小論文の分析的評価のための項目反応理論を用いた深層学習自動採点手法

Deep Neural Automated Essay Scoring Integrating Multidimensional Item Response Theory for Analytic Scoring

柴田 拓海 *1，宇都 雅輝 *1
Takumi Shibata*1, Masaki Uto*1
*1 電気通信大学
*1The University of Electro-Communications
Email: {shibata, uto}@ai.lab.uec.ac.jp

あらまし： 近年，深層学習を用いた小論文自動採点手法として，全体得点と複数の評価観点に対応する細目得点を同時に予測する手法が提案されている．しかし従来手法は，評価観点ごとに複雑なニューラルネットワーク層を持つため，得点予測の根拠について解釈性が低いという問題があった．この問題を解決するために，本研究では多次元項目反応理論を組み込むことで予測根拠の解釈性を高めた複数観点同時自動採点手法を提案する．

キーワード： 記述・論述試験，自動採点，深層学習，多次元項目反応理論，説明可能性

1 はじめに

近年，小論文試験の採点をコンピュータを用いて自動化する小論文自動採点（Automated Essay Scoring；AES）手法が注目されており，深層学習に基づいた手法が多数提案されている (e.g.,[1])．従来の自動採点モデルの多くは全体得点のみを予測するが，学習評価場面などで小論文試験を運用する場合，詳細なフィードバックを受検者に与えるために複数観点に基づく分析的評価を行いたい場面がある．このような自動採点を実現する手法として，全体得点だけでなく複数の評価観点に対応する得点も同時に予測できるモデルが近年提案されている．

現時点では Ridley ら [1] のモデルが最高精度を達成しているが，このモデルには解釈性の観点から次のような問題がある．(1) 評価観点ごとに複雑な多層ニューラルネットワークを持つため予測根拠を解釈することが難しい．(2) 一般に評価観点は，背後に測定したい能力尺度を想定し，それを測定できるように設計されるが，このモデルでは複数評価観点の背後に想定される能力尺度を解釈することができない．これらの問題を解決するために，本研究では項目反応理論（Item Response Theory；IRT）を組み込んだ解釈性に優れた複数観点同時自動採点モデルを提案する．

2 提案手法

提案モデルは，Ridley らが提案した複数観点同時自動採点モデルを基礎モデルとする．提案モデルの概念図を図 1 に示す．提案モデルは受検者 n の小論文を入力とし，評価観点 $m \in \mathcal{M} = \{1, 2, \ldots, M\}$ に対応する得点 $\hat{y}_{nm}$ を出力する．ここで M は評価観点数を表す．また，受検者 n の小論文は単語系列として，$\{w_{nsl} | s \in \{1, 2, \ldots, S\}, l \in \{1, 2, \ldots, l_s\}\}$ と表せる．w_{nsl} は受検者 n の小論文における s 番目の文の l 番目の単語であり，S はその小論文の文数，l_s は s 番目の文の単語数である．提案モデルは入力層から Concatenate 層まで評価観点数 $M = 1$ とした従来モデルと同じ構造

