電気通信大学情報理工学部
情報・通信工学科 情報数理工学プログラム卒業論文

Attentive Knowledge Tracing における ロジスティック関数を用いた忘却最適化

2022 年 1 月 30 日

情報数理工学プログラム

学籍番号 1511052

岡崎 哲

指導教員 植野 真臣

令和３年度 情報数理工学コース卒業論文概要

平成 27 年度 入学	学籍番号 1511052
指導教員 植野 真臣	氏名 岡崎 哲
題目	Attentive Knowledge Tracing における ロジスティック関数を用いた忘却最適化

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 Transformer を用いた Attentive Knowledge Tracing(AKT) が提案されている. AKT は過去の学習データを忘却し, 直近の学習に大きく関係するスキルを考慮して学習者の反応を予測する. これにより AKT は高い予測精度を示すことが報告されている. しかし AKT の忘却手法では過去の学習データから学習される Attention が一定値に収束し, 学習期間が長くなるほどノイズとして残る問題があった. これに対し関口ら (2021) は実際には影響を無視できるほど古い学習データを完全に忘却するため, 指定した長さのデータ数のみ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ような完全忘却型 AKT を提案し, それによって予測精度が向上することを示した. しかし, 完全忘却型 AKT では入力データ数 L が離散値であるため, L の最適化に膨大な時間コストが必要となる問題がある. また, 完全忘却型 AKT では, 設定された経過時間でステップ関数によりデータを完全忘却するため, 時間経過による忘却度変化が表現されていない問題もある. そこで本研究ではロジスティック関数を用いて以下の問題点を解決する新たな忘却最適化手法を提案する. 提案手法は経過時間に応じて Attention を 0 に漸近させる個音で過去の学習データを忘却し, 学習データが長くてもノイズを残さない. また, 忘却関数としてのロジスティック関数は二つの微分可能なパラメータを持ち, 勾配法を用いて高速に最適な忘却関数を学習できる. さらに, 最適な忘却関数としてのロジスティック関数に応じて過去の学習データを忘却することで, 経過時間による忘却度の変化を連続的に放言することができる. 本研究では従来の AKT と提案手法を用いて, 学習者の反応予測精度比較を行い, 提案手法の有効性を検証する.

目次

1.はじめに

近年世界的にオンライン教育が普及し，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．学習支援システムや人工知能分野では，これらの学習履歴データに基づいて項目解決に必要なスキルの習得状況を学習者ごとに把握し，適切な支援を与えることが課題となっている．機械学習手法を用いて学習履歴データ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手法を Knowledge Tracing(KT)と呼ぶ[Corbett 95, Piech 15, Chen 17, Minn18, Su18, Abdelrahman 19, Lee 19, Liu 21, Vie 19, Wang19]．未知の項目への反応を予測することにより，各学習者に最適な項目提供や支援を行うことが可能となる．

KT の代表的な手法として，確率モデルを用いた Bayesian Knowledge Tracing (BKT)と Item Response Theory (IRT)が知られている[Yudelson 13, Gonzalez 13, Käser 17, Baker 04]. BKT は隠れマルコフモデルに基づいて，学習者が項目解決に必要なスキルを習得しているかを推定し，学習者の未知の項目への反応を予測する．しかし BKT ではスキルの習得状態が 2 値で表されるため，スキルの習熟度変化を柔軟に表現することができない．一方 IRT は学習者のスキルの習得状態を表す能力パラメータと項目の難易度パラメータを推定し，学習者の各項目に対する正答確率を予測する．IRT では能力パラメータが連続量で表されるため，BKT に比べて表現力が高い．しかし，項目の局所独立性を仮定しているため，同じ項目に繰り返し取り組むような学習に用いることができない．さらに各スキルの独立性を仮定しているため，スキル間の関係性を考慮することができないという問題があった．

また，近年では深層学習モデルを用いた手法も開発されている．ディープラーニングアプローチの代表的な手法として，Recurrent Neural Network (RNN)を用いた Deep Knowledge Tracing (DKT)が提案されている[Piech 15]. DKT は Long-Short Term Memory (LSTM)を用いて学習者の潜在的な能力変化を表現し，学習者の各項目への反応を予測するモデルである．DKT では学習者のスキルごとの能力値が LSTM の隠れ層に圧縮されているとみなしており，各スキルをどの程度習得したかを表現することはできなかった．さらに反応予測精度を向上させるために，スキルごと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[Zhang 17].

DKVMN は高い反応予測精度を示すものの, DKT と同様に学習者の能力パラメータを持たないため, 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 が提案されている[Yeung 19]. Deep-IRT は学習者の能力パラメータや項目の難易度パラメータの解釈可能性をもつが, 推定される能力が項目の特性に依存しており, 同一スキル内の全ての項目が等質であると仮定しているために, 異なる困難度を持つ項目からの能力推定値が解釈できなかった. そこで Tsutsumi らは学習者の項目への反応を二つの独立な学習者ネットワークと項目ネットワークで表現し, 項目の特性に依存せずに能力値を推定する新たな Deep-IRT モデルを提案している[Tsutsumi 21].

