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

Deep-IRT の予測精度向上に寄与するデータの特性分析

2022 年 1 月 27 日

情報数理工学プログラム

学籍番号 1810433

鶴貝 昂也

指導教員 植野 真臣

令和３年度 情報数理工学プログラム卒業論文概要

平成 30 年度 入学	学籍番号 1810433
指導教員 植野 真臣	氏名 鶴貝 昂也
題目 Deep-IRT の予測精度向上に寄与するデータの特性分析	

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。

KT は現在様々な手法が存在し、これらはよく知られたベンチマークデータセットにより評価されてきたが、それぞれデータの特徴が異なり、各手法はそれらの条件を考慮して比較されてきたとはいえない。

本研究では、Deep-IRT（tsutsumi et. al 2021）の予測精度向上に寄与するデータの特性を分析する。具体的には Deep-IRT により算出した AUC スコアを目的変数とし、AUC スコアに影響を及ぼしたと思われる要素を説明変数として重回帰分析とランダムフォレストを行う。その結果、Deep-IRT は学習者一人あたりの解答数が AUC スコアに最も影響を与え、その数が大きい程高いスコアを出す傾向があること、「学習者の正答率」よりも「対象学習者がヒントを見たかどうか」や「その項目に挑戦した回数」といった「その項目に対する学習者履歴」の方が AUC スコアに影響を与えることがわかった。

1. はじめに

近年、オンライン学習システムが世界的に普及し、教育ビックデータが容易に入手できるようになった。教育現場では、個々の学習者の理解度を把握し、適切な教育支援を行うために、これらの教育ビックデータをいかに活用するかが課題となっている。人工知能分野では、機械学習手法を用いて教育ビックデータを解析することにより、過去の学習履歴から学習者の知識の習得状態を推定し、学習者の未知の課題への反応を予測する Knowledge Tracing が注目を集めている。Knowledge Tracing では知識の習得状態を学習者にフィードバックすることで教師や学習者が自ら得意・苦手分野を把握することができる。Knowledge Tracing は主に確率的アプローチと深層学習アプローチの二つに分けられる。確率的アプローチの主なモデルとして、Bayesian Knowledge Tracing(BKT)と Item Response Theory(IRT)がある[Yudelson 13,Gonzalez 17,Baker 04]。BKT では隠れマルコフモデルを用いて学習過程での学習者のスキルの習得状態を 2 値で推定し、未知の課題への反応を予測する。しかし、スキルの習得状態を 2 値で表現するために、習得状態の変化を柔軟に表現することができない。一方 IRT は、パラメータの解釈性が非常に高いモデルとして知られており、学習者の習熟度を表すパラメータと課題の特性パラメータを推定し、課題への正答率を予測する。しかし同じ課題に繰り返し取り組む学習に適応できず、更に複数のスキルの関係性を考慮した能力推定が出来ない。また、確率的アプローチは課題解決に必要な複数スキルの関係性を考慮した反応予測が出来ない。確率的アプローチの問題を解決するために、さまざまな深層学習アプローチが開発されている。Deep Knowledge Tracing(DKT)は Long Short Term Memory(LSTM)と呼ばれる回帰型ニューラルネットワークを用いて反応予測を行う[Piech 15, H.Sepp 97]。DKT は BKT より高い反応予測精度を示す。しかし全てのスキルに対する習熟度を単一の隠れ変数ベクトルで表現するため、各スキルをどの程度習得するか表現することができなかった。そこで、学習者の反応予測精度向上のために、外部に情報を記録する Memory Network を組み込んだ Dynamic Key-Value Memory Network(DKVMN)が開発された[Zhang 17]。しかし DKVMN は学習者の習熟度パラメータや課題の難易度を表す特性パラメータをもたないため、DKT と同様に解釈可能性が低いという問題があった。そこで、Yeung らは DKVMN と IRT を組み合わせ、習熟度パラメータと特性パラメータをもつ手法を開発した。Deep-IRT は DKVMN に比べて予測精度を落とすことなく、モデルの解釈性を向上させた[Yeung 19]。しかし、推定された習熟度パラメータが課題の特性に依存しており、特製の異なる課題に取り組む際には解釈性に制限があるという欠点があった。近年では更なる反応予測精度向上のため、自然言語処理の分野で用いられる transformer を用いた Knowledge Tracing 手法が複数提案されている。transformer は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。Pandey らは transformer を Knowledge tracing に応用し、Self Attentive Knowledge Tracing(SAKT)を開発した。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。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、直近の学習に依存すると主張し、Attentive Knowledge Tracing(AKT)を開発した[Ghosh 20]。AKT は過去の学習データを徐々に忘却させ、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。また、学習者が解答した課題とスキルの両方の特徴を考慮することにより、高い予測精度をした。しかし、AKT は解釈可能なパラメータをもたず、学習者の習熟度変化を表現することができなかった、最新の手法では、反応予測精度とパラメータの解釈性を両立するために新たな Deep-IRT が提案されている。Deep-IRT（tsutsumi et. al 2021）では、能力の時系列変化を表現する学習者ネットワークと独立な項目ネットワークを用いることにより反応予測精度を向上させた高いパラメータを実現した[tsutsumi et. al 21]。

これらはよく知られたベンチマークデータセットにより評価されてきたが、それぞれデータの特徴が異なり、各手法はそれらの条件を考慮して比較されてきたとはいえない。

本研究では、Deep-IRT（tsutsumi et. al 2021）の予測精度向上に寄与するデータの特性を分析することにより、Deep-IRT の特徴について考察する。具体的には Deep-IRT に対して複数のデータセットを用いて学習者の反応予測を行う。その後、算出した AUC スコアを目的変数とし、AUC スコアに影響を及ぼしたと思われる要素を説明変数として重回帰分析を行う。重回帰分析によって求められた t 値、P-値から目的変数である AUC スコアに影響を与えている説明変数を求める。

その後、重回帰分析と同じ条件でランダムフォレストによる分析を行う。まず二分木図を表示させることにより、AUC スコアが分岐する具体的な数値を確認する。次に feature importance を求めてグラフとして表示することにより、各要素の AUC スコアに対する重要度を分析する。

