

人間+AIクラウドにおけるマイクロタスク処理の効率化

小林 正樹^{1,a)} 若林 啓^{2,b)} 森嶋 厚行^{2,c)}

概要：本論文では、あるマイクロタスク集合が与えられた時に、公募によって集められた人間ワーカと AI ワーカが混在する「人間+AI クラウド (群衆)」における分担をどのように決めれば、品質をある程度保ったまま、より低コストで全てのタスクを処理できるのか、という問題において、分担の決定を自動化する手法を提案する。提案手法は、AI ワーカの部分的な分類性能を統計的検定に基づいて評価し、そのタスク処理結果を部分的に採用する。実験の結果、タスク結果の品質を大幅に下げること無く人間ワーカのタスク数を削減できること、また、品質とタスク数削減のトレードオフをパラメータで調整できることが示された。この結果は、性質がわからない AI ワーカと人間ワーカを適切に作業分担させる自動的な仕組みが可能である事を示している。

1. はじめに

本論文では、あるマイクロタスク集合が与えられた時に、公募によって集められた人間ワーカと AI ワーカが混在する「人間+AI クラウド (群衆)」における人間ワーカと AI ワーカの分担をどのように決めれば、一定以上の品質でより早く、低コストで全てのタスクを処理できるのか、という問題について議論する。

ここで AI ワーカとは、人間のワーカのように動作する性質不明の AI ソフトウェア^{*1}の事である。近年、Kaggle[1]や、クラウドソーシングサービスでリクルートされたプログラマ、自然災害時の IT ボランティアのように、AI の作成をアウトソースして活用する環境が揃ってきている。また、AI ワーカを募集するクラウドソーシングプラットフォームが出現するなど [2]、AI ワーカを受け入れるための準備も整っている。しかし、AI ワーカを入手したあと、それらを評価して問題解決をするまでのプロセスは通常は人手で行う必要がある。また、必ず AI が活用できるとも限らない。例えば、100 万件のデータに対して解を得たいときに 1 万件のラベル付けをクラウドソースしてコンペを行い、性能を見て最もよいモデルを入手しても、所定の品質に到達しない場合がある。この場合、追加のラベル付けを行った上で改めてコンペを実施したり、全てのタスク処

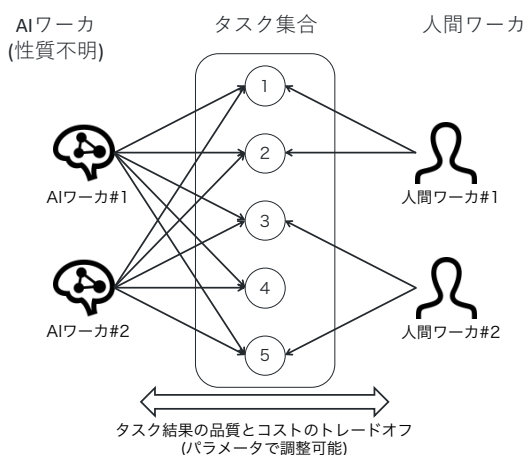


図 1 本論文では、人間ワーカと AI ワーカが適切に分担することで、マイクロタスク処理の効率化を目指す。

理を人間に依頼するといった判断が必要である。このような作業はスケールせず、大量の AI ワーカを活用出来る場合に、人間との適切な分担を見つけるのは困難である。

以上のように、問題解決のための AI を自分で作る事ができない依頼者であっても AI を活用できる世界が近づいているものの、それを上手に活用して問題を解くことは依然として困難である。これが容易になれば、より多くの人々が AI を活用した問題解決を行うことが可能になる。

本論文では、このような背景のもと、AI ワーカと人間ワーカの混在状況において、分担の決定を自動化するという問題を提起し、そのための手法を提案する。今回提案する手法では、人間は正しい結果を返すと仮定し (スパム除去や集約などの様々な手法で品質向上を行う必要があるであろう [3])、AI ワーカの性質は未知であるとする。このと