また, 近年新たなディープラーニングアプローチとして, Transformer と呼ばれる Attention モデルを KT に利用した SAKT が提案されている[Pandey 19]. Pandey らは RNN を用いた KT 手法はパラメータ推定に膨大な時間を必要とすること, またスパースデータに対して脆弱であるという問題を指摘した. Transformer は自然言語処理の分野で良く用いられており, 長期間で強い依存関係を持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[Vaswani 17]. SAKT では学習者の現在の反応が過去の反応データに大きく関係していることに注目し, 学習者の過去の学習データを全て用いて反応予測を行う. Ghosh らはこれに対し, 学習者の現在の反応は直近の短い期間の学習データに依存すると主張し, 新たに Attentive Knowledge Tracing(AKT)を提案した[Ghosh 20]. AKT では直近の学習に強く関係するスキルを考慮しながら過去の学習データを徐々に忘却する. この手法によって従来のディープラーニングアプローチに比べて学習者の反応予測精度が向上することが示された.

Attention 機構は自然言語処理の分野において, 単語間の依存関係を入力や出力の長さに依らずモデル化することを可能にした. つまり Attention 機構は文中の単語間の距離に依らず単語間の依存関係を計算するが, これを学習データに適用する場合, 学習過程における時間的な距離に関係なく学習データ間の依存関係を計算してしまう. これに対し Ghosh らは過去の学習データを忘却しつつ学習データ間の依存関係を計算するため, 単調減少関数を Attention 機構に組み込んだ. しかし, この手法では過去の学習データが完全には忘却されないため, 学習過程が長くなるほど過去の学習データがノイズとして残り, 反応予測精度を低下させている可能性があった.

そこで関口らは実際には影響を無視できるほど古い学習データを完全に忘却するため，指定した長さのデータ数のみ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ようなモデルを提案し，それによって予測精度が向上することを示した[関口 21]．本論文ではこの手法を以後完全忘却型 AKT と呼ぶ．

しかし，完全忘却型 AKT では反応予測に用いるデータ数Lが離散値であるため，最適なLの値を決定するためには複数の候補値についてそれぞれモデルの学習を行う必要があった．そのため，学習の期間に比例してLの最適値を決定するために必要な候補値の総数が増加し，長期間の学習ではモデルの学習に膨大な時間が必要となる問題がある．また，完全忘却型 AKT では時点tから$t-L$までの学習データの影響はそのままに，時点$t-L$より前の学習データを完全に忘却する．しかし実際の学習過程では，過去の学習の影響は経過時間に応じて小さくなると考えられる．したがって，完全忘却型 AKT では経過時間による忘却度合いの変化を表現できず，適切な忘却ができないため反応予測精度が低下している可能性がある．

そこで本研究ではロジスティック関数を用いて以下の問題点を解決する新たな忘却最適化手法を提案する．

- AKT において，過去の学習データから学習される Attention が一定値に収束し，学習過程が長くなるほどノイズとして残る問題
- 完全忘却型 AKT においてハイパーパラメータLの最適化に膨大な時間コストが必要となる問題
- 完全忘却型 AKT では，設定された経過時間でステップ関数によりデータを完全忘却するため，時間経過による忘却度変化が表現されていない問題

提案手法は以下の利点がある．(1)ロジスティック関数を用い，経過時間に応じて Attention を 0 に漸近させることで過去の学習データを忘却し，学習データが長くてもノイズを残さない．(2)忘却関数としてのロジスティック関数は二つの微分可能なパラメータを持ち，勾配法を用いて高速に最適な忘却関数を学習できる．(3)最適な忘却関数としてのロジスティック関数に応じて過去の学習データを忘却することで，経過時間による忘却度の変化を連続的に表現することができる．

最後に，本研究では従来の AKT と提案手法を用いて，学習者の反応予測精度比較を行い，提案手法の有効性を検証する．

2.関連手法

2.1. RNN を用いた手法

2.1.1. Deep Knowledge Tracing

DKT は Long-Short Term Memory(LSTM)を用いて，学習者の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[Hochreiter 97, Piech 15]. DKT ではスキル間の独立性が仮定されておらず，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，学習者の全てのスキルを単一の隠れ変数ベクトルで表現するため，学習者におけるスキルごとの習熟度を表現できない問題があった.

2.1.2. Dynamic Key-Value Memory Network

DKT の反応予測精度を向上させるために，スキルごとの能力値を保存する Memory Network を用いた DKVMN が考案された[Zhang 17]. DKVMN では学習過程にＮ個の潜在スキルを仮定し，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知られているが，DKT と同様に学習者の能力などを示す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

DKVMN のパラメータ解釈可能性を向上させるために，DKVMN と IRT を組み合わせた Deep-IRT が開発されている[Yeung 19]. Deep-IRT 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持つためモデルの解釈可能性が高く，反応予測精度も高い.

