人間+AI クラウドの相互作用によるタスク結果品質の管理手法

小林　正樹†　　若林　　啓††　　森嶋　厚行††

† 筑波大学 図書館情報メディア研究科　〒 305-8550 茨城県つくば市春日 1-2
†† 筑波大学 図書館情報メディア系　〒 305-8550 茨城県つくば市春日 1-2
E-mail: †makky@klis.tsukuba.ac.jp, ††kwakaba@slis.tsukuba.ac.jp, ††morishima-office@ml.cc.tsukuba.ac.jp

あらまし　本論文では，正解を返すとは限らない人間ワーカおよび計算機処理に基づくワーカ（AI ワーカ）が分担して大量のタスクを処理する状況における，タスク結果の品質を考慮したタスク割り当てに取り組む．人間ワーカのタスク結果は単なる成果物としてだけでなく，AI ワーカの学習と評価に利用するため，人間ワーカから得られるタスク結果品質がますます重要となる．しかしながら，現実のクラウドソーシングでは人間ワーカが実際に作業をするため様々な理由でタスク結果が不正確である可能性があり，これにより全体的なタスク結果品質の低下を引き起こすと考えられる．そこで本論文では，人間ワーカと AI ワーカの回答の不一致に着目して，人間ワーカに追加のタスクを割り当てることで，人間ワーカから得られるタスク結果品質を改善する手法を提案する．ベンチマークデータセットよる実験結果から，不確実な人間ワーカと AI ワーカが相互にタスク結果を共有することでタスク結果品質を改善しながら効率的にタスクを処理する仕組みが構築可能であることを示す．

キーワード　クラウドソーシング, タスク割り当て, 品質管理

1　はじめに

本論文では，クラウドソーシングで公募した不特定多数の人間ワーカおよび外部協力者等が開発した性能やアルゴリズムが不明である計算機処理に基づくワーカ（AI ワーカ）が，大量のタスクを分担して処理する際に，人間のタスク結果が正しいとは限らない事が，全体的なタスク結果品質の低下に繋がるという問題に取り組む．

近年では，AI プログラムの作成を外部協力者にクラウドソーシングする試みが普及しており，代表的な例として Kaggle[1]や AIcrowd[2]が挙げられる．このような仕組みを活用すれば，AI 技術に精通していない依頼者であっても AI 技術の恩恵を受けることが期待できる．しかしながら，AI プログラムの作成を依頼し，依頼者の課題解決に適用するまでのプロセスには手動での試行錯誤を伴うことが多い．例として，クラウドソーシングと機械学習モデルを組み合わせて，依頼者が要求する品質で 10,000 枚の画像を 10 クラス分類することを考える．依頼者はまず，クラウドソーシングを用いて一部の画像にラベルを付与して，機械学習モデルを作成するためのデータセットを構築する．次にデータセットを用いて，機械学習モデルの作成を Kaggle 等のプラットフォームで依頼する．このような手順を踏むことで，依頼者は機械学習モデルを入手することができるが，その機械学習モデルの出力が依頼者の要求する精度を満たすとは限らない．したがって，依頼者は学習およびテストデータを拡充した上で機械学習モデルの開発を再依頼するか，AI の利用を諦めて全ての画像の分類をクラウドソーシングで行うといった意思決定が求められる．このような工程を経て，人間による処理と計算機による処理を適切に組み合わせることは，専門家であっても容易ではない．