その結果、以下の知見を得た。

- 学習者一人あたりの解答数が AUC スコアに最も影響を与え、その数が大きければ大きい程高いスコアを出す傾向がある
- 「対象学習者がヒントを見たかどうか」や「その項目に挑戦した回数」といった学習者個人の「その項目に対する学習者履歴」データが AUC スコアに影響を与える

2. 先行研究

1. Item Response Theory (IRT)

IRT は過去の学習データから学習者の能力値と課題の特性パラメータを推定し、未知の課題への反応を予測する Knowledge Tracing 手法として用いられる[Baker 04]。ここでは最も一般的な 2 パラメータロジスティックモデルについて説明する。2 母数ロジスティックモデルでは、能力値$\theta_i \in (-\infty, \infty)$の学習者$i$が課題 j に正答する確率を次式で表わす。

$$P_j(\theta_i) = \frac{1}{1 + \exp\left(-a_j(\theta_i - b_j)\right)} \quad \cdots (1)$$

ここで、$a_j \in [0, \infty)$は課題 j の識別力パラメータ、$b_j \in (-\infty, \infty)$は課題 j の課題パラメータである。標準的な IRT では学習過程での学習者の能力は固定値であるため、能力の時系列変化は反映されていない。

2. Dynamic Key-Value Memory Networks (DKVMN)

DKVMN では N 個の潜在スキルを仮定しており、各課題と潜在スキルの関係を key memory $M^k \in R^{N \times d_k}$に保存し、時点 t の各潜在スキルに対する能力を value memory $M^k \in R^{N \times d_t}$に保存する[Zhang 17]。ここで、d_k, d_tはチューニングパラメータである。式(2)で入力q_jから生成される課題ベクトル$\beta_1^{(j)}$を用いて、時点 t で回答する課題と l 番目の潜在スキルの関係性の強さを表すアテンションを計算する。

$$\beta_1^{(j)} = W^{(\beta 1)} q_j + \tau^{(\beta 1)} \quad \cdots (2)$$

$$w_{tl} = softmax\left(M_l^k \beta_1^{(j)}\right) \quad \cdots (3)$$

ここで、M_l^kは key memory の l 行目を示す。なお、W, τはそれぞれニューラルネットワークの重みパラメータ、バイアスパラメータとする。

次に、アテンションを用いた value memory の重み付き和から学習者ベクトル$\theta_1^{(t)}$を計算し、$\beta_1^{(j)}$と組み合わせることで、時点 t の課題 j への正当確率p_{tj}を計算する。

$$\theta_1^{(t)} = \sum_{l=1}^{N} w_{tl} (M_{tl}^v)^\top \quad \cdots (4)$$

$$\theta_2^{(t)} = \tanh\left(W^{(\theta_2)}\right)\left[\theta_1^{(t)}, \beta_1^{(j)}\right] + \tau^{(\theta_2)} \quad \cdots (5)$$

$$p_{tj} = \sigma\left(W^{(y)}\theta_2^{(t)} + \tau^{(y)}\right) \cdots (6)$$

DKVMN は高い反応予測精度を示すことが知られているが、学習者の能力パラメータや課題困難度パラメータをもたないため、解釈可能性が低いという問題がある。

3.Deep-IRT

Deep-IRT は DKVMN に隠れ層を追加し、解釈可能な能力パラメータと課題困難度パラメータが得られるように設計されたモデルである[Yeung 19]。具体的には、以下のように時点 t で課題 j を回答するときの能力値$\theta_3^{(t,j)}$と課題 j の困難度$\beta_2^{(j)}$を DKVMN の式(2)と(5)を用いて算出する。

$$\theta_3^{(t)} = \tanh\left(W^{(\theta_3)}\theta_2^{(t)} + \tau^{(\theta_3)}\right) \cdots (7)$$

$$\beta_2^{(j)} = \tanh\left(W^{(\beta_2)}\beta_1^{(j)} + \tau^{(\beta_2)}\right) \cdots (8)$$

これを用いて、次のように正答確率を求める。

$$p_{tj} = \sigma\left(3.0 \times \theta_3^{(t,j)} - \beta_2^{(j)}\right) \cdots (9)$$

学習者の能力はM_t^vに圧縮されていると考えられるが、Deep-IRT では$\theta_3^{(t,j)}$を計算するために課題固有のアテンションw_tをもとにしたM_t^vの重み付き和を用いているため、得られる能力値が課題の特性に依存している。また、$\theta_2^{(t)}$を計算する際に、学習者ベクトル$\theta_3^{(t,j)}$と課題ベクトル$\beta_2^{(j)}$の両方を用いるため、能力値と困難度の特性を分離することができない。よって、Deep-IRT のもつパラメータ解釈性は十分ではないといえる。

4.Attentive Knowledge Tracing (AKT)

AKT では学習者が取り組んだ課題q_tとその課題への反応r_tを用いて、課題の特徴量ベクトルx_tと学習者の潜在的な能力ベクトルy_tを表現する[Ghosh 20]。t-1 時点までの学習データから$\{x_1, \cdots, x_t\}$と$\{y_1, \cdots, y_t\}$を計算し、時点 t で学習者がある課題に正答する確率を予測する。

まず時点 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 を求める。

$$\alpha_{t,\tau} = softmax\left(\frac{q_t^T k_\tau}{\sqrt{D_k}}\right) = \frac{\exp\left(\frac{q_t^T k_\tau}{\sqrt{D_k}}\right)}{\sum_{\tau'}\exp\left(\frac{q_t^T k'_{\tau'}}{\sqrt{D_k}}\right)} \cdots (10)$$

ここで、$q_t \in R^{D_k \times 1}$、$k_t \in R^{D_k \times 1}$は以下の式で求められる。

$$q_t = x_t W^Q \quad \cdots (11)$$

$$k_t = x_t W^k \quad \cdots (12)$$

W は重みを表す。これらを用いて新たな課題の特徴量ベクトルである$\hat{x}$を計算する。

$$\hat{x}_\tau = \sum_{t=1}^{\tau} \alpha_{t,\tau} v_\tau \quad \cdots (13)$$

$$v_t = x_t W^v \quad \cdots (14)$$

$\hat{y}$も同様に計算する。次に、$\hat{x}$と$\hat{y}$を用いて学習者が時点τで課題に正答する確率を求求める。AKT では学習者が取り組んだ課題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 Monotonic Attention を以下の式で計算する。

$$\alpha'_{t,\tau} = \frac{(s_{t,\tau'})}{\sum_{\tau'} \exp(s_{t,\tau'})} \quad \cdots (15)$$

$$s_{t,\tau} = \frac{exp\left(-\theta \cdot d(t,\tau)\right) \cdot q_t^T k_\tau}{\sqrt{D_k}} \quad \cdots (16)$$

