DAS3H: Modeling Student Learning and Forgetting for Optimally Scheduling Distributed Practice of Skills

Benoît Choffin, Fabrice Popineau, Yolaine Bourda & Jill-Jênn Vie

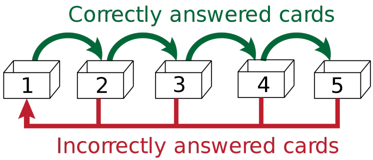
LRI/CentraleSupélec - University of Paris-Saclay | RIKEN AIP

EDM 2019, Montréal | July 4, 2019

Mitigating human forgetting with spaced repetition

- Human learners face a constant trade-off between **acquiring new knowledge** and **reviewing old knowledge**
- Cognitive science provides simple + robust learning strategies for improving LT memory
  - Spaced repetition
  - Testing
- Can we do better? **Yes**, by providing students with an *adaptive* and *personalized* spacing scheduler.

Mitigating human forgetting with spaced repetition



Ex. select the item whose memory strength is closest to a threshold θ [Lindsey, Shroyer, Pashler, and Mozer 2014] $\rightarrow$ "almost forgotten"

Beyond flashcard memorization

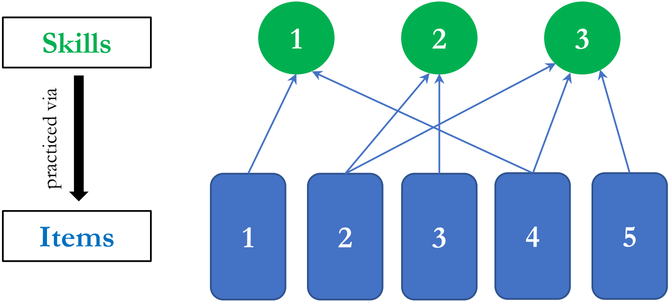
Problem: these algorithms are designed for optimizing *pure memorization* (of facts, vocabulary,. . .)

- In real-world educational settings, students also need to learn to master and remember a set of **skills**
- In that case, specific items are the only way to practice one or multiple skills because *we do not have to memorize the content directly*
- Traditional adaptive spacing schedulers are **not applicable for learning skills**

Extension to skill practice and review

Item-skill relationships require expert labor and are synthesized inside a binary q-matrix →

	skill 1	skill 2	skill 3
item 1	1	0	0
item 2	0	1	1
item 3	0	1	0
item 4	1	0	1
item 5	0	0	1



Limitations of student models

We need to be able to infer skill memory strength and dynamics, however in the student modeling literature:

- some models leverage item-skills relationships
- some others incorporate forgetting

But none does both!

Our contribution

We take a model-based approach for this task.

1. Traditional adaptive spacing algorithms can be extended to review and practice skills (not only flashcards).

2. We developed a new student *learning* and *forgetting* model that leverages item-skill relationships: **DAS3H**.
    - DAS3H outperforms 4 SOTA student models on 3 datasets.
    - Incorporating skill info + forgetting effect improves over models that consider one or the other.
    - Using precise temporal information on past skill practice + assuming different learning/forgetting curves for different skills improves performance.

Outline

1. DASH
2. Our model DAS3H
3. Experiments
4. Conclusion

DASH

$\rightarrow$ DASH = item **D**ifficulty, student **A**bility, and **S**tudent **H**istory

DASH [Lindsey, Shroyer, Pashler, and Mozer 2014] bridges the gap between *Factor Analysis models* and *memory models*:

$$\mathbb{P}\left(Y_{s,j,t} = 1\right) = \sigma(\alpha_s - \delta_j + h_\theta(\mathrm{t}_{s,j,1:l}, \mathrm{y}_{s,j,1:l-1}))$$

where:

- $Y_{s,j,t}$ binary correctness of student s answering item j at time t;
- σ logistic function;
- α_s ability of student s;
- δ_j difficulty of item j;
- h_θ summarizes the effect of the $l-1$ previous attempts of s on j at times $\mathrm{t}_{s,j,1:l-1}$ + the binary outcomes $\mathrm{y}_{s,j,1:l-1}$.

