INCREMENTAL CONTEXT AWARE ATTENTIVE KNOWLEDGE TRACING

Cheryl Sze Yin Wong, Guo Yang, Nancy F. Chen, Ramasamy Savitha

Institute for Infocomm Research (I2R), A-STAR
1 Fusionopolis Way, Singapore 138632

ABSTRACT

Knowledge Tracing is the prediction of the future performance of a learner, given the past performance. The existing knowledge tracing models represent the training data and does not generalize when there is a drift in the data distribution. We first empirically demonstrate an evolving Knowledge Tracing (eKT) scenario with distinct distribution of learner performances and diversity of questions from similar concepts. Next, we empirically characterize drift in the data and propose a task agnostic incremental context aware attentive knowledge tracing (iAKT) approach to learn incrementally from the eKT. The iAKT regularizes representations to learn from diverse learner performance distributions. Finally, we evaluate the ability of the proposed iAKT for knowledge tracing and study the effect of various regularization strategies on ranking difficulty of questions, using the ASSISTments 2017 data set. Performance results show that the iAKT adapts its representations to drift in data characteristics, while iAKT with EWC regularizer is better at ranking the difficulty of questions.

Index Terms— knowledge tracing, incremental learning, attention networks

1. INTRODUCTION

In AI for education, Knowledge Tracing (KT) is a key task, where the performance of a learner (responses to future questions) is predicted, given the past performance. Earlier works in KT can be divided into two groups, (a) Bayesian knowledge tracing (BKT) [1, 2] and (b) Item response theory (IRT) models [3, 4, 5, 6, 7]. These methods have excellent interpretability as they are able to measure the knowledge level of each learner on the various concepts [8]. Recent KT methods focused on using deep learning approaches [9, 10, 11, 8] to best capture the information found in large-scale learner response datasets. The Deep Knowledge Tracing (DKT) [9] and Dynamic Key-value Memory Networks (DKVMN) [10] are deep learning models that excel in predicting future performance, but lack interpretability. More recently, the use of attention networks in knowledge tracing has been explored through self-attentive knowledge tracing (SAKT) [11] and context-aware attentive knowledge tracing (AKT) [8]. The use of attention networks helps to provide some form of interpretability that was lacking in the previous approaches. Between SAKT and AKT, the newer AKT has a principled design of the algorithm that allows for good interpretability, alongside with better future prediction performance.

All the proposed algorithms have been designed for a traditional deep learning scenario, where the model is trained on all data that is available during training. However, learner population is increasing, which causes diversity in learner performances. Further, there is diversity of questions from similar concepts. These diversities cause severe drift in the data characteristics. Therefore, there is a need to develop models that are able to learn incrementally with drifting data, in order to generalize well to the learner population. In this paper, we first introduce the evolving knowledge tracing scenario (eKT), which is defined by multiple tasks with varying distributions under the pretext that data for the task are not available after training [12, 13, 14, 15]. Next, we characterise the drift in the data sets through training AKT models on independent distributions. We propose an Incremental Context Aware Attentive Knowledge Tracing (iAKT) to address the needs of the eKT, through regularization strategies. It is observed that all the regularization strategies offer better performance in learning incrementally. Finally, we compare the ability of iAKT with different regularization strategies to rank the difficulty of questions using Rasch embeddings [16], and observe that the Elastic Weight Consolidation (EWC) [17] is better than other regularization strategies. These observations imply that the iAKT with EWC provides both good learner performance prediction and ranks the difficulty of questions accurately.

2. RELATED WORKS

Deep Knowledge Tracing (DKT), using long short-term memory (LSTM) networks, is able to model complex, non-linear functions for predicting learning performance better than Bayesian knowledge tracing and other approaches. The DKVMN is an extension of the DKT by using an external memory matrix to characterise learner knowledge [10]. More recently, attention networks are used for knowledge tracing in SAKT and AKT. Attention mechanisms are more flexible

This work was supported by the Institute for Infocomm Research.

Data	No. of Users	No. of unique Questions	No. of unique Concepts
Task 1 (04/05)	689	1583	91
Task 2 (05/06)	1020	3189	93
Both	1709	3799	102

Table 1. Characteristics of the ASSISTment 2017 dataset in two disjoint tasks, suggesting drifts in their distributions.

than recurrent and memory-based neural networks and have shown better performance in Natural Language Processing (NLP) tasks [8]. Furthermore, attention networks are able to provide feature importance for a given prediction problem, and are desirable in Knowledge Tracing applications [11, 8]. AKT provides *context-aware representations* to model the learner's practise history and a novel *monotonic attention mechanism* to characterise time distance between questions that a learner has responded in the past. In addition, AKT uses a series of *Rasch model-based embeddings* to capture individual differences among questions in the same concept [8].