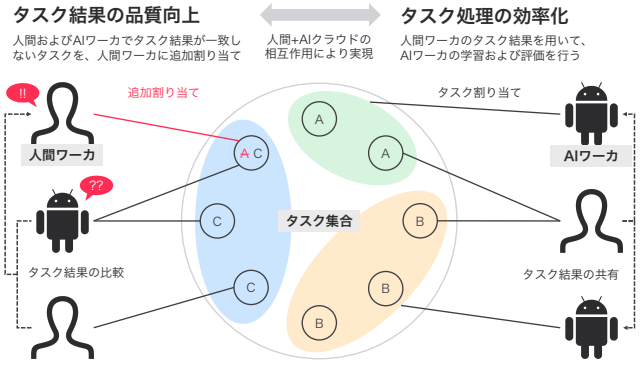

図 1　本論文では，人間ワーカと AI ワーカの相互作用を設計することで，全体的なタスク結果品質を改善しながらタスクを処理する仕組みを提案する

本研究では，前述の問題について，AI プログラムをブラックボックスの AI ワーカとしてモデル化することで取り組んでいる．先行研究では，依頼者の要求精度を満たすような人間ワーカおよび AI ワーカへのタスク割り当て問題を提起し，この問題に対するタスク割り当てアルゴリズムを提案した [1]．提案アルゴリズムは，人間ワーカから得たタスク結果を基準として AI ワーカの性能を統計的検定により評価する．AI ワーカが出力するラベルの種類ごと（タスククラスタ，3.2 節で説明する）に評価を行うことでより多くのタスクを AI ワーカに割り当てる．

しかしながら，クラウドソーシングによって人間ワーカから得られるタスク結果が正しいとは限らない．前述の割り当てアルゴリズムでは，人間ワーカのタスク結果を成果物の一部として利用するだけでなく，AI ワーカの学習や評価に用いるため，人間ワーカから得られるタスク結果の品質はより重要である．

1：https://www.kaggle.com
2：https://www.aicrowd.com

クラウドソーシングにおける品質管理の研究では，異なる人間ワーカに対して同一のタスクを割り当てて，それらの結果を統合することでより正確なタスク結果を得る多数決に基づく方法が広く用いられている [2]．全てのデータに対して多数決を適用することで，人間ワーカから得られるタスク結果品質の改善が期待できるが人間ワーカへ依頼するタスク数が膨大になることが懸念される．

そこで，本研究では人間ワーカと AI ワーカが相互にタスク結果を共有し合うことで，全体的なタスク結果を改善する仕組みを提案する（図 1）．提案手法では，人間ワーカと AI ワーカのタスク結果の不一致に基づいて，人間ワーカに追加のタスク割り当てを行うことで，人間ワーカから得られるタスク結果品質の信頼性を向上させる．高品質な人間ワーカのタスク結果を AI ワーカの学習や評価のデータセットとして用いることで，より多くのタスクを AI ワーカに割り当てることを目指す．実験結果から，提案手法が全てのタスクに多数決を適用する手法よりも少ない人間ワーカへの追加割り当て数で，同程度のタスク結果品質を得ることが出来ることが示唆された．

本研究の貢献は次の通りである：

- 先行研究 [1] で提案された人間＋ AI クラウドタスク割り当て問題のアルゴリズムの性能を，人間ワーカのタスク結果が不正確な状況設定で分析した．
- 既存のアルゴリズムと多数決を組み合わせることで，人間ワーカのタスク結果が不正確であっても要求精度を満たすタスク割り当てが可能であるが，人間ワーカへのタスク割り当てを削減できる余地があることを示した．
- 人間ワーカと AI ワーカの不一致に基づいた人間ワーカへの追加タスク割り当てと AI ワーカの評価を組み合わせたアルゴリズムを提案した．実験結果から，提案アルゴリズムが全タスクに多数決を適用する手法と比較して，要求精度を満たしながら人間ワーカへの追加割り当てを削減できることを示した．

2 関 連 研 究

クラウドソーシングにおける品質管理の問題に取り組む研究では，不特定多数の人間ワーカからより正確な成果物を得るためのさまざまな手法が提案されており，典型的な手法には複数のワーカの作業結果の統合，信頼性の高いワーカの検出，ワーカを訓練するためのタスクの導入，タスク設計の改良などが挙げられる [3]．本研究では，クラウドソーシングにおけるワーカとして，人間のワーカだけでなく AI ワーカが参加する設定において，タスク割り当ての問題に取り組んでいる．タスク結果の品質管理のために，本研究と既存手法を組み合わせることは有効な手段であると考えられる．

