

Using Logistic Regression to Trace Multiple Subskills in a Dynamic Bayes Net

YANBO XU and JACK MOSTOW
Carnegie Mellon University, United States

A challenge in estimating students' changing knowledge from sequential observations of their performance arises when each observed step involves multiple subskills. To overcome this mismatch in grain size between modelled skills and observed actions, we use logistic regression over each step's subskills in a dynamic Bayes net (LR-DBN) to model transition probabilities for the overall knowledge required by the step. Unlike previous methods, LR-DBN can trace knowledge of the individual subskills without assuming they are independent. We evaluate how well it fits children's oral reading fluency data logged by Project LISTEN's Reading Tutor, compared to other methods.

Key terms: dynamic Bayes net, logistic regression, knowledge tracing, multiple subskills, oral reading fluency

1. INTRODUCTION

Dynamic Bayes nets are often used to model skill acquisition, e.g., in Knowledge Tracing (KT) [Corbett and Anderson, 1995]. They model a skill as a hidden state of knowledge, and estimate the changing probability of this state by observing successive attempts to use the skill. However, KT does not model multiple subskills used in such an attempt.

Previous research has explored various approaches to this problem. One approach is to use a conjunctive model [Cen et al., 2008; Gong et al., 2010], which assumes that a student must master all of the subskills in order to perform the step correctly. If the subskills are independent, the probability of knowing them all can be estimated by multiplying the estimated probabilities of knowing the individual subskills. However, this product typically underestimates the probability of knowing all the subskills. The minimum of their estimated probability provides a less pessimistic estimate based on the assumption that the likelihood of a correct answer is dominated by the student's weakest subskill [Gong et al., 2010]. Alternatively, Koedinger et al. [2011] use techniques from Bayesian nets to avoid blaming each subskill equally in conjunctive models. By pre-specifying the relationship between a step and its subskills, all of these approaches make strong assumptions about independence of subskills. Performance Factors Analysis [Pavlik Jr. et al., 2009b] and Learning Factors Analysis [Cen et al., 2006; Pavlik Jr. et al., 2009a] use non-linear regression to estimate multiple subskills without making such assumptions, but statically: that is, they do not trace subskills over time.

This paper presents LR-DBN, a method that uses logistic regression to trace multiple subskills in a dynamic Bayes net student model without assuming they are independent. Section 2 explains how LR-DBN works, Section 3 evaluates it, and Section 4 concludes.

2. LOGISTIC REGRESSION IN A DYNAMIC BAYES NET

In a KT model, and in other dynamic Bayes net models for student modelling [Chang et al., 2006], we estimate the probability of a student knowing, learning, or forgetting the skill(s) required to perform each observed step. So, we use a latent variable to model a hidden knowledge state that changes over time, and infer it from sequential observations of the student's performance. If we know (or assume) which set of subskills a step

Authors' addresses: Y. Xu, E-mail: yanbox@cs.cmu.edu; J. Mostow, E-mail: mostow@cs.cmu.edu, Project LISTEN, RI-NSH 4103, 5000 Forbes Avenue, Carnegie Mellon University, Pittsburgh, PA 15213-3890, USA. This work was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A080628 to Carnegie Mellon University. The opinions expressed are those of the authors and do not necessarily represent the views of the Institute or U.S. Department of Education.

requires, it makes sense to estimate overall knowledge of the step as some function of the estimated knowledge of each individual subskill it requires.

Accordingly, we propose to model the probabilities of transitions between successive knowledge states using logistic regression over all of the subskills. We later prove that this approach is equivalent to modeling the knowledge probabilities themselves using logistic regression, but it provides a more convenient way to trace the individual subskills.

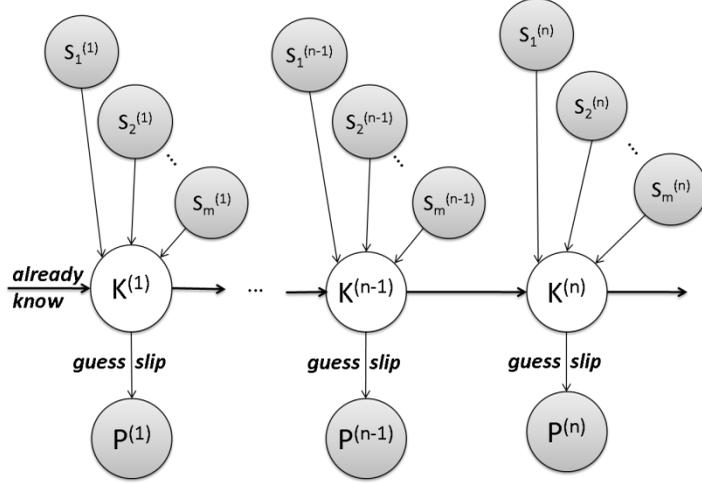


Fig. 1. A dynamic Bayes Architecture with Logistic Regression

Fig. 1 shows a dynamic Bayes architecture of KT framework model with binary variables:

- $S_j^{(n)}$: known indicator variable; 1 if step n requires subskill j , 0 otherwise.
- $K^{(n)}$: hidden; true iff the student has the knowledge step n requires.
- $P^{(n)}$: observed; true iff the student performs step n correctly.

