電気通信大学情報理工学域
I 類 (情報系) 情報数理工学プログラム卒業論文

Attention を用いた Knowledge Tracing モデルの忘却最適化

2021 年 1 月 28 日

情報数理工学プログラム

学籍番号 1610374

関口 昌平

指導教員 植野 真臣

令和２年度 情報数理工学プログラム卒業論文概要

平成 28 年度 入学	学籍番号 1610374
指導教員 植野 真臣	氏名 関口 昌平
題目	Attention を用いた Knowledge Tracing モデルの忘却最適化

概要

人工知能分野では教育ビックデータを用いて学習過程における学習者の能力値や知識状態を把握し，課題への反応予測を行う Knowledge Tracing(KT) が注目されている．最先端の KT 手法では Transfomer を用いた AKT が提案されている．AKT の特徴は過去の学習データを徐々に忘却し，さらに直近の学習に大きく関係するスキルを考慮して予測を行う．これにより，高い予測精度を示すことが報告されている．しかし，AKT では単調減少関数に従って徐々に過去の学習データを忘却するため，長時間の学習においては初期の学習データがノイズとして残ってしまう問題点があった．本研究では AKT において予測精度を最大化するように反応予測に用いるデータ数を最適化する新たな手法を提案する．評価実験では提案手法と既存手法を用いて反応予測精度比較を行い，提案手法の有効性を示す．

1　まえがき

近年，オンライン教育の普及に従い，大量の学習履歴データが容易に入手できるようになった．教育現場では学習者の発達を促すため，これらのデータに基づいて学習者ごとに項目解決に必要なスキル (例えば算数の学習であれば，加算・減算・乗算・除算など) の習得状態を把握し適切な支援を与えることが項目となっている．人工知能分野では機械学習手法を用いて過去の学習履歴から学習者の能力値を推定し，学習者の未知の項目への反応予測を行う Knowledge Tracing(KT) が注目を集めている [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]．項目への反応を予測することにより，教師が学習者の得意分野・苦手分野を把握し，個人に適した項目提供や支援を行うことが可能となる．

KT 手法には確率モデルを用いた確率的アプローチと深層学習モデルを用いたディープラーニングアプローチがある．確率的アプローチの代表的な手法には Bayesian Knowledge Tracing（BKT）[1, 11, 12, 13, 14] や Item response Theory（IRT）があり，学習履歴データから各スキルにおける学習者の潜在的な能力値を推定し，未知の項目への正答確率を予測することができる [15]．確率的アプローチは学習者の能力値の他に，項目の難易度を表すパラメータをもつため，パラメータ解釈性が高く，多くの学習支援システムで用いられている．しかし，確率的アプローチで推定される能力値はスキルごとに独立であり，多次元のスキルの関係性を考慮した能力推定ができない．また，IRT 手法は学習者の項目への反応は独立であることが仮定されているため，同じ項目に繰り返し取り組む学習では用いることができない．

また，予測精度向上のために多次元スキルにおける学習者の能力変化を考慮したディープラーニングアプローチが開発されている．ディープラーニングアプローチでは代表的な手法として Deep Knowledge Tracing (DKT)[2] が提案されている．DKT は Long-short term memory (LSTM)[16] を用いて学習者の能力変化を表現し，学習者の各項目への反応を予測するモデル である．DKT では LSTM の隠れ層に全てのスキル の能力値が圧縮されているとみなしており，BKT 手法と比較して反応予測精度が高いことが報告されている．さらに DKT の予測精度を向上させるために，各スキルの能力値を保存する Memory Network を用いた Dynamic Key-Value Memory Network (DKVMN) が提案されている [17]．DKVMN は高い反応予測精度を示すが，DKT と同様に学習者の能力

パラメータをもたず，モデルの解釈可能性が低いという問題があった．そこで，DKT や DKVMN のパラメータ解釈可能性を向上させた手法として，DKVMN と IRT を組み合わせた Deep-IRT[18] が提案されている．Deep-IRT は DKVMN の高い予測精度を保ちながら，学習者の能力パラメータや項目の難易度パラメータなどパラメータの解釈可能性もち，注目を集めている．