ハイパーパラメータθは$\theta > 0, \theta \in R$であり、$\hat{x}_\tau$は常に $\exp(-\theta \cdot d(t,\tau)) \leq 1$ の値を取るため、$d(t,\tau)$の値次第で Monotonic Attention は変化する。 $\tau \leq t$とすると$d(t,\tau)$は以下の式で求められる。

$$d(t,\tau) = |t - \tau| \cdot \sum_{t'=\tau+1}^{t} \gamma_{t,t'} \quad \cdots (17)$$

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

(18)式は時点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いる。このように AKT では、過去の学習課題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。

5.Deep-IRT（tsutsumi et. al 2021）

Deep-IRT（tsutsumi et. al 2021）では、Deep-IRT において反応予測精度を保ちなが

らパラメータ解釈性を向上させるために、学習者の習熟度と課題の特性をそれぞれ独立なネットワークにより推定する手法を提案した[tsutsumi et. al 21,Yeung 19]。学習者ネットワークでは IRT に基づき、以下のように深層ニューラルネットワークを計算し、課題 j に回答する際の能力値 θ を計算する。

$$\theta_1^{(t,j)} = \sum_{l=1}^{N} M_{tl}^{v} \quad \cdots (19)$$

$$\theta_k^{(t,j)} = \tanh\left(W^{(\theta_k)}\theta_{k-1}^{(t,j)} + \tau^{(\theta_k)}\right) \quad \cdots (20)$$

ここで、k は k={2,3,⋯,n}である。学習者ネットワークの層数 n は学習データに基づいて算出する。次に、項目ネットワークで課題 j の困難度β_{item}^{j}とその課題に必要なスキルに対する 困難度β_{skill}^{j}を求める。β_{item}^{j} は n 層のニューラルネットワークで以下のように推定する。

$$\beta_1^{(j)} = W^{(\beta_1)}q_j + \tau^{(\beta_1)} \quad \cdots (21)$$

$$\beta_1^{(j)} = \tanh\left(W^{(\beta_k)}\beta_{k-1}^{(j)} + \tau^{(\beta_k)}\right) \quad \cdots (22)$$

$$\beta_{item}^{(j)} = W^{(\beta_n)}\beta_n^{(j)} + \tau^{(\beta_n)} \quad \cdots (23)$$

同様に、ベクトル$sj \in R$ sからβ_{skill}^{j}を計算する。

$$\gamma_1^{(j)} = W^{(\gamma_1)}s_j + \tau^{(\gamma_1)} \quad \cdots (24)$$

$$\gamma_k^{(j)} = \tanh\left(W^{(\gamma_k)}\gamma_{k-1}^{(j)} + \tau^{(\gamma_k)}\right) \quad \cdots (25)$$

$$\beta_{skill}^{(j)} = W^{(\gamma_n)}\gamma_n^{(j)} + \tau^{(\gamma_n)}$$

最後に、習熟度と困難度の差から正答確率を予測する。

$$p_{tj} = \sigma\left(\theta_n^{(t,j)} - \left(\beta_{item}^{(j)} + \beta_{skill}^{(j)}\right)\right) \quad \cdots (27)$$

Deep-IRT（tsutsumi et. al 2021）では学習者の能力が解答する課題の特性に依存せず、複数のスキルに関する多次元の能力を表現できる。さらに、課題とスキルの双方向 の特徴を考慮した反応予測を行うことで、反応予測精度が向上することを示している。

3. Deep-IRT の予測精度の要因分析

本章では、Deep-IRT に対し複数のデータセットを用いて学習者の反応予測精度の要因分析を行う。

具体的には、まず 5 分割交差検証を用いて学習履歴データを訓練データ、検証データ、評価データに分割し、訓練データ、検証データから推定したパラメータを利用して評価データの反応予測の比較を行う。予測精度の評価指標には AUC スコアを算出する。

次にデータセットに含まれる要因の中から関連のありそうなものを選び、AUC スコアの説明変数として要因分析を行う。

本実験では、オンライン学習システムで収集された公開データセットを加工し、実験を行った。また、すべての条件において回答数が 5 問以下の学習者、10 人未満の学習者が取り組んだ課題を省いたデータセットを作成した。

1.重回帰分析

本節では、算出された AUC スコアを目的変数とし、以下の説明変数を所与として重回帰分析を行うことにより、データセットに含まれる要素がどの程度影響を与えているのかを算出する。

説明変数には一つ目に「学習者一人あたりの解答数の平均」を示す「Ave.No.student’s answers」を加える。これはデータが多いほど精度が安定する機械学習の特性を鑑みて、学習者一人あたりのデータ数を表す指標として登録する。二つ目に「項目一問あたりの解答数の平均」を示す「Ave.No.item's answers」を加える。先ほどの「Ave.No.student’s answers」が学習者一人あたりのデータ数だったのに対し、「Ave.No.item's answers」は項目の解答数の平均を示している。こちらも先ほどと同じ理由で登録する。三つ目に「学習者の正答率の二値エントロピー」を示す「Item Entropy」を加える。二値エントロピーを採用した理由は、学習者の正答率そのままのデータでは学習者の正答率が極端に高い場合、または低い場合にデータに同じ数字が並ぶこととなり、予測がしやすくなる。よって二値エントロピーを採用したことにより、学習者の正答率が 0.5 となる場合が高い値を示し、逆に 0 または 1 となる場合が最も低い値を示す。四つ目に「学習者全体の中でヒントを見た人の割合」を示すデータである「Rate.student.hint for item」を加える。これはヒントを見た学習者の割合を計測している。五つ目に「項目一問あたりの学習者の数の平均」を示す「Ave.No. students for an item」を加える。同じ項目に何度も挑戦した場合、AUC スコアに影響