人間によってラベル付けされたデータに基づいて機械学習モデルを学習する際に，機械学習モデルの出力によって次に人間によるラベル付けを必要とするデータを決定することで，より少ないラベル付データで高性能なモデルを学習できることが知られている．このような方式は能動学習と呼ばれ，コストの制約がある，複数の性能の異なる人間ワーカに問い合わせが可能であるといった状況における問い合わせ戦略が研究されている [4–7]．能動学習と本研究が取り組む問題は次の 2 つの観点から異なる．(1) 能動学習の目的は予算内で機械学習モデルの性能を最大化することであり，残りのデータを処理するためにどのタイミングで学習ループを停止するかを判断しない．本研究では，AI ワーカの性能を人間ワーカのラベルにより統計的検定を行い，未ラベルのタスクを AI ワーカに割り当てるかどうか判断する．(2) 能動学習は学習ループの実行を開始する前に，対象とする機械学習モデルを与える必要がある．本研究では，タスク処理の進行中に，ブラックボックスである複数 AI ワーカが増減することを許す．

複数の機械学習モデルが利用可能な状況では，それらの出力を統合（モデルアンサンブル）することによってより精度の高いモデルを構築できることが知られている [8]．しかしながら，機械学習モデルの候補やどのように組み合わせることが精度向上に繋がるかは自明ではなく，モデル設計者による実験が必要である．能動学習とアンサンブルを組み合わせる研究も存在する [9]．また，複数のモデル候補やパラメータ候補を与えることで，精度が高くなるようなモデルやパラメータの組み合わせを探索する AutoML [10] と呼ばれる研究が存在する．ベイズ最適化などの手法を導入することで，より少ない実験回数で最適な組み合わせに到達できることが知られている．

人間による作業と計算機処理を組み合わせて現実の問題に取り組むさまざまな方式の研究が存在する [11–18]．多くの研究では人間および計算機処理の役割を事前に設計する必要があるが，本研究では両者の分担を自動的に決定することに取り組む．

3 人間+AI クラウドタスク割り当て問題

先行研究で提案した人間＋ AI クラウドタスク割り当て問題 (Human+AI Crowd Task Assignment Problem, HACTAP) と割り当てアルゴリズム [1] について説明する．

3.1 問 題 設 定

与えられたタスク集合 T の各要素に対して，ラベル集合 A のいずれかの要素を付与することを考える．タスクの処理は人間ワーカもしくは AI ワーカに割り当てて行う．AI ワーカ集合 W には匿名の外部協力者によって実装された，アルゴリズムや性能がブラックボックスである AI ワーカが含まれる．AI ワーカ集合の各要素 $w_i \in W$ は関数 $w_i : T \to \mathbb{N}$ であり，入力されたタスクがどのタスククラスタに属するかを返すものとする．各 AI ワーカについて，任意のデータセットによる訓練と与えられたタスクに対するラベルの予測の機能のみを呼び出し可能とする．問題の簡略化のため，個々の人間ワーカ毎のスキルや能力を考慮しないため，人間ワーカを単に ‘h’ と表記する．

HACTAP は q を満たし，かつより多くのタスクを AI ワーカに割り当てるようなタスク割り当て集合 S を求める問題である．ここで，依頼者が設定する全体的なタスク結果の要求精度を $0 \leqq q \leqq 1$ とする．どのタスクをどのワーカが処理したかというタスク割り当ての集合 S は，タスクとワーカのペア

(t_j, w_i) で構成される．ただし，$w_i \in W \cup \{\text{`h'}\}$ とする．全タスクへの割り当てが完了した時 $|S| = |T|$ となる．

3.2 タスククラスタ

先行研究で提案されたタスク割り当てアルゴリズムは，類似タスクからなるタスククラスタを構成し，AI ワーカをタスククラスタ毎に評価することで，より多くのタスクを AI ワーカに割り当てるというアイデアに基づく．

q を満たすような AI ワーカが存在する場合，その AI ワーカに未処理のタスク全てを割り当てることで目的を達成できる．しかしながら，そのような理想的な AI ワーカが存在するとは限らず，しばしば入手できるまでに時間を要する．

