1Q-03

強化学習を用いた最適指導法提示について

久保谷 善記 † 福原 吉博 † 森島 繁生 ‡
† 早稲田大学 ‡ 早稲田大学理工学術院総合研究所

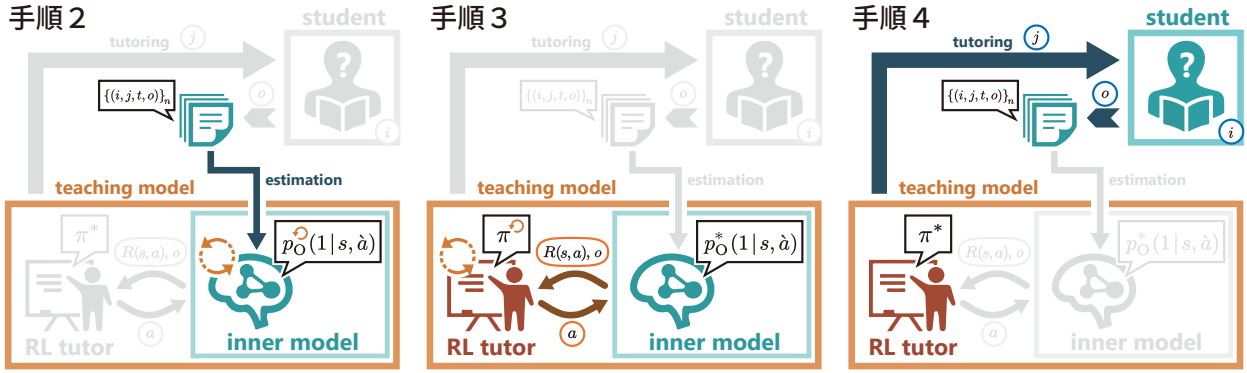

図 1: 提案手法の概要図

1. はじめに

COVID-19 によって社会全体でオンライン化が進む中，教育分野においても e-learning の普及が急務となっている．教育現場で一般的な問題形式として，英単語学習など問題に対して答えが一意に決まるフラッシュカード形式のものがあるが，これは出題と採点を自動化できるため e-learning システムと相性が良い．こうした学習形式において学習者の理解度を把握するため，近年では時間変化する学習者の知識状態を学習履歴から推定する Knowledge Tracing（KT）[3] が注目を集めている．しかし学習者が実際に効率的な学習を行うためには，追従された知識状態をもとにどのような指導を行うか（e.g. どの問題を出題するか）を検討することも必要である．

そこで本稿ではフラッシュカード形式の学習指導を想定して，(1)KT により推定した学習者の情報を元に (2) 強化学習を用いて最適指導を与える (3) 実用的なフレームワークを提案する．提案手法では指導対象の認知モデルを内部に構成することで，対象との相互作用の回数を抑えつつ指導法の最適化を目指す．

2. 関連研究

2.1 Knowledge Tracing

Knowledge Tracing（KT）[3] とは，時間変化する学習者の知識状態を学習履歴から推定するタスクのことである．ベイズ推定を用いた研究を起源として近年多様な手法が提案されつつあるが，ここでは次式 (1), (2) で定義される Choffin ら [1] の DAS3H モデルを紹介する．

$$\mathbb{P}(O_{i,j,l}=1) = \sigma\left(\alpha_i - \delta_j + \sum_{k\in \mathrm{KC}_j} \beta_k + h_\theta(t_{i,j,1:l}, o_{i,j,1:l-1})\right) \quad (1)$$

$$h_\theta(t_{i,j,1:l}, o_{i,j,1:l-1}) = \sum_{k\in \mathrm{KC}_j} \sum_{w=0}^{W-1} \theta_w \ln(1+c_{i,j,w}) + \phi_w \ln(1+n_{i,j,w}) \quad (2)$$

Modeling Personalized Teaching Methods Using Deep Reinforcement Learning:
Yoshiki Kubotani†, Yoshihiro Fukuhara†, and Shigeo Morishima‡ (†Waseda University, ‡Waseda Research Institute for Science and Engineering)

