The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

効率的な知識獲得予測に向けた知識の表現方法に関する分析

Analysis of Machine-Friendly Knowledge Representation toward More Efficient Knowledge Tracing

中川 大海 *1 Hiromi Nakagawa　岩澤 有祐 *1 Yusuke Iwasawa　那須野 薫 *1 Kaoru Nasuno　松尾 豊 *1 Yutaka Matsuo

*1東京大学工学系研究科
The University of Tokyo

Recent advancements in computer-assisted learning systems have increased research in the area of *knowledge tracing*. Knowledge tracing is a task to estimate student proficiency based on their past interaction with the learning systems, and it is reported that leveraging neural networks enables efficient estimation. However, such a development of neural network-based knowledge tracing methods suggests the necessity to review the definition of "knowledge", which previously has been designed by human experts and treated as given. In this context, recently a method to automatically learn efficient knowledge representation from student exercise logs has been proposed, and it is becoming important to designing more machine-friendly knowledge representation, which enables efficient performance of knowledge tracing. In this paper, we analyze the properties of knowledge representation learned to maximize the performance of knowledge tracing, and investigate the important factors for machines to efficiently perform knowledge tracing. Using a math open dataset, we empirically validated that the learned representation has more information-efficient structure than the existing representation, and that they have mutually complementing character to improve the performance of knowledge tracing. These results provide useful insights for reviewing the definition of knowledge, which previously has been treated as given, and designing machine-friendly knowledge representation, which could help improve the learning experience of students in more diverse environments.

1. はじめに

近年，オンライン教育サービスの発展に伴い，知識獲得予測 (knowledge tracing) [Corbett 94] の研究が活発化している．知識獲得予測は，生徒の学習行動ログに基づいて，各生徒が知識をどの程度習熟しているかを推定するもので，個人の習熟状況に即した教材推薦などに活用することで，生徒の学習効率を向上させることが可能である．歴史的に様々な手法が提案されてきたが，[Piech 15] が提案した Deep Knowledge Tracing(DKT) という手法は，深層学習の Recurrent Neural Networks(RNN) [Hochreiter 97] を活用することで，従来の手法が必要としていた，人間による複雑な特徴量の設計を必要とせずに，モデルが自動的に知識間の潜在的な関係性を学習できることが報告されている．

このような深層学習による知識獲得予測の発展は，従来は人手で設計されてきた「知識」自体の見直しを要請するものでもある．例えば，DKT を含む従来の知識獲得予測の手法は，人間の専門家によって事前に定義された，演習問題を解く上で必要となる知識を表す「知識タグ」の存在を前提としていたが，[Nakagawa 18] はこの知識タグ自体も生徒の回答ログから学習することを提案し，このタグを用いることで既存の知識タグと比較して知識獲得予測の精度が向上することを報告している．すなわち，従来人手で設計されてきた知識は，人間にとって解釈しやすく，取り扱いやすいものではあるものの，機械で知識獲得予測を行う上では必ずしも最適な構造ではない可能性が示唆されており，今後，知識獲得予測に基づいた学習内容の個人最適化が一般的になっていく中で，機械にとってより予測しやすい知識表現を設計することの重要性が高まっている．

本研究では，[Nakagawa 18] の研究を発展させ，知識獲得の予測性を最大化するように回答ログから学習された知識タグの性質を分析することで，機械にとって予測しやすい知識構造についての分析を行う．まず，[Nakagawa 18] で提案されたタグ学習モデルによって，DKT における予測性を最大化するようなタグ表現を学習する．次に，学習された知識タグ（以下，学習タグ）と既存の知識タグ（以下，既存タグ）が，それぞれ DKT の予測性能に与える影響を計測することで，両者の関係性を分析し，また各タグの構造を定量的に比較することで，両者の性質の違いを検証する．実験の結果，学習タグは，既存タグと組み合わせることでより高い予測性能を発揮し，かつ既存タグと相互に補完し合う性質を有することが確認され，また，既存タグに比べて生徒の問題回答の情報を効率的に表現する構造を獲得していることが検証された．これらの結果を踏まえ，既存タグと学習タグそれぞれが有する補完的な性質について考察し，機械にとって予測しやすい知識表現を設計する上での要件について議論した．

