ピアアセスメントの精度を最適化する自動グループ構成システム

Automated grouping system to maximize peer assessment accuracy

宇都雅輝 *1，Nguyen Duc Thien*1，植野真臣 *1
Masaki Uto*1, Nguyen Duc Thien*1，Maomi Ueno*1
*1 電気通信大学
*1The University of Electro-Communications
Email: uto@ai.lab.uec.ac.jp

あらまし： 近年，MOOCs に代表される大規模 e ラーニングの普及に伴い，ピアアセスメントを学習者の能力測定に用いるニーズが高まっている．一般に，MOOCs のように学習者数が多い場合のピアアセスメントは，評価の負担を軽減するために学習者を複数のグループに分割してグループ内のメンバ同士で行わせることが多い．しかし，この場合，学習者の能力測定精度がグループ構成の仕方に依存する問題が残る．この問題を解決するために，本研究では，項目反応理論を用いて，学習者の能力測定精度を最大化するようにグループを構成するシステムを提案する．しかし，実験の結果，ランダムにグループを構成した場合と比べ，提案法が必ずしも高い能力測定精度を示すとは限らないことが明らかとなった．そこで，本研究では，グループ内の学習者同士でのみ評価を行うという制約を緩和し，各学習者に対して少数のグループ外評価者を割り当てる外部評価者選択システムを提案する．本システムにより外部評価者を数名追加することで，能力測定精度を大幅に改善できることを実験から示す．

キーワード： ピアアセスメント，項目反応理論，グループ構成，評価者選択，能力測定精度，最適化問題．

1 はじめに

近年，社会構成主義に基づく学習評価法として，学習者同士の相互評価法を表すピアアセスメントが注目されている．これまで，学習場面におけるピアアセスメントは，学習者同士でフィードバックを行わせることによる学習支援ツールとして活用されることが一般的であった．一方，近年では，MOOCs に代表される大規模 e ラーニングの普及に伴い，ピアアセスメントを学習者の能力測定に用いるニーズが高まっている．学習者数が大幅に増加すると，少数の教師がすべての学習者を評価することは困難になる．しかし，ピアアセスメントでは，学習者を複数のグループに分割してグループ内のメンバ同士で評価を行わせることで，教師や学習者の負担を大幅に増加させることなく評価を実施できる．また，社会構成主義の考え方に基づくと，能力とは，同一コミュニティのメンバが判断するものと解釈できるため，ピアアセスメントによる能力測定は妥当であるといえる．以上から，本研究では，ピアアセスメントを学習者の能力測定に用いる場合に着目する．

ピアアセスメントに基づく能力測定の課題として，その測定精度が評価者の特性（評価の厳しさ等）に依存する問題が知られている [1]．この問題を解決する手法の一つとして，数理モデルを用いたテスト理論の一つである項目反応理論（Item Response Theory: IRT）に対し，評価者特性を表すパラメータを付与したモデルが多数提案されてきた [2]．これらの IRT モデルでは評価者特性を考慮して学習者の能力を推定できるため，素点の合計や平均といった単純な得点化法に比べて高精度な能力測定が可能である [1, 2]．

一方で，上述のように，学習者数が多い場合のピアアセスメントは，評価の負担を軽減するために学習者を複数のグループに分割してグループ内のメンバ同士で行わせることが多い．しかし，この場合，IRT による能力測定精度がグループ構成の仕方に依存する問題が残る．例えば，評価の一貫性が低い評価者で構成されたグループでは，そのグループ内の学習者に対して高精度な能力測定は期待できない．

この問題を解決するために，本論文では，評価者特性パラメータを付与した IRT モデルを用いて，各学習者に対する能力測定精度を最大化するようにグループを構成するシステムを提案する．しかし，実験の結果，ランダムにグループを構成した場合と比べて，提案システムが必ずしも能力測定精度を改善できるとは限らないことが明らかとなった．これは，グループ内の学習者同士でのみ評価を行う場合，全ての学習者に情報量の高い評価者を割り当てるグループ構成は困難であることを示唆する．

そこで，本論文では，この制約を緩和し，グループ外から数名の外部評価者を導入し各学習者に割り当てる外部評価者選択システムを提案する．具体的には，学習者に対する外部評価者のフィッシャー情報量の下限を最大化する整理計画問題として外部評価者選択問題を定式化する．実験から，提案システムを用いて外部評価者を数名追加することで，グループ内の学習者のみによる評価に比べて，学習者の能力測定精度が大幅に改善されることが示された．一般に，外部評価は，評価の精度や客観性の改善に有効であることが知られており，本研究の結果もそれを支持したと解釈できる．