2.2. Attention を用いた手法

2.2.1. A Self-Attentive model for Knowledge Tracing

最新の KT 手法として，Transformer と呼ばれる Attention 機構を用いる新たな手法が開発されている[Pandey 19, Ghosh 20]. Transformer は自然言語処理分野で用いられる手法で，長期間で強い依存関係を持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[Vaswani 17]. Pandey らは学習過程において，学習者の現在の項目に対する反応に過去の全ての学習データが強く関係していることに注目し，Transformer を用いて現在の項目に対する反応と過去の全ての

学習データの依存関係を推定することで反応予測を行う SAKT を開発した[Pandey 19]. 彼らは RNN を用いた従来の手法はパラメータ推定に膨大な時間がかかり，またスパースデータに対して脆弱である可能性があると指摘し，SAKT を用いることで改善されること示した. Transformer は逐次的に計算を行う RNN に比べて並列計算に適しているため，SAKT は従来手法に比べ学習に必要な計算時間が少ない.

3. Attentive Knowledge Tracing

Ghosh らは学習過程における学習者の反応は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である Attentive Knowledge Tracing(AKT)を提案した[Ghosh 20]. AKT では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 を計算することにより，従来のディープラーニングアプローチと比べて反応予測が向上することが示された.

AKT では時点tにおける学習者iについて，学習者が回答した項目をq_t^i，その項目が扱うスキルをc_t^i，および回答の正誤をr_t^iとする．ここで$q_t^i \in \mathbb{N}^+, c_t^i \in \mathbb{N}^+, r_t^i \in \{0,1\}$である．ただし，学習者$i$が項目$q_t^i$に正答した場合$r_t^i = 1$，誤答した場合$r_t^i = 0$である．AKT は学習者$i$の学習履歴を$\{(q_1^i, c_1^i, r_1^i), \ldots, (q_{t-1}^i, c_{t-1}^i, r_{t-1}^i)\}$とするとき，現在の時点$t$におけるスキル$c_t^i$に関する項目$q_t^i$に対する学習者$i$の反応$r_t^i$を予測する.

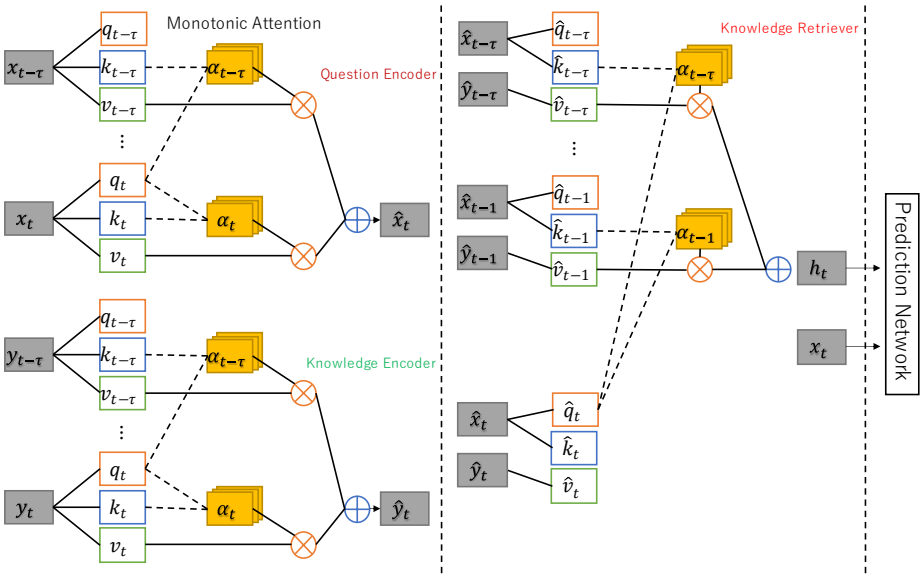

図 1. AKT のモデル図

3.1. AKT におけるモデルの構造

AKT は Question Encoder, Knowledge Encoder と呼ばれる二つのエンコーダ部分，Knowledge Retriever と呼ばれる部分，および feed-forward response prediction model と呼ばれる反応予測モデルからなる．(図 1)二つのエンコーダ部分はそれぞれ学習過程を考慮した項目と学習者の潜在的な能力ベクトルを学習し，Knowledge Retriever は学習者の知識状態を計算する．

Question Encoder は項目の特徴量ベクトル$\{x_1, \ldots, x_t\}$を入力として受け取り，Monotonic Attention 機構を用いてスキルと項目の関係性を考慮した新たな項目の特徴量ベクトル$\{\hat{x}_1, \ldots, \hat{x}_t\}$を計算する．同様に Knowledge Encoder はエンベディングした学習者の反応ベクトル$\{y_1, \ldots, y_{t-1}\}$を入力として，Monotonic Attention 機構を用いてスキルの習得状態を考慮した新たな潜在的な能力ベクトル$\{\hat{y}_1, \ldots, \hat{y}_{t-1}\}$を計算する．(Monotonic Attention 機構についてはセクション 3.2 を参照)

Knowledge Retriever は$\{\hat{x}_1, \ldots, \hat{x}_t\}, \{\hat{y}_1, \ldots, \hat{y}_{t-1}\}$を入力として，Monotonic Attention 機構を用いて現在の項目に対する知識状態h_tを計算する．また，反応予測モデル部分はh_tとx_tを入力として時点tでの学習者の反応を予測する．

