Modeling Speed-Accuracy Tradeoff in Adaptive System for Practicing Estimation

Juraj Nižnan
Masaryk University Brno
niznan@mail.muni.cz

ABSTRACT

Estimation is useful in situations where an exact answer is not as important as a quick answer that is good enough. A web-based adaptive system for practicing estimates is currently being developed. We propose a simple model for estimating student's latent skill of estimation. This model combines a continuous measure of correctness and response-times. The advantage of the model is its simple update method which makes it directly applicable in the developed adaptive system.

1. INTRODUCTION

Estimation is a very useful skill to possess. Particularly in situations where an exact answer is not as important as being able to quickly come up with an answer that is good enough (e.g., total amount on a bill in a restaurant, number of people in a room, total of the coins in a wallet, number of cans of paint needed for painting a room, converting between metric and imperial units). It was shown that estimation ability correlates with the ability to solve computational problems [2, 9, 8]. Because estimation is so useful, we have decided to develop a computerized adaptive system that will let its users practice estimating by solving various tasks.

The adaptive system will include exercises for practicing numerical estimation (results of basic arithmetic operations, converting between imperial and metric units, converting between temperature units, currencies and exchange rates) and visual estimation (counting the number of objects in a scene).

In order to provide adaptive behavior of the system, we need a way of inferring student's ability of estimation. In our setting, the binary-valued correctness-based modeling approach is not suitable. We do not expect the users to input exact responses, we expect them to input their best estimates. So our model should work with some measure of the quality of an answer. Another important point is the speed-accuracy tradeoff. Figure 1A shows a hypothetical tradeoff curve for one user with fixed ability. User can answer a task very quickly but it will probably be a very rough estimate. Or he/she can decide to spend more time on the task and respond with a more precise answer. Therefore, response-time should be a vital part of our model.

The system should be able to detect prior skill (i.e., how good the user was at estimation before he started using the system) which can be deduced from the first interactions of the user with the system. The goal of the developed system is to enable the user to get better at estimating. Therefore, the proposed model should also take into account user's improvement (or learning) over time. Figure 1B illustrates answers of several users on one task as red dots. Ideally, the system will help its users to learn to perform near the green mark, to be fast and accurate.

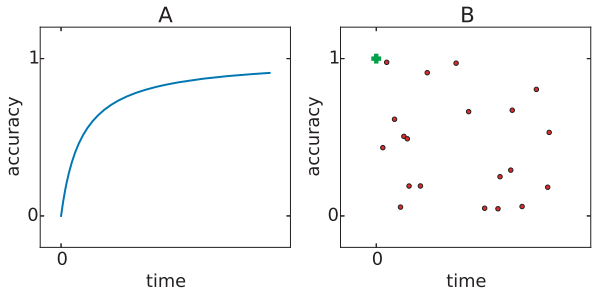

Figure 1: A) hypothetical speed-accuracy tradeoff curve, B) goal of the system

The value of the system will also be in the data that will be collected. It can be used to answer some interesting research questions. Does the speed-accuracy tradeoff curve have the same shape for converting between EUR and USD as for estimating the number of displayed objects? How do the learning curves look? Can estimation tasks in one area be learned more quickly than in another area? How close to the perfect mark can users push their performance? What is the influence of a countdown timer on user's performance? What is the appropriate level of challenge that motivates the users? The last question was addressed in [3], where the authors were trying to validate the *Inverted-U Hypothesis* (i.e., we most enjoy challenges that are neither too easy, neither too hard) on data collected from online estimation game called *Battleship Numberline*. They found out that the

easier the game was, the longer users played the game.

2. MODELS

In this section, we present a few existing models for combining correctness and response-times in Item Response Theory (IRT) and a model for tracking learning currently used in our other adaptive practice system. We then propose a simple model that could be used in the system for practicing estimates. The described models use a logistic function $\sigma(z) = (1 + e^{-z})^{-1}$. Users of the system (or students) are indexed by j. The items (or tasks, problems, questions) that the users solve are indexed by i.

2.1 Models from IRT

A typical example of an approach to the modeling of both correctness and response-times in Item Response Theory is from van der Linden [10]. The approach uses two models, one for correctness (binary) and the other one for response-times (distributed lognormally). The probability of success of a student j on item i can expressed by the 3PL model:

$$p_{ij} = c_i + (1 - c_i) \cdot \sigma(a_i(\theta_j - b_i))$$

where parameter θ_j is the skill of student j and a_i, b_i, c_i are the discrimination, difficulty and pseudo-guessing parameters for the item i. The logarithm of a response-time t_{ij} can be predicted by:

$$\ln \hat{t}_{ij} = \beta_i - \tau_j \quad (1)$$

where β_i represents the amount of labor required to solve item i and τ_j the speed of student j. The disadvantage of this model is that it does not model the speed-accuracy tradeoff explicitly.