があるのかを確認するために登録をする。最後に「項目を解くために費やした時間」を示す「Ave.item response time」を加える。項目を解く時間が早い場合、または遅い場合、何かしらの要因、具体的には項目を解く時間が速い理由は解いた学習者が優秀だったから、解いた項目が簡単だったから、などとこの項目から複数の情報が読み取れると思われるため登録する。

以上の説明変数と AUC スコアを目的変数に代入し、重回帰分析を行い、その結果を表 2 にまとめた。

表 1.説明変数

説明変数	内容
Ave.No.student's answers	学習者一人あたりの解答数の平均
Ave.No.item's answers	項目一問あたりの解答数の平均
Item Entropy	学習者の正答率の二値エントロピー
Rate.student.hint for item	全ての学習者の中でヒントを見た学習者の割合
Ave.No.students for an item	項目一問あたりの学習者の数の平均
Ave.item response time	項目を解くために費やした時間(/秒)

まず表 2 の t 値から確認する。t 値は一般的に絶対値が 2 以上の時、その説明変数は目的変数に影響を与えるといわれている。表 2 より「Ave.No.item's answers」のみ値が 1.47 と 2 を下回り、その他の説明変数の値は 2 以上であることがわかる。また t 値はその数値が大きいほど与える影響も大きい。t 値の大きい順に説明変数を並べていくと、「Rate.student.hint for item」が最も影響度が高く、その後は「Ave.No. students for an item」、「Item Entropy」「Ave.No.student's answers」「Ave.item response time」の順に数値が大きい。よって AUC スコアに与える影響もその順に大きいと考えられる。

次に P-値に注目する。P-値は一般的に 0.05 以下の場合、その説明変数は目的変数と関連があるといわれている。用意した説明変数の中で「Ave.No.item's answers」のみ 0.05 を上回っており、他の説明変数はすべて 0.05 以下である。

重回帰分析の結果だけを見ると、「Ave.No.item's answers」のみ AUC スコアに影響がなく、その他の説明変数は先程挙げた順に AUC スコアと関わりを持っているように思える。しかし説明変数の中には相関関係のあるものもあり、重回帰分析では多重共線性の恐れがある。事実、「Ave.No.item's answers」と相関があると思われる「Ave.No. students for an item」は高い影響力を持っている。よって変数間の相関関

係を許容する分析方法でも分析を行う必要がある。そこで本実験では相関関係を許容するアプローチとしてランダムフォレストを用いて分析を行う。

表 2. Deep-IRT の重回帰分析結果

	係数	標準誤差	t	P-値
切片	112.0017	11.63171	9.628999	5.95E-09
Ave.No.student's answers	-0.01489	0.005478	-2.71845	0.013232
Ave.No.item's answers	-0.03356	0.020468	-1.63959	0.116727
Item Entropy	30.61972	8.320867	3.679871	0.001485
Rate.student.hint for item	8.279776	1.474404	5.615677	1.7E-05
Ave.No.students for an item	-46.1514	9.811428	-4.70385	0.000136
Ave.item response time	-0.32772	0.134306	-2.44013	0.024116

2.ランダムフォレスト

2-1.二分木図

本節ではランダムフォレストによる分析を行う。ランダムフォレストは線形モデルである重回帰分析とは異なり非線形モデルであり、パターン予測に基づいて予測を行う。

最初に二分木図を表示することにより、各要素が AUC スコアに影響を与える具体的な数値を確認すると共に、分岐先の結果から数値によってどのような影響を与えるかを確認する。

まず図 2 の Ave.No.student's answers に注目する。Ave.No.student's answers は 5.2、37.91、45.12、89.81、318.115 の五つの数値で分岐しており、またおおよその分岐先で値が高い方が AUC スコアが高くなっている。このことから Ave.No.student's answers では数値が大きい方が高い AUC スコアになると推測できる。この推測が正しいのか確かめるために、Ave.No.student's answers の値ごとにおける AUC スコアの平均を求め、図 1 に載せる。その結果 Ave.No.student's answers と AUC スコアの間に比例関係があることが確認できた。よって、Ave.No.student's answers の値が高いほど AUC スコアも高くなる傾向があるといえる。

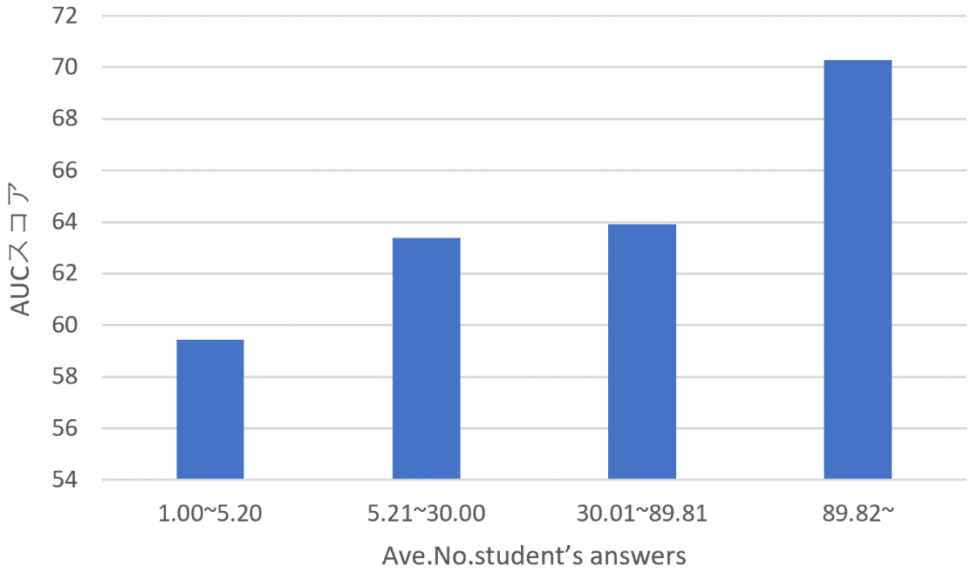


図 1. Ave.No.student's answers **の値と AUC スコアの関係**

次に、Rate.student.hint for item に注目すると、0.194、0.231 の二つで分岐が起きており、両方で Rate.student.hint for item の値が大きい方が AUC スコアの高い分岐となっている。よって Rate.student.hint for item は値が大きいほど AUC スコアが高くなると考えられる。