3.2. Monotonic Attention 機構

AKT は Attention を計算する際に Scaled Dot-Product Attention[Vaswani 17]を経過時間の影響も考慮できるように改変した Monotonic Attention と呼ばれる機構を用いる．

3.2.1. Scaled Dot-Product Attention

Scaled Dot-Product Attention では，まず項目の特徴量ベクトル$\{x_1, \ldots, x_t\}$について，各項目とスキルの関係性を表す Attention αを計算し，新たな項目の特徴量ベクトル$\{\hat{x}_1, \ldots, \hat{x}_t\}$を求める．時点$t$の入力$x_t$から時点$\tau$の入力$x_\tau$までの学習データにおける Attention $\alpha_{t,\tau}$は次の式で求める．

$$\alpha_{t,\tau} = softmax\left(\frac{q_t^T k_\tau}{\sqrt{D_k}}\right) = \frac{\exp\left(\frac{q_t^T k_\tau}{\sqrt{D_k}}\right)}{\sum_{\tau'} \exp\left(\frac{q_t^T k'_{\tau'}}{\sqrt{D_k}}\right)} \tag{1}$$

ここで，$q_t \in \mathbb{R}^{D_k \times 1}, k_\tau \in \mathbb{R}^{D_k \times 1}$は$x_t, x_\tau$から以下で求められる．

$$q_t = x_t W^Q, k_\tau = x_\tau W^V \tag{2}$$

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．また，$\hat{y}$も同様に計算する．

$$\hat{x}_\tau = \sum_{t=1}^{\tau} \alpha_{t,\tau}, v_\tau \tag{3}$$

$$v_\tau = x_\tau W^V \tag{4}$$

次に$\hat{x}, \hat{y}$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．

3.2.2. Monotonic Attention

Scaled Dot-Product Attention は学習過程における時間的な距離に依らず学習データ間の依存関係を計算するため，学習過程における忘却を考慮できない．そこで AKT で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するための Attention である Monotonic Attention を用いる．Monotonic Attention は以下のように表される．

$$\alpha'_{t,\tau} = \frac{\exp(s_{t,\tau})}{\sum_{\tau'} \exp(s_{t,\tau'})} \tag{5}$$

$$s_{t,\tau} = \frac{\exp(-\theta \cdot d(t,\tau)) \cdot q_t^T k_\tau}{\sqrt{D_k}} \tag{6}$$

ここでハイパーパラメータθは$\theta > 0, \theta \in \mathbb{R}$である．$d(t,\tau)$は時点$t,\tau$間の時間差を用いて以下のように表される．ただし$\tau \leq t$である．

$$d(t,\tau) = |t - \tau| \cdot \sum_{t'=\tau+1}^{t} \gamma_{t,t'} \tag{7}$$

$$\gamma_{t,t'} = \frac{\exp\left(\frac{\mathrm{q_t^T k'_t}}{\sqrt{D_k}}\right)}{\sum_{1 \leq \tau' \leq t} \exp\left(\frac{\mathrm{q_t^T k'_{\tau'}}}{\sqrt{D_k}}\right)} \tag{8}$$

式(8)は時点τから時点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いる．また，AKT はそれぞれθを持つH個の独立した Attention head

を用いて Monotonic Attention を計算し，最終的な出力を$(D_v \cdot H) \times 1$のベクトルに連結して次の層へ渡す Multi-head attention を取り入れている．これによって AKT は過去の複数の項目に対して多面的に注目することを可能にしている．

Ghosh らは$d(t,\tau) = |t-\tau|$として Monotonic Attention を計算したモデルとも精度比較を行っているが，式(7)を用いた場合の方が優れた予測精度を示すことが報告されている．以上のことから，AKT では項目への反応予測を行う際，過去の項目からの経過時間に応じた忘却と，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といえる．

3.3. 完全忘却型 AKT [関口 21]

これに対し関口らは，長期間の学習では AKT の Monotonic Attention が一定値に収束していることを示し，情報量のない過去のデータを用いて反応予測を行うために予測精度が低下する可能性を指摘した[関口 21]．(図 2) そこで関口らは反応予測に用いるデータ数を直近の数問に制限し，情報量の少ない項目による予測精度の低下を防ぐ新たな AKT 手法を提案した．この手法ではハイパーパラメータ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反応予測を行う．すなわち，時点$t-L$より前の学習データを完全に忘却する．

このモデルは以下のように定式化され，Lの値はデータセットに応じて最適値を決定する．

$$\gamma_{t,t'} = \frac{\exp\left(\frac{q_t^T k_t'}{\sqrt{D_k}}\right)}{\sum_{t-L\le\tau'\le t}\exp\left(\frac{q_t^T k_{\tau'}'}{\sqrt{D_k}}\right)} \tag{9}$$

しかし実際の学習過程では過去の学習の影響は経過時間に応じて小さくなると考えられるため，この手法では経過時間による忘却度合いの変化を表現できず，過剰な忘却によって予測精度が低下している可能性がある．