本論文の貢献は以下の 2 点である．

- 生徒の問題回答ログから学習された知識タグ表現は，DKT による知識獲得予測を行う上で有効な表現を獲得しており，また，既存の知識タグと相互に補完し合う性質を持つことを示した．
- 学習されたタグ表現の構造や，既存の知識タグと補完し合う性質について定量的に分析することで，知識獲得の予測性を向上させる知識タグの設計に関する洞察を得た．

2. Deep Knowledge Tracing (DKT)

DKT の構造は伝統的な RNN の構造に基づいており，時刻 t における入力を $\mathbf{x}_t$，隠れ層を $\mathbf{h}_t$，出力を $\mathbf{y}_t$ とすると，その関係性は以下の式で表される．

$$\begin{aligned} \mathbf{h}_t &= f(\mathbf{x}_t, \mathbf{h}_{t-1}) \\ \mathbf{y}_t &= g(\mathbf{h}_t) \end{aligned}$$

連絡先: 中川大海, 東京大学工学系研究科技術経営戦略学専攻, nakagawa@weblab.t.u-tokyo.ac.jp

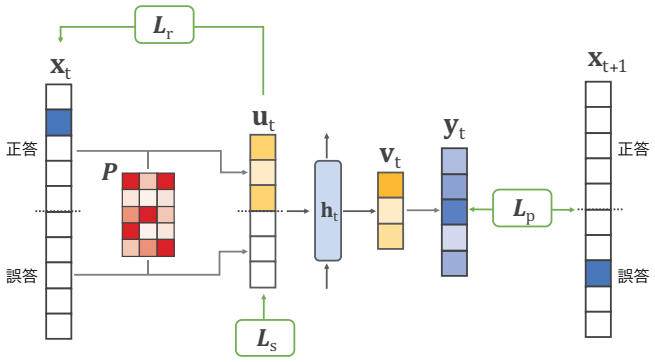

図 1: タグ学習モデルの構造

モデルは関数 f と g によって定義されており，LSTM [Hochreiter 97] や GRU[Cho 14] など様々な活性化関数を利用できる．入力 $\mathbf{x}_t$ は，全部で N 個ある既存の知識タグが紐付いた問題のうち，正答したタグに対応する要素のみ 1 で他は 0 の長さ N のベクトルと，誤答したタグに対応する要素のみ 1 で他は 0 の長さ N のベクトルを結合した，長さ $2N$ の二値のベクトルとして表現される．

最適化は式 (1) で表される回答正誤予測に関する損失関数 L_p を最小化することにより行われる．

$$L_p = \sum_t l(\mathbf{y}_t^T \tilde{\delta}(\mathbf{q}_{t+1}), \mathbf{a}_{t+1}) \tag{1}$$

ここで $\mathbf{y}_t$ は時刻 t の出力である問題正答予測ベクトルであり，$\tilde{\delta}(\mathbf{q}_{t+1})$ は実際の時刻 $t+1$ の問題回答ベクトルであり，$\mathbf{a}_{t+1}$ は時刻 $t+1$ に対応する問題で正答したか否か（1 か 0）のベクトルであり，l は交差エントロピー誤差関数である．

学習済みの DKT のモデルからは，知識獲得の時系列性や知識間の関係性を考慮した，包括的な知識間の関係性を表すネットワークを抽出することができる．問題 i と j のすべての有向ペア (i, j) のうち，問題 i が出現した後に残りの問題系列の中で問題 j が出現する系列数が問題 i が出現する問題系列数全体の V% 以上であるものに対して，式 (2) の影響度 J_{ij} を割り当てる．

$$J_{ij} = \frac{y(j|i)}{\sum_k y(j|k)} \tag{2}$$

ここで $y(j|i)$ は，ある生徒が最初に問題 i に正答した場合に，次の時刻に問題 j に正答する確率である．

3. タグ学習モデル

3.1 モデルの構造

一般的な DKT では，生徒の問題に関する回答は，専門家が定義したルールに基づいて，知識タグに関する回答に変換されてモデルに入力される．この処理は，問題の数が M，知識タグの数が N である場合に，問題回答を表現する長さ M の二値ベクトルに，問題とタグの対応を表す $M \times N$ の大きさの二値の行列 $\mathbf{P}$ をかけることで，タグに対する回答を表現する長さ N の二値ベクトルを得る処理と定式化できる．ここで，$\mathbf{P}$ は，各行が問題を，各列がタグを表し，各問題に紐づくタグに対応する要素が 1 でそれ以外が 0 の行列である．