Ave.item response time に注目すると、40.725、47.862、49.115、52.524 の四つの数値で分岐が起きており、その半分で Ave.item response time の値が大きい方が AUC スコアの高い分岐となっている。このことから AUC スコアは Ave.item response time の値に因らないと推測できる。

Ave.No.item's answers に注目すると、8.775、8.85、30.31、35.12、52.145 の五つの数値で分岐が起きており、その内の 4 つで Ave.No.item's answers の値が大きい方が AUC スコアの高い分岐となっている。このことから Ave.No.item's answers は値が大きい方が AUC スコアが高くなると考えられる。

Item Entropy では 0.79 で分岐が起き、値が小さい方が AUC スコアの高い分岐となっている。

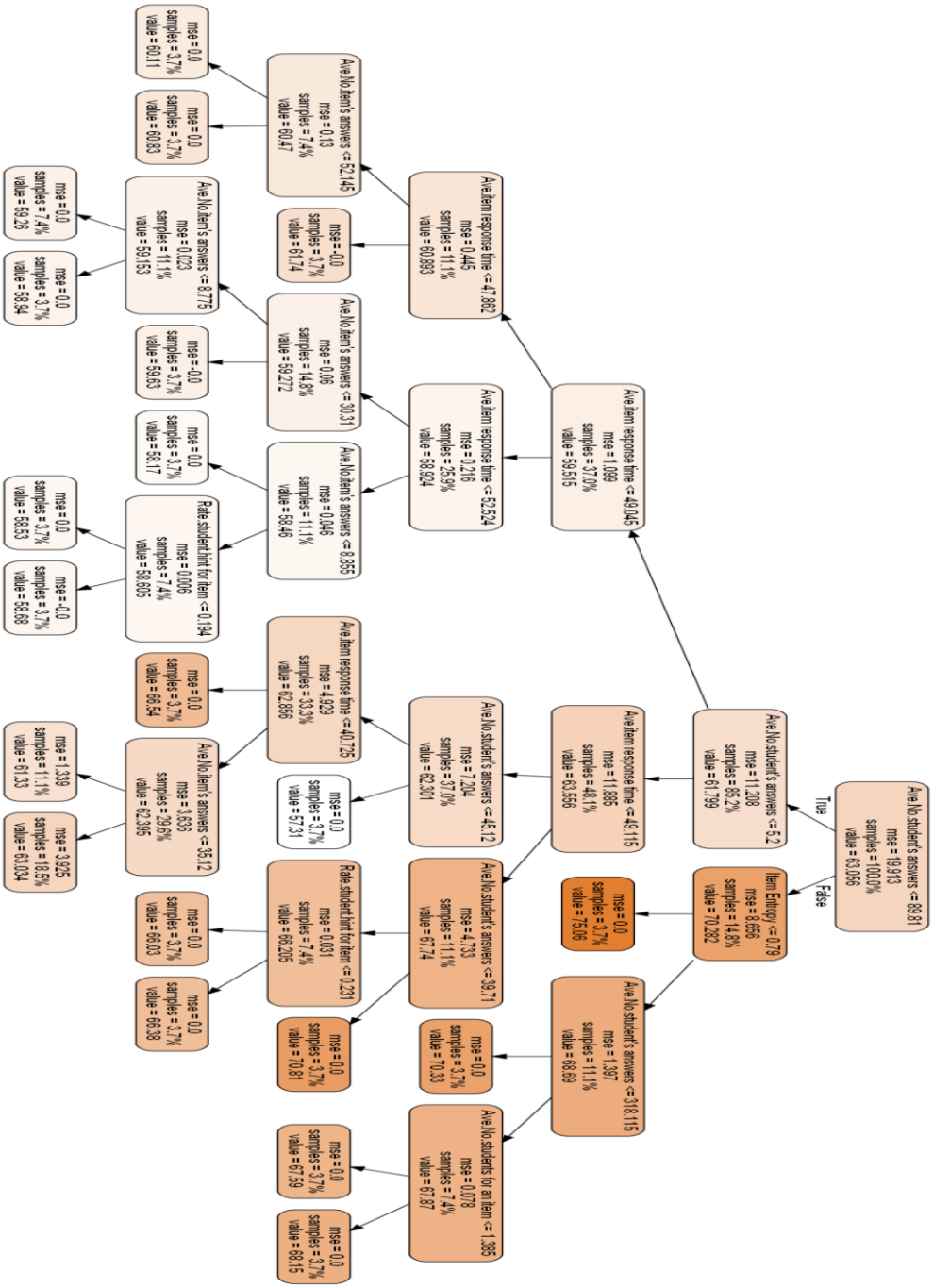


図 2. Deep-IRT の AUC スコアによる二分木

Ave.No. students for an item では 1.3875 で分岐が起きており、値が大きい方が AUC スコアの低い分岐となっている。

しかしこの二つはそれぞれ分岐が一つしか存在せず、これだけで何かを推測することはできない。

2-2.feature importance

本節では、ランダムフォレストにより特徴量の重要性を定量化する feature importance を求め、各要素の AUC スコアに対する重要度を求める。

ランダムフォレストを用いて Deep-IRT での各要素の AUC スコアに対する重要度を求めるために、「1.重回帰分析」と同様に AUC スコアを目的変数、その他の要素を説明変数として分析を行った。feature importance を求めて図にしたことにより、どの要素がどれくらい AUC スコアに影響を与えているのかを視覚的に確認することが出来る。