Besides, we denote *already know* at initial state as $P(K^{(0)} = \text{true})$. Then we use logistic regression to model the $(1 - \text{already know})$ and transition probabilities from the knowledge state $K^{(n-1)}$ at step $n-1$ to the knowledge state $K^{(n)}$ at step n over m subskills:

$$1 - \text{already know} = P(K^{(0)} = \text{false}) = \frac{\exp(-\sum_{j=1}^m \beta_j^{(0)} S_j^{(0)})}{1 + \exp(-\sum_{j=1}^m \beta_j^{(0)} S_j^{(0)})} \quad (1)$$

$$P(K^{(n)} = \text{false} \mid K^{(n-1)} = \text{false}) = \frac{\exp(-\sum_{j=1}^m \beta_j S_j^{(n)})}{1 + \exp(-\sum_{j=1}^m \beta_j S_j^{(n)})} \quad (2)$$

$$P(K^{(n)} = \text{false} \mid K^{(n-1)} = \text{true}) = \frac{\exp(-\sum_{j=1}^m \gamma_j S_j^{(n)})}{1 + \exp(-\sum_{j=1}^m \gamma_j S_j^{(n)})} \quad (3)$$

Here $\beta_j^{(0)}$ is the coefficient fit for skill j at the initial state, where $S_j^{(0)} = 1$, and β_j and γ_j represent skill j 's respective contributions (when involved) to the transition probabilities in (2) and (3) from $K^{(n-1)}$ to $K^{(n)}$. We assume that β_j and γ_j do not vary over time. These equations imply that the more subskills a step requires, or the harder they are to learn, the lower the probability of knowing or learning the step. We now show how the model can trace individual subskills, first in a simple case and then in the general case.

Tracing subskills in a simple case: Consider a simple scenario in which a student repeatedly practices a single step that involves multiple subskills. The transition probabilities in equations (1), (2), and (3) correspond to $(1 - \text{already know})$, $(1 - \text{learn})$, and *forget* in KT. For now we assume *forget* is 0, a standard assumption in KT. Thus:

$$\text{already know} = P(K^{(0)} = \text{true}) = 1 - P(K^{(0)} = \text{false}) = \frac{1}{1 + \exp(-\sum_{j=1}^m \beta_j^{(0)} S_j)} \quad (4)$$

$$\begin{aligned} \text{learn} &= P(K^{(n)} = \text{true} \mid K^{(n-1)} = \text{false}) = 1 - P(K^{(n)} = \text{false} \mid K^{(n-1)} = \text{false}) \\ &= \frac{1}{1 + \exp(-\sum_{j=1}^m \beta_j S_j)} \end{aligned} \quad (5)$$

$$\text{forget} = P(K^{(n)} = \text{false} \mid K^{(n-1)} = \text{true}) = 0 \quad (6)$$

Note that the variables S_j 's that indicate which subskills are used in a step do not have superscripts of n because their assignments are determined by the step, so they remain constant for repeated practice of the same step. Now we compute $P(K^{(1)} = \text{true})$:

$$\begin{aligned} P(K^{(1)} = \text{true}) &= 1 - P(K^{(1)} = \text{false}) \\ &= 1 - [P(K^{(1)} = \text{false} \mid K^{(0)} = \text{true}) \times P(K^{(0)} = \text{true}) \text{ (By Bayes rule)} \\ &\quad + P(K^{(1)} = \text{false} \mid K^{(0)} = \text{false}) \times P(K^{(0)} = \text{false})] \\ &= 1 - 0 - \frac{\exp(-\sum_{j=1}^m \beta_j S_j)}{1 + \exp(-\sum_{j=1}^m \beta_j S_j)} \times \frac{\exp(-\sum_{j=1}^m \beta_j^{(0)} S_j)}{1 + \exp(-\sum_{j=1}^m \beta_j^{(0)} S_j)} \quad (\text{By eq. 6, 4, and 5}) \\ &= 1 - \frac{\exp(-\sum_{j=1}^m (\beta_j^{(0)} + \beta_j) S_j)}{1 + \exp(-\sum_{j=1}^m \beta_j S_j) + \exp(-\sum_{j=1}^m \beta_j^{(0)} S_j) + \exp(-\sum_{j=1}^m (\beta_j^{(0)} + \beta_j) S_j)} \end{aligned} \quad (7)$$