さらに近年，新たなディープラーニングアプローチとして従来の KT 手法のように Recurrent Neural Network(RNN) を用いず，Transformer と呼ばれる Attention のみを用いるモデル SAKT が開発されている [19]．Pandey らは従来のディープラーニングアプローチにはパラメータ推定に膨大な時間がかかることとスパースデータに対して脆弱である問題を指摘し，反応予測に Transformer の導入を提案している．Transformer は自然言語処理の分野でよく用いられる手法であり，長期間で強い依存関係をもつ言語データの予測に対して有効であることが知られている [20]．Pandy らの SAKT において現在の学習者の反応には過去の反応データが大きく関係していることに注目しており，学習者の過去の全ての学習データを用いて反応予測を行う．これに対し，Ghosh らは学習者の現在の反応は全ての過去の学習データに依存するのではなく，直近の短い期間の学習に依存することを主張している [21]．そこで，彼らは SAKT に過去の学習データを徐々に忘却し，さらに直近の学習に大きく関係するスキルをより考慮するように Attention を計算する新たな Attention モデル，Attentive Knowledge Tracing(AKT) を提案している．この手法により，従来のディープラーニングアプローチと比較して学習者の反応予測精度が向上することが示された．

しかし AKT では単調減少関数に従って徐々に過去の学習データを忘却するため，長時間の学習においては初期の学習データがノイズとして残ってしまう問題点があった．

この問題を解決するために，本研究では AKT の反応予測に用いるデータ数を制限することでノイズによる予測精度低下を防ぐ新たな AKT 手法を提案する．指定した長さのデータ数を用いて予測を行い，それ以外のデータは完全に忘却される最適なデータ数を推定することにより反応予測精度の向上が期待できる．

本研究では，実際にオンライン学習システムで収集された学習履歴データを用いて，従来の KT 手法である DKT，DKVMN，SAKT，AKT と提案手法を用いて，学習者の反応予測精度比較を行う．学習期間，学習者数，項目数の異なる様々な学習データを用いて実験を行い，提案手法の有効性を示す．

2　学習者の能力推定モデル

本章では既存の KT 手法を紹介する．

2.1　RNN を用いた手法

2.1.1　Deep Knowledge Tracing

DKT は時系列の深層学習モデルである LSTM(Long term short memory)[16] を用いて過去の学習データから未知の項目への反応を予測するモデルである．DKT ではスキル間の独立性を仮定せず，LSTM の隠れ層に学習者のスキルの能力値を多次元かつ連続量で格納できる．しかし，DKT は学習者のすべてのスキルを単一の隠れ変数ベクトルで表現しているため，学習者が各スキルに関する知識をどの程度習得したかを表現できない問題があった．

2.1.2　Dynamic Key-Value Memory Network

DKT の反応予想精度を向上させるため各スキルの能力値を保存する Memory Network を用いた DKVMN が考案されている [17]．DKVMN では学習過程に N 個の潜在スキルに対応する能力があると仮定し，各項目と潜在スキルの関係を推定しながら反応予測を行う．DKVMN は高い反応予測精度を示すことが報告されているが，DKT と同様に学習者の能力値を表現するパラメータを持たないため，解釈性が低いという問題があった．

2.1.3　Deep-IRT

DKT，DKVMN 手法においてパラメータの解釈性が低いという問題を解決するために DKVMN と IRT を組み合わせた Deep-IRT が開発されている．[18]．学習者の能力パラメータを項目の難易度パラメータをもち，高い解釈性と反応予測精度を示すことが報告されている．Deep-IRT では，学習者の項目への予測正答確率 p_t を学習者の能力パラメータ θ と項目の難易度パラメータ β を用いて以下の式から求める。

$$p_t = \sigma(3.\ 0 \times \theta - \beta) \tag{1}$$

2.2 Attention を用いた手法

2.2.1 A Self-Attentive model for Knowledge Tracing

一方, 新たな KT 手法として従来の KT 手法のように RNN を用いず, Transformer と呼ばれる Attention のみを用いる新たな KT 手法が開発されている [19, 21]．Transformer は自然言語処理の分野でよく用いられる手法であり，長期間で強い依存関係をもつ言語データの予測に対して有効であることが知られている [20]．Pandey らは学習過程においても現在の学習者の反応には過去の反応データが大きく関係していることに注目し，Transformer 用いて学習者の過去の全ての学習データにおける依存関係を推定し，反応予測を行う SAKT 手法を提案した [19].