求めた feature importance を図 3 にグラフとしてまとめる。

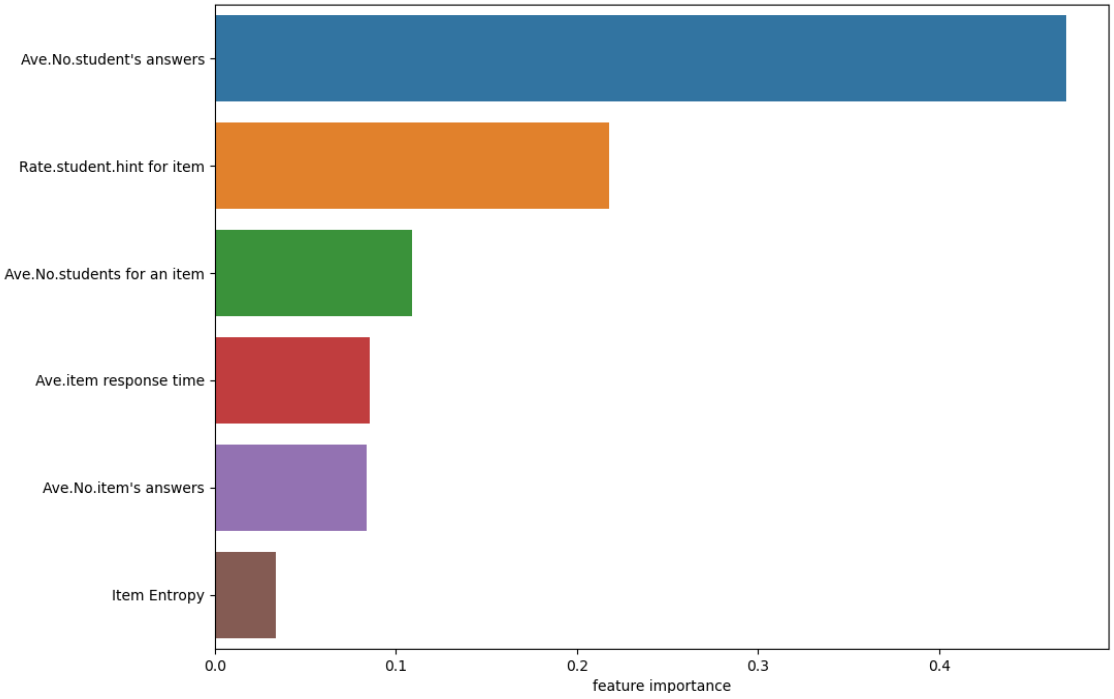


図 3. Deep-IRT の feature importance

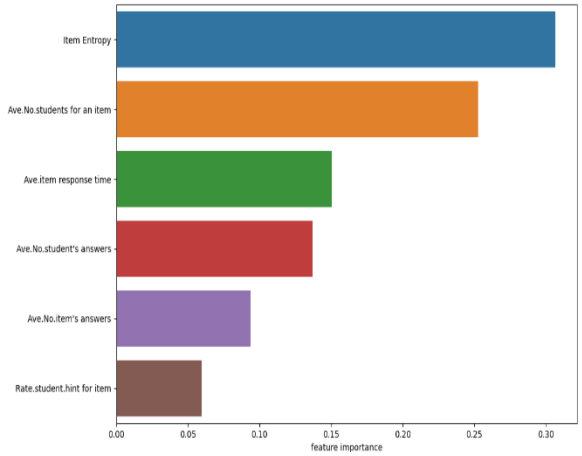
図 3 より、Ave.No.student's answers の AUC スコアに対する重要度が最も高いこと

が分かる。その大きさは二番目以下の要素と比べて大きく差を付けている。また、Ave.No.item's answers の値がそれほど高くないことから、Deep-IRT では、同じデータ数に関わる要因でも「項目一問あたりの解答数」よりも「学習者一人あたりの解答数」が重要なことが分かった。

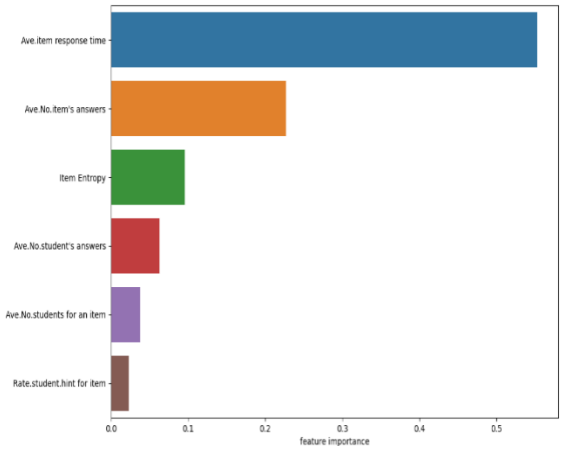
反対に Item Entropy の AUC スコアに対する重要度がとても小さいことがわかる。線形モデルである重回帰分析では大きな影響を与えていた Item Entropy が、なぜ非線形モデルであるランダムフォレストでは影響が小さいのかを考察する。

まず Item Entropy の AUC スコアに対する重要度が常に小さいのか、今回分析した四つのデータセットごとに再度 feature importance を求めた。すると以下の結果になった。

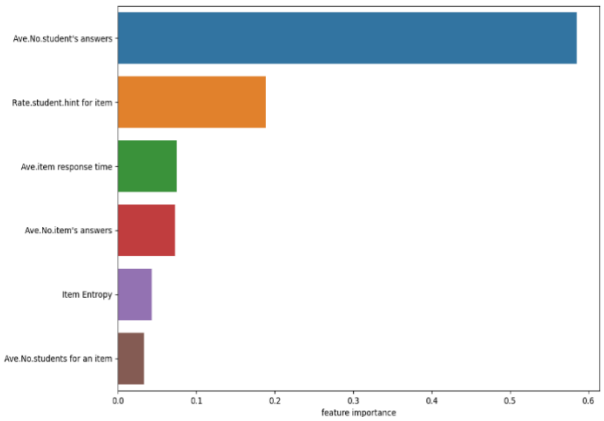
データセット 1



データセット 2



データセット 3



データセット 4

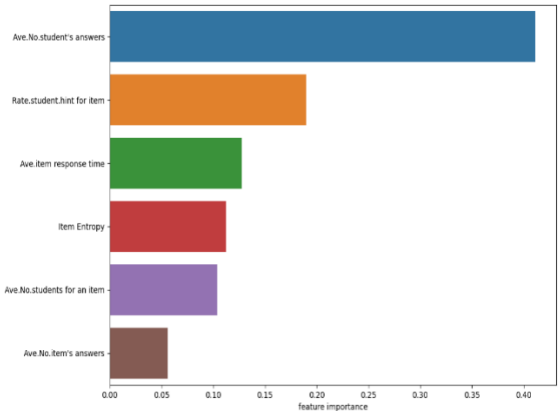


図 4.各データセットの feature importance

図 4 のデータセット 1 から常に Item Entropy の AUC スコアに対する重要度が小さいとは限らないことがわかる。では、どのような場合に Item Entropy の AUC スコアに対する重要度が大きくなるかを探るために、Item Entropy の標準偏差などをまとめた表を用意した。