ここで，$\alpha_i, \delta_j, \beta_k$ はそれぞれ学習者 i の能力，問題 j の難易度，知識 k の習熟度を意味するパラメータであり，$O_{i,j}$ は学習者 i の問題 j に対する解答の正誤を表すバイナリ応答の確率変数である．また，シグモイド関数を $\sigma(\cdot)$ で略記した．h_θ は過去の学習履歴を反映した項であり，$n_{i,j,w}$ は学習者 i が問題 j に対して解答を試みた回数を，$c_{i,j,w}$ はそのうち正答した回数を時間窓 τ_w ごとにカウントしたものを意味する．時間窓 τ_w とは，記憶忘却の時間スケールを表すパラメータであり，心理学分野で提案されている複数の認知モデルを参考に導入された因子である [4]．$\tau_w < \tau_{w+1}$ を満たす離散時間スケールごとにカウントを分けることで，過去の学習の時間的分布を考慮した記憶率推定が可能となる．

2.2 指導最適化

最適な指導を獲得を目指す研究で主流なアプローチとして存在するのが復習間隔の最適化である．反復学習が記憶強化にもたらす効果は心理学の分野で古くから議論されており，経験則的に有効と知られる手法として，Leitner [7] の提案したライトナーシステムがある．

またヒューリスティックな手法ではなく，対象に最適化するアルゴリズムを提案することで，より適応的な指導法の獲得を目指す研究も存在する．Rafferty ら [5] は生徒指導を部分観測マルコフ決定過程 (POMDP) として定式化し，最適化を試みている．Reddy ら [6] はこれを受けて次のような POMDP を考え，強化学習の一手法である Trust Region Policy Optimization（TRPO）を用いて最適指導方策 π^* を求めている．

Reddy らによる POMDP の定義

- $\mathcal{S}$ ： 状態集合．学習者の知識状態 s を含む．
- $\mathcal{A}$ ： 行動集合．復習する問題 a を含む．
- p_T ： 状態遷移確率関数 $p_\mathrm{T}(\acute{s}|s,a)$．記憶の強化を表現．
- R ： 報酬関数 $R(s,a)$．学習者の記憶率の関数．
- $\mathcal{O}$ ： 観測集合．学習者のバイナリ応答 o を含む．
- p_O ： 観測確率関数 $p_\mathrm{O}(o|s,\grave{a})$．問題への応答を表現．

しかし，強化学習による最適化は膨大な量の相互作用を要するため，上記の手法は学習者の数理モデル化を前提としていた．従って，実際の人間を対象とする場合には

適用できないという実用上の問題があった．

3. 提案手法

本稿では，指導対象との相互作用を現実的な回数に抑えつつ，強化学習を用いて個々人に最適な指導法を獲得するフレームワーク（図 1）を提案する．提案手法は，対象生徒の記憶忘却モデル（内部モデル）を持ち，そのモデルに対して TRPO の派生手法である Proximal Policy Optimization（PPO）を用いて方策の最適化を行う．内部モデルとしては Choffin ら [1] の DAS3H モデルを採用し，POMDP の設定は Reddy ら [6] に準拠した．また内部モデルは事前にマスデータを用いて学習しておき，指導を通して対象生徒に fine-tuning する．具体的には以下の手順に沿って最適な指導を提供する．

手順 1. マスデータを用いて内部モデルを事前学習

$$p_{\mathrm{O}}(1\,|\,s,\grave{a}) \xrightarrow[\text{pre-train}]{\{(i,j,t,o)\}_N} p^{\checkmark}_{\mathrm{O}}(1\,|\,s,\grave{a})$$

手順 2. 学習者の解答情報を元に内部モデルを更新

$$p^{\checkmark}_{\mathrm{O}}(1\,|\,s,\grave{a}) \xrightarrow[\text{update}]{\{(i,j,t,o)\}_n} p^{*}_{\mathrm{O}}(1\,|\,s,\grave{a})$$

手順 3. 内部モデルに対して強化学習的に方策最適化

$$\pi^* = \underset{\pi\in\Pi}{\operatorname{argmax}}\ \underset{\tau\sim\pi}{\mathbb{E}}\left[\mathcal{L}^{\mathrm{PPO}}\,\middle|\,\pi\right]$$

手順 4. 獲得した方策を用いて対象を指導

$$\pi^* \xrightarrow{j} \text{  } \xrightarrow{o} \{(i,j,t,o)\}_n$$

手順 5. 手順 2 – 手順 4 を繰り返す

内部モデルを介する（手順 2）ことで，指導方策学習時（手順 3）に強化学習エージェントが学習者と直接相互作用することなく最適化できるようにした．

4. 実験

4.1 実験概要

提案したフレームワークによる指導の有効性を確かめるため，他の指導法との比較実験を行った．なお，学習者には実際の人間の代わりに Choffin ら [1] の DAS3H モデルを用い，内部モデルの事前学習は世界最大の教育データセット（$N \simeq 9.0 \times 10^7$）である EdNet [2] を用いて行った．図 2 に実験の概要を示す．