彼らは RNN を用いた従来のディープラーニングアプローチにはパラメータ推定に膨大な時間がかかることとスパースデータに対して脆弱である可能性を指摘し，SAKT を用いることで改善することを示した.

また Attention を用いることで項目間の関連度を求めることができ，項目をスキル毎にクラスタリングすることが可能である．また，Transformer では逐次的に計算を行う RNN よりも並列計算に適しているため，SAKT では従来手法に比べ学習にかかる計算時間が少ないことが示されている.

2.3 Attentive Knowledge Tracing

さらに，Ghosh らは学習過程における学習者の反応は，全ての過去の学習データに依存するのではなく，直近の短い期間の学習に依存すると仮定し，SAKT において過去の学習データを忘却する新たな手法 Attentive Knowledge Tracing(AKT) を提案している [21]．AKT は過去の学習データを徐々に忘却すると同時に，直近の学習に大きく関係するスキルを重視するように Attention を計算する．この手法により，従来のディープラーニングアプローチと比較して反法予測精度が向上することが示された．本章では AKT 手法について詳しく説明する.

AKT では学習者が取り組んだ項目 q_t と項目への反応 r_t を用いて，項目固有の特徴量ベクトル x_t と学習者の潜在的な能力ベクトル y_t を表現する．$t-1$ 時点までの学習データから $\{x_1, ..., x_t\}$ と $\{y_1, ..., y_{t-1}\}$ を計算し，時点 t で学習者がある項目に正答する確

率を予測する．

AKT モデルは大きく 2 つのステップに分かれている．まず項目の特徴量ベクトル $\{x_1, ..., x_t\}$ について，各項目とスキルの関係性を表すアテンション α を計算し，新たな項目の特徴量ベクトル $\{\hat{x_1}, ..., \hat{x_t}\}$ を求める．同様に $\{y_1, ..., y_{t-1}\}$ とアテンション α からスキルの関係性を考慮した新たな能力ベクトル $\{\hat{y_1}, ..., \hat{y_{t-1}}\}$ を求める．具体的には，時点 t の入力 x_t から時点 τ の入力 x_τ までの学習データにおけるアテンション α は次の式で求める

$$\alpha_{t,\tau} = softmax(\frac{q_t^T k_\tau}{\sqrt{D_k}}) = \frac{exp(\frac{q_t^T k_\tau}{\sqrt{D_k}})}{\sum_{\tau'} exp(\frac{q_t^T k'_\tau}{\sqrt{D_k}})} \tag{2}$$

ここで，$q_t \in \mathbb{R}^{D_k \times 1}$，$k_t \in \mathbb{R}^{D_k \times 1}$ は x_t，x_τ から以下で求められる。

$$q_t = x_t W^Q, k_\tau = x_\tau W^K \tag{3}$$

W は重みを表す．これらを用いて以下のように $\hat{x}$ を計算する．$\hat{y}$ も同様に計算する．

$$\hat{x_\tau} = \sum_{t=1}^{\tau} \alpha_{t,\tau} v_\tau \tag{4}$$

$$v_\tau = x_\tau W^V \tag{5}$$

次に，$\hat{\boldsymbol{x}}$ と $\hat{\boldsymbol{y}}$ を用いて学習者が時点 t で項目に正答する確率を求める．AKT は学習者が取り組んだ項目数に応じて過去の学習データを徐々に忘却させるためのアテンションをもつ．このアテンションを Monotonic Attention と呼ぶ．Monotonic Attention は以下の式で表される

$$\alpha'_{t,\tau} = \frac{exp(s_{t,\tau})}{\sum_{\tau'} exp(s_{t,\tau'})} \tag{6}$$

$$s_{t,\tau} = \frac{exp(-\theta \cdot d(t,\tau)) \cdot q_t^T k_\tau}{\sqrt{D_k}} \tag{7}$$