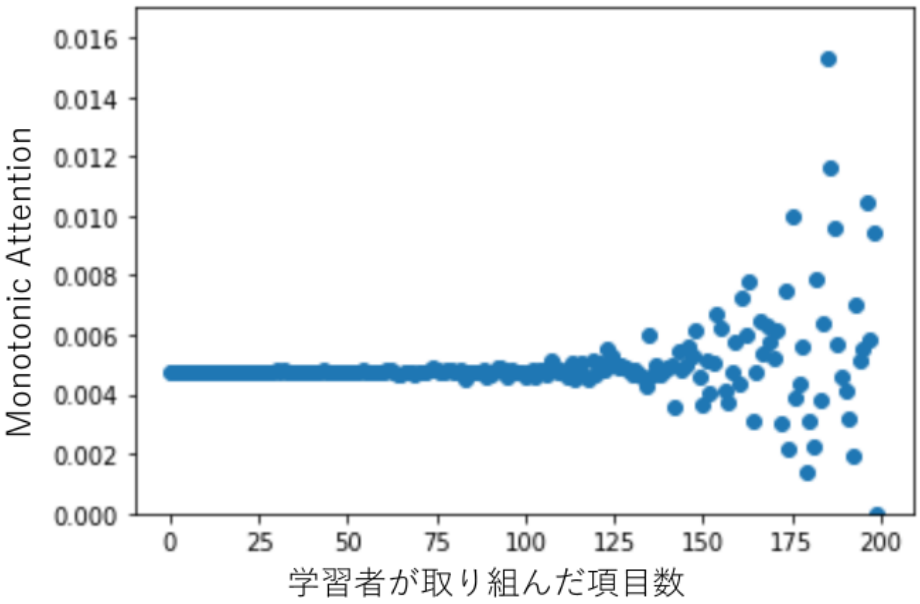

図 2. ある学習者における過去の各項目に対する Monotonic Attention

4. 提案手法

前章では,学習過程での学習者の反応予測を行う KT 手法として深層学習手法を紹介した. 本研究では最先端の手法である AKT の予測精度を向上させるため，ロジスティック関数を用いて時間差に応じて徐々に過去のデータを忘却する新たな AKT における忘却最適化手法を提案する. 提案手法のモデル図を図 3 に示す. 具体的には ATK において時点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 Attention αを計算する際，ロジスティック関数を用いて経過時間$|t-\tau|$に応じて Attention αを 0 に漸近させる.

提案モデルにおける Attention は以下で計算する. ただしa, bはハイパーパラメータであり，$a, b \in \mathbb{R}$である.

$$\alpha_{t,\tau} = \frac{1}{1+e^{a(|t-\tau|-b)}} \cdot \frac{\exp(s_{t,\tau})}{\sum_{\tau'} \exp(s_{t,\tau'})} \tag{10}$$

式(10)においてaはロジスティック関数の勾配，bは変曲点を決定する. また，経過時間に応じて Attention を減衰させるため$a > 0$である. a, bはそれぞれ標準正規分布に従う乱数から初期値を生成する. 提案手法では全ての学習可能なパラメータは Adam optimizer [Diederik 15]を用いて学習を行った. Adam はニュ

ーラルネットワークの学習においてよく用いられる最適化手法の一つであり，各パラメータに対して適切に重みを更新する．

また，提案手法は Knowledge Retriever の出力h_tと時点tの項目の特徴量x_tを全結合層に格納し，学習者iが時点tの項目に正答する予測確率$\hat{r}_t^i$を計算する．提案手法における全ての学習パラメータは，以下の全ての学習者の反応についての二値クロスエントロピーを最小化するように学習される．

$$l = \sum_i \sum_t -\left(r_t^i log \hat{r}_t^i + \left(1 - r_t^i\right) \log\left(1 - \hat{r}_t^i\right)\right) \tag{11}$$

ただし，$r_t^i \in [0,1]$は学習者iの時点tの項目に対する実際の反応を示す．

図 4 に時点tにおける完全忘却型 AKT と提案手法の忘却曲線を示す．縦軸は忘却率β，横軸は経過時間を示す．ただし$t > \tau$であり，忘却率βはそれぞれのモデルにおいて以下を満たすものとする．

$$\alpha_{t,\tau} = \beta \cdot \frac{\exp\left(s_{t,\tau}\right)}{\sum_{\tau'} \exp\left(s_{t,\tau}\right)} \tag{12}$$

完全忘却型 AKT では時点$t - L$より過去の学習データを完全に忘却するが，Lが離散値であることから，Lの最適値を決定するためには複数の候補値について個別にモデルを学習し，反応予測精度を比較する必要がある．これに対し提案手法ではロジスティック関数を用いるため，関数の概形を最適化することで忘却最適化を行うことができる．このとき，ロジスティック関数の概形は二つの微分可能なハイパーパラメータによって決定されるため，勾配法を用いてそれぞれを高速に最適化することが可能である．また，提案手法は完全忘却型 AKT と異なり，経過時間による忘却度合いの変化を表現することができる．

従来の AKT では$\{\hat{x}_1, \ldots, \hat{x}_t\}, \{\hat{y}_1, \ldots, \hat{y}_{t-1}\}$および学習者が時点$t$で項目に正答する確率を求める際にそれぞれ Multi-head attention を用いて Monotonic Attention を計算している．また，完全忘却型 AKT では全ての Attention に共通のLを与えているため，本実験では提案モデルにおいて全ての Monotonic Attention について個別にハイパーパラメータa, bを与えて最適化する場合(提案_個別)と，共通のa, bを与えて最適化する場合(提案_一括)についてそれぞれ精度比較を行った．

