N-012

ディープラーニングを用いた Knowledge Tracing による模擬試験受験者の知識状態評価の試み
Estimating Knowledge State of Mock Exam Takers using Deep Learning in Knowledge Tracing

伊藤 彰† 倉本 卓† 市川 尚† 對間 博之†
Akira Itoh Taku Kuramoto Nao Ichikawa Hiroyuki Tsushima

1. はじめに

本学をはじめとした診療放射線技師養成校にとって、学生を国家試験合格に導くことは重要な役割の一つである。国家試験対策にあたっては学習の達成度を測るための模擬試験を複数回実施する。その結果を基に国家試験合格に向けての学習達成度を評価するが、模擬試験の素点はそれぞれのテストの難易度に影響され、受験者の真の能力を示すとは限らない。模擬試験の難易度の統制は困難であり、根拠をもって異なる実施回の試験結果を比較できないのが現状である。このように、ある難易度のテストで測られた結果を受験者の能力と見なす枠組みは古典テスト理論と呼ばれる。これに対して、テストの難易度と受験者の内在的な能力を分離して評価する枠組みに項目反応理論(Item Response Theory: IRT)がある。IRT を用いると、異なるテストを受験した受験者を同じ尺度で評価することができる。その際、等化やリンキングを行うが、困難度を測定した項目バンクの構築や受験者間で共通項目を課すなどテストのデザインに制約がある。大学単位で実施する国家試験受験のための模擬試験では規模の制約から問題バンクが未構築である。また、受験者が固定されているため共通問題を使った等化も非現実的である。さらに、受験者の知識状態が国家試験本番に向けて経時的に変化していくことも等化を困難にする。

オンライン教育の分野等では過去の学習の履歴から新たな項目への反応を予測する Knowledge Tracing(KT)が注目されている。これは知識状態の変化をモデル化するもので、国家試験対策のように時間経過とともに学習者の能力パラメータが変化する状況によく適応すると予想される。KT には確率モデルや因子分析、深層学習まで様々なモデルが試みられているが、近年は深層学習モデルが注目されている[1]。リカレントニューラルネットワーク(RNN)によって学習者の知識状態をモデル化する Deep Knowledge Tracing(DKT)は、学習者の学習履歴を元にした推論が可能である。

本研究では RNN や記憶セルを用いた KT 実装の前段階として、深層学習を利用した deep-IRT モデルの評価を行った。このモデルは受験者ネットワークと項目ネットワークの 2 つのニューラルネットワークをもち IRT と同様の解釈性をもつ。受験者ネットワークが出力する能力パラメータは項目ネットワークから出力する困難度と独立している。試験の難易度と独立して受験者の能力パラメータを評価できるかを評価し、RNN ベースのモデルに与えるパラメータとして適切かを検討した。

2. 背景

2.1 国家試験対策模試における IRT の利用状況

IRT は大規模テストにけるテスト評価に用いられるようになっているが、診療放射線技師教育で行われてきたテスト評価は古典テスト理論の適用に留まり IRT の利用は見られない。

これは IRT の適用に制約があることが一因である。IRT による安定したテスト分析には一般に数百から数千の被験者が必要とされる。各養成校の模擬試験の受験者は 100 名程度であり、サンプルサイズが不十分である。また、既存の模試問題は IRT による問題バンク化を考慮せずに作成、実施されており、試験結果同士を横並びで比較するための等化処理の適用が困難である。以上の背景から、等化処理を必要としない Deep IRT に着目した。

2.2 Deep IRT の概要

木下[2]らや Tsutsumi[3]らによって提唱された Deep IRT は、受験者ネットワークと項目ネットワークの二つの独立したニューラルネットワークをもつ。それぞれの出力を組み合わせることで受験者が項目に正答する確率を求める。受験者ネットワークは他の受験者の反応も利用して受験者の能力を推定する構造をとっており、従来の IRT が求める同一母集団からランダムサンプリングの必要がないとされる。

3. 実験の目的

問題バンク化を考慮せずに作成された模擬試験問題に対する Deep IRT の精度評価を行う。従来の IRT と比較し、少ない受験者数における精度について評価する。

4. 実験の方法

4.1 使用したコード

Deep IRT は Tsutsumi ら[3]による Github の公開リポジトリのコード用いた。コードはサポートが終了した Chainer フレームワーク上で書かれているため、PyTorch に移植した。ネットワーク構造やハイパーパラメータは変更していない。poor_student, excelent_student パラメータはそれぞれ 1 σ となるよう得点率を設定した。diff_questions, easy_questions パラメータも同様に正答率を設定した。