An example of a model that directly combines binary correctness with response-time is Roskam's model [7]:

$$p_{ij} = \sigma(\theta_j + \ln t_{ij} - b_i)$$

Here, an increase in item difficulty (or decrease in student's ability) can be always compensated by spending more time on a problem. This tradeoff is called an increasing conditional accuracy function.

2.2 Model for factual knowledge

Here, we present a model that is currently used in a popular adaptive system for practicing geographical facts [4]. This model consists of two parts, one (Elo) estimates the prior knowledge of a student and the second one (PFAE) models student learning. A big advantage of this model is that it uses fast online methods of parameter estimation which makes it suitable for use in an interactive adaptive practice system.

The prior knowledge of a student is modeled by the Rasch (1PL) model. The probability that a student j answers item i correctly is modeled by the likelihood $p_{ij} = \sigma(\theta_j - b_i)$. The parameters are estimated using Elo rating system [1]. Elo was originally developed for rating chess players, but the process of student answering an item can be interpreted as a "match" between the student and the item. After each "match", the parameters are updated as follows:

$$\begin{aligned} \theta_j &:= \theta_j + U(n_j) \cdot (\mathit{correct} - p_{ij}) \\ b_i &:= b_i + U(n_i) \cdot (p_{ij} - \mathit{correct}) \end{aligned}$$

where $U(n)$ is the uncertainty function $U(n) = \frac{\alpha}{1+\beta n}$ and n is the number of updates of the parameter and α and β are metaparameters. The variable *correct* takes value 1 if the student has answered correctly and value 0 otherwise. This model is used for predicting– and trained on–first responses.

After the first interaction of a student j with item i has been observed, we can set student's skill in that particular item to $\theta_{ij} = \theta_j - b_i$. An extended version of Performance Factors Analysis [5] called PFAE is used to model learning and predicting the following interactions of the student with the item. Likelihood of a correct answer is $p_{ij} = \sigma(\theta_{ij})$. The update to student's knowledge of item θ_{ij} after observation is:

$$\theta_{ij} := \begin{cases} \theta_{ij} + \gamma \cdot (1 - p_{ij}) & \text{if the answer was correct} \\ \theta_{ij} + \delta \cdot p_{ij} & \text{if the answer was incorrect} \end{cases}$$

where γ and δ are metaparemeters. The reason for two different metaparameters is that the student learns also during an incorrect response.

2.3 Proposed model for estimates

Here, we propose a model that can be used in the adaptive practice system for estimates. The model combines Roskam's model and the update scheme from Elo and PFAE.

A simple extension of the correctness-based modeling to the setting of practicing estimates is to use a measure of correctness, or a *score* – a rational number ranging from 0 to 1. The way of scoring of an answer could be based on the domain being practiced by the user. For example, for the scenario where the user is estimating the number of objects in a scene, the exact answer would get a score of 1, deviating by one object a score of 0.8, etc.

The model assumes the same parameters and relationship as Roskam's model, but instead of expressing a probability of a correct answer it specifies the expected score:

$$\hat{s_{ij}} = \sigma(\theta_j + \ln t_{ij} - b_i)$$

Figure 2 shows how the score changes as a function of time for different values of user's skill θ_j (with fixed $b_i = 0$). It nicely demonstrates the speed-accuracy tradeoff.

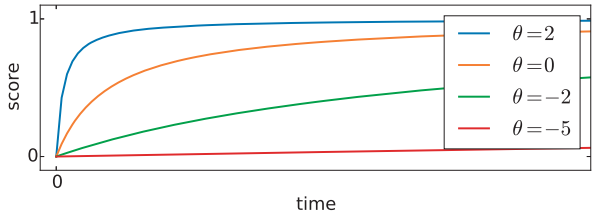

Figure 2: Score function for different values of skill

After observing score s_{ij} that user j obtained for answering item i and response-time t_{ij}, we can update model's beliefs

in the parameters:

$$\begin{aligned}\theta_j &:= \begin{cases}\theta_j + \gamma \cdot (s_{ij} - \hat{s_{ij}}) & \text{if } s_{ij} \geq \hat{s_{ij}} \\ \theta_j + \delta \cdot (\hat{s_{ij}} - s_{ij}) & \text{if } s_{ij} < \hat{s_{ij}}\end{cases} \\ b_i &:= b_i + U(n_i) \cdot (\hat{s_{ij}} - s_{ij})\end{aligned}$$

Note, that the model uses a single parameter θ_j for the student. This is different from the approach taken in PFAE, where the student has a parameter for each item θ_{ij}. While that approach is suitable for modeling the knowledge of facts – where it is reasonable to assume that the knowledge of one fact is independent of the knowledge of another – it is not suitable here. Student's ability to convert 2 miles to kilometers is surely dependent on his ability to convert 3 miles to kilometers.