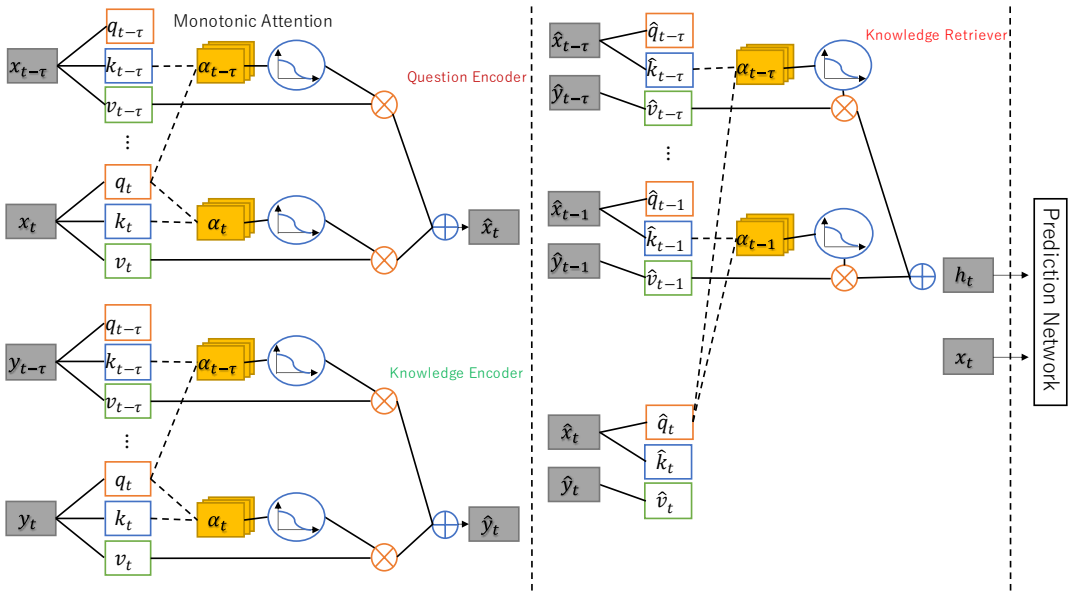

図 3. 提案手法のモデル図

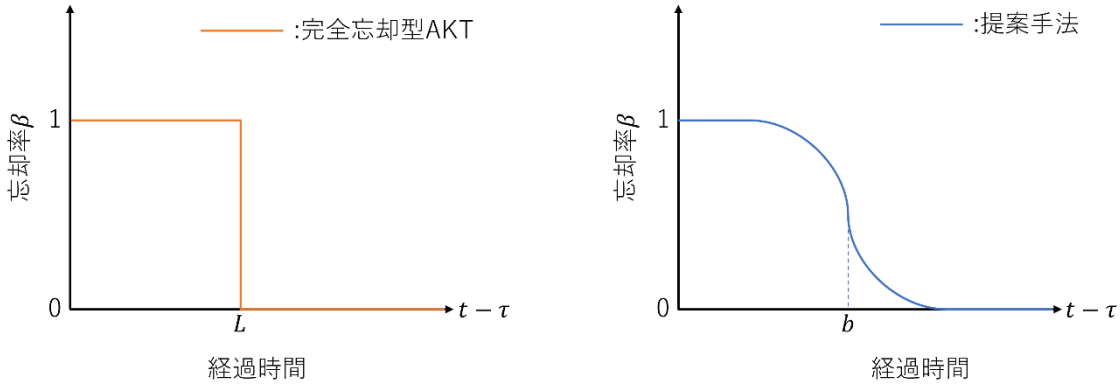

図 4. 完全忘却型 AKT と提案手法における忘却曲線

5. 予測精度評価

5.1. データセット

本実験では大規模オンライン学習システムで収集され，一般に公開されているデータセットから Statics2011, ASSISTments2009, ASSISTments2015, ASSISTments2017 を用いた．これらのデータセットの概要を表 1 に示す．各データセットには項目番号およびスキル番号が付与されているが, Statics2011 では項目番号が付与されていないため，項目数は N/A と表記した．また，入力する学習データの偏りを防ぐために入力に用いる学習データの上限を学習者 1 人あたり 200 問とした[Yeung 19].

表 1. データセットの詳細

Dataset	学習者数	スキル数	項目数	平均回答項目数	平均正答率
Statics2011	333	1,223	N/A	168.1	79.8%
ASSISTments2009	4,151	110	16,891	52.1	63.6%
ASSISTments2015	19,840	100	N/A	33.9	73.2%
ASSISTments2017	1,709	102	3,162	187.7	39.0%

5.2. 評価実験

従来の KT 手法と AKT[Ghosh 20]，完全忘却型 AKT[関口 21]および提案手法を用いて学習者の反応予測精度比較を行う．具体的には 5 分割交差検証法を用いてデータセットの 60%をトレーニングデータ, 20%をバリデーションデータ, 20%をテストデータとして行った．評価指標には KT 手法の精度比較において一般に用いられる Area Under the Curve(AUC)スコアを用いた．また，Multi-head attention の head 数は$H = 8$とした[Ghosh 20].