従来の DKT ではこの行列が暗黙的に所与のものとされてきたが，タグ学習モデルでは，この写像行列を生徒の問題回答ログから学習する．モデルの構造を図 1 に示す．タグ学習モデルでは，問題に関する $2M$ 次元の回答ベクトルをそのまま入力 $\mathbf{x}_t$ とした上で，問題に対する M 次元の正答予測 $\mathbf{y}_t$ を出力する．また，問題空間を低次元のタグ空間に変換する写像行列 $\mathbf{P}$ を学習するために，N' 次元のタグの存在を仮定し，入力 $\mathbf{x}_t$ と隠れ層の $\mathbf{h}_t$ の間に $2N$' 次元の $\mathbf{u}_t$，および $\mathbf{h}_t$ と出力層 $\mathbf{y}_t$ の間に N' 次元の $\mathbf{v}_t$ という中間層を設ける．タグの次元数 N' は任意に設定することが可能であるが，生徒の学習過程を個別の問題よりも抽象的な粒度で捉えるという目的を踏まえると，問題の次元 M よりも小さい次元に設定することが適切である．特に本研究では，既存の知識タグとの比較を行うという背景から，比較を公平に行い差分を明確に検証するために，$N' = N$ として既存の知識タグと同じ次元数に設定する．以上を踏まえて，モデルは以下の式で定義される．

$$\begin{aligned}
\mathbf{x}_t &= [\mathbf{x}_{t_pos}, \mathbf{x}_{t_neg}] \\
\mathbf{u}_t &= [\mathbf{P}\mathbf{x}_{t_pos}, \mathbf{P}\mathbf{x}_{t_neg}] \\
\mathbf{h}_t &= \varphi(\mathbf{u}_t, \mathbf{h}_{t-1}, \theta) \\
\mathbf{v}_t &= \sigma(\mathbf{W}_{hv}\mathbf{h}_t + \mathbf{b}_v) \\
\mathbf{y}_t &= \sigma(\mathbf{W}_{vy}\mathbf{v}_t + \mathbf{b}_y)
\end{aligned}$$

ここで，$\mathbf{x}_{t\_pos}$, $\mathbf{x}_{t\_neg}$ はそれぞれ $\mathbf{x}_t$ の前半と後半の長さ M のベクトルで，問題の正答・誤答を表現しており，$\mathbf{W}_{hv}$，$\mathbf{W}_{vy}$ は重み行列を指し，$\mathbf{b}_v$，$\mathbf{b}_y$ はバイアス項を指し，φ は活性化関数を表す．なお，θ は任意のパラメータを指し，用いる活性化関数によって異なる．

3.2 目的関数

タグ学習モデルでは，より良質な写像行列 $\mathbf{P}$ を学習するために，式 (1) の既存の損失関数 L_p に加え，2 つの損失関数を導入する．

1 つは式 (3) で表される「再構成誤差 (L_r)」である．これは，タグ空間の各要素が，問題の背後に潜む知識間の依存関係などを適切に学習できていれば，生徒の問題に対する正答はそれに対応するタグ空間の状態から予測できるはずであるという前提を反映することを目的としたものであり，以下の式で定義される．

$$\begin{aligned}
g(\mathbf{u}_t) &= \sigma(\mathbf{W}_{vy}\mathbf{u}_t + \mathbf{b}_y) \\
L_r &= \sum_t l(\mathbf{x}_t, g(\mathbf{u}_t))
\end{aligned} \tag{3}$$