3. EVOLVING KNOWLEDGE TRACING

In this section, we first introduce the knowledge tracing problem setup. At each discrete time step t, the learner i would answer question index $q_t^i \in \mathbb{N}^+$ that covers concept index $c_t^i \in \mathbb{N}^+$. Their graded response is recorded as $r_t^i \in \{0, 1\}$, where 0 indicates an incorrect response and 1 a correct response. Hence, the performance record of learner i would consist of a tuple (q_t^i, c_t^i, r_t^i). For each learner i, given their past history $\{(q_1, c_1, r_1), \ldots, (q_{t-1}, c_{t-1}, r_{t-1})\}$, the objective is to predict their response r_t to question q_t on concept c_t at the current time step t. This setup is similar to the problem definition in [8]. We introduce a incremental knowledge tracing scenario using the ASSISTments 2017 data set [18], where the data from the past is not available all at once. To this end, we divide the ASSISTments data set into two tasks, according to the academic year (2004/2005 and 2005/2006). It must be noted that there is no overlap of student population between these years.

3.1. Dataset: ASSISTments 2017

The dataset is divided by the school year which the ASSISTment dataset was being used: 2004/2005 and 2005/2006.

Table 1 presents the characteristics of the two disjoint tasks in the ASSISTment 2017 dataset. However, the two tasks shared some of the questions and most of the concepts. It is to note that the ASSISTments data has questions with the same *problemId* (question) with different *skill* (concept) assigned to them. This could be because in a different year a new problem could be tagged with the same *problemId* and this can be identified through the *assistmentId* feature. Therefore, we have given these questions another *problemId* to identify them as unique questions. Thus, we have a total of 3799 unique questions instead of 3162 reported in [8]. As the learners in Task 1 and Task 2 are different, there is a diversity in the performances of learners in the two tasks. Moreover, as shown in Table 1, there are new questions attempted in Task 2, as compared to Task 1, thus causing diversity of questions from similar concepts in different tasks. These diversities cause severe drift in the data characteristics.

The objective of an incremental context-aware knowledge tracing model is to learn the representation for the data in 2004/2005 (Task 1) and then continually adapt the model to learn the representations from the data in 2005/2006 (Task 2), without forgetting knowledge in the Task 1 and not having access to data of Task 1.

3.2. Algorithm: Incremental Context-Aware Attentive Knowledge Tracing

The incremental Context-aware attentive knowledge tracing model is based on the Context-Aware Attentive Knowledge Tracing (AKT) [8], which is a methodological approach for knowledge tracing, consisting of four components:

- Question self-attentive encoder: models contextualised representation of each question, given the sequence of questions the learner has previously practised on
- Knowledge acquisition self-attentive encoder: models contextualised representations of the knowledge the learner acquired while responding to past questions
- Single attention-based knowledge retriever: retrieves knowledge acquired in the past that is relevant to the current question using an attention mechanism
- Feed-forward response prediction model: predicts the learner's response to the current question using the retrieved knowledge

As the AKT allows us to use the *Rasch model-based embeddings* to capture relationships between questions in a given concept, which enables ranking questions according to their difficulty, we choose the AKT to develop our iAKT. For complete details on AKT model, one should refer to [8].

In order to train the AKT incrementally, we regularize representations across tasks, drawing inspiration from continual learning strategies [17]. Regularisation strategies in incremental learning are aimed at learning new distributions without catastrophically forgetting the past tasks. To this end, these strategies introduce an additional loss to prevent the weights in the neural network from adapting too quickly to the new task. In this paper, the AKT model is regularized through the L2 penalty (Eq. (1)) and elastic weight consolidation (EWC) penalty (Eq. (2)) [17].

$$\mathcal{L}(\theta) = \mathcal{L}_s(\theta) + \sum_j \lambda(\theta_j - \theta^*_{s-1,j})^2 \qquad (1)$$

Eq. (1) shows the loss function of iAKT, with L2 regularization. It must be noted the loss function has two terms. The first term $\mathcal{L}_s(\theta)$ refers to the cross entropy loss to ensure efficient representation of the current task s and the second term sums the square of difference between the current parameters θ_j and the optimal parameters $\theta^*_{s-1,j}$ found at the end of the previous task $s-1$. λ is a hyperparameter that controls the impact of the regularisation, depending on the importance of the older tasks compared to the current one. The loss function of iAKT with EWC regularisation is 2.

$$\mathcal{L}(\theta) = \mathcal{L}_s(\theta) + \sum_j \frac{\lambda}{2} F_j(\theta_j - \theta^*_{s-1,j})^2 \qquad (2)$$

Similar to the L2 regularisation, the first term $\mathcal{L}_s(\theta)$ refers to the cross entropy loss in the current task s. In the second term, F represents the Fisher information matrix of task $s-1$. F_j helps to determine the important parameters j for each task, so that the weights can be regularized such that more significant weights are less adapted and vice-versa. The number of EWC penalties depend on the number of tasks. More information about the EWC regularisation can be found in [17].

