ピアアセスメントにおける項目反応理論

Item response theory for peer assessment

宇都雅輝 *1， 植野真臣 *1
Masaki Uto*1, Maomi Ueno*1
*1 電気通信大学大学院情報システム学研究科
*1Graduate School of Information System, University of Electro Communication
Email: uto_masaki@ai.is.uec.ac.jp

あらまし： ピアアセスメントにおける課題として，評価の信頼性が評価者の特性に依存する問題が指摘されている．評価者の特性を考慮した評価のために，これまで評価者パラメータを付加した項目反応理論が提案されている．しかし，多数の評価者が存在し，評価者と学習者の数が同程度となるピアアセスメントでは，パラメータ数に対するデータ数が少ないため，パラメータ推定の頑健性が保証されずこれらを利用することは困難である．そこで，本論では，通常の項目反応理論について，できる限りパラメータ推定の頑健性を維持しながら，評価者パラメータを付加した，ピアアセスメントのための新たな項目反応理論を提案する．
キーワード： ピアアセスメント，項目反応理論，評価者バイアス，モデル選択，パラメータ推定

1 はじめに

近年，CSCL(Computer Supported Collaborative Learning) などの協調学習において，学習者同士による学習成果物の相互評価を指すピアアセスメント [1] と呼ばれる評価手法の利用が注目されている．ピアアセスメントには，多くの利点が報告されており [2], 支援システムも多数開発されている．

一方，ピアアセスメントにおける課題のひとつとして，評価の信頼性が評価者の特性に依存する問題が指摘されている (例えば，[2])．具体的には，(1) 評価者間で評価の甘さ/辛さが存在すること，(2) 評価者間あるいは評価者内で評価基準が一貫している保証がないことなどが，信頼性の低下を引き起こすことが報告されている [3].

これらの影響を考慮した評価のために，論述式テストなどの評価において，評価者パラメータを付加した項目反応理論が提案されてきた [3][4]．しかし，多数の評価者が存在し，評価者と学習者の数が同程度となるピアアセスメントでは，パラメータ数に対するデータ数が少ないため，パラメータ推定の頑健性が保証されずこれらを利用することは難しい．

ピアアセスメントのための項目反応理論としては，Ueno et al.[2] があるに留まる．ここでは，パラメータ推定の問題解決を目標としているが簡易法であり推定精度が悪い．

そこで，本論では，通常の項目反応理論について，できる限りパラメータ推定の頑健性を維持しながら，評価者パラメータを付加した，ピアアセスメントのための新たな項目反応理論を提案する．提案手法の特徴として，(1) パラメータ数が改善されたことで，既存手法より頑健なパラメータ推定が可能である点，(2) 評価者特性として評価の一貫性と厳しさの影響を反映した学習者の能力推定が可能である点，(3) これらの結果，学習者の正確な能力推定が期待できる点，などが挙げられる．本論では，シミュレーション実験により提案手法の有効性を示す．

2 ピアアセスメントにおける項目反応理論

本論では，課題 $i(i = 1,\cdots,I)$ に対する学習者 $j(j = 1,\cdots,J)$ の成果物に対し，評価者 $r(r = 1,\cdots,R)$ が与える評価カテゴリ $k(k_i = 1,...,K)$ の集合をデータとする．

Patz et al.[4] は，論述式テストの評価のために，多値型項目反応理論の一つである一般化部分採点モデル (Generalized Partial Credit Model) を拡張した以下のモデルを提案している．

$$P_{ijrk} = \frac{\exp\sum_{m=0}^{k}[\alpha_i(\theta_j - \beta_{im} - \rho_{ir})]}{\sum_{l=0}^{K}\exp\sum_{m=0}^{l}[\alpha_i(\theta_j - \beta_{im} - \rho_{ir})]}$$

ここで，θ_j は学習者 j の能力パラメータ，a_i は項目 i の識別力，β_{ik} は項目 i におけるカテゴリ $k-1$ からカテゴリ k への遷移の難しさを表すステップパラメータ (ただし $\beta_{i0} = 0$)，ρ_{ir} は項目 i における評価者 r の評価の厳しさを表す．