もう 1 つは式 (4) で表される「スパース正則化項 (L_s)」である．タグ学習モデルで学習される $\mathbf{P}$ の要素はそのままでは連続値であるため，学習後に一定の条件に基づき $\mathbf{P}$ の各値を 0 または 1 に離散化することでタグを生成するが，初めから $\mathbf{P}$ が離散表現に近いほうが離散化する際に情報量の損失が少ない．よってこのスパース正則化項を導入することにより，$\mathbf{P}$ を事前に 0 または 1 に近づくように学習させる．

$$L_s = \sum_t (0.5 - |\mathbf{u}_t - 0.5|)^2 \tag{4}$$

最終的に，モデル全体の損失関数 L は以下の式 (5) によって定められ，この損失関数を最小化するようにモデルが最適化される．

$$L = \alpha L_p + \beta L_r + \gamma L_s \tag{5}$$

ここで，α，β, γ は任意の非負の実数である．

表 1: データセットの統計量

生徒数	問題数	既存タグ数	ログ数
1,132	3,439	193	592,407

表 2: 各モデルにおける実験結果

入力→予測	AUC	フロー階層性	GRC	σ
既存→既存	0.810	0.72	0.70	9,701
学習→学習	**0.823**	**0.88**	**0.87**	**3,674**
学習→既存	0.816	-	-	-
既存+学習→既存	0.826	-	-	-

4. 実験

4.1 実験設定

実験には，KDDCup の Educational Data Mining Challenge で使用された「Bridge to Algebra 2006-2007」[Stamper 10] という，小学校から高校レベルの算数・数学に関するオンライン教育サービス上の生徒の問題回答ログからなるデータセットを用いた．Bridge to Algebra 2006-2007 は，「Calculate area of overlap」や「Identify GCF」など，専門家によって設計された知識タグ (以下，既存タグ) の情報を有するため，タグ学習モデルによって学習したタグ（以下，学習タグ）との比較検証が可能である．まず，[Nakagawa 18] と同様の条件に基づき，実験に用いるデータを抽出した．データセットの統計量を表 1 に示す．

次に，生徒の問題に対する回答ログにタグ学習モデルを適用した．RNN の部分には GRNN を用い，隠れ層のユニット数は 400 とした．ドロップアウトは $\mathbf{u}_t$ から $\mathbf{h}_t$ の方向にドロップアウト率 0.5，$\mathbf{h}_t$ から $\mathbf{y}_t$ の方向にドロップアウト率 0.2 で適用した．最適化は，式 (5) の L を目的関数として Adam[Kingma 14] によって最小化した．評価指標は AUC を採用した．訓練：検証：テスト $= 8:1:1$ となるようにユーザを分け，訓練ユーザのデータでモデルを構築し，検証ユーザのデータでハイパーパラメータを調整し，検証ユーザのデータで精度が最も高かったモデルを採用した．

最後に，タグ学習モデルによって学習された写像行列 $\mathbf{P}$ を以下の式に基づいて離散化し，タグ表現を獲得した．

$$\mathbf{P'}_{i,j} = \begin{cases} 1 & \text{if } \mathbf{P}_{i,j} = \max(\mathbf{P}_i) \text{ or } \mathbf{P}_{i,j} \geq \theta \\ 0 & \text{else} \end{cases} \quad (6)$$

ここで，i, j は $\mathbf{P}$ の行と列に対応する．θ はハイパーパラメータであり，学習されたタグを DKT に適用した際に検証ユーザのデータで最も高い精度を発揮する値を探索した．

4.2 実験結果

まず，[Nakagawa 18] と同様に，各問題を既存タグに変換して DKT に適用した場合（以下，「既存→既存」）と，学習タグに変換して DKT に適用した場合（以下，「学習→学習」）の精度を比較した結果を表 2 に示す．精度の高いものを太字で示した．タグ学習モデルのハイパーパラメータは，$\alpha = 1.0, \beta = 0.5, \gamma = 1.0, \theta = 0.7$ とした．実験の結果，既存タグを用いる場合と同等以上の精度で知識獲得予測を行うことのできる学習タグを抽出できた．

