回答における偏りを考慮したクラウドソーシングの早期終了手法

五十嵐健夫† 佐藤一誠†

概要. 本稿では，クラウドソーシングによる意思決定において，利用するクラウドワーカーの数を動的に調整する方法を提示する． 基本的なアイデアは，これまでのクラウドワーカーの反応が十分に偏っていた場合，クラウドワーカーの募集を終了する，というものである． 本稿では，クラウドワーカーの募集をいつ終了するかを決定する基準を提示し，数値分析と実際のクラウドソーシングによる実験によってその有効性を検証する．これらの結果により，提案手法を用いることで，単純な終了基準を使用する標準的な方法と比較して，精度を維持しながら利用するワーカーの数を大幅に減らすことができることを示す．

1 はじめに

マイクロタスククラウドソーシングは，複数の候補の中から最良のビジュアルデザインを選択するなどの意思決定に使用することができる．特に，細かい意思決定（タスク）を大量に行わなければならない場合に効果的である．標準的なアプローチは，各タスクに対して固定数（例えば，n = 10）のクラウドワーカーを募集し，その多数を取ることで結論をだすというものである．しかし，最初のいくつかの回答がすでに片方に偏っている場合，より多くのクラウドワーカーを募集する必要はないと考えられる．本稿では，一対比較タスク（A または B を選択）に関して，上記の考え方に基づいた動的なクラウドワーカー数決定手法を提示する．

システムは，クラウドワーカーからの回答が得られる度に，これまでの回答の偏りを調べ，結論が既に明確かどうかを決定する．回答が十分に偏っていれば，システムは募集を終了し，これまでの回答に基づいて結論を導き出す．回答が十分に偏っていない場合，システムは次のワーカーを募集する．

単純なアプローチとして，A または B のいずれかがすでに予定している数の半分以上の回答を集めている場合に，ワーカーの募集を早期終了するという手法が考えられる[4]．このアプローチは，早期終了せずに予定数を集めてから過半数を取った場合と同一の結果を返すことが保証されている．この論文では，この簡単な早期終了手法を，「標準的手法」と呼ぶ．本稿では，標準的手法よりもさらにワーカーの数を減らすことのできる，よりアグレッシブかつ詳細な基準を提案する．

提案手法の有効性の検証として，数値解析と実際のクラウドソーシングによる実験評価を行った．それらの結果により，提案された方法が，結論の正確さを維持しながら，標準的手法に比べてクラウドワーカーの数を減らすことができることを示す．提案された方法は，汎用的でありかつ実装が容易なので，クラウドソーシングを意思決定に使用する場合に広く利用することができると考えられる．

2 関連研究

Crowdsourcing は，人間の知覚と判断を必要とする意思決定のための便利なプラットフォームを提供する．典型的なアプローチは，各タスクに対して一定数のクラウドワーカーを募集し，全ての回答を得た後に多数を取ることである． Gurari と Grauman [3]は，仕事の内容に基づいてクラウドワーカーの人数を変更するよう提案し，その方法が品質を失うことなく時間とコストを節約できることを示した．我々の手法は，このような手法と同じような目的を持っているが，数を節約する手がかりとしてタスクの内容の代わりに動的なクラウドワーカーの回答の偏りを使用するという，より一般的な方法となっている．

人工知能の研究コミュニティでは，ノイズを多く含むクラウドワーカーからの真のラベルを推論する方法が研究されている[2,8,10,9,6,5,11]．これらは，最適化プロセスにおいて各集団作業者の信頼性を考慮に入れることで，単純多数決よりもより正確な結論をだすことを実現している．しかし，これらの方法では，クラウドワーカーが複数のラベル付け作業の

† 東京大学

ために繰り返し募集され，すべての作業に対する作業者応答の分布を取ることによって，各作業のもっともらしいラベルを計算することを前提としている．したがって，これらの手法は，解くべき問題が単一の独立した問題であり，各ワーカーがその問題だけを処理するような状況には適用できない．