表 3. Item Entropy **の各データ**

	平均	標準偏差	最大値	最小値
全体	0.83	0.16	1.0	0.56
データセット 1	0.85	0.14	0.99	0.63
データセット 2	0.80	0.18	1.0	0.56
データセット 3	0.84	0.16	1.0	0.63
データセット 4	0.83	0.18	1.0	0.58

図 5 と表 3 より、標準偏差が小さい時に Item Entropy の AUC スコアに対する重要度が大きい。このことから Item Entropy の標準偏差を小さくする要因となる要素は、AUC スコアに対する重要度が増すのではないかと考えることができる。

これらを確かめるために、重回帰分析とランダムフォレストの両方で AUC スコアに対する重要度が大きい Rate.student.hint for item と Ave.No. students for an item に対して、Item Entropy との関係を確かめる。

まず Rate.student.hint for item と Item Entropy との関係を確かめる。そのために「ヒントを見た場合における学習者の正答率の平均」と「ヒントを見ていない場合における学習者の正答率の平均」を表にしてまとめた。

表 4.ヒントを見た場合と見ていない場合の学習者の正答率

	ヒントを見た	ヒントを見ていない
全体	0.767	0.010
データセット 1	0.703	0.037
データセット 2	0.797	0.004
データセット 3	0.782	0.011
データセット 4	0.778	0.001

表 4 より学習者がヒントを見た場合、学習者の正答率が大きく下がっていることがわかる。また、図 4 において Rate.student.hint for item の AUC スコアに対する重要

度が小さかったデータセット 1 では、他のデータセットと比べてヒントを見ていない場合の学習者の正答率が高いこともわかる。以上より、Rate.student.hint for item の AUC スコアに対する重要度が高い理由は、学習者の正答率に偏りを生じさせ、Item Entropy の標準偏差を小さくするからであると考えることができる。

次に Ave.No. students for an item に注目する。こちらは全体のデータと Ave.No. students for an item の値が 3 以上であるデータに分けて学習者の正答率を求める。求めたものを表 5 にまとめた。

表 5. Ave.No. students for an item の値が 3 以上である学習者の正答率

	全体	Ave.No. students for an item の値が 3 以上
全体	0.691	0.005
データセット 1	0.699	0.015
データセット 2	0.705	0.002
データセット 3	0.684	0.001
データセット 4	0.678	0.001

表 5 より Ave.No. students for an item の値が 3 以上の場合、Rate.student.hint for item と同様に学習者の正答率に明確に差が出来ることが確認できた。このことから Ave.No. students for an item においても学習者の正答率を偏らせることにより、Item Entropy の標準偏差を小さくしていることがわかる。

以上の結果より、重回帰分析では高い影響を与えていた Item Entropy がランダムフォレストによる分析において AUC スコアに対する重要度が低かった理由は、「学習者の正答率」よりも「ヒントを見たかどうか」や「その項目に挑戦した回数」といった各項目に対する「学習者履歴」の方が AUC スコアに対して重要度が高くなっているからだと考えられる。

次に、「Ave.No.item's answers」の結果を見る。「Ave.No.item's answers」の feature importance における AUC スコアに対する重要度は高くなく、この結果と重回帰分析結果の結果から「Ave.No.item's answers」は AUC スコアに対してあまり影響を与えていないと考えられる。

最後に「Ave.item response time」に注目する。図 4 を見ると、「Ave.item response time」は全体的に AUC スコアに対して中程度の重要度を持つように見えるが、データセット 2 においては高い重要度があることがわかる。なぜこのような結果になったのかを考える。

まず、「Ave.item response time」が「Ave.No.student's answers」のようにAUCスコアと比例関係にあるかを確かめる。そのために「Ave.item response time」とAUCスコアの関係を図で示す。なお、「Ave.item response time」のグループ分けを行う数値は二分木図を参考にしている。