ハイパーパラメータ θ は $\theta > 0, \theta \in \mathbb{R}$ であり，式 (4) は常に $exp(-\theta \cdot d(t,\tau)) \leq 1$ の値を取るため，$d(t,\tau)$ の値によって Monotonic Attention は変化する．

$\tau \leq t$ とすると $d(t,\tau)$ は

$$d(t,\tau) = |t-\tau| \cdot \sum_{t'=\tau+1}^{t} \gamma_{t,\ t'}, \tag{8}$$

$$\gamma_{t,\ t'} = \frac{exp(\frac{q_t^T k'_t}{\sqrt{D_k}})}{\sum_{1\leq\tau'\leq t} exp(\frac{q_t^T k_{\tau'}}{\sqrt{D_k}})} \tag{9}$$

で求められる．式 (8) は時点 τ から t までの学習のうち，特に重視するスキルについての重みを計算してる．Ghosh らは $d(t,\tau) = |t-\tau|$ として Monotonic Attention を求めたモデルでも精度比較を行なっているが，式 (7) を用いた場合の方が高い予測精度を示すことが報告されている．すなわち，AKT では項目への反応予測を行う際に，過去の学習項目からの経過時間による忘却と関連するスキルの重みを考慮することにより，予測精度を向上させることができる．

しかし，AKT における忘却方法では，学習過程が長くなるほど関連性の低いスキルの重みがノイズとして残ってしまう問題がある．図 1 はある学習者の 200 問目の回答を予測する際，過去の各項目との Monotonic Attention のスコアを図示したグラフである．グラフの縦軸は Monotonic Attention の値であり，横軸は学習者が取り組んだ項目番号を表す．図より，直近の 100〜200 項目では 200 問目との関連度が分散して推定されているが，それ以前の項目は一定値に収束している．この結果から，AKT はほとんど関連性のない過去のデータを用いて反応予測を行なっていることがわかる．これらの関連性の低い過去のデータがノイズとなり，反応予測精度を低下させている可能性がある．

3　提案モデル

前章では AKT は従来手法と比較して高い反応予測精度を示すものの，長時間の学習においては初期の学習データの関連性が低くなりノイズとして残る問題を指摘した．この問題を解決するために，本研究では AKT において反応予測に用いるデータ数を直近の数問に制限し，ノイズによる予測精度低下を防ぐ新たな AKT 手法を提案する．具体的には図 2 のように入力データ数 L を設定し，時点 $t-L$ から時点 t までの学習データを用いて時点 $t+1$ での反応を予測する．すなわち，時点 $t-L$ 以前の学習データは完全に忘却する．この仕組みによって AKT において情報量の少ないノイズデータを除去することが可