学習者の反応予測精度比較を行った結果を表 2 に示す．提案モデルはこれらのデータセットにおいて従来の AKT よりも高い予測精度を示したが，全てのデータセットにおいて完全忘却型 AKT が最も良い精度を示した．

表 2. 学習者の反応予測精度比較

Dataset	DKT	DKVMN	Deep-IRT	SAKT	AKT	完全忘却型AKT	提案_個別	提案_一括
Statics2011	0.8233	0.8195	0.8086	0.8029	0.8285	**0.8300**	0.8296	0.8290
ASSISTments2009	0.8170	0.8093	0.8126	0.7520	0.8219	**0.8312**	0.8255	0.8255
ASSISTments2015	0.7310	0.7276	0.7246	0.7212	0.7293	**0.7643**	0.7295	0.7294
ASSISTments2017	0.7263	0.7124	0.7187	0.7073	0.7558	**0.7693**	0.7596	0.7625

5.3. 考察

提案手法は DKT, DKVMN, Deep-IRT, SAKT, AKT から反応予測精度を向上させた．しかし，完全忘却型 AKT を上回ることはなかったが，ほぼ同程度の反応予測精度を示した．以下でこの結果について考察する．

完全忘却型 AKT においてLの最適値はデータセットに応じて決定するが，Lは離散値であるため最適値を決定するためには 1 つのデータセットに対して複数の候補値を試行する必要がある．したがって，データセットの学習期間に比例して最適な忘却度合いの決定に必要な時間も長くなってしまう問題があった．これに対し，提案手法はロジスティック関数のパラメータa, bも含めて微分可能なパラメータのみで構成されるため，勾配法を用いて同時に最適化を行うことができる．そのため, 1 度の学習で最適な忘却度合いを決定することができ，学習時間の大幅な削減が可能となった．

さらに, Statics2011, ASSISTments2009, ASSISTments2017 に関しては完全忘却型 AKT とほぼ同精度の推定値を得ることができる．ただし，推定するパラメータ数が増えた為にパラメータの推定が難しくなり，トレーニングデータへのオーバーフィッティングが起きている可能性がある．そのため，完全忘却型 AKT を上回ることができなかったと考えられる．一方, ASSISTments2015 に関しては反応予測精度が完全忘却型 AKT より低い結果となった．これは提案手法が短い学習期間のデータセットに対してロジスティック関数を最適化できない可能性がある．このため，学習データを忘却しない従来の AKT と同程度の精度となったと考えられる．

また，完全忘却型 AKT ではハイパーパラメータLを設定し，時点$t-L$以前の学習データを完全に忘却する．このとき，Lは全ての Attention に共通の値が用いられる．提案手法では完全忘却型 AKT により近い設定とするため，計算される全ての Attention に対して共通のロジスティック関数を用いた場合(提案_一

括)でも精度比較を行ったが，全ての Attention に個別のロジスティック関数を用いた場合(提案_個別)に比べて反応予測精度が向上することはなかった

6. おわりに

本研究では，AKT においてロジスティック関数を用いて忘却最適化を行う手法を提案した．

従来の AKT では過去の項目ほど Attention が一定値に収束してしまう問題があり，完全忘却型 AKT は反応予測に用いる学習データ数を制限し，ある時点より過去の学習データを完全に忘却することで反応予測精度を向上させた．しかし，反応予測に用いる学習データ数Lを最適化するためには複数の候補値に対してモデルの学習を行う必要があり，長期間の学習ほど最適値の決定に膨大な時間を要するという問題があった．また，経過時間による忘却度合いの変化を表現できないため，過剰な忘却により反応予測精度が低下している可能性があった．

提案手法はロジスティック関数を用いて Attention を 0 へ漸近させることで過去の学習データを忘却する．評価実験ではこの忘却最適化手法により全てのデータセットにおいて AKT に比べて反応予測精度が向上したことを示した．また提案手法はロジスティック関数のパラメータを含む全ての学習可能なパラメータを勾配法により最適化することで，一度のモデル学習で忘却最適化を行うことを可能にした．これにより完全忘却型 AKT がモデルの学習に膨大な時間を必要とする問題を解決した．

また，提案手法はロジスティック関数を用いることで，経過時間による忘却度合いの変化を表現できるが，完全忘却型 AKT に対して反応予測精度を向上させることはできなかった．提案手法は実験に使用した Statics2011, ASSISTments2009, ASSISTments2017 では完全忘却型 AKT と同程度の反応予測精度を示したが，これは推定するパラメータ数が増えたことでパラメータの推定が難しくなり，トレーニングデータにオーバーフィッティングしてしまった可能性がある．また，ASSISTments2015 では完全忘却型 AKT よりも低く，従来の AKT と同程度の反応予測精度となったことから，提案手法は短い学習期間のデータセットに対して適切に忘却を最適化できない可能性が考えられる．

KT では反応予測精度だけでなくモデルの解釈性の高さも重要であり，従来の

KT 手法から解釈性を高めたモデルも存在する[Tsutsumi 21]. 提案手法では学習者の能力値が得られないため，モデルの解釈性が低いという問題がある. Tsutsumi らは Deep-IRT において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. 提案手法においても同様の手法を取り入れることで解釈性を向上させることが可能であると考えられる.