DASH

Lindsey et al. chose:

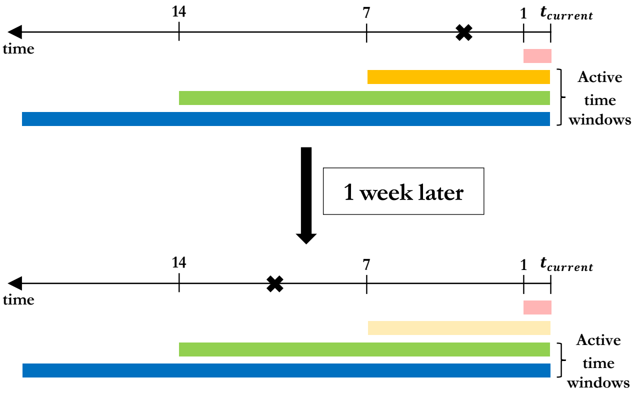
$$h_\theta(\mathrm{t}_{s,j,1:l}, \mathrm{y}_{s,j,1:l-1}) = \sum_{w=0}^{W-1} \theta_{2w+1} \log(1 + c_{s,j,w}) - \theta_{2w+2} \log(1 + a_{s,j,w})$$

where:

- w indexes a set of expanding time windows;
- $c_{s,j,w}$ number of correct answers of s on j in time window w;
- $a_{s,j,w}$ number of attempts of s on j in time window w;
- θ is *learned* by DASH.

DASH

Assuming that the set of time windows is $\{1, 7, 14, +\infty\}$:



DASH

DASH:

- accounts for both *learning* and *forgetting* processes;
- induces diminishing returns of practice inside a time window (log-counts);
- has a time module h_θ inspired by ACT-R [Anderson, Matessa, and Lebiere 1997] and MCM [Pashler, Cepeda, Lindsey, Vul, and Mozer 2009].

From DASH to DAS3H

- DASH
  - outperforms a hierarchical Bayesian IRT on Lindsey et al. experimental data (vocabulary learning).
  - was successfully used to adaptively personalize item review in a real-world cognitive psychology experiment.
- However, DASH
  - does not handle multiple skill item tagging → useful to account for knowledge transfer from one item to another.
  - assumes that memory decays at the same rate for every KC.

Our model DAS3H

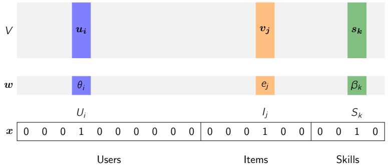
We extend DASH in **3 ways**:

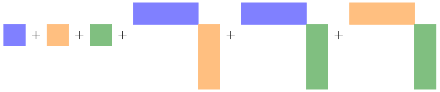
1. Extension to handle multiple skills tagging: new temporal module h_θ that also takes the multiple skills into account.
   - Influence of the temporal distribution of past attempts and outcomes can differ from one skill to another.
2. Estimation of easiness parameters for *each* item j and skill k;
3. Use of KTMs [Vie and Kashima 2019] instead of mere logistic regression for multidimensional feature embeddings and pairwise interactions.

Knowledge Tracing Machines (KTMs)

Just pick features (ex. user, item, skill) and you get a student model

Each feature k is modeled by bias w_k and embedding $\boldsymbol{v_k}$.





$$\operatorname{logit} p(\boldsymbol{x}) = \mu + \underbrace{\sum_{k=1}^{N} w_k x_k}_{\text{logistic regression}} + \underbrace{\sum_{1 \leq k < l \leq N} x_k x_l \langle \boldsymbol{v_k}, \boldsymbol{v_l} \rangle}_{\text{pairwise relationships}}$$

Jill-Jênn Vie and Hisashi Kashima (2019). "Knowledge Tracing Machines: Factorization Machines for Knowledge Tracing". In: *Proceedings of the 33th AAAI Conference on Artificial Intelligence*, to appear. URL: http://arxiv.org/abs/1811.03388

Our model DAS3H

$\rightarrow$ DAS3H = item **D**ifficulty, student **A**bility, **S**kill and **S**tudent **S**kill practice **H**istory

For an embedding dimension of $d = 0$, DAS3H is:

$$\mathbb{P}(Y_{s,j,t} = 1) = \sigma(\alpha_s - \delta_j + \underbrace{\sum_{k \in KC(j)} \beta_k}_{\text{skill easiness biases}} + h_\theta(\mathrm{t}_{s,j,1:l}, \mathrm{y}_{s,j,1:l-1})).$$

We choose:

$$h_\theta(\mathrm{t}_{s,j,1:l}, \mathrm{y}_{s,j,1:l-1}) = \sum_{k \in KC(j)} \sum_{w=0}^{W-1} \theta_{k,2w+1} \log(1 + c_{s,k,w}) - \theta_{k,2w+2} \log(1 + a_{s,k,w}).$$

$\rightarrow$ Now, h_θ can be seen as a sum of *skill* memory strengths!