予算制約の下でタスクの数を最大にする手法なども提案されている[7]．この研究では，クラウドワーカーが動的に雇用される状況のもとで，発注者が予算制約およびタスク完了期限を考慮して，価格設定のプロセスを自動化する動的なメカニズム設計手法を提案している．

3 提案手法

クラウドワーカーの募集を終了するための基準をどのように設定するかについては，唯一絶対の基準はなく，いくつかの可能性が考えられる． 我々は，いくつかの選択肢を検討し，妥当性と実用性の観点から，もっとも適切と思われるものを選択した．

まず，以下の仮説を立てる．「クラウドワーカーの A と B の選択には偏りがない」

システムは，現在までのクラウドの反応を調べ，この仮説の下で，その反応が高い確率で成り立つ事象であるかどうかの判断を行う．答えが「高い確率で起こる事象である」の場合，まだ十分に偏っていないと判断され，システムはクラウドワーカーの募集を継続する． それ以外の場合は，仮説の下で「低い確率で起こる事象が発生した」ということなので，仮説の正しさが疑われ(十分偏っていると判断され)募集を停止する．

定式化：クラウドの反応は，独立同分布に従う確率変数であり，p_A と p_B はワーカーが A と B を選択する確率であると仮定する．なお，$p_A = 1 - p_B$ である．r_i を i 番目のクラウドワーカーの応答とする．

まず以下の仮定を置く．「クラウドの回答は完全にランダムである．($p_A = p_B = 1/2$)」

その仮定のもとで，$|N_i^A - N_i^B|$ を計算する．ここで N_i^A と N_i^B はこれまで得られている回答$\{r_1, r_2, \ldots, r_i\}$の中で，A が選ばれた回数と，B が選ばれた回数である．この時，事前にユーザが設定する意思決定の信頼度 δ に対して，以下の確率不等式が成立する．

$$\Pr\left\{|N_i^A - N_i^b| \leq \lambda_i\right\} > \delta$$

$$\lambda_i = \frac{1}{3}\left(1 + \sqrt{1 + 18i\left(\log\frac{1}{1-\delta}\right)^{-1}}\right)\log\frac{1}{1-\delta}$$

この式の導出については 付録に示す．

もし $|N_i^A - N_i^B|$ が閾値 λ_i 以下なら，現在の仮説の下で δ より高い確率で現在の観測が起こることを意味する（その仮説はおそらく正しい）．よって，システムはワーカーの募集を継続する．システムは，あらかじめ定められた最大数に達するまで上記の手順を繰り返し，多数決を適用する．しかし， $|N_i^A - N_i^B|$ が閾値 λ_i より大きい場合，仮説が間違っている可能性がある，つまり $p_A \neq p_B$ という可能性があると考えられる．この場合，システムはワーカーの募集を終了し，現在までに得られている回答において，過半数を取ることで結論を出す．

Algorithm 1 は，提案手法の擬似コードを示す．ここに示されているように，提案手法は，標準的手法（5-7 行目）に対して，さらに一つ条件分岐（8-11 行目）を追加するだけの簡単な実装となっている．

Algorithm 1 Majority Vote with Dynamic Termination

```
1:  procedure DYNAMIC(n : max_workers, δ : confidence)
2:      X ← {}
3:      for i = [1...n] do
4:          X.append(get_crowd_response())
5:          if X.count(A) > n/2 ∨ X.count(B) > n/2 then
6:              Break
7:          end if
8:          Compute λ_i with i and δ
9:          if |X.count(A) − X.count(B)| > λ_i then
10:             Break
11:         end if
12:     end for
13:     return (X.count(A) − X.count(B))?A : B
14: end procedure
```

この方法を使う場合には，信頼度 δ を手動で設定する必要がある． 高い信頼性を設定することは，より慎重な決定を行うことに対応しており，より多くの応答を収集しながら，より高い精度を実現する．低い信頼度を設定することは，より積極的な決定を行うことに対応し，この場合，利用するワーカーの数は少なくて済むが，精度が低下する可能性がある．