¹ 筑波大学 図書館情報メディア研究科, 茨城県つくば市春日 1-2

² 筑波大学 図書館情報メディア系, 茨城県つくば市春日 1-2

a) makky@klis.tsukuba.ac.jp

b) kwakaba@slis.tsukuba.ac.jp

c) mori@slis.tsukuba.ac.jp

^{*1} 単純なアルゴリズムやルールベース, 機械学習なども含む広義の AI プログラム

きに、タスク集合に対してどのように分担して割り当てると、一定以上の品質を保ちながら、より早く、低コストで結果を返すための割当てを動的に計算する(図1)。

今回提案する手法は次の2つの特徴を持つ。第1に、AIワーカーの全体の性能を見るのではなく、AIワーカーの得意分野に注目して、それを元にした人間との分担を行う事である。すなわち、AIが得意なものはAIに任せ、苦手なことは人間が行う、という原則に基づいた分担を行う。第2に、品質とコストはトレードオフの関係にあるが、それについてはパラメータで調整できる。すなわち、品質は犠牲にしても安く・早く終了させたいのか、もしくは、より品質を高めるためにコストや時間を掛けても良いのか、を利用者が選択できる。以上のような特徴をもって、人間+AIクラウドの状況にあわせて、自動的に分担を決定するのである。具体的には、個々のAIワーカーの部分的な分類性能を統計的検定に基づいて評価する仕組みを導入することで、AIワーカーの得意なタスクを割り当てる。この検定は、人間ワーカーの結果と一致するかどうかに基づく。

シミュレーション実験の結果から、AIワーカーにそれぞれが得意な一部のタスクを依頼することで、最終的なタスク結果の品質を大幅に下げること無く、人間ワーカーに依頼するタスク数を削減できること、また、品質とコストのトレードオフはパラメータで調整できることが確認できた。この結果は、人間ワーカーと性質が未知であるAIワーカーの混在状況において、これらに適切に作業を分担させる自動的な仕組みが可能である事を示している。

2. 関連研究

公募によるAI開発の例として、データ分析コンペティションプラットフォームのKaggle[1]が挙げられる。Kaggleでは依頼者が、AI開発者等に予測モデル等(AI)の開発をコンペ形式で依頼し、AI開発者は提案するAIの性能を競う。KaggleはAIの入手が目的であるため、入手したAIを利用するためには依頼者が手動でシステムに組み込む必要がある。一方で本研究では、タスク結果の入手が目的であり、提案手法によりタスク処理に参加するAIの能力をシステムが判断し、その性能が統計的検定により認められれば自動的にAIからのタスク結果を受け入れる。

人間と計算機が協力して課題解決をする仕組みが提案されている。能動学習[4]では、AIが性能を改善するために、正解ラベルを必要とするデータに対して、専門家への依頼やクラウドソーシングによってラベルを入手する。能動学習では人間がAIの要求に対応するが、本研究ではAIが担当できるタスクは人間ワーカーが担当しない点異なる。人間が担当するタスクとAIが担当するタスクを自動的に判断することでこれを実現する。

アンサンブル学習は、複数のAIを組み合わせることで、全体としての推論結果の品質を改善する手法である[5]。ア

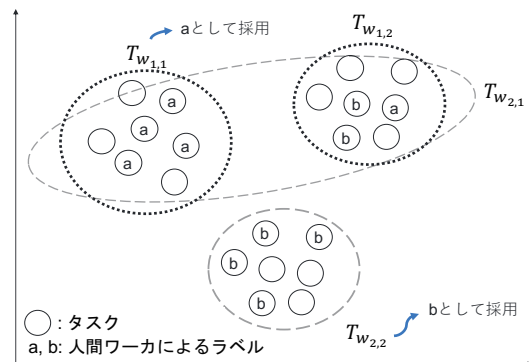


図2 提案手法の概要。タスククラスタを統計的検定により評価することで、タスククラスタに含まれるタスクの回答を採用する。

ンサンブル学習では与えた入力に対して何らかの出力を行うが、本研究では1つ以上のAIの出力クラスタをそれぞれ評価し、性能が認められた一部の出力のみを採用する。どのAIの出力も採用されなかった場合には、人間クラウドがタスク処理を担当する。