Define $C = \exp(-\sum_{j=1}^m \beta_j S_j) + \exp(-\sum_{j=1}^m \beta_j^{(0)} S_j)$. We can always find some $\delta_j^{(1)}$'s such that $\frac{1}{1+C} = \exp(-\sum_{j=1}^m \delta_j^{(1)} S_j)$, e.g., by choosing $\delta_j^{(1)}$'s to be equal, based on the assumption that the subskills change equally. Then:

$$\begin{aligned} P(K^{(1)} = \text{true}) &= 1 - \frac{\exp(-\sum_{j=1}^m (\beta_j^{(0)} + \beta_j) S_j)}{1 + C + \exp(-\sum_{j=1}^m (\beta_j^{(0)} + \beta_j) S_j)} \\ &= 1 - \frac{\frac{1}{1+C} \exp(-\sum_{j=1}^m (\beta_j^{(0)} + \beta_j) S_j)}{1 + \frac{1}{1+C} \exp(-\sum_{j=1}^m (\beta_j^{(0)} + \beta_j) S_j)} \\ &= 1 - \frac{\exp(-\sum_{j=1}^m (\beta_j^{(0)} + \beta_j + \delta_j^{(1)}) S_j)}{1 + \exp(-\sum_{j=1}^m (\beta_j^{(0)} + \beta_j + \delta_j^{(1)}) S_j)} \\ &= \frac{1}{1 + \exp(-\sum_{j=1}^m (\beta_j^{(0)} + \beta_j + \delta_j^{(1)}) S_j)} \\ &= \frac{1}{1 + \exp(-\sum_{j=1}^m (\beta_j^{(0)} + \Delta\beta_j^{(1)}) S_j)} \end{aligned} \quad (8)$$

Here we define $\Delta\beta_j^{(1)} = \beta_j + \delta_j^{(1)}$. More generally, at step n we have:

$$P(K^{(n)} = \text{true}) = \frac{1}{1 + \exp(-\sum_{j=1}^m (\beta_j^{(0)} + \Delta\beta_j^{(1)} + \dots + \Delta\beta_j^{(n)}) S_j)} \quad (9)$$

Therefore, we can use the proposed model to trace subskill j to step n by using $\Delta\beta_j^{(1)}, \dots, \Delta\beta_j^{(n)}$. Meanwhile, we also showed that the probability of having all the knowledge for step n is a logistic regression over coefficients $\beta_j^{(0)} + \Delta\beta_j^{(1)} + \dots + \Delta\beta_j^{(n)}$ for each subskill $j = 1, \dots, m$.

Tracing subskills in the general case: What if steps require different sets of subskills?

$$\begin{aligned} P(K^{(1)} = \text{true}) &= P(K^{(1)} = \text{true} \mid K^{(0)} = \text{true}) \times P(K^{(0)} = \text{true}) \\ &\quad + P(K^{(1)} = \text{true} \mid K^{(0)} = \text{false}) \times P(K^{(0)} = \text{false}) \quad (\text{By Bayes rule}) \\ &= \frac{1}{1 + \exp(-\sum_{j=1}^m \nu_j S_j^{(1)})} \frac{1}{1 + \exp(-\sum_{j=1}^m \beta_j^{(0)} S_j^{(0)})} \end{aligned} \quad (\text{By eq. 1, 2 and 3})$$

$$+ \frac{1}{1 + \exp(-\sum_{j=1}^m \beta_j S_j^{(1)})} \frac{\exp(-\sum_{j=1}^m \beta_j^{(0)} S_j^{(0)})}{1 + \exp(-\sum_{j=1}^m \beta_j^{(0)} S_j^{(0)})} \quad (10)$$

(Note that *forget* is no longer 0 in the general case.)