IRT は統計ソフト R の irtyo パッケージ(v.0.2.2)に含まれる est 関数を用いた。使用したモデルは 1PL モデルである。得点率 5%以下と 95%以上の問題はオミットした。

4.2 研究方法

ある実施回(テスト A)の国家試験模試に対する受験者の反応データを二分割し、Deep IRT, IRT それぞれで受験者能力値 θ を算出した。算出した値を平均 0、SD1 に正規化し、

二乗平均平方根誤差(Root mean square error: RMSE)で評価した。受験者の分割は乱数を用いて 13 通り行い、13 組の乱数シードを使用した。Deep IRT は python3, IRT は R と異なったプラットフォームで実行したが、R で作成したリストを Python に渡すことで分割を統制した。

試行に対して最低 10 回結果が得られた。これに対して Wilcoxon の順位和検定を行った。

4.3 使用したデータ

本学における国家試験模試に対する反応データを用いた。

4.4 倫理的配慮

本研究は研究課題名「国家試験模試の評価への項目反応理論の応用」について、神戸常盤大研究倫理委員会の承認を得ている。(承認番号: 神常大研倫第 24-21 号)

5. 結果

国家試験模試(テスト A)について Deep IRT と IRT で解析を行った。Deep IRT は全ての seed の分割法について解析が成立し、平均値は 0.175, SD は 0.038 であった。IRT は 10 通りの分割法について解析が成立し、平均値は 0.240、SD は 0.039 であった。

Wilcoxon の順位和検定を実施したところ、p 値は 0.00151 となり、二群の中央値は等しいとの帰無仮説は棄却された。

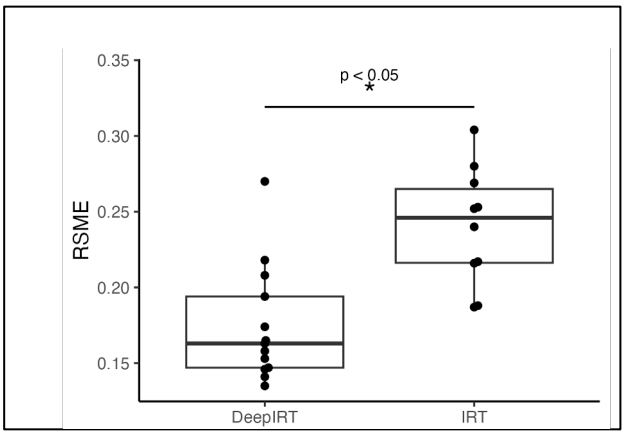

図 1　Deep IRT と IRT の精度比較(テスト A)

表 1 Deep IRT と IRT の精度比較(テスト A)

	seed1	seed2	seed3	seed4	seed5	seed6	seed7	seed8	seed9	seed10	seed11	seed12	seed13
IRT	0.216	0.187	0.24	0.188	0.304	0.253	0.28	0.217	0.252	fail	0.269	fail	fail
DeepIR	0.141	0.147	0.165	0.163	0.208	0.194	0.27	0.146	0.218	0.135	0.174	0.153	0.158

6. 考察

Deep IRT 法によって項目に対する学習者の反応を予測するには毎回学習が必要で比較的計算コストが高い。しかし、IRT による 1pl でのモデル化が失敗したケースでも Deep IRT は高い精度を保っている(表 1)。

IRT は正答確率をロジスティック関数と仮定しているため、この前提から外れた反応に対しては適切なモデルを提示できず解析が失敗する。Deep IRT 法は深層学習モデルなのでこの問題から逃れられると考えられる。現実のテストにおいて、正答確率の分布関数がシグモイド状になるとは限らないため、Deep IRT 法の有用性は高い。

7. おわりに

対象者数が少なく問題に特段の配慮がない場合でも Deep IRT 法が有用であることがわかった。引き続き KT の実装を試みる。

謝辞

研究にご協力いただいた研究参加者の皆様に深く御礼申し上げます。

参考文献

[1] Yeung, C.-K., “Deep-IRT: Make Deep Learning Based Knowledge Tracing Explainable Using Item Response Theory”, Proc. 12th Int. Conf. Educational Data Mining, EDM(2019).
[2] 木下涼, 植野真臣, “深層学習によるテスト理論：Item Deep Response Theory”, 電子情報通信学会論文誌 D J103-D, 314-329 (2020).
[3] Tsutsumi, E., Kinoshita, R. & Ueno, M, “Deep Item Response Theory as a Novel Test Theory Based on Deep Learning”, Electronics, Vol.10, No.9, 1020(2021)

†神戸常盤大学　Kobe-Tokiwa University