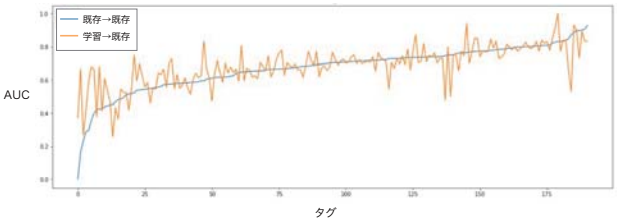

図 2: 各モデルにおける各タグの予測精度

次に，学習タグが予測精度向上に与える寄与や，学習タグと既存タグとの関係性を検証するために，モデルの出力は問題を既存タグに変換したベクトルとした上で，モデルの入力でも問題を既存タグに変換したベクトルを用いた場合（「既存→既存」），モデルの入力では問題を学習タグに変換したベクトルを用いた場合（「学習→既存」），モデルの入力に，問題を既存タグに変換したベクトルと学習タグに変換したベクトルを結合したものを用いた場合（「既存+学習→既存」）の精度を比較した．結果を表 2 に示す．まず，「学習→既存」の精度が「既存→既存」よりも高いことから，学習タグが，既存タグに基づいた回答行動を予測するという，従来のタスクを解く上でもより効率的な表現となっていることが示された．また，「既存+学習→既存」の精度が最も高かったことから，既存タグと学習タグが相互に補完的な性質を有することが示唆された．この補完性について検証するために，「既存→既存」と「学習→既存」のモデルにおいて，モデルが各既存タグに対して出力する正答予測確率を元に，各タグについて個別に AUC を算出し，図 2 に示した．青の線が「既存→既存」を，オレンジの線が「学習→既存」を表しており，各点を 1 つのタグとしてプロットし，「既存→既存」における AUC スコアのの昇順に左から並べた．図より，「既存→既存」と「学習→既存」それぞれで，予測精度の高いタグと低いタグが存在することがわかり，こうした性質を補完し合うことにより「既存+学習→既存」において最も高い精度を発揮できたと考えられる．

さらに，[Piech 15] に基づき，「既存→既存」と「学習→学習」の各モデルで知識間影響ネットワークを作成し，その階層性を計測した．階層性の指標としてフロー階層性 [Luo 11] と GRC[Mones 12] を採用した．結果を表 2 に示す．値が大きいものを太字で示した．表より，学習タグのネットワークは既存タグのネットワークに比べて高い階層性を有することがわかった．このことから，学習タグが，教育学的に「手続き的知識」として階層性が高いといわれている [Anderson 82, Anderson 90] 数学の学問的な特性を反映していることがわかる．また，階層的にカリキュラム設計が生徒の学習効率の向上を促すことも報告されており [Block 71, Cohen 79]，そのような効率的な構造を反映した階層的な構造になっている可能性も考えられる．

最後に，既存タグと学習タグについて，各タグが回答ログに出現する回数の分布を比較した．DKT においては，入出力の各要素はタグに該当するため，各タグが回答ログに出現する回数，すなわちモデルに入力される回数は，DKT の学習の過程で各ユニットに関する重みの学習されやすさに直接影響する．例えば，頻繁に回答されるタグに対応するユニットは十分に学習が進む一方，あまり回答されないタグに対応するユニットは十分に学習できない可能性が高い．既存タグと学習タグにおいて，各タグが回答ログに出現する回数の分布の標準偏差 (σ) を表 2 に示す．値の小さいものを太字で示した．表より，既存タグはタグ間で回答頻度の分散が大きい一方，学習タグは分

散が小さいことがわかる．これは，学習タグはニューラルネットワークの最適化の過程で学習されたことで，入出力層の各ユニットに相当する各タグが効率よく情報を有することができるように，情報が均等に分布するよう学習された一方，既存タグは，学問の伝統的な背景や，人間にとっての取扱やすさなどを考慮して設計されており，回答の情報を均等化することは考慮されていないためだと考えられる．