謝辞

本研究の遂行にあたって，終始ご指導，ご助言を賜りました植野真臣教授に深く感謝いたします．また，研究に関する議論や論文執筆についてご指摘いただきました先輩方，研究室の皆様に感謝いたします．

7. 参考文献

[Abdelrahman 19] Abdelrahman, Ghodai; WANG, Qing. 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)

[Baker 04] Baker, Frank B.; KIM, Seock-Ho (ed.). Item response theory: Parameter estimation techniques. CRC Press (2004)

[Chen 17] Chen, Y., Liu, Q., Huang, Z., Wu, L., Chen, E., Wu, R., Su, Y., and Hu, G.: 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)

[Corbett 95] Corbett, A. T. and Anderson, J. R.: Knowledge tracing: Modeling the acquisition of procedural knowledge, User Modeling and User-adapted Interaction, Vol. 4, No. 4, pp. 253–278 (1995)

[Diederik 15] Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In Proc. International Conference on Learning Representations.

[Ghosh 20] Ghosh, A., Heffernan, N. T., and Lan, A. S.: Context-Aware Attentive Knowledge Tracing, in Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 2330–2339 (2020)

[Gonzalez 13]Gonzalez-Brenes, Jose, Yun Huang, and Peter Brusilovsky. "Fast: Feature-aware student knowledge tracing." Proceedings of NIPS 2013 Workshop on Data Driven Education. University of Pittsburgh, (2013)

[Hochreiter 97] Hochreiter, S. and Schmidhuber, J.: Long short-term memory, Neural Computation, Vol. 9, No. 8, pp. 1735–1780 (1997)

[Käser 17]Käser, T., Klingler, S., Schwing, A. G., & Gross, M. :Dynamic Bayesian networks for student modeling. IEEE Transactions on Learning Technologies, 10(4), pp.450-462 (2017)

[Lee 19] Lee, J. and Yeung, D.-Y.: 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)

[Liu 21] Liu, Q., Huang, Z., Yin, Y., Chen, E., Xiong, H., Su, Y., and Hu, G.: EKT: Exercise-Aware Knowledge Tracing for Student Performance Prediction, IEEE Transactions on Knowledge and Data Engineering, Vol. 33, No. 1, pp. 100–115 (2021)

[Minn18] Minn, S., Yu, Y., Desmarais, M. C., Zhu, F., and Vie, J.-J.: Deep Knowledge Tracing and Dynamic Student Classification for Knowledge Tracing, in 2018 IEEE International Conference on Data Mining (ICDM), pp. 1182–1187 (2018)

[Pandey 19] Pandey, S. and Karypis, G.: A self-attentive model for knowledge tracing, in 12th International Conference on Educational Data Mining, EDM 2019, pp. 384–389 (2019)

[Piech 15] Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L., and Sohl-Dickstein, J.: 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)

[Su18] Su, Y., Liu, Q., Liu, Q., Huang, Z., Yin, Y., Chen, E., Ding, C. H. Q., Wei, S., and Hu, G.: Exercise-Enhanced Sequential Modeling for Student Performance Prediction., in AAAI, pp. 2435–2443 (2018)

[Vaswani 17] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I.: Attention is All You Need, in Proceedings of the 31st International Conference on Neural Information Processing Systems, Vol. 30, pp. 5998–6008 (2017)

[Vie 19] Vie, J.-J. and Kashima, H.: Knowledge Tracing Machines: Factorization Machines for Knowledge Tracing, Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, No. 1, pp. 750–757 (2019)

[Wang19] Wang, Z., Feng, X., Tang, J., Huang, G. Y., and Liu, Z.: Deep Knowledge Tracing with Side Information., in International Conference on Artificial Intelligence in Education, pp. 303–308 (2019)

[Yeung 19] Yeung, C.-K.: Deep-IRT: Make Deep Learning Based Knowledge Tracing Explainable Using Item Response Theory., in EDM (2019)

[Yudelson 13]Yudelson, Michael V.; KOEDINGER, Kenneth R.; GORDON, Geoffrey J. Individualized bayesian knowledge tracing models. In: International conference on artificial intelligence in education. Springer, Berlin, Heidelberg, pp. 171-180 (2013)

[Zhang 17] Zhang, J., Shi, X., King, I., and Yeung, D.-Y.: 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)

[関口 21] 関口昌平, 堤瑛美子, and 植野真臣. "Attentive Knowledge Tracing における過去データの忘却最適化." 人工知能学会全国大会論文集 第 35 回全国大会 (2021). 一般社団法人人工知能学会, 2021

[Tsutsumi 21] Emiko Tsutsumi, Ryo Kinoshita, Maomi Ueno: Deep-IRT with independent student and item networks, Proceedings of the 14th International Conference on Educational Data Mining (EDM), 2021