3. 提案手法

提案手法である、AIワーカーの得意分野に注目することで、それに基づいた人間ワーカーとの分担を実現する方法について説明する。本論文では、AIワーカーの出力が同じとなるタスクのグループ(教師無しの場合には同じクラスタに分類されるタスクの集合、教師有りの場合には同じラベルがつくタスクの集合)をタスククラスタと呼ぶ。タスククラスタを人間ワーカーによるタスク結果を用いて統計的検定により評価することで、個々のタスククラスタに含まれるタスクの回答を受け入れるかを判断する。

入力。N個の選択肢(ラベル)から回答するタスク集合を $T = \{t_1, \dots, t_M\}$ と表す。これらのタスクに付与される選択肢(ラベル)の候補の集合を $A = \{a_1, \dots, a_N\}$ とする。また、利用可能なAIワーカーの集合を W と表す。 W には教師あり学習AIワーカーや教師なし学習AIワーカーなど、多種多様な手法に基づく不特定多数のAIワーカーが含まれることを想定する。AIワーカー $w_i \in W$ について、 $C_{w_i} : \mathcal{P}(T \times A) \rightarrow \mathcal{P}(\mathcal{P}(T))$ を w_i の振る舞いを表す関数とする。この関数は、タスクとラベルのペアの集合 D を引数とし、タスククラスタの集合 $\{T_{w_{i,1}}, T_{w_{i,2}}, \dots\}$ を返す。ただし、タスククラスタ $T_{w_{i,j}}$ の直和集合が T となる。つまり、 $T = T_{w_{i,1}} \oplus T_{w_{i,2}} \oplus \dots$ である。

D をこれまで人間が回答したタスクとラベルのペア集合とすると、AIワーカー w_i が教師無し学習器の場合、 $C_{w_i}(D)$ は常に同じタスククラスタ集合を返す。一方で教師あり学習器の場合、タスク実行が進んで人間ワーカーの回答が D に追加されるに従い、 $C_{w_i}(D)$ を再計算する。以降の議論では、 $\bigcup_{w_i \in W} C_{w_i}(D)$ を $C(D)$ と表記する。タスククラスタの評価にあたり、タスククラスタの採用されやすさを調整

Algorithm 1 アルゴリズム**Input:** T, A, W, C_{w_i}, p **Output:** Z

```

1: Let  $D = \{(t, a) | \exists t \in T \text{ ans}(t) = (a, 'h')\}$ 
2: for all  $t \in T$  s.t.  $\text{ans}(t) \neq (\phi, \phi)$  do
3:   for all  $T_{w_{i,j}} \in C(D)$  do
4:     if  $\text{statistical\_test}(T_{w_{i,j}}, D, p)$  then
5:       let  $\hat{a}$  be the label for  $T_{w_{i,j}}$ 
6:       update  $\text{ans}$  so that for all  $t' \in T_{w_{i,j}}, \text{ans}(t') = (\hat{a}, w_i)$ 
7:     end if
8:   end for
9:    $a' \leftarrow \text{humanworker\_result}(t)$ 
10:  update  $\text{ans}$  so that  $\text{ans}(t) = (a', 'h')$ 
11: end for
12: return  $\{(t_1, \text{ans}(t_1).result), (t_2, \text{ans}(t_2).result), \dots\}$ 

```

するためのパラメータ p を用いる。

出力。提案手法の出力は、タスクと回答のペアの集合 $\{(t_1, a_{k1}), \dots, (t_M, a_{kM})\}$ である。ただし、 $a_{ki} \in A$ である。

