle is training the models on real data. In this case there won't be a ground truth learning rate, but rather the variance of learning rate estimates on different sub-sets of the data will be compared. On real data it will also become apparent if the usage of the additional variable actually improves the found learning rate.

4 Research Priorities

Below are my priorities for the research.

Must Use eIRT model and knowledge tracing model. Study their performance both on artificial and real data.

Should Incorporate attention into the used models, create new artificial data to train these models and the simpler models on. Apply models to real data.

Could Include performance factor analysis model and compare to the other two models

Would have Include Bayesian model and/or poks model or alternatively obtain more real world datasets

5 Method

The two different factors and two different angles described in section 3 divide the research into 6 different parts.

1. Consider current models and specifically look at how amount of data influences parameter estimates
   (a) Train models on generated data to inspect how well parameters are retrieved. Also cross check how models perform on data generated by other models to study the influence of model mismatch.
   (b) Train models on real data. Inspect parameter variances to form an idea of how well these parameters are represented in reality.
2. Extend the models to include an attention variable.
   (a) Train the extended model on generated data to inspect how well parameters are retrieved. Also cross check simpler models on the same data to study the influence of 'omitting' a variable from the model.
   (b) Add noise to the engagement variable to see effect on parameter values
   (c) Train the extended model on real data. Inspect influence of additional variable on variance of parameters and actual performance.
   (d) Use model results rather than ground truth values as engagement variable. Possibly slowly mixing model data in with ground truth.

5.1 Model Performance

The most common performance measure used is some measure of accuracy of model predictions on next-item student performance. Although this research looks into the values of model parameters and is less concerned with other measures of performance such a measure will still be used. Maximizing this accuracy is what models are build to do and thus plays an important role in the fitting of parameters. It will be interesting to see how this accuracy, which is generally the only measure looked at in modelling, behaves compared to how well parameters are estimated.

The specific accuracy measure used for this research will be A'. For this measure two items are represented to the model, one which was answered correctly and one which was answered incorrectly. The model is used to determine which is which. The advantage of this method is that "values of A' are statistically comparable across models and data sets" [DB12].

Since the learner models are also used as an approximation of how students learn and as such influence decisions made on what exercises are given to a student, what exercises are included in general and what topics require more attention, the validity of the specific parameters of the model need to taken into account. In this research I will not look at external validity (e.g. see if the parameter settings seem realistic), but rather check to see if the found parameters are consistent over multiple trainings of the model and (in the case of generated data) whether the training algorithm recovers the used parameters correctly.

The IRT, KT and PFA models all use learning rates that can relatively easily be compared to each other and over multiple runs. Bayesian and POKS-based learner models provide more of a challenge in comparing their results. These models do not only contain parameters that are fitted during training, but also structures that are build in training. Thus in comparing the results from different runs simply looking at the value of particular parameters will not do. In comparing different models to each other (or maybe just models built during different runs) inspecting the proficiency in particular skills at a given moment might provide a good approach. This will be further explored if these models are actually used in the research.

5.2 Data Folding

Doing multiple training runs for every model means that the real data will be split into parts or that the runs will be done on partially overlapping data. Some consideration needs to be spend on balancing the number of runs, the possible amount of data per run and overlap of data between runs and what the consequences will be (e.g. variance of parameter values trained on large data sets may be less accurate or lower due to using less runs or overlapping data respectively)

5.3 Parameter values

In generating the data there are many other factors that might play a role in the performance of the models. Among those factors are the values that the parameters will be given. To make the generated data experiments resemble reality, the experiments (or at least some) on real data should be performed first to use values from those experiments for the generated data.

5.4 Extended Models

For the extended models in 2 it is important to compare the simple model and the extended model on data generated by the extended model. That way the (possible) distortion of leaving out an important variable becomes clear. Also it can become clear whether trying this idea on real data has any chance of success.

Training on real data would start by using the observed 'engagement' (i.e. ground truth) rather than any model. This is to check whether the idea has any merit in reality at all and also to approach the question of what effect omitting a variable more directly. If this is successful, some noise can be added to this variable to see the effect that imperfections in the modeling of engagement will have on the extended model. Finally a model of engagement can be used to provide the values of the engagement variable. A good candidate for a model to use could be [Bak07]. Results here would give more of an indication whether current engagement models can easily be incorporated into learner models.

6 Models

Give a short description of the models and also of how to generate data using these models. One idea for obtaining parameters to experiment with is to start training the models on the real data and take ranges of parameters as a good indication of what would be realistic to run the simulations with.

6.1 Knowledge Tracing

6.2 Item Response Theory

6.3 Performance Factor Analysis

Performance factor analysis is actually quite similar to IRT and thus it would be interesting to see how they perform on each others generated data-sets. Due to it's different structure comparing the learning rates will be a bit less straightforward, but manageable by taking the weighted average according to the fraction of items answered correctly. I would expect that the effect of including attention to this model would be quite interesting.

6.4 Existing Attention Models

[Bec05] [JW06]

6.5 Bayesian Model

These last two are less standardized, so there is more of a question on how to compare separate models and what method to use exactly.

6.6 Partial Order Knowledge Structure

7 Outstanding issues

- Describe terms and use those consistently
- 1-skill per item setting or multiple skill per item setting?
- find a good engagement data-set...

References

[Bak07] R.S.J. Baker. Modeling and understanding students' off-task behavior in intelligent tutoring systems. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 1059–1068. ACM, 2007.

[BD12] B. Beheshti and M. Desmarais. Improving matrix factorization techniques of student test data with partial order constraints. *User Modeling, Adaptation, and Personalization*, pages 346–350, 2012.

[Bec05] Joseph E. Beck. Engagement tracing: using response times to model student disengagement. *Artificial intelligence in education: Supporting learning through intelligent and socially informed technology*, 125:88, 2005.

[Bec07] J.E. Beck. Difficulties in inferring student knowledge from observations (and why you should care). In *Educational Data Mining: Supplementary Proceedings of the 13th International Conference of Artificial Intelligence in Education*, pages 21–30, 2007.

[DB12] M.C. Desmarais and R.S.J. Baker. A review of recent advances in learner and skill modeling in intelligent learning environments. *User Modeling and User-Adapted Interaction*, pages 1–30, 2012.

[DG06] M. Desmarais and M. Gagnon. Bayesian student models based on item to item knowledge structures. *Innovative Approaches for Learning and Knowledge Sharing*, pages 111–124, 2006.

[GB11] Y. Gong and J. Beck. Items, skills, and transfer models: which really matters for student modeling. In *Proceedings of the 4th International Conference on Educational Data Mining*, pages 81–90, 2011.

[JW06] J. Johns and B. Woolf. A dynamic mixture model to detect student motivation and proficiency. In *Proceedings of the National Conference on Artificial Intelligence*, volume 21, page 163. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006.

[KMS] K.R. Koedinger, E.A. McLaughlin, and J.C. Stamper. Automated student model improvement.

[PCK09] P.I. Pavlik, H. Cen, and K.R. Koedinger. Learning factors transfer analysis: Using learning curve analysis to automatically generate domain models. In *Proceedings of the 2nd International Conference on Educational Data Mining*, pages 121–130, 2009.

[RJF12] Diederik M. Roijers, Johan Jeuring, and Ad Feelders. Probability estimation and a competence model for rule based e-tutoring systems. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, LAK '12, pages 255–258, New York, NY, USA, 2012. ACM.