そこで，AI ワーカからの出力の部分集合に注目する．$R_{w_i} = \{(t_j, t_k) \mid w_i(t_j) = w_i(t_k)\}$ をある AI ワーカ w_i が 2 つのタスクに対して同一の種類の出力をするという二項関係とする．AI ワーカ w_i から得られるタスククラスタ集合は商集合 $C_{w_i} = T/R_{w_i} = \{T_{w_i,1}, T_{w_i,2}, ...\}$ で表される．タスククラスタ単位での精度の評価を行うことにより，AI ワーカからの出力を部分的に採用するか否かの判断が可能になる．これにより，特性が多種多様であり，全体的な性能が十分とは限らない複数の AI ワーカにタスクを割り当てることが期待できる．AI ワーカ集合 W から得られる全てのタスククラスタは $C = \bigcup_{w_i \in W} C_{w_i}$ と表記する．

先行研究 [1] の評価実験ではタスククラスタを導入することで，要求精度 q を満たしながらより多くの AI ワーカへのタスク割り当てを実現できることが報告されている．

3.3 タスク割り当てアルゴリズム

先行研究 [1] で提案されたタスク割り当てアルゴリズムの 1 つである Clusterwise Test-based Assignment (CTA) について説明する．CTA は人間ワーカのラベルを用いて，AI ワーカから得たタスククラスタを統計的検定により評価する．統計的検定では，対象のタスククラスタに含まれる人間ワーカラベルについて最頻出ラベルの出現割合が q を上回るか評価する．タスククラスタが採用されると，タスククラスタ中の未ラベルのタスクに対して最頻出ラベルを付与する．この処理は AI ワーカからのタスク結果を受け入れることに相当する．

CTA のアルゴリズムをアルゴリズム 1 に示す．CTA の入力は，タスク集合 T，AI ワーカ集合 W，要求精度 q および統計的検定の有意水準 α である．CTA が出力するのは，タスクとワーカのペアで構成されるタスク割り当ての集合 S である．CTA は全てのタスクに割り当てが決定されるまで，ランダムな順序でタスクの選択を繰り返す（1 行目）．ここで，$ans : T \to \{A \times (W \cup \{\text{`h'}\})\} \cup \{(\emptyset, \emptyset)\}$ はタスクを入力するとラベルとワーカのペアを返す更新可能な関数である．選択されたタスクは人間ワーカに割り当てられ（2 行目），そのタスク結果を用いて ans 関数を更新する（3 行目）．次に，現時点で利用可能なタスククラスタの集合 C の各要素に対して処理を行う（4 行目）．5 行目では，あるタスククラスタに対して統計的検定を行う．先行研究では統計的検定手法として二項検定を用い，タスククラスタ中の人間ワーカラベルのうち最頻出ラベル（6 行目）を正解と仮定し，最頻出ラベルが占める割合が要求精度を上回るかを評価する．最頻出ラベルの占める割合が要求精度と同程度であるという帰無仮説が棄却された場合，タスククラスタ中の未ラベルであるタスクの ans 関数を $(\hat{a}, w_i)$ で更新する（7 行目）．

Algorithm 1 Clusterwise Test-based Assignment (CTA)

Input: A set T of tasks, a set W of AI workers, the accuracy requirement q, and the significance level α.
Output: A sequence of pair (task, worker)s.