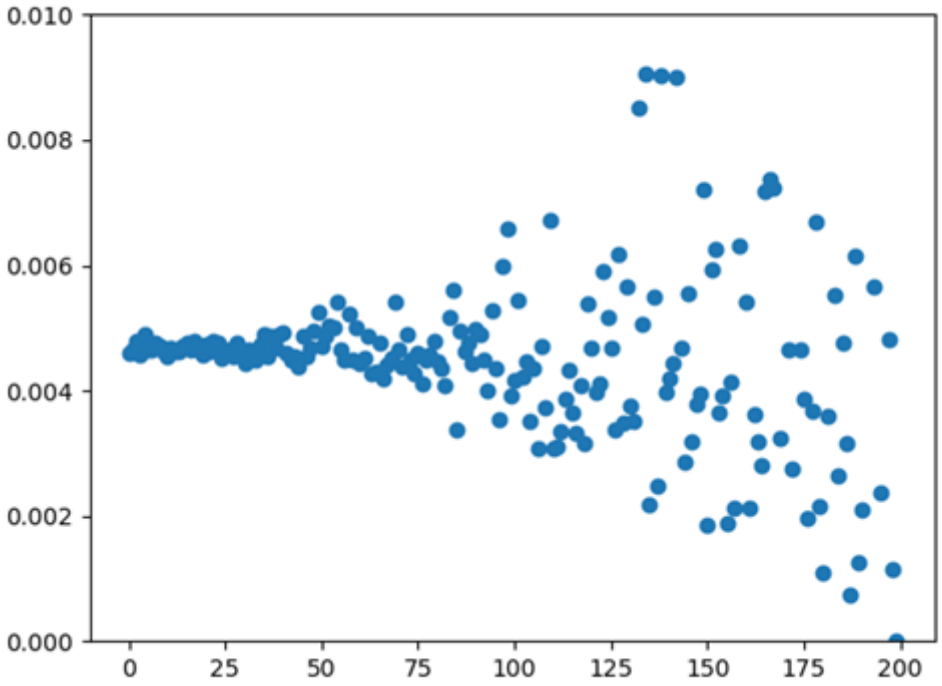

図 1　ある学習者における過去 200 問の各項目との関連度

能となる．

提案モデルは以下のように定式化される。

$$d(t,\tau) = |t-\tau| \cdot \sum_{t'=\tau+1}^{t} \gamma_{t,\ t'}, \tag{10}$$

$$\gamma_{t,\ t'} = \frac{exp(\frac{q_t^T k'_t}{\sqrt{D_k}})}{\sum_{t-L\leq\tau'\leq t} exp(\frac{q_t^T k_{\tau'}}{\sqrt{D_k}})} \tag{11}$$

L が大きい場合は広範囲の項目を扱い，L が小さい場合は小さな範囲の項目を扱うモデルとなる．提案モデルではこの L をハイパーパラメータとしてデータセットに応じて最適値を決定する．

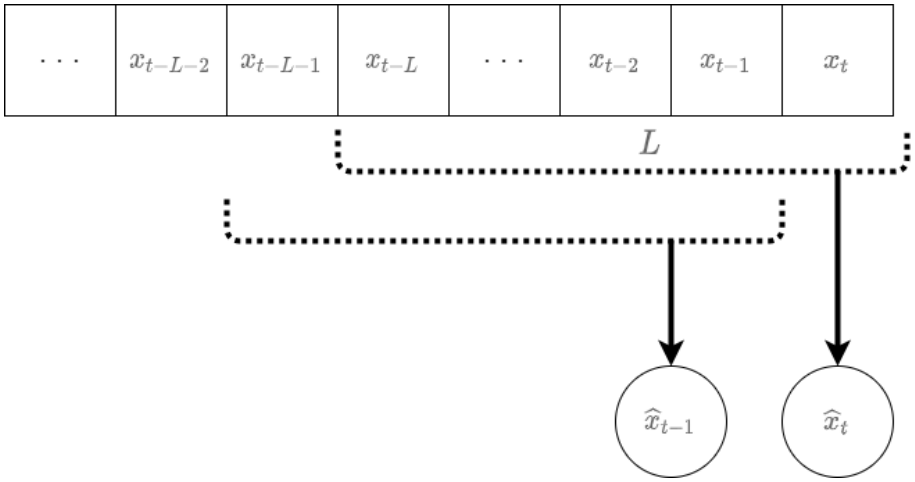

図 2　モデルの概要図

表 1　データセットの詳細

Dataset	学習者数	スキル数	項目数	平均解答項目数	平均正解率
Statics2011	333	1,223	NA	168.1	79.8%
ASSISTments2009	4,151	110	16,891	52.1	63.6%
ASSISTments2015	19,840	100	NA	33.9	73.2%
ASSISTments2017	1,709	102	3,162	187.7	39.0%

4　予想精度評価

4.1　データセット

本実験では一般に公開されている大規模なオンライン学習システムで収集された Statics2011，ASSISTments2009，ASSISTments2015，ASSISTments2017 を用いた．これらのデータの概要を表 1 に示す．データセットには各項目に項目番号とスキル番号が付与されているが，Stratics2011 と ASSIST2015 では項目番号がないため項目数は NA と表記した．また，本研究では入力する学習データの偏りを避けるために入力する学習データの上限を学習者 1 人につき 200 問とした [18].