The iAKT algorithm is summarized in Algorithm 1. Thus,

Algorithm 1: iAKT

Data: $\mathcal{D}_s$; $s = 1, \ldots, l$
for $s = 1, \ldots, l$ **do**
  Training Phase
  Input : (q^i_t, c^i_t, r^i_t) for $i \in \mathcal{D}^{train}_s$
  if *s=1* **then**
    Initialise and train AKT model to estimate $\boldsymbol{\theta}^*_1$ using cross entropy loss.
  else
    Update the AKT model ($\boldsymbol{\theta}^*_{s-1}$) with cross entropy and regularization using loss function in Eq 1 or Eq 2.
  end
  Testing Phase
  for *any sample* $\in$ *task* $\{1, \ldots, s\}$ **do**
    Input : (q^i_t, c^i_t, r^i_t) for $i \in \mathcal{D}^{test}_p$
    Output: $\hat{r}^i_{t+1}$
    Make predictions $\hat{r}^i_{t+1}$ with current model with parameters $\boldsymbol{\theta}^*_s$.
  end
end

it can be seen that the proposed iAKT is capable of task-agnostic incremental learning of a sequence of tasks.

4. EXPERIMENTS

We first characterize the drift in the data by training the model on data set for independent task, and predicting the performance on all tasks. We then compare the performance of iAKT with AKT model without the incremental learning strategies. To this end, we train the AKT model with all the available data in a Joint AKT (J-AKT) and with data for individual tasks in a Disjoint AKT (D-AKT). It must be noted that the task is defined according to the academic year (Section 3.1), where data from 2004/2005 and 2005/2006 are labelled as Task 1 and Task 2, respectively. We randomly divide 60%,20%,20% number of users in each task for training, validation and testing. At the end of the incremental training, we evaluate the accuracy of the model on the test data sets of both the Task 1 and Task 2, under 5 different random seeds. We use the accuracy and the area under the receiver operating characteristics curve (AUC) at the end of training the model with each task, as the evaluation metrics. AUC is typically used in the evaluation of knowledge tracing methods on predicting the future performance of the learner.

Scenario	Train Set	Test Set	Test AUC	Test ACC
J-AKT	Both	Task 1	76.5 ± 0.12	72.3 ± 0.18
		Task 2	75.0 ± 0.26	69.4 ± 0.22
D-AKT	Task 1	Task 1	75.9 ± 0.22	72.0 ± 0.25
		Task 2	62.4 ± 0.19	62.4 ± 0.45
D-AKT	Task 2	Task 1	65.8 ± 0.31	67.0 ± 0.59
		Task 2	74.0 ± 0.21	68.8 ± 0.27
Incremental Learning				
iAKT-EWC, $\lambda = 100$	Task 1	Task 1	75.9 ± 0.22	72.0 ± 0.25
	Task 2	Task 1	72.9 ± 0.44	69.4 ± 0.41
		Task 2	74.6 ± 0.57	69.1 ± 0.40

Table 2. Performance of AKT in joint (J-AKT) and disjoint (D-AKT) scenarios, iAKT-EWC in incremental learning scenario. The performance of D-AKT enumerates the drift in the data. D-AKT trained on Task 1 has lower accuracies for Task 2 and vice-versa.

4.1. Empirical Drift Characterization and Performance Study

Table 2 provides the performance of AKT in the joint, disjoint scenarios and iAKT-EWC in task-agnostic incremental learning scenario. The results of J-AKT, where all data is available, is considered the upper bound. It can be noted that when the model is trained only on Task 1 data in the D-AKT, it is not able to predict the performance of the learners in Task 2 and vice-versa, efficiently. There is at least a 12% drop in accuracy on D-AKT, in comparison to the J-AKT. This observation shows that there is considerable drift in the data distribution, across the population between Task 1 and Task 2. On the other hand, it can be observed that the performance of iAKT-EWC on both the tasks is better than the D-AKT, which indicates that the iAKT-EWC has learnt the drifted distribution of data. The performance of iAKT-EWC is almost similar to that of J-AKT. However, after incrementally training