1: **for all** $t \in T$ s.t. $ans(t) = (\emptyset, \emptyset)$ in a random order **do**
2: $\quad a' \leftarrow task\_result(assign(t, \text{`h'}))$
3: $\quad$ update ans so that $ans(t') = (a', \text{`h'})$
4: $\quad$ **for all** $T_{w_i,j} \in C$ **do**
5: $\quad\quad$ **if** $statistical\_test(T_{w_i,j}, q, \alpha)$ **then**
6: $\quad\quad\quad \hat{a} = \arg\max_{a \in A} |\{t \mid t \in T_{w_i,j}, ans(t) = (a, \text{`h'})\}|$
7: $\quad\quad\quad$ for $t' \in T_{w_i,j}$ s.t. $ans(t') = (\emptyset, \emptyset)$, $assign(t', w_i)$ and update ans so that $ans(t') = (\hat{a}, w_i)$
8: $\quad\quad$ **end if**
9: $\quad$ **end for**
10: **end for**

CTA では，人間ワーカから得られるタスク結果が正しいと仮定して，人間ワーカのタスク結果を AI ワーカの学習と評価に用いる．そのため，本論文の実験結果が示すように，人間ワーカのタスク結果に誤りが含まれる場合に，要求精度を満たす割り当てを得られない．本論文では，人間ワーカのタスク結果が正しいとは限らない設定におけるアルゴリズムを提案する．

4 提案手法

4.1 問題設定

本研究で扱う問題は，全体的なタスク結果品質が要求精度 q を満たすようなタスク割り当て S を求めることである．ここで，分類タスクの集合 T，AI ワーカの集合 W，要求精度 q および，h の確率で正解を返す人間ワーカが与えられる．ただし，同一のタスクに対して人間ワーカを v 回割り当て，その結果の多数決を人間ワーカからのラベルとして利用することを許す．

4.2 本研究のアプローチ

本論文で提案するタスク割り当てアルゴリズムを説明する前に，アルゴリズムを構成する 2 つの要素について説明する．

4.2.1 人間ワーカへの追加タスク割り当て

本論文の問題設定では，人間ワーカに割り当てたタスクの結果を最終的なタスク結果および AI ワーカの学習・評価のために用いる．そのため，人間ワーカから得られるタスク結果品質を高めることが，全体的なタスク結果品質の向上に繋がる．ここで，タスク結果品質とは具体的にはタスク結果の精度とする．

人間ワーカから得られるタスク結果品質の管理手法として最も単純なのは，人間ワーカに割り当てる全タスクに多数決を適用することである．多数決により一部の回答が不正確であって

図 2　個々のワーカの正答率と多数決する人数ごとの正答率の関係

も全体としては信頼性の高いタスク結果を得ることができる．多数決を行う人数と個々の人間ワーカの正答率の組み合わせにおける，多数決結果の正答率は図 2 の関係にある．

一方，人間ワーカに割り当てる全てのタスクに多数決を適用することは，タスク数が膨大な状況では現実的でない．そこで，本論文では AI ワーカを用いて多数決を必要とするタスクを選択することを提案する．あるタスクに対して，複数の人間ワーカを割り当てて多数決を適用することを追加割り当てと呼ぶ．

追加割り当てを行うタスクを選択するにあたり，人間ワーカと AI ワーカのタスク結果に不一致が生じているタスクに注目する．不一致が生じているタスクは人間ワーカもしくは AI ワーカのいずれかが誤りであるという仮説のもと，(1) まずは人間ワーカのタスク結果品質を改善し，(2) そのタスク結果を用いて AI ワーカの訓練と評価を行い，これらの処理を繰り返すことで，より正確に全てのタスクを処理することを目指す．

4.2.2　活用と探索

HACTAP ではより多くのタスクを AI ワーカに割り当てることが求められるが，一方で AI ワーカの訓練と評価を正確に行うために人間ワーカから得られるタスク結果品質が重要であり，多数決で品質管理を行うためには十分な票数が必要である．そこで，AI ワーカへのタスク割り当てを活用，人間ワーカへの追加割り当てを探索とし，両者の処理のバランスをとることを検討する．活用と探索の定義は次の通りである．

- 活用: タスククラスタ集合 C を評価して，AI ワーカに対してタスク割り当てを行うこと．ϵ の確率でこの操作を行う．
- 探索: 人間ワーカに対して追加タスク割り当てを行い，人間ワーカから得られるタスク結果の品質を改善すること．$1-\epsilon$ の確率でこの操作を行う．