手続き。Algorithm1 に、本手法の手順を示す。ここで、関数 $\text{ans} : T \rightarrow A \times (W \cup \{h'\}) \cup \{(\phi, \phi)\}$ は、タスク t が与えられると結果ラベルと回答したワーカーを返す関数である。ここで h' は人間ワーカーを表す。初期値では、全ての $t \in T$ に対して $\text{ans}(t) = (\phi, \phi)$ を返すが、タスクの進行に応じて返す値は更新される。本手法は $\text{ans}(t)$ がすべての t に対して (ϕ, ϕ) を返さなくなると処理を終える。

AI のタスク結果を採用するか否かは、既に人間ワーカーが返したタスク結果の集合である D を正解と見なし、各タスククラス $T_{w_{i,j}} \in C(D)$ に含まれる全てのタスクの結果が特定の回答 $\hat{a}$ (例えば a や b) である可能性を調べる (図 2)。その可能性が高いと認められた場合、全てのタスク $t \in T_{w_{i,j}}$ の $\text{ans}(t)$ のラベルを $\hat{a}$ へと更新する。一方でタスクの結果が全て特定の回答 $\hat{a}$ である可能性が高いと認められない場合は、 $T_{w_{i,j}}$ は採用しない。Ans(とそれに伴う D) の更新は、人間ワーカーがタスク t にラベル (例えば、 a or b) を与える ($\text{humanworker_result}(t)$) 毎に行われる。

タスククラス $T_{w_{i,j}}$ を採用するかどうかの評価は、統計的検定に基づく手法で行う。具体的には、タスククラスに含まれるタスクを対象に、これまで人手で行われたタスクの回答として最も多数に得られた回答 $\hat{a}$ を求め、 $T_{w_{i,j}}$ 内の全てのタスクの回答が $\hat{a}$ である、と言えるかどうかを検定する。例えば、片側二項検定を用いる場合、成功回数は $m = |\bigcup_{t \in T_{w_{i,j}}} \text{ans}(t) = (\hat{a}, 'h')|$ 、試行回数 n は $n = |\bigcup_{t \in T_{w_{i,j}}} \text{ans}(t) = (*, 'h')|$ ($*$ は何らかの回答)、成功確率を p として検定を行う。ただし、 $m \leq n$ である。

4. 実験

AI ワーカーと人間ワーカーが利用可能である状況において、提案手法を用いることで与えられたタスクを効率よく分業出来ることを検証するシミュレーション実験を行った。

4.1 設定

タスク。手書き文字画像データセットである MNIST[6] を用いて、10 クラス分類タスクを作成した。ラベル付きデータから無作為に 10,000 件を抽出し、これらのデータにラベルを付与することを本実験での課題とした。実験で新たに付与したラベルと正解ラベルを照合することで、提案手法によるタスク結果の正答率を評価した。

人間ワーカー。実験では、人間ワーカーは割り当てられたタスクに対して必ず正解ラベルを付与するものとした。実験では、人間ワーカーは一度に 200 タスクにラベルを付与した。

AI ワーカー。実験では、教師あり学習による AI ワーカーとしてロジスティック回帰を、教師なし学習による AI ワーカーとして K-means を用いた。ロジスティック回帰 (LR) では、人間ワーカーによるラベル付きデータは学習用とテスト用に分割して用いた。K-means では、訓練には処理対象の全てのデータの特徴を用いた。学習は最初に実施し、以降では検定に利用するタスク数が変化するがモデルの更新は行わなかった。実験は、AI ワーカーとしてロジスティック回帰と K-means の両者を用いた場合、K-means のみを用いた場合、ロジスティック回帰のみを用いた場合で実施した。複数の AI ワーカーが同一のタスクに回答可能な場合には、先に採用された AI ワーカーの回答を受け入れた。