たとえば，n＝19[1]の場合には，δ ＝0.8 とすることで，標準的手法と比較して同程度の精度を維持しつつコストを削減することができる（次節参照）．大きな n の場合に，標準的手法に匹敵する精度を得るには，より大きな δ を使用する必要がある． 適切な δ は個々のユーザのニーズ(どの程度精度とコストを重視するか) に応じて異なってくるので， 利用しようとしている n に応じて，次節に示すような分析を実行することが推奨される．

4 数値分析

ここでは，数値解析によって，提案手法の性能を分析する． クラウドワーカーは確率 p で正解を，確率 1 - p で誤答を返すと仮定する． これは単純なベルヌーイ試行であるため，与えられた応答列の出現確率を 2 項分布として解析的に計算することができる．

$$P[s] = \binom{n}{k} p^k (1-p)^{n-k}$$

ただし，s はバイナリ応答列であり，n が s の長さ，k が s 中の正解の数を表す．

動的手法の期待精度の計算は，以下のように行う．まず，最大ワーカー数 (n) に等しい長さのバイナリ応答列をすべて列挙する（2^n通り）． 次に，各列に動的手法を適用する． 精度の期待値は，動的手法が正しい決定を返す列の確率を合計することによって与えられる．使用するクラウドワーカーの数の期待値も同時に計算することができる．Algorithm 2 に，この計算過程を示す. なお，ここでの DYNAMIC は，Algorithm 1 に示されているものに加えて，終了した時の変数 i の値を count として返すものとする．

Algorithm 2 Numerical Analysis of the Dynamic Method

```
for s ∈ all_the_possible_sequences_with_length_n do
    {answer, count} ← DYNAMIC(n, δ, s)
    if answer = correct_answer then
        accuracy+ = probability_of_s
    end if
    worker_count+ = count × probability_of_s
end for
```

同様の方法で標準的手法の期待精度と期待ワーカー数を計算する．具体的には， DYNAMIC (n, δ , s）の中身を，アルゴリズム 1 の 8-11 行目を削除して得られる，標準的手法に対応するものに置き換える．

図 1 に，異なるワーカー信頼度（各クラウドが正解を返す確率）における，精度の期待値（縦軸）と，ワーカー数の期待値（横軸）のトレードオフの関係を示す．青が標準的手法の結果を，紫が動的手法の結果を示す．これらの結果は，動的手法が，いずれのワーカー信頼度においても，より良い精度 - コストのトレードオフを達成することを示している．これらの結果に基づいて， n = 19 の標準的手法と同等の精度を達成するデフォルトの閾値として δ = 0.8 を選択することとする．

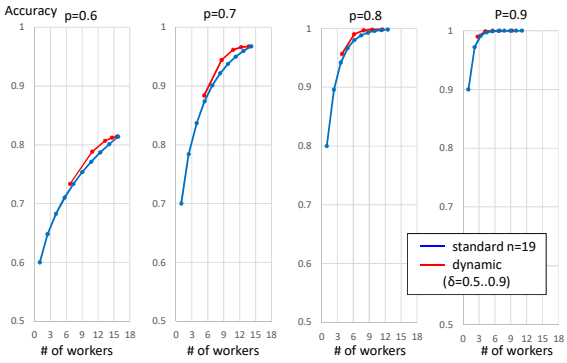

図 1: 標準的手法（n = 1-19）と動的手法（δ = 0.5-0.9）の正確さのトレードオフ． p はワーカーの信頼度を示す．

図 2 左は，異なるワーカー信頼度に対応する，標準的手法（n = 19）と動的手法（ δ = 0.7, 0.8 ）の期待精度を示している．これは，閾値 δ = 0.8 を持つ動的数値法の期待精度が， n = 19 （精度低下は 0.18 - 0.001%）の標準数値法の精度に非常に近いことを示している．閾値 δ = 0.7 の動的手法は，標準的手法よりわずかに低くなっている．

図 2 右は，異なるワーカー信頼度に対応する，標準的手法（n = 19）および動的手法（ δ = 0.7,

[1] 本論文では，提案手法の適用例として n=19 を主に例として扱う．主な理由としては，本手法がターゲットとしている，クラウドからのレスポンスによって処理を行うようなシステムで使われている数字が n=10 程度であること，一方で n が小さい場合（n=10 以下）では，7 節で述べるように本手法の効果があまり大きくないことなどから設定した数字である．20 にしなかったのは，偶数だと判断が割れる場合があり扱いが面倒なためである．