Using the same trick as for (8), let $C' = \exp(-\sum_{j=1}^m \beta_j S_j^{(1)}) (1 + \exp(-\sum_{j=1}^m \beta_j^{(0)} S_j^{(0)}))$, and let $C'' = \exp(\sum_{j=1}^m \beta_j^{(0)} S_j^{(0)}) (1 + \exp(\sum_{j=1}^m \beta_j S_j^{(1)})) + \exp(\sum_{j=1}^m \gamma_j S_j^{(1)})$. We choose $\delta_j^{(1)}$ and $\epsilon_j^{(1)}$ such that $\frac{1}{1+C'} = \exp(-\sum_{j=1}^m \delta_j^{(1)} S_j^{(0)} \bar{S}_j^{(1)})$ and $1 + C'' = \exp(-\sum_{j=1}^m \epsilon_j^{(1)} \bar{S}_j^{(0)} S_j^{(1)})$, where $\bar{S} = 1 - S$. We then rewrite (10) as:

$$P(K^{(1)} = \text{true}) = \frac{1}{1 + \exp(-\sum_{j=1}^m (\beta_j^{(0)} S_j^{(0)} + (\beta_j + \gamma_j) S_j^{(0)} S_j^{(1)} + \Delta\beta_j^{(1)} S_j^{(0)} \bar{S}_j^{(1)} + \Delta\gamma_j^{(1)} \bar{S}_j^{(0)} S_j^{(1)}))} \quad (11)$$

Here $\Delta\beta_j^{(1)} = \beta_j + \gamma_j + \delta_j^{(1)}$ and $\Delta\gamma_j^{(1)} = \beta_j + \gamma_j + \epsilon_j^{(1)}$. More generally, at step n :

$$P(K^{(n)} = \text{true}) = \frac{1}{1 + \exp(-\sum_{j=1}^m (\beta_j^{(0)} S_j^{(0)} + \sum_{k=1}^n (\beta_j + \gamma_j) S_j^{(k-1)} S_j^{(k)} + \Delta\beta_j^{(k)} S_j^{(k-1)} S_j^{(k)} + \Delta\gamma_j^{(k)} S_j^{(k-1)} S_j^{(k)}))} \quad (12)$$

Thus we can trace subskill j at step n by using $\Delta\beta_j^{(n)}$ if $S_j^{(n-1)} = 0$ and $S_j^{(n)} = 1$, $\Delta\gamma_j^{(n)}$ if $S_j^{(n-1)} = 1$ and $S_j^{(n)} = 0$, and $\beta_j + \gamma_j$ if $S_j^{(n-1)} = 1$ and $S_j^{(n)} = 1$. To save space, we only showed here how to update subskill knowledge independent of observed performance on the step. To condition on performance, we can further derive $P(K^{(2)} = \text{true} | P^{(1)})$, and therefore derive $P(K^{(n)} = \text{true} | P^{(1)}, \dots, P^{(n-1)})$ in (9) and (12).

3. EXPERIMENTS

We implemented LR-DBN in the Bayes Net Toolbox for Matlab [Murphy, 2006]. Specifically, we defined the knowledge K given m subskills S_j as a ‘‘softmax’’ node in the toolbox. We used LR-DBN to model the growth of children’s oral reading fluency, where performance P denotes whether the student read a word fluently. Our data was recorded by Project LISTEN’s Reading Tutor [Mostow and Aist, 2001] during the 2005-2006 school year. We scored each word as fluent if read without help or hesitation and accepted by the automated speech recognizer.

We assume that whether a student read a word fluently depended on whether the student knew the grapheme-to-phoneme mappings in the word. So in our experiment, the subskills required in a student’s reading word step are the word’s grapheme-to-phoneme mappings. We modeled 27 children who read a total of 5,078 distinct word types with 332 unique grapheme-phoneme mappings. To evaluate our models, we fit them separately for each student on the first half of all of that student’s data, tested on the second half, and averaged the test results across students. The test set contains a total of 32,122 read words, out of which 23,222 were fluent. For comparison, we also applied the original KT model and estimated the probability of knowing a word as the minimum probability of knowing all of its grapheme-to-phoneme mappings, based on assuming that the student’s weakest subskill determined whether he read the word fluently.

Table I shows the results. The values in parentheses show 95% confidence intervals based on standard error calculated from the unbiased weighted sample variance of individual students’ accuracies. Since the data is unbalanced (72.3% of the words were fluent), we also report within-class accuracies. Table I shows that LR-DBN significantly outperformed the weakest-subskill KT model, especially on non-fluent words. High within-class accuracy on unbalanced data is often hard for KT [Zhang et al., 2008].