表 2　学習者の反応予測精度

	DKT	DKVMN	Deep-IRT	SAKT	AKT	提案モデル
Statics2011	0.8233	0.8195	0.8086	0.8029	0.8265	**0.8300**
ASSISTments2009	0.8170	0.8093	0.8126	0.7520	0.8300	**0.8312**
ASSISTments2015	0.7310	0.7276	0.7246	0.7212	0.7296	**0.7643**
ASSISTments2017	0.7263	0.7124	0.7187	0.7073	0.7561	**0.7693**

4.2　評価実験

提案手法と既存の KT 手法を用いて学習者の反応予測精度の比較を行う．具体的にはデータセットの 60% をトレーニングデータ，20% をバリテーションデータ，20% をテストデータとして 5 分割交差検証を行なった．評価指標には一般に KT 手法の精度比較に用いられる AUC スコアを採用した．

予想精度の実験結果を表 2 に示す．提案モデル全てのデータセットにおいて他のモデルよりも高い予測精度を示した．特に，提案手法は学習者の平均解答数が少ない Assistments2009 と Assistments2015 においても AKT 手法を上回っており，忘却するデータ数を最適化した提案手法が有効であったことがわかる．

4.3　入力データ数と AUC の関係

図 3 に各学習データにおいて提案モデルの入力データ数 L 変更したときの予測精度を示す．縦軸は AUC，横軸は L を表す。

Statics2011 では L のサイズが小さくなるにつれて AUC の値が改善していることから残存するデータの影響が特に大きいと思われる．Monotonic Attention は現在の時点 t と学習者が過去の項目に取り組んだ時点 τ の差に応じて減衰するため Statics2011 のように項目の平均数が長いほどノイズの影響力が大きいと考えられる．提案モデルでは $L = 37$ で最も良い精度を示し，比較的少ない項目数でも十分に高い精度で予測できる事が分かった．

ASSISTments2015 では L の数が大きくなるほど精度が向上している．ASSISTments2015 の平均解答項目数は比較的小さいため，過去の項目の影響力がノイズにな