0.8 ）における，利用ワーカー数の期待値を示している．予想される通り，両方の手法において，信頼性が低い場合は多くのワーカーが必要であり，信頼性が高い場合は少ないワーカーで済むことが示されている．いずれの場合においても，動的手法は，標準的手法 n = 19 よりも少ないワーカー数となっていることが示されている（δ = 0.8 において 7 - 38% の削減）．

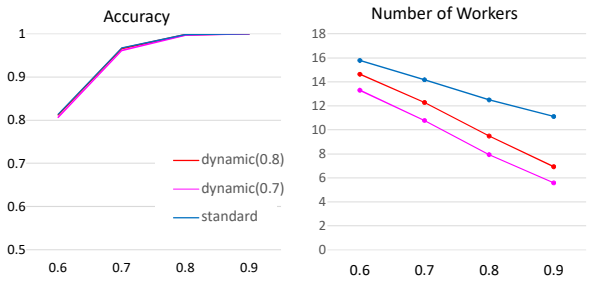

図 2: 標準的手法と動的手法の期待精度と期待ワーカー数． δ = 0.8 の動的手法の精度は，標準的手法の精度に非常に近い．

5 実験

ここでは，実際のクラウドソーシングで得られたデータを用いて，動的手法の有効性の検討を行う．得られる結果は，前節での分析から推測できるものといえるが，実際のデータを示すことも重要であると考えて分析を行った．

課題は，一対の画像を見て，どちらが好ましいかを回答するというものである．全部で 9 つのタスク（ペア）を用意し，各ワーカーは 9 回のタスクすべてを行った(図 3)．50 人のワーカーを雇用し，合計 450 件の回答を集めた．各ワーカーへの報酬は 100 円である．各タスクについて，すべての 50 件の回答に対して多数決をとったものを，正解とした． 9 つのタスクにおいて，正解を返したワーカーの割合は，それぞれ 0.84, 0.98, 0.98, 0.66, 0.52, 0.58, 0.72, 0.94, 0.94 であった．これは各タスクにおけるワーカーの信頼度と見ることができる．

図 3 実験に利用した画像．左側が ID1-5, 右側が ID6-9 に対応する．

各タスクについて，得られた 50 件の応答から 19 件の応答を無作為に抽出して 1 個の応答列を作り（例えば，最初のタスクであれば，1 応答あたり確率 p=0.84 で正解を返すような応答を 19 件並べることに相当する），その応答列に対して，標準的手法および動的手法（δ = 0.8）を用いて得られた結果の精度と効率を調べた．具体的には，各タスクについて，10 万個の応答列を上記の方法で生成し，平均精度と平均ワーカー数を計算した．

図 4 にその結果を示す．図 4 上は，2 つの方法の精度を比較しており，精度がほぼ同一であることを示している．図 4 下は，2 つの方法で使用されたクラウドワーカーの数を示しており，動的手法がクラウドワーカーの数を大幅に削減できていることを示している．予想される通り，ワーカーの信頼度が低い場合は，クラウドワーカーが多くなっている．これらの結果は，実際のクラウドソーシングにおける意思決定において，標準的手法とほとんど区別できない精度を維持しながら，動的方法を使用することによってコストを減らすことができることを示している．特に，ワーカーの信頼度 p を事前に設定する（推測する）ことなく，安定した改善を得られることが重要である．