宇佐美 [3] は，評価者内/評価者間で評価が一貫している保証がないことを指摘し，これに対応する評価者パラメータを加えた以下のモデルを提案している．

$$P_{ijrk} = \frac{\exp\sum_{m=0}^{k}[\alpha_i\alpha_r(\theta_j - (\beta_i + \beta_r) - d_{im}d_r)]}{\sum_{l=0}^{K}\exp\sum_{m=0}^{l}[\alpha_i\alpha_r(\theta_j - (\beta_i + \beta_r) - d_{im}d_r)]}$$

ここで，α_r は評価者 r の評価の一貫性，β_i は項目 i の位置パラメータ，β_r は評価者 r の位置パラメータ，d_{jk} は項目 i におけるカテゴリ k の閾値パラメータ，d_r は評価者 r の閾値パラメータを表す．ただし，パラメータの識別性のために，$\Pi_r\alpha_r = 1$，$\sum_r \beta_r = 0$，$\Pi_r d_r = 1$ を仮定する．

一方，Ueno et al.[2] は，ピアアセスメントのために，段階反応モデル (Graded Response Model) を拡張した以下のモデルを提案している．

$$P_{ijrk} = P^*_{ijrk-1} - P^*_{ijrk}$$

ただし，$P^*_{ijrk} = [1 + \exp(-a_i\theta_j + b_i + \varepsilon_{r,k})]^{-1}$ $(k = 1,\cdots,K-1)$，$P^*_{ijr0} = 1$，$P^*_{ijrK} = 0$ である．b_i は課題 i の難易度，$\varepsilon_{r,k}$ は評価者 r による評点 k への厳しさを表す．ただし $\varepsilon_{r,0} < \varepsilon_{r,1} < \cdots < \varepsilon_{r,K-1}$.

しかし，これらのモデルは，多数の評価者が存在し，評価者と学習者の数が同程度となるピアアセスメントでは，パラメータ数に対するデータ数が少ないため，パラメータ推定の頑健性が保証されず利用できない．

3 提案モデル

上記の問題を解決するために，ここでは，通常の項目反応理論に対し，できる限りパラメータ推定の頑健性を維持しな

がら，評価者パラメータを付加した以下の項目反応モデルを提案する．

$$P_{ijrk} = P^*_{ijrk-1} - P^*_{ijrk}$$

ただし，$P^*_{ijrk} = [1 + \exp(-a_i a_r(\theta_j - b_{i,k} - \varepsilon_r))]^{-1}$ $(k = 1, \cdots, K-1)$ とする．$b_{i,k}$ は課題 i において評点 k を得る難易度，ε_r は評価者 r の評価の厳しさを表す．ただし $b_{i,0} < b_{i,1} < \cdots < b_{i,K-1}$.

ここで，提案モデルのパラメータの解釈を示すために，評価者特性の異なる 2 人の評価者による反応曲線を図 1 に示す．ここでは，課題パラメータを $a_i = 1.5, b_{i0} = -1.5, b_{i1} = -0.5, b_{i2} = 0.5, b_{i3} = 1.5$ とし，評価者パラメータを，Rater1(左図) は $a_r = 1.5, b_r = 1.0$，Rater2(右図) は $a_r = 0.8, b_r = -1.0$ とした．

図 1 では，横軸に学習者の能力 θ_j，縦軸に各評価者がそれぞれの評価カテゴリを付与する確率を表す．図より，どちらの評価者も，能力が低い学習者には低い評点を与える確率が高く，能力が高い学習者には高い評点を与える確率が高いことがわかる．一方，評価者間の差異として，評価の一貫性が高い Rater1 は，Rater2 に比べ，学習者の能力の小さな差異により各評価カテゴリへの反応確率が大きく変動しており，能力の違いを精度良く識別することがわかる．更に，Rater1 は，評価の厳しさパラメータが大きく，反応曲線が全体として右に移動しており，高い評点を与えるために必要な能力が高くなっている．以上より，Rater1 は一貫した厳しい基準で評価を行っており，Rater2 は，評価の一貫性があまり保たれておらず，評価も甘い傾向があることがわかる．ここでは，評価者特性の分析例を示したが，課題特性についても同様の分析が可能である．