Experiments

1. Experimental setting
2. Contenders
3. Datasets
4. Main results
5. Further analyses

Experimental setting

How to compare ML models?

Train the models on one part of the dataset
Test on the other part
Gather prediction metrics, compare the models

- **5-fold cross-validation** at the student level: predicting binary outcomes on unseen students (*strong generalization*)
- Distributional assumptions to avoid overfitting:
  - When $d = 0$: L2 regularization/$\mathcal{N}(0,1)$ prior
  - When $d > 0$: hierarchical distributional scheme
- Same time windows as Lindsey et al.: $\{1/24, 1, 7, 30, +\infty\}$

Contenders

5 contenders:

- DAS3H
- DASH [Lindsey, Shroyer, Pashler, and Mozer 2014]
- IRT/MIRT [Linden and Hambleton 2013]
- PFA [Pavlik, Cen, and K. R. Koedinger 2009]
- AFM [Cen, K. Koedinger, and Junker 2006]

Every model was cast within the KTM framework → 3 embedding dimensions (0, 5 & 20) + sparse feature encoding.

	users	items	skills	wins	fails	attempts	tw [KC]	tw [items]
DAS3H	x	x	x	x		x	x	
DASH	x	x		x		x		x
IRT/MIRT	x	x						
PFA			x	x	x			
AFM			x			x		

Datasets

- 3 datasets: ASSISTments 2012-2013, Bridge to Algebra 2006-2007 & Algebra I 2005-2006 (KDD Cup 2010)
  - Data consists of logs of student-item interactions on 2 ITS
  - Selected because they contain *both* timestamps and items with multiple skills $\rightarrow$ rare species in the EDM datasets fauna
- Preprocessing scheme: removed users with < 10 interactions, interactions with `NaN` skills, duplicates

Dataset	Users	Items	Skills	Interactions	Mean correctness	Skills per item	Mean skill delay	Mean study period
assist12	24,750	52,976	265	2,692,889	0.696	1.000	8.54	98.3
bridge06	1,135	129,263	493	1,817,427	0.832	1.013	0.83	149.5
algebra05	569	173,113	112	607,000	0.755	1.363	3.36	109.9

Table 2: Datasets characteristics

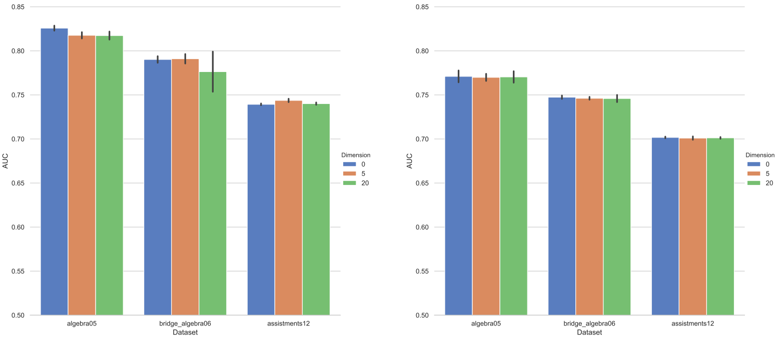
Main results

model	algebra05	bridge06	assist12
DAS3H	**0.826** $\pm$ 0.003	**0.790** $\pm$ 0.004	**0.739** $\pm$ 0.001
DASH	0.773 $\pm$ 0.002	0.749 $\pm$ 0.002	0.703 $\pm$ 0.002
IRT	0.771 $\pm$ 0.007	0.747 $\pm$ 0.002	0.702 $\pm$ 0.001
PFA	0.744 $\pm$ 0.004	0.739 $\pm$ 0.003	0.668 $\pm$ 0.002
AFM	0.707 $\pm$ 0.005	0.692 $\pm$ 0.002	0.608 $\pm$ 0.002

Table 3: AUC comparison between the different student models for an embedding dimension $d = 0$ (all datasets, 5-fold cross-validation).

$\rightarrow$ On every dataset, **DAS3H outperforms** the other models (between $+0.04$ and $+0.05$ AUC compared to DASH).

Main results



(a) DAS3H (b) IRT

Figure 1: AUC comparison on two models for $d = 0, 5$ and 20 (all datasets, 5-fold cross-validation).

$\rightarrow$ The impact of the multidim feature embeddings is small and not consistent across datasets and models (+ unstable sometimes).