タスククラスタの評価 (統計的検定)。統計的検定に基づいてタスククラスタを評価するために、片側二項検定を用いた。実験では有意水準を 5% とし、これを満たす場合にタスククラスタの結果を採用した。

4.2 結果

図 3 は、AI ワーカーとしてロジスティック回帰 (LR) と K-means を用いた結果 (左)、K-means のみを用いた結果 (中央)、ロジスティック回帰のみを用いた結果 (右) を表している。グラフの横軸は、人間ワーカーが処理したタスク数で、縦軸は全体での完了タスク数を表している。線の種類は、それぞれの実験で変化させたパラメータ (左: p , 中央: K-means のクラス数, 右: AI ワーカーの訓練時のテストデータの比率) ごとの結果および、人間ワーカーが全てのタスク処理を担当した場合の結果を表している。図 4 は、それぞれの実験での各パラメータでの採用タスククラス数および AI ワーカーが担当したタスクの正答率を示している。

5. 考察

図 3 左では、タスククラスタの評価のパラメータである、 p を変化させた。 p を高くすると、AI によるタスク結果の品質は向上するが、採用されたタスククラス数は減少した。このように AI からのタスク結果を採用する基準のパラメータにより、人間ワーカーにタスクを依頼するコストと結果品質のトレードオフを調整できることが示された。

図 3 中央の K-means のみを用いた場合と図 3 右のロジ

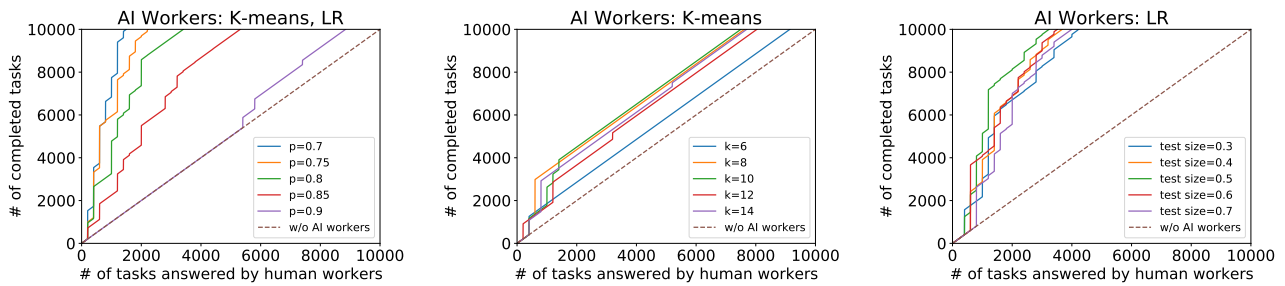


図3 人間ワーカーが処理したタスク数と全体での完了タスク数の関係。(左) K-means ($k=20$) と LR ワーカー (test size=0.5) が参加。(中央) および (右) では $p=0.8$ を用いた。