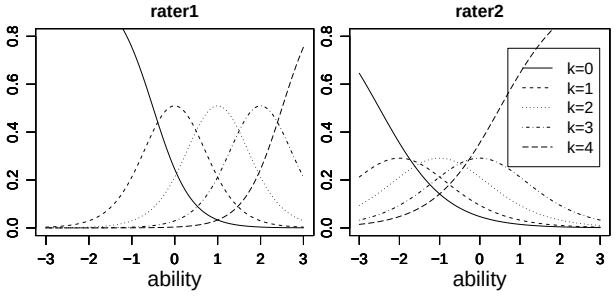

図 1　異なる評価者による提案モデルの反応関数

次に，既存モデルと提案モデルのパラメータ数を比較する．提案モデルは，$K = 5$ の場合，$2R > 3I$ かつ $I > 2$ の条件を満たすと，パラメータ数が最小となる．これらの条件式は，課題数に対し評価者数が多いピアアセスメントでは，一般に満たされる．このことは，パラメータ推定の頑健性において提案モデルが最も優れていることを意味している．

4　モデル評価

本章では，シミュレーションデータを用いた実験により提案モデルの有効性を評価する．実験手順は次の通りである．(1)$I = 5$，$K = 5$ の設定において最も複雑な Patz et al.[4] のモデルを用いて，$R = J = 10, 20, 50$ のデータを発生させた．(2) 生成したデータを用いて，各モデルのパラメータ推定を，MCMC の一種であるギブス内メトロポリスヘイスティング [4] により行った．(3) 能力パラメータの推定精度を評価するために，データの発生に用いた学習者の能力パラメータの真値 θ^* と各モデルにおける推定値 $\hat{\theta}$ の平均平方二乗誤差 (RMSE) を算出した．(4) モデルの適切性を評価するために，モデル選択基準の一つとして知られる BIC(Bayesian Information Criterion) を各モデルについて算出した．結果を表 1 に示す．

BIC では値が小さいモデルを最適モデルとみなす．表 1 より，データ数が少ないときは，提案モデルが最適モデルとして選択されており，データ数が増大すると，複雑なモデルの当てはまりが次第に良くなり，最終的に真のモデルである Patz が最適モデルとして推定されている．ピアアセスメントでは少数データからの推定が重要であることから，提案モデルが有効であることがわかる．

更に，表 1 より，提案モデルによる学習者の能力パラメータの推定結果が，真のモデルとほぼ同等の精度を与えていることがわかる．このことから，本論の目的である学習者の正確な能力推定においても，提案モデルが優れていることが示された．

表 1　モデルの評価結果

J=R		Patz	Usami	Ueno	Proposed
10	BIC	775	746	786	738
	RMSE	0.213	0.432	0.349	0.262
20	BIC	2691	2565	2831	2653
	RMSE	0.091	0.373	0.459	0.189
50	BIC	15009	14982	16375	15421
	RMSE	0.211	0.257	0.548	0.226

5　まとめと今後の課題

本研究では，ピアアセスメントにおける新たな項目反応モデルを提案した．評価実験により，少数データからの推定において，提案モデルが最も頑健なパラメータ推定を実現でき，学習者の能力の推定精度に優れることを示した．今後は，実データを用いた実験により，提案モデルの有効性を確認していく．

参考文献

[1] Topping K.J, Smith E.F, Swanson I, Elliot A, “Formative Peer Assessment of Academic Writing between Postgraduate Students.”Assessment & Evaluation in Higher Education, vol.25, no.2, p149-169, 2000

[2] Ueno, M, Okamoto, T, “Item Response Theory for Peer Assessment.” Advanced Learning Technologies. ICALT ’08 , 2008

[3] 宇佐美 慧. “採点者側と受験者側のバイアス要因の影響を同時に評価する多値型項目反応モデル：MCMC アルゴリズムに基づく推定.” 教育心理学研究 58(2), 163-175, 2010

[4] Patz, R.J., Junker, B.W. “Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses.” Journal of Educational and Behavioral Statistics, 24, pp.342-366, 1999