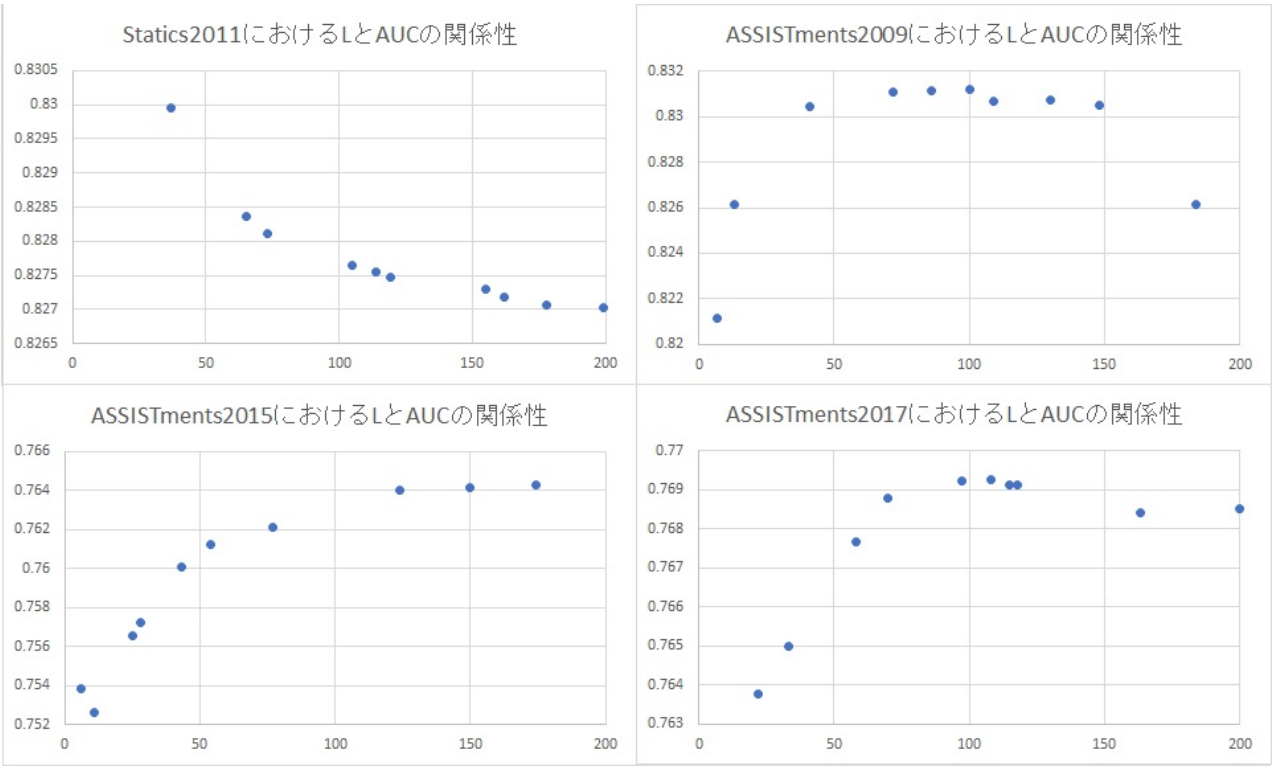

図 3　L と AUC の関係

りにくいと考えられる．しかしその場合でも $L = 174$ で AKT よりも良い予測精度を示した．

図 3 より，ASSISTments2009，ASSISTments2017 では AUC が単峰分布であり，最大の予測精度を示す最適な L が存在することを示している．ASSISTments2009 は L=100，ASSISTments2017 は L=97 で最も良い精度を示した．AKT では項目を 200 個用いていたがそれに比べると約半数になっており，不要なデータを忘却することで予測精度が向上することがわかった．

以上の例から AUC を最大にする L を探索し，適切にデータを忘却することが重要であることがわかった。

5　おわりに

本研究では AKT におけるデータ忘却を最適化するために，反応予測に用いるデータ数を制限することで情報量の少ないデータを完全に忘却する新たな AKT 手法を提案した．評価実験ではすべてのデータセットにおいて学習者の反応予測精度が向上した．

ただし，提案モデルは入力長 L の最適値を探索するために他モデルに比べ学習時間が非常に長くなっている．また Attention を用いた KT 手法は計算量が多いため RNN を用いた既存手法に比べ多くの計算時間が必要である．提案モデルはこういった欠点に対し改善の余地があると考えている．AKT ではすべての項目間に対し Attention を計算している．しかし入力データ数 L を設定すれば各項目は L 個の項目に対してアテンションを計算すれば良いため計算量が減らせると考える．モデルに L の値に応じてアテンションの計算量を減らす変更を行うことで AKT の問題の 1 つであった計算時間を減らす事が可能であると考える．

また最適な L を探索するアルゴリズムであるが，現在のアルゴリズムでは必ずしも最良の値を見つけることができないという欠点がある．Attention を用いたモデルは学習時間が多くかかってしまうため闇雲に様々な L の値を試すと学習時間が比例して長くなってしまう．よって他の効率的なアルゴリズムを導入することで学習時間を減らしたり，最良の L の値を見つけられるようなモデルになると考える．

KT では予測精度だけではなく解釈性の高さも重要であり，従来の KT 手法よりも解釈性を高めたモデルである Deep-IRT[17] が考案されている．提案モデルでは学習者の項目への正答確率を予測しているため，学習者の能力値が得られず解釈性が低いという問題がある．Deep-IRT では学習者の能力パラメータと項目の難易度パラメータを求めることで高い解釈性を実現している．提案モデルにも同様の手法を取り入れることで高い解釈性を得ることが可能だと考える．

参考文献

[1] Albert T. Corbett and John R. Anderson. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-adapted Interaction*,

Vol. 4, No. 4, pp. 253–278, 1995.