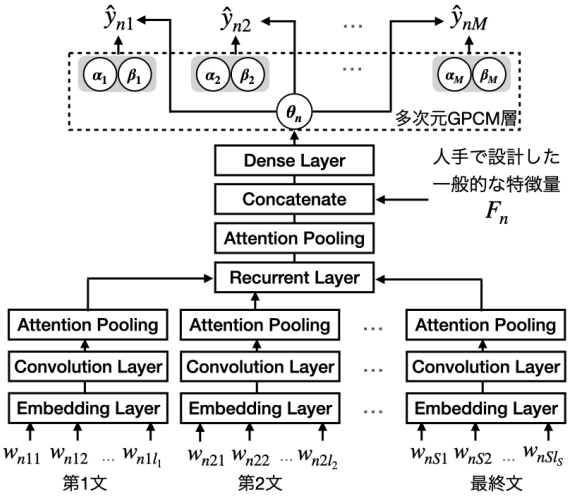

図 1 提案モデルの概念図

を持ち，これらの層を用いて受検者 n の文章単位の分散表現 $\boldsymbol{h}_n$ を生成する．各層の詳細は文献 [2] を参照されたい．

提案モデルでは，この文章単位の分散表現 $\boldsymbol{h}_n$ から各評価観点の予測得点を計算する出力層として，代表的な多次元多値型 IRT モデルである多次元一般化部分採点モデル（Generalized Partial Credit Model；GPCM）[3] を用いる．ここでは各評価観点を項目とみなして多次元 GPCM を適用する．具体的には受検者 n が評価観点 m において，得点 $k \in \{1, 2, \ldots, K_m\}$ を得る確率を次式で与えるモデルを適用する．

$$P_{nmk} = \frac{\exp(k\boldsymbol{\alpha}_m^T \boldsymbol{\theta}_n + \sum_{u=1}^{k} \beta_{mu})}{\sum_{v=1}^{K_m} \exp(v\boldsymbol{\alpha}_m^T \boldsymbol{\theta}_n + \sum_{u=1}^{v} \beta_{mu})} \quad (1)$$

ここで，$\boldsymbol{\theta}_n = (\theta_{n1}, \theta_{n2}, \ldots, \theta_{nd})$ は受検者 n の d 次元の能力を表すパラメータベクトルであり，ベクトルの各要素は各次元の能力値を表す．$\boldsymbol{\alpha}_m = (\alpha_{m1}, \alpha_{m2}, \ldots, \alpha_{md})$ は $\boldsymbol{\theta}_n$ に対応した評価観点 m の d 次元識別力，β_{mu} は評価観点 m においてカテゴリ $u-1$ から u に遷移する困難度を表すパラメータである．K_m は，評価観点 m における得点段階数を表す．なお，モデルの識別性のために，$\beta_{m1} = 0 : \forall m$ を所与とする．

表 1　課題別の平均 QWK スコア

モデル	課題番号									p 値		
	1	2	3	4	5	6	7	8	Avg.	提案-1dim	提案-2dim	提案-3dim
従来モデル	**0.685**	**0.655**	**0.660**	0.720	**0.706**	**0.750**	0.694	0.568	**0.680**	0.009	0.699	0.014
提案-1dim	0.656	0.617	0.620	0.713	0.689	0.731	0.638	0.549	0.652	-	0.180	0.378
提案-2dim	0.666	0.631	0.637	**0.722**	0.699	0.732	**0.704**	**0.576**	0.671	-	-	1.000
提案-3dim	0.679	0.633	0.642	0.704	0.698	0.734	0.696	0.553	0.667	-	-	-

提案モデルでは，Concatenate 層で得られた分散表現 $\boldsymbol{h}_n$ に対して全結合層を適用することで多次元 GPCM の能力パラメータ $\boldsymbol{\theta}_n$ を求め，それを用いて，式 (1) を計算することで，各評価観点 $m \in \mathcal{M}$ に対する得点の出力確率を計算する．得点予測の際には，期待得点 $\sum_{k=1}^{K_m} kP_{nmk}$ を予測得点とする．損失関数には，多クラス交差エントロピー誤差を用いる．なお，モデルの各種ハイパーパラメータは先行研究 [1] に合わせ，最適化アルゴリズムには学習率を 0.001 に設定した RMSProp を用いる．

3　実験

本研究では実データとして，AES 研究の分野で広く利用される Automated Student Assessment Prize (ASAP) と ASAP++ を用いる．これらのデータセットには 8 つの小論文課題に関する答案が含まれており，それぞれの答案に対して全体得点と 4 から 6 種類の評価観点別の得点が付与されている．小論文数の課題ごとの平均は約 1622，平均単語数は 275 である．