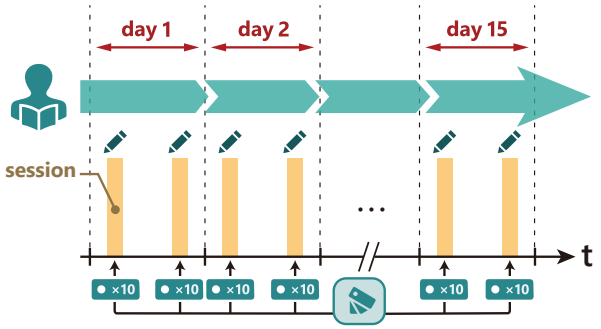

図 2: 実験概要

実験は，学習者が 15 日間で 30 問の問題を学習するという想定で行った．より現実的な設定とするため，集中的に問題へ取り組む時間（以下セッションと呼ぶ）を 1 日に 2 回設け，各セッションで 10 問の問題を学習することとした．また比較手法としてランダムな指導法とライトナーシステムを用意し，同様の実験を行った．

4.2 結果

各指導法で実験を行った結果を図 3 に示す．異なるシード値で 5 回実験を行い，学習者代わりに用いた数理モデルの記憶率のセッション平均を対ランダム指導法比でプロットした．実線は平均値を，色付きの帯部分は標準偏差帯を意味している．提案手法による指導が，他の二手法に比べてセッション全体を通して学習者の記憶率を高く維持しており，記憶強化に貢献していることが読み取れる．セッションが進むにつれ各手法間の結果の差が小さくなっているのは，ランダムな問題提示によっても記憶が強化されることに起因しており，同程度の差であっても実験が後半になるにつれてその対ランダム指導法比は小さくなっていくためである．

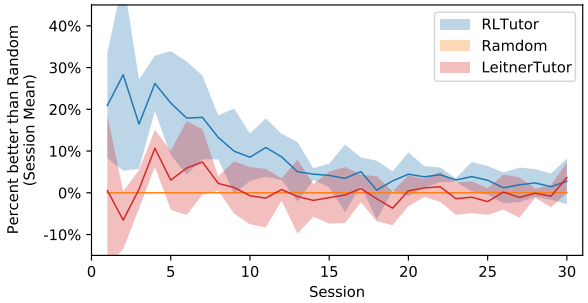

図 3: 実験結果

5. おわりに

本稿では推定された学習者の知識状態を元に，強化学習的に最適指導を提供する実用的なフレームワークを提案した．また，数理モデルを用いた実験を行い提案手法の有効性を評価した．今後は，問題数や実験期間などの設定を変えた場合の提案手法の有効性を確かめるほか，異なる認知モデルに対して同様の実験を行うことを考えている．また，比較する指導手法の数を増やし，最終的には実際の人間を用いた実験を行いたい．

謝辞

本研究は、JST ACCEL (JPMJAC1602), JST 未来社会創造事業 (JPMJMI19B2), JSPS 科研費 (JP19H01129) の補助を受けています。

参考文献

[1] Choffin B. *et al.* “DAS3H: Modeling Student Learning and Forgetting for Optimally Scheduling Distributed Practice of Skills”. *arXiv preprint arXiv:1905.06873*, 2019.

[2] Choi Y. *et al.* “Ednet: A large-scale hierarchical dataset in education”. In *International Conference on Artificial Intelligence in Education*, pp. 69–73, 2020.

[3] Corbett A. T. *et al.* “Knowledge tracing: Modeling the acquisition of procedural knowledge”. *User modeling and user-adapted interaction*, pp. 253–278, 1994.

[4] Lindsey R. V. *et al.* “Improving students’ long-term knowledge retention through personalized review”. *Psychological science*, pp. 639–647, 2014.

[5] Rafferty A. N. *et al.* “Faster teaching via pomdp planning”. *Cognitive science*, pp. 1290–1332, 2016.

[6] Reddy S. *et al.* “Accelerating human learning with deep reinforcement learning”. In *NIPS workshop on teaching machines, robots, and humans*, 2017.

[7] Leitner S. *So lernt man lernen.* Herder, 1974.