活用と探索のバランス制御は ϵ によって決定する．ただし，$0 \leqq \epsilon \leqq 1$ である．実験では次式により ϵ をタスク割り当ての進行に応じて変動させる．

$$epsilon\_linear = \frac{|t \mid t \in T, ans(t) == (\emptyset, \emptyset)|}{|T|} \quad (1)$$

$$epsilon\_sigmoid = \frac{1}{2}(1 + \tanh(\frac{1}{2}\alpha epsilon\_linear)) \quad (2)$$

ただし，$\alpha = 10$ とする．完了タスク数における ϵ の変化を図 3 に示す．図には実験で比較する $\epsilon = 0.2, 0.5, 0.8$ も含む．

式 1 と式 2 を用いた ϵ の制御は，固定値の場合と比較して，全体的なタスク結果品質を下げることなく AI ワーカへの割り当てを増やすことが期待できる．理由は，タスク割り当ての前半には集中的に人間ワーカへの追加割り当てを行い，後半では AI ワーカの評価を頻繁に行うからである．一方で，要求精度が低い状況では，AI ワーカの性能が十分であるにもかかわらず AI ワーカの評価ではなく人間ワーカへの追加割り当てが行われ，結果として AI ワーカにタスクを割り当てるタイミングが遅れることが予想される．

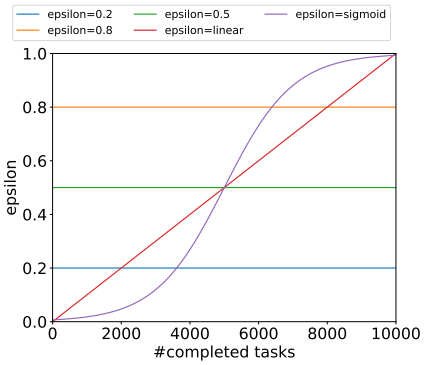

図 3　タスク割り当ての進行と各探索方策における ϵ の関係

4.3　提案アルゴリズム

提案アルゴリズムである Interactive CTA (アルゴリズム 2) について説明する．Interactive CTA では，人間ワーカと AI ワーカのタスク結果の不一致に基づいて，人間ワーカに追加タスク割り当てを行うことで，人間ワーカから得られるタスク結果品質の向上を図りながら AI ワーカを活用する．4 行目で ϵ と乱数を比較し，活用（5 - 10 行目）を行うか，探索（12 - 20 行目）を行うか判定する．13 - 16 行目では全てのタスククラスタについて，人間ワーカと AI ワーカのタスク結果が不一致であるタスクを列挙し，17 行目で最も不一致であるタスクを選択，18 行目でそのタスクを人間ワーカに割り当てる．

5　評 価 実 験

5.1　実 験 設 定

くずし字データセットである KMNIST [19] を用いて，10 クラス分類タスクを作成した．訓練データから 10,000 件の画像とラベルのペアを無作為抽出し，タスク集合として用いた．

AI ワーカ集合として，scikit-learn 0.23.1 に実装されている機械学習モデルの中から，次のクラス名で定義されたアルゴリズムを初期パラメータの設定で用いた: MLPClassifier, ExtraTreeClassifier, LogisticRegression, KMeans, DecisionTreeClassifier, SVC.

実験では，既存手法である CTA と提案手法である Intaractive CTA を比較する．各アルゴリズムの統計的検定には二項検定を利用し，有意水準は 5%とした．

次の 2 つの評価指標を用いて実験結果を分析した．

- AI ワーカへのタスク割り当ての割合：タスク割り当ての結果における AI ワーカへのタスク割り当ての割合を評価す

Algorithm 2 Intaractive Clusterwise Test-based Assignment

Input: A set T of tasks, a set W of AI workers, a function C_{w_i} that returns task clusters, the accuracy requirement q, the significance level α, intaractive threshold ϵ.