Table I. LR-DBN vs. KT Models of Children's Reading Fluency Growth

	Accuracy	Accuracy on fluent words	Accuracy on non-fluent words
LR-DBN	88.8% ($\pm 1.7\%$)	92.0% ($\pm 2.7\%$)	80.5% ($\pm 9.5\%$)
KT of weakest skill	72.6% ($\pm 3.6\%$)	94.5% ($\pm 5.8\%$)	19.3% ($\pm 8.4\%$)

4. CONCLUSION

This paper describes and evaluates LR-DBN, a novel student modeling method to trace hidden subskills by using logistic regression within a dynamic Bayes net. We used oral reading data from 27 children to compare LR-DBN to conventional knowledge tracing of weakest subskills. LR-DBN performed significantly better overall, thanks to similar accuracy (92.0% \pm 2.7% vs. 94.5% \pm 5.8%) on fluent words combined with 4-fold higher accuracy on disfluent words (80.5% \pm 9.5% vs. 19.3% \pm 8.4%). We later [Xu and Mostow, 2011] tested both models on a published data set [Koedinger et al., 2010] from 123 students working on a geometry area unit of the Bridge to Algebra Cognitive Tutor[®]. LR-DBN fit this data significantly better too, with only half as many prediction errors on unseen data. Future tests should use data from more students and tasks, and compare LR-DBN to other baselines, such as conjunctive modeling [Koedinger et al., 2011].

REFERENCES

- CEN, H., KOEDINGER, K. and JUNKER, B. 2006. Learning Factors Analysis – A General Method for Cognitive Model Evaluation and Improvement. In *Proceedings of the 8th International Conference on Intelligent Tutoring Systems*, Jhongli, Taiwan, June 26-30, 2006, 164-175.
- CEN, H., KOEDINGER, K.R. and JUNKER, B. 2008. Comparing Two IRT Models for Conjunctive Skills. In *Ninth International Conference on Intelligent Tutoring Systems*, Montreal, 2008, 796-798.
- CHANG, K.-M., BECK, J.E., MOSTOW, J. and CORBETT, A. 2006. A Bayes Net Toolkit for Student Modeling in Intelligent Tutoring Systems. In *Proceedings of the 8th International Conference on Intelligent Tutoring Systems*, Jhongli, Taiwan, June 26-30, 2006, K. ASHLEY and M. IKEDA, Eds., 104-113.
- CORBETT, A. and ANDERSON, J. 1995. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction* 4, 253-278.
- GONG, Y., BECK, J. and HEFFERNAN, N.T. 2010. Comparing Knowledge Tracing and Performance Factor Analysis by Using Multiple Model Fitting Procedures. In *Proceedings of the 10th International Conference on Intelligent Tutoring Systems*, Pittsburgh, PA, 2010, V. ALEVEN, J. KAY and J. MOSTOW, Eds. Springer Berlin / Heidelberg, 35-44.
- KOEDINGER, K., PAVLIK, P.I., STAMPER, J., NIXON, T. and RITTER, S. 2011. Fair Blame Assignment in Student Modeling. In *Fourth International Conference on Educational Data Mining*, Eindhoven, NL, July 6-8, 2011.
- KOEDINGER, K.R., BAKER, R.S.J.D., CUNNINGHAM, K., SKOGSHOLM, A., LEBER, B. and STAMPER, J. 2010. A Data Repository for the EDM community: The PSLC DataShop. In *Handbook of Educational Data Mining*, C. ROMERO, S. VENTURA, M. PECHENIZKIY and R.S.J.D. BAKER, Eds. CRC Press, Boca Raton, FL, 43-55.
- MOSTOW, J. and AIST, G. 2001. Evaluating tutors that listen: An overview of Project LISTEN. In *Smart Machines in Education*, K. FORBUS and P. FELTOVICH, Eds. MIT/AAAI Press, Menlo Park, CA, 169-234.
- MURPHY, K. 2006. Bayes Net Toolbox for Matlab.
- PAVLIK JR., P.I., CEN, H. and KOEDINGER, K.R. 2009a. Learning factors transfer analysis: Using learning curve analysis to automatically generate domain models. In *Proceedings of the 2nd International Conference on Educational Data Mining*, Cordoba, Spain, 2009a, T. BARNES, M. DESMARAIS, C. ROMERO and S. VENTURA, Eds., 121-130.
- PAVLIK JR., P.I., CEN, H. and KOEDINGER, K.R. 2009b. Performance Factors Analysis - A New Alternative to Knowledge Tracing. In *Proceedings of the 14th International Conference on Artificial Intelligence in Education (AIED09)*, 2009b, 531-538.
- XU, Y. and MOSTOW, J. 2011. Logistic Regression in a Dynamic Bayes Net Models Multiple Subskills Better! In *4th International Conference on Educational Data Mining* Eindhoven, Netherlands, July 6-8, 2011.
- ZHANG, X., MOSTOW, J. and BECK, J.E. 2008. A Case Study Empirical Comparison of Three Methods to Evaluate Tutorial Behaviors. In *9th International Conference on Intelligent Tutoring Systems*, Montreal, June 23-27, 2008, B. WOLFF, Ed., Springer-Verlag, 122-131.