2 ピアアセスメントデータ

本研究では，講義開講期間中に I 個の課題 $i \in \{1, \cdots, I\}$ が提示され，個々の課題が完了する度にピアアセスメントを実施する場合を考える．ピアアセスメントは，学習者を G 個のグループ $g \in \{1, \cdots, G\}$ に分割して行うとし，グルーピングは課題 i ごとに変更すると仮定する．ここで，x_{igjr} を，課題 i において学習者 $j \in \{1, \cdots, J\}$ と評価者 $r \in \{1, \cdots, J\}$ が同一のグループ g に属する場合に 1，そうでない場合に 0 をとるダミー変数とすると，課題 i におけるグループは $\boldsymbol{X}_i = \{x_{igjr} \mid x_{igjr} \in \{0, 1\}\}$ と定義できる．

また，ピアアセスメントデータ $\boldsymbol{U}$ は，課題 i における学習者 j の成果物に評価者 r が与える評価カテゴリ $u_{ijr} \in \{1, ..., K\}$ で構成され，$\boldsymbol{U} = \{u_{ijr} | u_{ijr} \in \{-1, 1, \cdots, K\}\}$ と定義できる．ここで，$u_{ijr} = -1$ は欠測データを表し，グループ内の学習者同士でのみ評価を行う場合，$\sum_{g=1}^{G} x_{igjr} = 0$ となる j と r に対応する評価データは欠測データとなる．

本研究では，評価データ $\boldsymbol{U}$ から IRT により高精度に学習者の能力を測定できるように，グループ $\boldsymbol{X} = \{\boldsymbol{X}_1, \cdots, \boldsymbol{X}_N\}$ を最適化することを目指す．

3 ピアアセスメントにおける項目反応理論

本研究では，ピアアセスメントにおける IRT モデルとして，最も能力測定精度が高いことが報告されている Uto and Ueno[1] のモデルを用いる．このモデルでは，課題 i における学習者 j の成果物に，評価者 r が評価カテゴリー k を与える確率 P_{ijrk} を次式で定義する．

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

$$P^*_{ijrk} = [1 + exp[-\alpha_i\alpha_r(\theta_j - \beta_{ik} - \epsilon_r)]]^{-1} \quad (2)$$

ただし，$P^*_{ijr0} = 1$，$P^*_{ijrK} = 0$ とする．ここで，θ_j は学習者 j の能力，α_i は課題 i の識別力，β_{ik} は課題 i においてカテゴリ k を得る困難度（ただし，$\beta_{i1} < \cdots < \beta_{iK-1}$ と制約），α_r は評価者 r の評価の一貫性，ϵ_r は評価者 r の厳しさを表す．パラメータの識別性のために $\alpha_{r=1} = 1$ と $\varepsilon_1 = 0$ を仮定する．

IRT における能力測定精度は，フィッシャー情報量により評価されることが多い．ここで，課題 i において，能力 θ_j の学習者 j に対して評価者 r が与えるフィッシャー情報量を $I_{ir}(\theta_j)$ とすると，$I_{ir}(\theta_j)$ が大きくなる評価者 r は学習者 j を精度よく評価できると解釈できる．複数の評価者による情報量は，個々の評価者による情報量の和として計算できる．すなわち，各学習者に対する，同一グループ内の評価者情報量の総和が最大になるようにグループを構成することで，グループで行うピアアセスメントの能力測定精度を最適化できると期待できる．

4　グループ構成最適化手法

本研究では，各課題 i におけるグループ構成の最適化問題を，学習者に対する情報量の下限 y_i を最大化する以下の整数計画問題として定式化する．

$$\begin{aligned}
\text{maximize}: \quad & y_i \\
\text{subject to}: \quad & \sum_{\substack{r=1 \\ r\neq j}}^{J}\sum_{g=1}^{G} I_{ir}(\theta_j)x_{igjr} \geq y_i && \forall j \\
& \sum_{g=1}^{G} x_{igjj} = 1 && \forall j \\
& \sum_{g=1}^{G}(1 - x_{igjj})\sum_{r=1}^{J} x_{igjr} = 0 && \forall j \\
& n_l \leq \sum_{j=1}^{J} x_{igjj} \leq n_u && \forall j \\
& n_l \leq \sum_{g=1}^{G} x_{igjj}\sum_{r=1}^{J} x_{igjr} \leq n_u && \forall j
\end{aligned}$$

ここで，提案システムの有効性を確認するために，次のシミュレーション実験を行った．1）$J = 30$, $K = 5$, $I \in \{3, 5\}$ として，発生させた IRT モデルの真パラメータを所与として評価データをサンプリングした．2）$G \in \{4, 5\}$ について，提案法（*OptG* と呼ぶ）とランダム法（*RndG* と呼ぶ）でグループを構成し，異なるグループに属する学習者間の評価データを欠測させた．3）評価者と課題パラメータの真値を所与として，欠測データから学習者の能力を推定し，能力真値との RMSE を計算した．4）以上を 10 回繰り返し，RMSE の平均を求めた．