Output: A sequence of pair (task, worker)s.

for all $t \in T$ s.t. $ans(t) = (\emptyset, \emptyset)$ in a random order **do**
  $a' \leftarrow task\_result(assign(t, `h'))$
  update ans so that $ans(t') = (a', `h')$
  if $Math.random() < \epsilon$ **then**
    for all $T_{w_{i,j}} \in C$ **do**
      if $statistical\_test(T_{w_{i,j}}, q, \alpha)$ **then**
        let $\hat{a}$ be the label for $T_{w_{i,j}}$
        for $t' \in T_{w_{i,j}}$ s.t. $ans(t') = (\emptyset, \emptyset)$, $assign(t', w_i)$ and update ans so that $ans(t') = (\hat{a}, w_i)$
      end if
    end for
  else
    $Q = \{\}$
    for all $T_k \in C$ **do**
      let $\hat{a}$ be the label for $T_{w_{i,j}}$
      $Q \leftarrow Q \cup \{t \mid t \in T_k, ans(t) = (*, `h'), t \neq \hat{a}\}$
    end for
    $t' = most\_frequent(Q)$
    $a' \leftarrow majority\_vote([ans(t'), task\_result(assign(t', `h'))])$
    update ans so that $ans(t') = (a', `h`)$
  end if
end for

る．AI ワーカに多くのタスクを割り当てるほど優れた割り当てアルゴリズムであると判断する．

- 要求精度達成率：同じ条件設定での複数回の実験において要求精度を満たした割合のことを要求精度達成率と呼ぶ．

5.2 結　　果

以下に示すパラメータで，10 試行ずつ行った実験結果を示す．

- 多数決の人数 $v = \{1, 3, 5, 7\}$.
- 人間ワーカの正答率 $h = \{0.8, 0.9, 1.0\}$.
- 要求精度 $q = \{0.8, 0.85, 0.9, 0.95\}$.
- $\epsilon = \{0.2, 0.5, 0.8, epsilon\_progress, epsilon\_sigmoid\}$
- 人間ワーカへの追加割り当ての基準: 人間ワーカと AI ワーカが一致，人間ワーカと AI ワーカが不一致，ランダム選択

5.2.1 人間ワーカの正答率を変えた実験

人間ワーカが正答を返すとは限らない設定 ($h = \{0.8, 0.9, 1.0\}$) における CTA アルゴリズムの振る舞いを明らかにするための比較実験を行った (図 4)．上段が人間ワーカによるタスク結果数と完了タスク数の関係を，下段が全体的なタスク結果品質および AI ワーカが処理した部分の精度と要求精度パラメータの関係を表している．上段の図は，横軸に対して縦軸の値が大きいということは，AI ワーカにタスクを割り当てたことを意味する．

実験結果から，h が低い条件では要求精度を達成できる割合が低下することが分かる．具体的には q が h を上回る条件では要求精度を満たさなかった．$h = 0.9$ における $q = 0.8, 0.85$ の設定では，AI ワーカが処理した部分のタスク結果品質が h を上回った．これは，ノイズを含む訓練データを学習した AI ワーカが真の正解に近い分類を行ったことを意味する．一方で，h が高い設定の方が人間ワーカへの割り当てが少ない状況で AI ワーカへの割り当てが行われる傾向が見られた．

5.2.2 多数決の人数を変えた実験

人間ワーカから 1 つのタスクに対して複数のタスク結果が得られる設定で，CTA アルゴリズムの挙動を明らかにするための実験を行った．実験では人間ワーカの正答率を $h = 0.8$ とし，多数決の人数 v を変化させた．$v = \{3, 5, 7\}$ の設定における実験結果を図 5 に示す．図の構成は図 4 と同様である．

実験結果から，v が高いほど，要求精度達成率が高くなることが分かる．$v = 3$ の設定では $q = 0.8, 0.85, 0.9$ の設定でのみ要求精度を満たしたが，一方で $v = 5$ の設定では全ての q で要求精度を満たした．$v = 3$ の設定では，$q = 0.95$ のときに要求精度を達成することが出来なかったが，$v = 5, 7$ の設定では全ての設定で要求精度を満たした．図 4 の $h = 0.8$ の結果は $v = 1$ の設定とみなすことができ，一連の結果から要求精度達成率は v の増加に伴い向上したと言える．この結果は図 2 の関係からも裏付けられる．