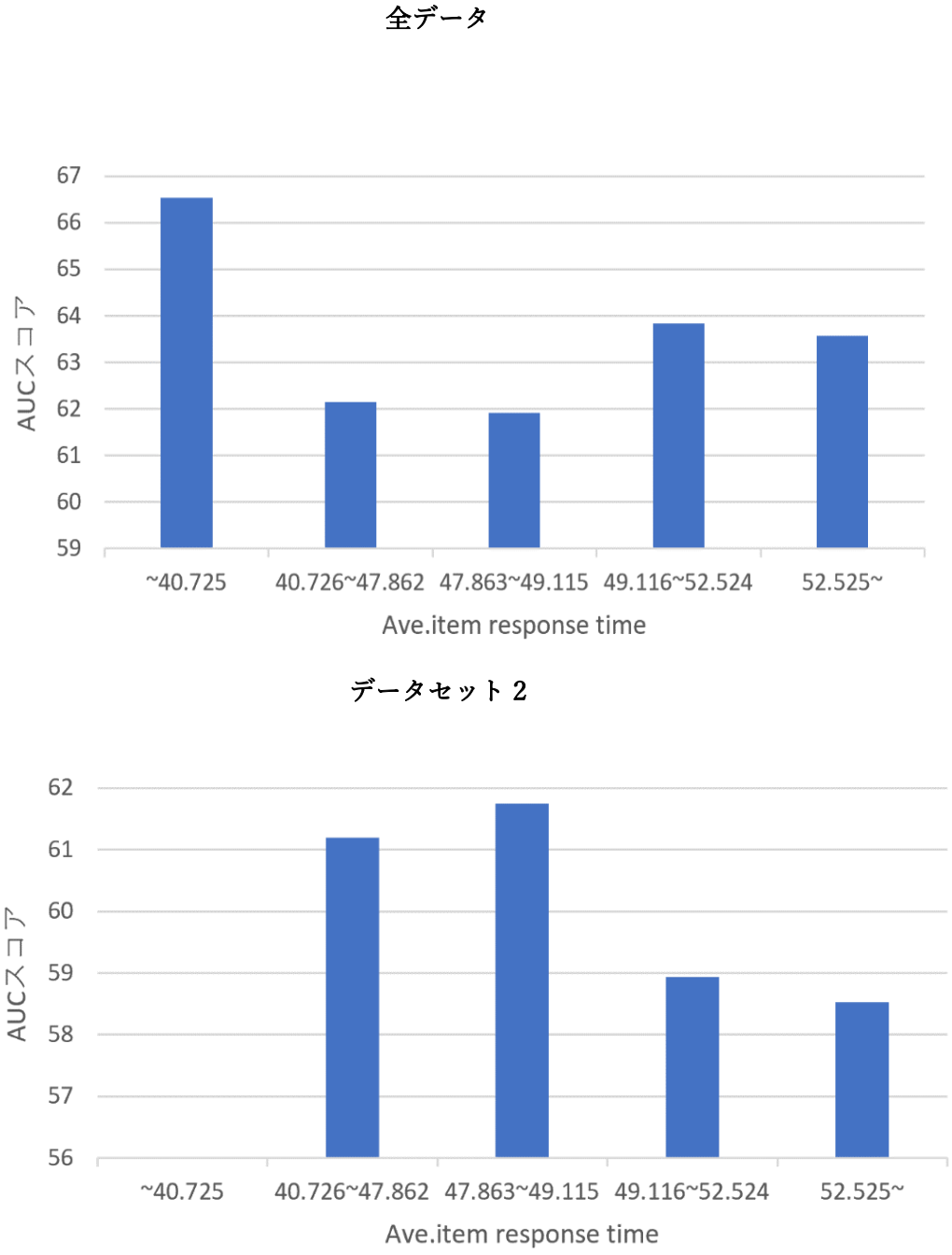


図 5. Ave.item response time と **AUC スコアの関係**

図 5 より、二分木図から推測した通り「Ave.item response time」と AUC スコアの間に相関関係はなく、それぞれが独立していることがわかる。また、全体とデータセット 2 を比較したところ、データの凹凸の形が異なるが、AUC スコアと関係がありそうな関係は見られなかった。

次に、「Item Entropy」の時のように、標準偏差が小さい時に AUC スコアに対する重要度が大きくなるかを確かめるために「Ave.item response time」の標準偏差などを求め、表 6 にまとめた。表 6 よりデータセット 2 の平均点、標準偏差が共に他のデータセットと比べて数値が低いことがわかる。このことから、「Ave.item response time」は「Item Entropy」同様に、標準偏差が小さい時に AUC スコアに強い影響を与えるのではないかと考えられる。その理由としては、解答までに掛かった時間が近いということは項目の難易度が近く、学習者の正答率に若干の偏りが生じたからだと考えることができる。

表 6. Ave.item response time **の各データ**

	平均	標準偏差	最大値	最小値
全体	48.11	4.09	53.94	40.473
データセット 1	48.28	4.91	53.94	41.58
データセット 2	47.38	3.48	53.03	41.50
データセット 3	48.26	3.84	52.62	40.473
データセット 4	48.51	3.99	53.885	40.976

3.まとめ

本章では、Deep-IRT に対して複数のデータセットを用いて学習者の反応予測を行い、その結果に対して重回帰分析とランダムフォレストによる分析を行った。ランダムフォレストによる分析では、二分木図を表示することにより分岐する条件を具体的な数値で表した。これらの分析により、まずこのモデルでは学習者一人当たりの解答数が AUC スコアに大きな影響を与えること、またその数値が大きいほど AUC スコアへ与える影響が大きいことがわかった。一方で「項目一問あたりの解答数の平均」は AUC スコアにあまり影響を与えないことがわかった。

更に学習者の正答率の二値エントロピーの AUC スコアに対する重要度が、重回帰分析とランダムフォレストで異なることを発見した。その理由は、全体の傾向を見る線形モデルである重回帰分析と異なり、パターン予測に基づいて予測を行う非線形モデルであるランダムフォレストでは、「学習者の正答率」そのものよりも、「ヒントを

見た回数」や「項目一問あたりの学習者の数」といった各項目に対する「学習者履歴」の方が AUC スコアに対して高い重要度を見出しているからだと考える。

　また「項目を解くために費やした時間」は、「学習者の正答率の二値エントロピー」の時と同様に標準偏差が小さいほど AUC スコアに対して高い重要度があると考えられる。

4.むすび

本研究では、Deep-IRT に適したデータセットの形を探るために複数のデータセットに対する結果から重回帰分析とランダムフォレストを用いて分析を行い、このモデルの特徴を分析した。

その結果、Deep-IRT の特徴として、「学習者一人あたりの解答数」の影響が大きく、更に学習者が解いた項目が多ければ多い程高い AUC スコアになる傾向があることがわかった。また、「ヒントを見た回数」や「項目一問あたりの学習者の数」といった、学習者個人の「各項目に対する学習者履歴」データが AUC スコアに対して高い重要度を示していることがわかった。よって Knowledge Tracing を比較する際には、学習者履歴も公平なものを用意する必要がある。

一方、「ヒントを見た回数」がなぜ Deep-IRT に影響を与えるのかは分かっておらず、今後の課題となる。

今回の研究のように各モデルの特徴を調べていくことにより、データセットを解析する際にどのモデルが最も適しているのかが分かるようになり、場面に適したモデルを使用することが出来るようになる。更には、それぞれの長所と短所を把握することにより、各モデルの更なる改良の余地も発見することが出来るのではないかと考える。

5.参考文献

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

[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)

[H.Sepp 97]H. Sepp and S. Jurgen, “Long short-term memory,” Neural Computation, vol.14, pp.1735–1780, 1997.

[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)

[tsutsumi et. al 21] 堤瑛美子, 木下涼, and 植野真臣. "独立な学習者・項目ネットワークを もつ Deep-IRT." 電子情報通信学会論文誌 D, 104(7), pp.596-608 (2021)

[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)