Algorithm	Test Set	Test AUC	Test ACC
No strategy	Task 1	72.3 ± 0.63	68.9 ± 0.51
	Task 2	74.6 ± 0.52	68.8 ± 0.74
L2, $\lambda = 0.01$	Task 1	74.0 ± 0.35	70.5 ± 0.28
	Task 2	73.9 ± 0.83	68.7 ± 0.56
EWC, $\lambda = 100$	Task 1	72.9 ± 0.44	69.4 ± 0.41
	Task 2	74.6 ± 0.57	69.1 ± 0.40

Table 3. Ablation study: Analysis of different regularisation strategies in iAKT. Performance of iAKT is similar regardless of regularization strategy.

on Task 2, its performance on Task 1 drops by $\approx 3\%$, due to forgetting.

4.2. Comparison of incremental learning methods

Table 3 presents results of iAKT with (1) no regularisation strategy (iAKT-NR), (2) L2 regularization (iAKT-L2) and (3) EWC regularization (iAKT-EWC) for the incremental knowledge tracing scenario. It can be observed that the prediction performance of iAKT is similar, regardless of the regularization strategy. Hence, all the regularization strategies are suitable for learning incrementally towards learner performance prediction.

Next, we rank the difficulty of questions in Task 1 and Task 2, using Rasch embeddings of the iAKT. It must be noted that the AKT model is capable of ranking the questions accurately [8]. Hence, we compare the rankings of the iAKT with the various regularization strategies, with those of the J-AKT, as J-AKT has access to the entire dataset.

$$\delta_c = \sum_q ||\gamma_{q,c} - \gamma_{q,c}^{joint}|| \tag{3}$$

where δ_c represents the sum of differences in rankings for concept c in the iAKT, in comparison with J-AKT. The $\gamma_{q,c}$ and $\gamma_{q,c}^{joint}$ are the rank of question q of concept c in the iAKT and J-AKT, respectively. The smaller the value of δ_c, the more similar are the rankings of iAKT and J-AKT.

We choose two concepts, namely, '*addition*' and '*area*' for this study. These concepts are chosen as they are starkingly different in the two tasks. Specifically, the number of attempts on questions in concept '*addition*' is higher in the Task 1 (10666) compared to the Task 2 (1579). On the other hand, the number of attempts to questions in concept '*area*' in Task 1 is lower (14085) than Task 2 (20223). This indicates that the concept '*addition*' has more samples in Task 1 than in the Task 2, and vice-versa for the concept '*area*'.

Table 4 presents the δ_c for these concepts under various incremental learning scenarios, (a) without any regularisation strategy, (b) with L2 regularisation and (c) with EWC regularisation with $\lambda = 100, 1000$. From the table, it can be observed that all the iAKT-L2 and iAKT-EWC are better in ranking difficulty of questions in concept '*addition*'. As the L2 regularizer favours preserving representations for past task, its ability to rank difficulty of questions in '*addition*', where there are more samples in Task 1, is better than that of '*area*'. On the other hand, as EWC is aimed at incremental learning of sequence of tasks while preserving representations of the past tasks through relative adaptation of weights, its ability to rank difficulty of questions in both the concepts ('*addition*' and '*area*') is better, regardless of the uneven distribution of samples across task.

Concept, c	No strategy	L2 $\lambda = 0.01$	EWC $\lambda = 100$	EWC $\lambda = 1000$
Addition Task 1: 10666 Task 2: 1579	130	84	70	68
Area Task 1: 14085 Task 2: 20223	378	424	332	368

Table 4. Ablation Study: Analysis of question difficulty ranking of regularization strategies for iAKT. Smaller δ_c indicates that the ranking order is similar to J-AKT.

Finally, we study the effect of the regulaization constant λ, which is used to control adaptation of weights, in the EWC. We train the iAKT-EWC with two different values for λ, $\lambda = 100$ and $\lambda = 1000$. With higher lambda, the δ_c of iAKT-EWC is lower for the concept '*addition*' with larger sample set in Task 1. This shows that a higher values of λ emphasizes on remembering the previous task. In comparison, iAKT-EWC with a lower values of λ has lower δ_c for the concept '*Area*', which has more samples in Task 2 than Task 1. In general, it must be noted that the λ must be chosen depending on (a) the task that the user is interested most in and (b) the number of samples in each task.