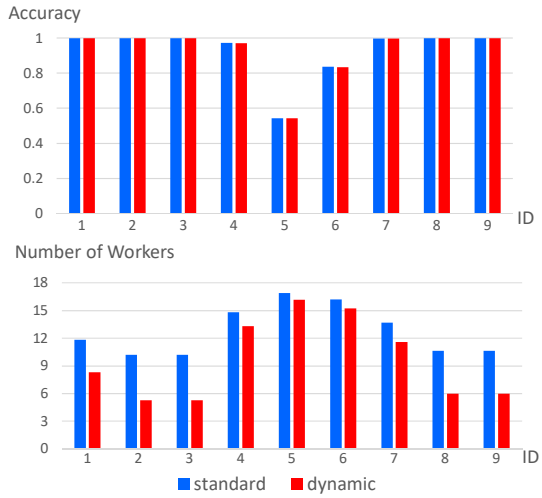

図 4　クラウドソーシングによる実験の結果．上は回答の精度を示し，下には利用したワーカーの数を示す．横軸はタスク ID を示す．

6　今後の課題

もともと少ない数のワーカーしか使わない場合には，標準的手法に対する動的手法の優位性はあまり大きくない．たとえば，$n=9$ の場合，正確性を維持しようとすると，ワーカーの削減数は 1 未満であった．提案手法の有用性は，より多くのワーカーに適用されるときにより明確になる．

現状，提案手法を使うにあたっては，対象とする最大ワーカー数 n に応じて，精度とコストのトレードオフを調べて適切な信頼度 δ を手動で設定する必要がある．ただ，「どの程度の信頼度 δ を与えると，標準的手法に対してどの程度精度を維持しながらどの程度ワーカー数を削減できるのか」は数値的に計算することができるので，そのような計算を自動で行って結果を返すようなインタラクティブな機能を実装することが考えられる．また別の問題として，現在の方法は，単純な選択を多数決で決定するために設計されており，応答の分布を知りたいような場合には直接適用できないという点が挙げられる．

提案手法の拡張として，A か B を決定するだけでなく，「決定できない」という結果も含むようなより細かい意思決定をサポートすることが考えられる．他の拡張としては，2 つ以上の選択肢の中からの選択を効率化することが考えられる．また，選択だけでなく，離散的または連続的な「数値スコア」を扱えるようにもしていきたいと考えている．

7　議論

本手法は，「マイクロタスククラウドソーシングを利用したビジュアルデザイン」のような，クラウドワーカーという人間をシステムとして一部含むようなインタラクティブなシステムの設計にあたって，その効率を向上する手法を提案するものである．

マイクロタスククラウドソーシング，および，ビジュアルデザイン支援，は，HCI コミュニティにおいて頻繁に議論される中心的なトピックの一つであり，その効率化を実現する手法の提示は，当該コミュニティに対する重要な貢献であると考えられる．

8　結論

本稿では，多数決に基づく意思決定を行うにあたって，クラウドワーカーの数を動的に調整する方法を提案した．提案手法では，ワーカーからの回答が得られる度にそれまでに得られている回答の分布を調べ，分布がすでに十分偏っている場合には，ワーカーの募集を停止して結論を出す．同じように早期に終了する，より単純な方法も知られているが，提案手法では，よりアグレッシブかつ詳細な終了判断基準を利用する．実際のクラウドソーシングを利用した実験により，本手法を用いることで，単純な方法と比較して，精度を維持しながらコストを大幅に削減できることを確認している．

9　付録

本文中における λ の式の導出をここに示す．
まず以下のような定理が知られている．

定理 1　(Bernstein Inequality [1] 2.8 節)
$\{X_k\}$が独立ゼロ平均確率変数であり，$|X_k|<=R$ がほとんど確実に成り立っている場合，以下の関係が成り立つ．

$$\forall t \geq 0,\ \lambda > 0 \quad \Pr\left\{\left|\sum_{k=1}^{t} X_k\right| > \lambda\right\} \leq \exp\left(-\frac{\lambda^2/2}{\sum_{k=1}^{t}\mathbb{E}[X_k^2] + R\lambda/3}\right)$$

また，以下のような δ を導入しておく．

$$1-\delta = \exp\left(-\frac{\lambda^2/2}{\sum_{k=1}^{t}\mathbb{E}[X_k^2] + R\lambda/3}\right)$$

$$\Leftrightarrow \lambda = \frac{R}{3}\left(1 + \sqrt{1 + 2\left(\frac{3}{R}\right)^2 \left(\log\frac{1}{1-\delta}\right)^{-1} \sum_{k=1}^{t}\mathbb{E}[X_k^2]}\right)\log\frac{1}{1-\delta}$$