3.1　得点予測精度の評価実験

ここでは提案モデルの次元数を 1，2，3 と変化させて得点予測精度を評価する実験を行う．モデルの性能評価は，課題ごとに独立して 5 分割交差検証で行う．エポック数は全てのモデルで 30 としている．評価指標には，2 次の重み付きカッパ係数（Quadratic Weighted Kappa；QWK）を用いる．実験結果を表 1 に示す．表 1 では観点ごとに QWK スコアを計算し，その平均スコアを課題ごとに示している．各条件で最も精度が高い手法の結果を太字で示してある．表 1 より，提案モデルについては 2 次元の能力を仮定した場合が，最も平均精度が高いことがわかる．精度が最も高いのは従来モデルであるが，提案モデルと大きな差はないことが読み取れる．ここで各モデルの平均スコアに有意な差があるかを定量的に測定するため，ボンフェローニ法による多重比較検定を行った．結果を表 1 の「p 値」列に示す．表から最適な次元数を持つ 2 次元の提案モデルと従来モデルには有意な差が見られないことがわかった．このことから提案モデルは高々 2 次元で従来モデルと比較して精度を落とさずに得点予測ができたことがわかる．

3.2　評価観点パラメータの解釈

ここでは，提案モデルで推定された評価観点パラメータの解釈について述べる．例として表 2 に課題 2 のデータにおいて，能力次元数を 2 次元としたときの評価観点パラメータの推定結果を示した．

表 2　提案モデル（2 次元）を用いて推定した課題 2 の評価観点パラメータ

評価観点	α_{21}	α_{22}	β_{22}	β_{23}	β_{24}	β_{25}	β_{26}
全体得点	3.24	0.20	-6.56	-4.29	0.09	4.64	6.12
Content	2.27	2.13	-6.21	-2.02	0.81	2.97	6.28
Organization	2.48	2.00	-5.72	-1.58	1.52	3.30	7.26
Word Choice	1.75	2.86	-6.05	-2.06	1.18	3.82	6.88
Sentence Fluency	1.15	3.09	-6.39	-3.27	0.32	3.63	6.75
Conventions	1.00	2.90	-5.57	-2.09	0.90	3.75	7.04

表 2 に示した識別力値 $(\alpha_{21}, \alpha_{22})$ を分析することで，各観点がそれらの能力をどの程度の精度で測定できるか解釈できるとともに，各次元がどのような能力を測定しているかを把握することができる．例えば，全体得点，Content，Organization は 1 次元目の識別力値が 2 次元目の値に比べて高く，他の観点では 2 次元目の識別力値の方が高いことが読み取れる．このことから，1 次元目は，全体得点，Content，Organizaiton に対応する能力を表しており，2 次元目はその他の観点に共通する能力を表現していると解釈できる．観点の内容を考慮すると，1 次元目は内容面に重視した評価軸であり，2 次元目は文章表現に重視した評価軸と解釈できる．また，観点別に識別力値を確認すると，例えば，全体得点は 1 次元目の能力はよく測定できるが，2 次元目の能力測定には全く寄与しないことや，Content は両方の能力測定に寄与していること，などが読み取れる．また困難度パラメータ $(\beta_{22}, \ldots, \beta_{26})$ からは，各観点における各得点の出現分布を解釈することができる．例えば，上記の特性値から全体得点は得点に中心化傾向があることがわかる．このように提案モデルでは得点予測の背後にある構造を解釈できることがわかる．より詳しい解釈については，文献 [2] を参照されたい．

4　まとめ

本研究では全体得点と同時に観点別得点も予測できる自動採点手法に，多次元項目反応理論を組み込んだ手法を提案した．また本研究で使用した実データは高々 2 次元の能力尺度しか測定していない可能性が示唆された．

参考文献

[1] Robert Ridley, Liang He, Xin-yu Dai, Shujian Huang, and Jiajun Chen. Automated cross-prompt scoring of essay traits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35, pp. 13745–13753, 2021.

[2] 柴田拓海, 宇都雅輝. 深層学習と多次元項目反応理論を用いた小論文の観点別自動採点. 日本行動計量学会第 50 回大会, 2022.

[3] Lihua Yao and Richard D. Schwarz. A multidimensional partial credit model with associated item and test statistics: An application to mixed-format tests. *Applied Psychological Measurement*, Vol. 30, No. 6, pp. 469–492, 2006.