5. CONCLUSION

This paper introduces the incremental knowledge tracing scenario and proposes a task-agnostic incremental context aware attentive knowledge tracing algorithm, with different regularization strategies, for learning a sequence of knowledge tracing tasks. The drift in knowledge tracing tasks are characterized, and the effectiveness of the iAKT in learning the sequence of knowledge tracing task incrementally, is demonstrated using the ASSISTment 2017 dataset. We study the effect of individual regularization strategies that enable incremental learning of AKT, and enumerate their ability to rank difficulty of questions for different concepts. Performance studies and the question ranking difficulty studies show that the iAKT-EWC is a better candidate for incremental knowledge tracing problems. Future studies can involve learning incrementally from different learning centers, and exploring other strategies of incremental learning.

6. REFERENCES

[1] Zachary A. Pardos and Neil T. Heffernan, "Modeling individualization in a bayesian networks implementation of knowledge tracing," in *User Modeling, Adaptation, and Personalization*, Paul De Bra, Alfred Kobsa, and David Chin, Eds., Berlin, Heidelberg, 2010, pp. 255–266, Springer Berlin Heidelberg.

[2] Michael V. Yudelson, Kenneth R. Koedinger, and Geoffrey J. Gordon, "Individualized bayesian knowledge tracing models," in *Artificial Intelligence in Education*, H. Chad Lane, Kalina Yacef, Jack Mostow, and Philip Pavlik, Eds., Berlin, Heidelberg, 2013, pp. 171–180, Springer Berlin Heidelberg.

[3] Frederic M Lord, *Applications of item response theory to practical testing problems*, Routledge, 2012.

[4] Andrew S. Lan, Christoph Studer, and Richard G. Baraniuk, "Time-varying learning and content analytics via sparse factor analysis," in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, 2014, KDD '14, p. 452–461, Association for Computing Machinery.

[5] Hao Cen, Kenneth Koedinger, and Brian Junker, "Learning factors analysis – a general method for cognitive model evaluation and improvement," in *Intelligent Tutoring Systems*, Mitsuru Ikeda, Kevin D. Ashley, and Tak-Wai Chan, Eds., Berlin, Heidelberg, 2006, pp. 164–175, Springer Berlin Heidelberg.

[6] Benoît Choffin, Fabrice Popineau, Yolaine Bourda, and Jill-Jênn Vie, "DAS3H: modeling student learning and forgetting for optimally scheduling distributed practice of skills," *CoRR*, vol. abs/1905.06873, 2019.

[7] Robert V. Lindsey, Jeffery D. Shroyer, Harold Pashler, and Michael C. Mozer, "Improving students' long-term knowledge retention through personalized review," *Psychological Science*, vol. 25, no. 3, pp. 639–647, 2014, PMID: 24444515.

[8] Aritra Ghosh, Neil Heffernan, and Andrew S. Lan, "Context-aware attentive knowledge tracing," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery amp; Data Mining*, New York, NY, USA, 2020, KDD '20, p. 2330–2339, Association for Computing Machinery.

[9] Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas J Guibas, and Jascha Sohl-Dickstein, "Deep knowledge tracing," in *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, Eds. 2015, vol. 28, Curran Associates, Inc.

[10] Jiani Zhang, Xingjian Shi, Irwin King, and Dit-Yan Yeung, "Dynamic key-value memory network for knowledge tracing," *CoRR*, vol. abs/1611.08108, 2016.

[11] Shalini Pandey and George Karypis, "A self-attentive model for knowledge tracing," *CoRR*, vol. abs/1907.06837, 2019.

[12] Andrea Cossu, Antonio Carta, Vincenzo Lomonaco, and Davide Bacciu, "Continual learning for recurrent neural networks: an empirical evaluation," *CoRR*, vol. abs/2103.07492, 2021.

[13] Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Ales Leonardis, Gregory G. Slabaugh, and Tinne Tuytelaars, "Continual learning: A comparative study on how to defy forgetting in classification tasks," *CoRR*, vol. abs/1909.08383, 2019.

[14] Kibok Lee, Kimin Lee, Jinwoo Shin, and Honglak Lee, "Overcoming catastrophic forgetting with unlabeled data in the wild," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 312–321.

[15] Gido M van de Ven, Hava T Siegelmann, and Andreas S Tolias, "Brain-inspired replay for continual learning with artificial neural networks," *Nature communications*, vol. 11, no. 1, pp. 1–14, 2020.

[16] G. Rasch, *Probabilistic Models for Some Intelligence and Attainment Tests*, Mesa Press, 1993.

[17] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell, "Overcoming catastrophic forgetting in neural networks," *Proceedings of the National Academy of Sciences*, vol. 114, no. 13, pp. 3521–3526, 2017.

[18] Thanaporn Patikorn, Douglas Selent, N. Heffernan, Biao Yin, and Anthony F. Botelho, "Assistments dataset for a data mining competition to improve personalized learning," 2017.