ここで，クラウドソーシングによる回答に対応するものとして確率変数 X_t を導入する．

$$X_t = \begin{cases} +1 & \text{if select item } A \text{ with probability } \frac{1}{2} \text{ at step } t \\ -1 & \text{if select item } B \text{ with probability } \frac{1}{2} \text{ at step } t \end{cases}$$

この確率変数の合計 $\sum_{k=1}^{t} X_k$ は，A と回答した数と B と回答した数の差を表す．各ステップにおいて $|X_k|<1$ かつ $\Sigma_{k=1}^{t}\mathbb{E}[X_k^2]=t$ である．これを Bernstein inequality に当てはめると，

$$\Pr\left\{\left|\sum_{k=1}^{t} X_k\right| > \lambda\right\} \le \exp\left(-\frac{\lambda^2/2}{t+\lambda/3}\right)$$

となる．δ を使って書き換えると

$$\Pr\left\{\left|\sum_{k=1}^{t} X_k\right| > \lambda\right\} \le 1-\delta$$

$$\lambda_t = \frac{1}{3}\left(1+\sqrt{1+18t\left(\log\frac{1}{1-\delta}\right)^{-1}}\right)\log\frac{1}{1-\delta}$$

が得られる．

謝辞

本研究は JST CREST JPMJCR17A1 の支援を受けたものである．

参考文献

[1] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. 2013. Concentration Inequalities: A Nonasymptotic heory of Independence. Oxford University Press.

[2] Alexander Philip Dawid and Allan M Skene. 1979.Maximum likelihood estimation of observer error-rates using the EM algorithm. Applied statistics (1979), 20–28.

[3] Danna Gurari and Kristen Grauman. 2017. CrowdVerge: Predicting If People Will Agree on the Answer to a Visual Question. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. ACM, 3511-3522.

[4] Greg Little, Lydia B Chilton, Max Goldman, and Robert C Miller. 2009. Turkit: tools for iterative tasks on mechanical turk. In Proceedings of the ACM SIGKDD workshop on human computation. ACM, 29-30.

[5] Chao Liu and Yi-Min Wang. 2012. TrueLabel + Confusions: A Spectrum of Probabilistic Models in Analyzing Multiple Ratings. In Proceedings of the 29th International Conference on Machine Learning.

[6] Vikas C. Raykar and Shipeng Yu. 2011. Ranking annotators for crowdsourced labeling tasks. In Advances in Neural Information Processing Systems 24. 1809-1817.

[7] Yaron Singer and Manas Mittal. 2013. Pricing mechanisms for crowdsourcing markets. In Proceedings of the 22nd international conference on World Wide Web. 1157–1166.

[8] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast–but is it good?: evaluating non-expert annotations for natural language tasks. In Proceedings of the Conference on Empirical Methods in Natural Language Processing. 254-263.

[9] Peter Welinder, Steve Branson, Serge Belongie, and Pietro Perona. 2010. The Multidimensional Wisdom of Crowds. In Advances in Neural Information Processing Systems 23. 2424-2432.

[10] Jacob Whitehill, Paul Ruvolo, Ting fan Wu, Jacob Bergsma, and Javier Movellan. 2009. Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise. In Advances in Neural Information Processing Systems 22. 2035–2043.

[11] Yuchen Zhang, Xi Chen, Denny Zhou, and Michael I Jordan. 2014. Spectral methods meet EM: A provably optimal algorithm for crowdsourcing. In Advances in neural information processing systems. 1260–1268.

未来ビジョン

本研究は，人間の判断を計算プロセスに取り込むというマイクロタスククラウドソーシングによるヒューマンコンピュテーションにかかわるものである．今後，ますます多くの人間がスマートフォンのような情報機器と日常的に密着し，常にネットワークにつながった生活を行うようになるにつれて，このようなヒューマンコンピュテーションはより現実的な技術として広く使われるようになっていくものと予想される．そのような状況になったときに，本研究で提案しているような効率化手法は，大きなインパクトを持つようになると考えている．