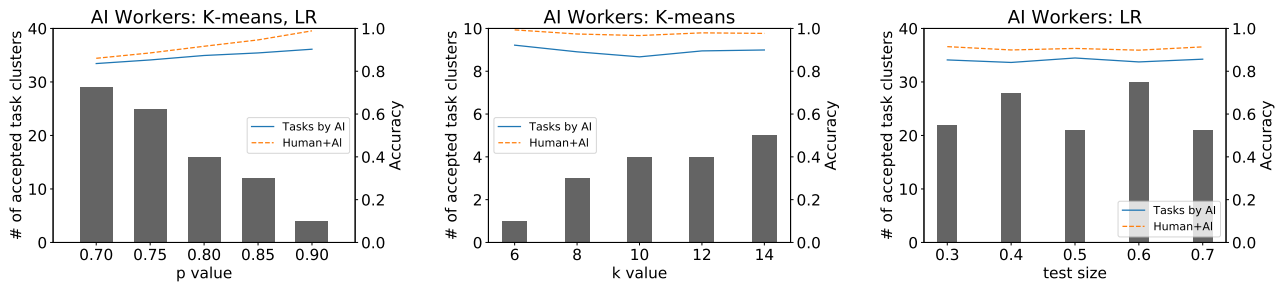


図4 参加した AI ワーカーごとの採用タスククラス数と AI ワーカーが担当したタスクの正答率。

スティック回帰のみを用いた場合の実験では、AI ワーカー側のパラメータを変化させた。本研究で扱う問題では、AI ワーカーの性質を事前に予測することは出来ず、また性能を制御することも出来ないのが前提である。しかし、AI ワーカーの開発者の立場で考えた場合、多くのタスククラスが採用されるような工夫を導入することで、AI ワーカーが活用される可能性が高くなると考えられる。

K-means ワーカーのクラス数数を大きくすることで、より多くのタスククラスが採用されたが、タスク結果の品質に大きな差は見られなかった。これは、提案手法により認められたタスククラスのみを採用するためであると考えられる。 k を大きな値とした場合、タスククラスに含まれるタスク数が少なく、統計的検定に用いるためのサンプル数が不足すると考えられる。

ロジスティック回帰ワーカーのテストデータの割合を変化させたところ、テストサイズと採用されるタスククラス数の間に関係は見られなかった。また、採用されたクラスによるタスク結果の品質も大きく変化しなかった。テストデータを多く確保すれば、タスククラスの評価のためのサンプル数を確保できるが、一方で学習データ数が不十分となる。取り扱う課題や AI ワーカーによっては適切に調整することで AI ワーカーの活用につながると考えられる。

6. まとめと今後の課題

本稿では、AI ワーカーと人間ワーカーが混在する状況において、タスク処理の分担を自動的に決定するという問題に対して、人間ワーカーの解答に基づいて統計的検定を行い AI ワーカーの得意分野を発見することで、AI ワーカーからのタ

スク結果を部分的に採用する手法を提案した。シミュレーション実験の結果、提案手法により全体のタスク結果品質を大幅に下げることもなく、人間ワーカーが担当するタスク数を削減できることが示された。この結果は、AI ワーカーと人間ワーカーが適切に作業を分担することで、効率的なタスク処理を実現できる可能性を示唆するものであり、大規模な AI ワーカーを想定した検証が今後の課題として挙げられる。

謝辞 本研究の一部は JST CREST (#JPMJCR16E3) の支援による。

参考文献

- [1] Goldbloom, A. and Hamner, B.: Kaggle: Your Home for Data Science, Kaggle Inc (online), available from (<https://www.kaggle.com>) (accessed 2019-07-01).
- [2] Crowd4U: Crowd4U における AI ワーカーの参加方法について, Crowd4U (オンライン), 入手先 (https://crowd4u.org/events/ai_open_call) (参照 2019-07-01).
- [3] Daniel, F., Kucherbaev, P., Cappiello, C., Benatallah, B. and Allahbakhsh, M.: Quality Control in Crowdsourcing: A Survey of Quality Attributes, Assessment Techniques, and Assurance Actions, *ACM Comput. Surv.*, Vol. 51, No. 1, pp. 7:1–7:40 (online), DOI: 10.1145/3148148 (2018).
- [4] Settles, B.: Active learning literature survey, Technical report, University of Wisconsin-Madison Department of Computer Sciences (2009).
- [5] Dietterich, T. G.: Ensemble Methods in Machine Learning, *Multiple Classifier Systems*, Berlin, Heidelberg, Springer Berlin Heidelberg, pp. 1–15 (2000).
- [6] Lecun, Y., Bottou, L., Bengio, Y. and Haffner, P.: Gradient-based learning applied to document recognition, *Proceedings of the IEEE*, Vol. 86, No. 11, pp. 2278–2324 (online), DOI: 10.1109/5.726791 (1998).