5. 考察

実験結果より，DKT において既存タグよりも高い予測性能を発揮した学習タグは，既存タグに比べて情報が効率よく分配され，高い階層性を有するものであることが示され，深層学習を用いた知識獲得予測の予測性を向上させる上で重要な素性が示唆された．一方，図 2 の分析からわかるように，全てのタグの予測において学習タグが既存タグを上回っているわけではなく,「既存→既存」「学習→既存」に比べ「既存+学習→既存」が最も高い精度を発揮したことからも，既存タグと学習タグにはそれぞれ異なる点で優れた性質を持ち，相互に補完し合うことが可能なものであることが示された．これは，人間の解釈しやすさなどに基づいて既存タグを設計される過程で失われた，個別の問題回答が有する豊富な情報を，学習タグが効果的に学習できた一方，学習タグが問題回答ログのみからでは十分に学習できなかった，複雑な知識の特性や教育環境などの前提を，人間の設計した既存タグは暗黙的に反映できていた可能性が考えられる．より多様なデータセットで同様の実験を行い，かつ既存タグと学習タグでそれぞれ予測が優れている点を詳細に比較することで，既存タグと学習タグそれぞれが反映できている情報，反映できていない情報などに関してより詳細な分析が可能になり，知識の設計を考える上でより深い洞察を得られると考えられ，今後の研究課題である．

6. 結論

本研究では，近年の深層学習による知識獲得予測の発展に伴い生じた，従来は人手で設計されてきた「知識」自体の見直しの必要性に着目し，知識獲得の予測性を最大化するように回答ログから学習された知識タグの性質を分析することで，機械にとって予測しやすい知識構造についての分析を行った．実験の結果，学習された知識タグは，既存の知識タグと組み合わせることでより高い予測性能を発揮し，かつ既存タグと相互に補完し合う性質を有することが確認され，また，既存タグに比べて生徒の問題回答の情報を効率的に表現できる構造を獲得していることが検証された．これらの結果は，従来人手で設計されてきた知識は，人間にとって解釈しやすく，取り扱いやすいものではあるものの，機械で知識獲得予測を行う上では必ずしも最適な構造ではないという可能性を示唆するものであることに加え，機械にとって予測しやすい知識を設計する上で，有益な示唆を与えるものである．より多様な教科やデータセットにおいて同様の検証を行うことで，今後さらに普及が進むと考えられる知識獲得予測に基づく学習内容の個人最適化を，さらに効果的なものにしていく上で重要な「知識」を適切に定義する上での要件が明らかにされると考えられる．本研究が，そうした新たな「知識」の解明につながると信じている．

謝辞

本研究は JSPS 科研費 JP25700032, JP15H05327, JP16H06562 の助成を受けたものです．

参考文献

[Anderson 82] John R Anderson: Acquisition of cognitive skill, Psychological review 89, 4 (1982), 369.

[Anderson 90] John R Anderson: The adaptive character of thought, Psychology Press (1990).

[Block 71] James H Block and Peter W Airasian: Mastery learning: Theory and practice. Holt Rinehart & Winston (1971).

[Cho 14] Kyunghyun Cho, Bart Van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 (2014).

[Cohen 79] S Alan Cohen and Joan S Hyman. Learning for Mastery: Ten Conclusions after 15 Years and 3,000 Schools. Educational Leadership 37, 2 (1979), 1049.

[Corbett 94] Albert T Corbett and John R Anderson: Knowledge tracing: Modeling the acquisition of procedural knowledge, User modeling and user-adapted interaction 4, 4 (1994), 253278.

[Hochreiter 97] Sepp Hochreiter and Jurgen Schmidhuber: Long short-term memory, Neural computation 9, 8 (1997), 17351780.

[Kingma 14] Kingma, Diederik and Ba, Jimmy: Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).

[Luo 11] Jianxi Luo and Christopher L Magee: Detecting evolving patterns of self-organizing networks by flow hierarchy measurement, Complexity 16, 6 (2011), 5361.

[Mones 12] Enys Mones, Lilla Vicsek, and Tamas Vicsek: Hierarchy measure for complex networks, PloS one 7, 3 (2012), e33799.

[Piech 15] 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. 505513 (2015).

[Stamper 10] J. Stamper, A. Niculescu-Mizil, S. Ritter, G.J. Gordon, and K.R Koedinger. Bridge to Algebra 2006-2007, Development data set from KDD Cup 2010 Educational Data Mining Challenge, http://pslcdatashop.web.cmu.edu/KDDCup/downloads.jsp. (2010).

[Nakagawa 18] 中川 大海, 那須野 薫, 岩澤 有祐, 上野山 勝也, and 松尾 豊: Deep Knowledge Tracing の拡張による擬似知識タグの生成, 人工知能学会論文誌 33, 3 (2018).