We propose using separate model for each concept (e.g., estimating the number of objects, conversion lb to kg, conversion EUR to USD). It is true that student's ability to estimate items corresponding to one concept tells us something about his ability to estimate the other concepts. However, if the user does not know the conversion rate from EUR to USD then being able to estimate well the other concepts will not help him.

The model can be easily extended by adding a discrimination parameter a or a guessing parameter c (similarly to the IRT model): $\hat{s_{ij}} = c + (1-c) \cdot \sigma(a(\theta_j + \ln t_{ij} - b_i))$. These added parameters could be either metaparameters of the model or parameters of the item i. The guessing parameter may be useful for the scenario where the user has to select a value on a numberline.

As we mentioned earlier, this model suffers from the issue that increasing the time spent on an item increases the expected score. This may hold true for the instance where the user knows the underlying concept (e.g., the conversion rate from EUR to USD) but it does not hold when he does not know it. But the model uses the logarithm of response-time and the time a student is willing to spend on an item is limited. Therefore, the model should have reasonable behavior for the time interval of interest, as is demonstrated in Figure 2 by the curve corresponding to $\theta_j = -5$.

3. DISCUSSION

The model works with the response-time as a parameter. Therefore, it cannot be used for predicting response-times directly. A model similar to (1) can be used for that. Predicted time and score can be used for item selection (i.e., which item to offer the user next). This can be done by setting a target score and recommending an item with predicted score close to the target.

Does the model perform better than a simple 1PL model that does not use response-times at all? Does it make sense to add more parameters to the model? How does the model fare against more complicated models? To be able to answer these questions, we need to somehow evaluate the performance of the model. The choice of metric is interesting because a model can predict both score and response-time. When considering only the predicted score, a standard metric like RMSE can be used [6]. When we have a measure of performance, we can explore if the model is well-calibrated with respect to response-times or if the model works similarly well for all the domains (concepts).

Other question that we could ask is how well does the speed-accuracy tradeoff curve that the model assumes correspond to reality.

Acknowledgements

We thank Radek Pelánek (for guidance and useful suggestions) and Roman Orlíček (for actively working on developing the application).

4. REFERENCES

[1] A. E. Elo. *The rating of chessplayers, past and present*, volume 3. Batsford London, 1978.

[2] S. A. Hanson and T. P. Hogan. Computational estimation skill of college students. *Journal for Research in Mathematics Education*, pages 483–499, 2000.

[3] D. Lomas, K. Patel, J. L. Forlizzi, and K. R. Koedinger. Optimizing challenge in an educational game using large-scale design experiments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 89–98. ACM, 2013.

[4] J. Papoušek, R. Pelánek, and V. Stanislav. Adaptive practice of facts in domains with varied prior knowledge. In *Educational Data Mining (EDM)*, pages 6–13, 2014.

[5] P. I. Pavlik, H. Cen, and K. R. Koedinger. Performance factors analysis-a new alternative to knowledge tracing. In *Proc. of Artificial Intelligence in Education (AIED)*, volume 200 of *Frontiers in Artificial Intelligence and Applications*, pages 531–538. IOS Press, 2009.

[6] R. Pelánek. A brief overview of metrics for evaluation of student models. In *Approaching Twenty Years of Knowledge Tracing Workshop*, 2014.

[7] E. Roskam. Toward a psychometric theory of intelligence. *Progress in mathematical psychology*, 1:151–174, 1987.

[8] P. M. Seethaler and L. S. Fuchs. The cognitive correlates of computational estimation skill among third-grade students. *Learning Disabilities Research & Practice*, 21(4):233–243, 2006.

[9] R. S. Siegler, C. A. Thompson, and M. Schneider. An integrated theory of whole number and fractions development. *Cognitive psychology*, 62(4):273–296, 2011.

[10] W. J. Van Der Linden. Conceptual issues in response-time modeling. *Journal of Educational Measurement*, 46(3):247–272, 2009.