表 1　シミュレーション実験による能力測定精度の比較結果

				$n^R = 1$		$n^R = 2$	
I	G	*RndG*	*OptG*	*RndEx*	*OptEx*	*RndEx*	*OptEx*
3	4	0.368	0.360	0.343	0.297	0.325	0.287
	5	0.438	0.408	0.374	0.321	0.333	0.304
5	4	0.252	0.264	0.253	0.235	0.230	0.216
	5	0.298	0.307	0.299	0.253	0.259	0.241

結果を表 1 に示す．表より，提案システムが必ずしも能力測定精度を改善できていないことがわかる．これは，グループ内の学習者同士でのみ評価を行う場合，全ての学習者に情報量の高い評価者を割り当てるようなグループ構成は困難であることを示唆する．そこで，本研究ではさらに，この制約を緩和し，グループ外から数名の外部評価者を導入し各学習者に割り当てる外部評価者選択手法を提案する．

5　外部評価者選択手法

本研究では，外部評価者選択問題を，課題 i におけるグループ構成 $\boldsymbol{X}_i$ を所与として，各学習者に対する情報量の下限を最大化する整数計画問題として次のように定式化する．

$$\begin{aligned}
\text{maximize}: \quad & y_i \\
\text{subject to}: \quad & \sum_{r\in\boldsymbol{C}_{ij}} I_{ir}(\theta_j)z_{ijr} \geq y_i && \forall j \\
& \sum_{r\in\boldsymbol{C}_{ij}} z_{ijr} = n^R && \forall j \\
& \sum_{j=1}^{J} z_{ijr} \leq n^J && \forall r \\
& z_{ijj} = 0 && \forall j
\end{aligned}$$

ここで，z_{ijr} は，課題 i において学習者 j に評価者 r が割り当てられた場合に 1，そうでない場合に 0 をとる変数である．$\boldsymbol{C}_{ij}$ は課題 i における学習者 j の外部評価者集合を表す．また，n^R は各学習者に割り当てる外部評価者数，n^J は一人の評価者が担当するグループ外学習者数の上限を表す．

提案手法の有効性を評価するために，4 章と同様のシミュレーション実験を行った．ただし，本実験では，4 章の実験手順 2 において *OptG* で構成されたグループを所与とし，$n^R = \{1, 2\}$ 人の外部評価者を，提案法（*OptEx* と呼ぶ）とランダム法（*RndEx* と呼ぶ）で割り当てた．

実験結果を表 1 に示す．表 1 から，提案システムを用いて少数の外部評価者を導入することで，外部評価者を導入しない場合（*OptG*）やランダムに導入した場合（*RndEx*）に比べて測定誤差を大幅に減らすことができたことがわかる．

6　実データ実験

本研究では，被験者実験により収集した実際のピアアセスメントデータを用いて，シミュレーションと同様の実験を行った．ここでは，被験者数 $J = 34$，課題数 $I = 4$，カテゴリ数 $K = 5$ とした．なお，本実験ではパラメータ真値は未知のため，真値の代わりに完全データを用いた推定値を用いた．

表 2　実データ実験による能力測定精度の比較結果

			$n^R = 1$		$n^R = 2$	
G	*RndG*	*OptG*	*RndEx*	*OptEx*	*RndEx*	*OptEx*
4	0.241	0.259	0.236	0.210	0.226	0.197
5	0.287	0.323	0.295	0.255	0.272	0.206

結果を表 2 に示す．表 2 から，シミュレーション実験と同様の傾向が確認できる．すなわち，内部評価のみではグループ構成を最適化しても能力測定精度は必ずしも改善できないが，提案システムにより最適な外部評価者を導入することで精度を大幅に改善できた．

7　まとめ

紙面の都合上割愛したが，本研究では，学習者あたりの評価者数が一定となるように外部評価者を導入する手法や真パラメータが未知の場合の運用法なども開発し，より詳細な実験も行なっている．これらの詳細は [3] を参照されたい．

参考文献

[1] Masaki Uto and Maomi Ueno. Item response theory for peer assessment. *IEEE Transactions on Learning Technologies*, Vol. 9, No. 2, pp. 157–170, 2016.

[2] 宇都雅輝, 植野真臣. パフォーマンス評価のため項目反応モデルの比較と展望. 日本テスト学会誌, Vol. 12, No. 1, pp. 55–75, 2016.

[3] Masaki Uto, Nguyen Duc Thien, and Maomi Ueno. Group optimization to maximize peer assessment accuracy using item response theory. In *Proc. International Conference on Artificial Intelligence in Education*, 2017.