[2] Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas Guibas, and Jascha Sohl-Dickstein. Deep knowledge tracing. In *NIPS'15 Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, Vol. 28, pp. 505–513, 2015.

[3] Yuying Chen, Qi Liu, Zhenya Huang, Le Wu, Enhong Chen, Runze Wu, Yu Su, and Guoping Hu. Tracking knowledge proficiency of students with educational priors. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pp. 989–998, 2017.

[4] Sein Minn, Yi Yu, Michel C. Desmarais, Feida Zhu, and Jill-Jenn Vie. Deep knowledge tracing and dynamic student classification for knowledge tracing. In *2018 IEEE International Conference on Data Mining (ICDM)*, pp. 1182–1187, 2018.

[5] Yu Su, Qingwen Liu, Qi Liu, Zhenya Huang, Yu Yin, Enhong Chen, Chris H. Q. Ding, Si Wei, and Guoping Hu. Exercise-enhanced sequential modeling for student performance prediction. In *AAAI*, pp. 2435–2443, 2018.

[6] Ghodai Abdelrahman and Qing Wang. Knowledge tracing with sequential key-value memory networks. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 175–184, 2019.

[7] Jinseok Lee and Dit-Yan Yeung. Knowledge query network for knowledge tracing: How knowledge interacts with skills. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, pp. 491–500, 2019.

[8] Qi Liu, Zhenya Huang, Yu Yin, Enhong Chen, Hui Xiong, Yu Su, and Guoping Hu. Ekt: Exercise-aware knowledge tracing for student performance prediction. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 33, No. 1, pp. 100–115, 2021.

[9] Jill-Jenn Vie and Hisashi Kashima. Knowledge tracing machines: Factorization machines for knowledge tracing. *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, No. 1, pp. 750–757, 2019.

[10] Zhiwei Wang, Xiaoqin Feng, Jiliang Tang, Gale Yan Huang, and Zitao Liu. Deep knowledge tracing with side information. In *International Conference on Artificial Intelligence in Education*, pp. 303–308, 2019.

[11] Michael V. Yudelson, Kenneth R. Koedinger, and Geoffrey J. Gordon. Individualized bayesian knowledge tracing models. In *International Conference on Artificial Intelligence in Education*, Vol. 7926, pp. 171–180, 2013.

[12] Jose Gonzalez-Brenes, Yun Huang, and Peter Brusilovsky. Fast: Feature-aware student knowledge tracing. 2013.

[13] Tanja Kaser, Severin Klingler, Alexander G. Schwing, and Markus Gross. Dynamic bayesian networks for student modeling. *IEEE Transactions on Learning Technologies*, Vol. 10, No. 4, pp. 450–462, 2017.

[14] Radek Pelánek. Conceptual issues in mastery criteria: Differentiating uncertainty and degrees of knowledge. In *International Conference on Artificial Intelligence in Education*, pp. 450–461, 2018.

[15] Frank B. Baker and Seock-Ho Kim. *Item Response Theory: Parameter Estimation Techniques*. 2004.

[16] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, Vol. 9, No. 8, pp. 1735–1780, 1997.

[17] Jiani Zhang, Xingjian Shi, Irwin King, and Dit-Yan Yeung. Dynamic key-value memory networks for knowledge tracing. In *WWW '17 Proceedings of the 26th International Conference on World Wide Web*, pp. 765–774, 2017.

[18] Chun-Kit Yeung. Deep-irt: Make deep learning based knowledge tracing explainable using item response theory. In *EDM*, 2019.

[19] Shalini Pandey and George Karypis. A self-attentive model for knowledge tracing. In *12th International Conference on Educational Data Mining, EDM 2019*, pp. 384–389, 2019.

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Vol. 30, pp. 5998–6008, 2017.

[21] Aritra Ghosh, Neil T. Heffernan, and Andrew S. Lan. Context-aware attentive knowledge tracing. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2330–2339, 2020.