Importance of time windows

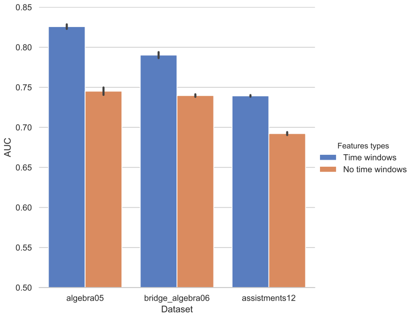


Figure 2: AUC comparison on DAS3H *with* and *without* time windows features (all datasets, 5-fold cross-validation).

Without time windows, h_θ counts past wins and attempts in DAS3H. $\rightarrow$ Using temporal distribution of past skill practice instead of simple win/fail counters improves AUC performance: the ***when*** matters.

Importance of different learning/forgetting curves per skill

	d	bridge06	algebra05	assist12
DAS3H	0	**0.790** $\pm$ 0.004	**0.826** $\pm$ 0.003	0.739 $\pm$ 0.001
	5	**0.791** $\pm$ 0.005	0.818 $\pm$ 0.004	**0.744** $\pm$ 0.002
	20	0.776 $\pm$ 0.023	0.817 $\pm$ 0.005	0.740 $\pm$ 0.001
DAS3H_{1p}	0	0.757 $\pm$ 0.003	0.789 $\pm$ 0.009	0.701 $\pm$ 0.002
	5	0.757 $\pm$ 0.005	0.787 $\pm$ 0.005	0.700 $\pm$ 0.001
	20	0.757 $\pm$ 0.003	0.789 $\pm$ 0.006	0.701 (<1e-3)

Table 4: AUC comparison between DAS3H and DAS3H_{1p} (all datasets, 5-fold cross-validation).

$\rightarrow$ Assuming **different learning and forgetting curves for different skills** in DAS3H consistently yields better predictive power: some skills are easier to learn and slower to forget.

In a nutshell

- Human forgetting is *ubiquitous* but luckily:
  - Cognitive science gives us efficient and simple learning strategies
  - ML can build us tools to **personalize these strategies** and further improve LT memory retention
- Adaptive spacing algorithms have been focusing on *pure memorization* (e.g. vocabulary learning)
  - They can be used for optimizing practice and retention of skills
- Our student model **DAS3H**
  - incorporates information on *skills* **and** *forgetting* to predict learner performance
  - shows higher predictive power than other SOTA student models
  - fits our model-based approach for optimally scheduling skill review

Thanks for your attention!

Our paper is already available at:

`https://arxiv.org/abs/1905.06873`

Python code is freely available on my GitHub page:

`https://github.com/BenoitChoffin/das3h` !

To send me questions about our paper or my research work:

`benoit.choffin@lri.fr`

Anderson, John R, Michael Matessa, and Christian Lebiere (1997). "ACT-R: A theory of higher level cognition and its relation to visual attention". In: *Human-Computer Interaction* 12.4, pp. 439–462.

Cen, Hao, Kenneth Koedinger, and Brian Junker (2006). "Learning factors analysis–a general method for cognitive model evaluation and improvement". In: *International Conference on Intelligent Tutoring Systems*. Springer, pp. 164–175.

Linden, Wim J van der and Ronald K Hambleton (2013). *Handbook of modern item response theory*. Springer Science & Business Media.

Lindsey, Robert V, Jeffery D Shroyer, Harold Pashler, and Michael C Mozer (2014). "Improving students' long-term knowledge retention through personalized review". In: *Psychological science* 25.3, pp. 639–647.

Pashler, Harold, Nicholas Cepeda, Robert V Lindsey, Ed Vul, and Michael C Mozer (2009). “Predicting the optimal spacing of study: A multiscale context model of memory”. In: *Advances in neural information processing systems*, pp. 1321–1329.

Pavlik, Philip I., Hao Cen, and Kenneth R. Koedinger (2009). “Performance Factors Analysis - A New Alternative to Knowledge Tracing”. In: *Proceedings of the 14th International Conference on Artificial Intelligence in Education, AIED 2009*, pp. 531–538.

Vie, Jill-Jênn and Hisashi Kashima (2019). “Knowledge Tracing Machines: Factorization Machines for Knowledge Tracing”. In: *Proceedings of the 33th AAAI Conference on Artificial Intelligence*, to appear. URL: http://arxiv.org/abs/1811.03388.