5.2.3 活用と探索のパラメータを変えた実験

ϵ が AI ワーカへのタスク割り当て数および要求精度達成率に与える影響を明らかにする実験を行った．実験では，人間ワーカの正答率を $h = 0.8$，各タスクの人間ワーカへの最大割り当て数 $v = 5$ とし，活用と探索のパラメータごと ($\epsilon = 0.2, 05, 0.8, epsilon\_linear, epsilon\_sigmoid$) の結果を比較した．実験の結果を図 6 に示す．

$\epsilon = 0.5, 0.8$ の設定では，図 4 における $h = 0.8$ の設定と概ね同じ結果が得られた．これは，人間ワーカへの追加割り当てが十分に行われなかったことにより，AI ワーカの学習や訓練が十分に行われなかったことが原因であると考えられる．

一方で，$\epsilon = 0.2$ の設定では，全ての設定で要求精度を満たすことができた．AI ワーカへのタスク割り当て数を図 5 における $v = 5$ の設定と比較すると，$\epsilon = 0.2$ の方がより多くのタスクを AI ワーカに割り当てた．この結果から，ϵ を適切に調整することにより，全てのタスクに対して人間ワーカへの追加割り当てを行うことなく，全体として十分な品質をもたらすタスク割り当てが可能であることを示している．

ϵ をタスク割り当ての進行に伴って変更する $\epsilon = epsilon\_linear, epsilon\_sigmoid$ の設定では，$\epsilon = 0.5$ では要求精度を満たすことが出来なかった $q = 0.85$ の設定で要求精度を満たすことが出来た．さらに，$q = 0.85, 0.9$ の設定で AI ワーカにタスクを割り当てることが出来た．この結果は，人間ワーカへの追加割り当ての回数は同程度であったとしても，追加割り当てを行うタイミングを調整することによりより多くのタスクを AI ワーカに割り当てることが可能であることを示している．しかしながら，$q = 0.9, 0.95$ の設定では要求精度を満たすことが出来なかったことから，タスクへの追加割り当ての許容回数や ϵ の調整手法に改良の余地がある．

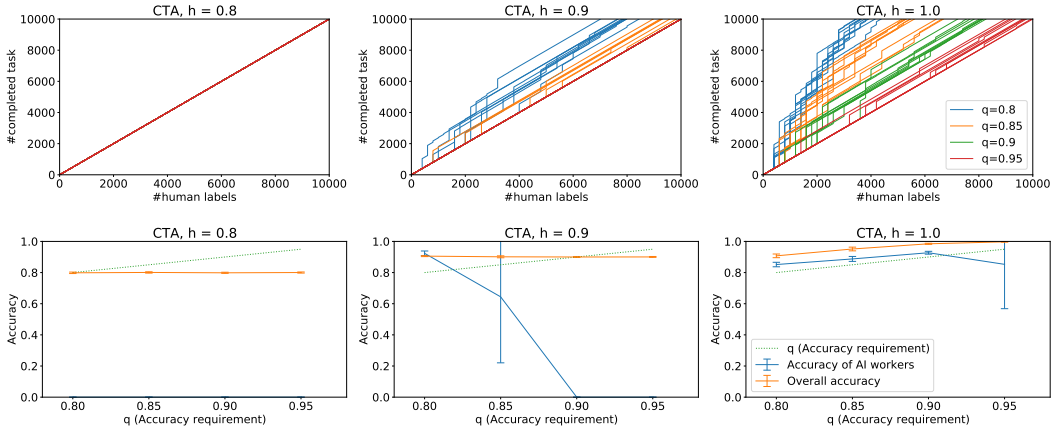

図 4　異なる人間ワーカの正答率での実験結果: 人間ワーカへのタスク割り当て数と完了タスク数の関係 (上)，要求精度と実際のタスク結果の精度の関係 (下)

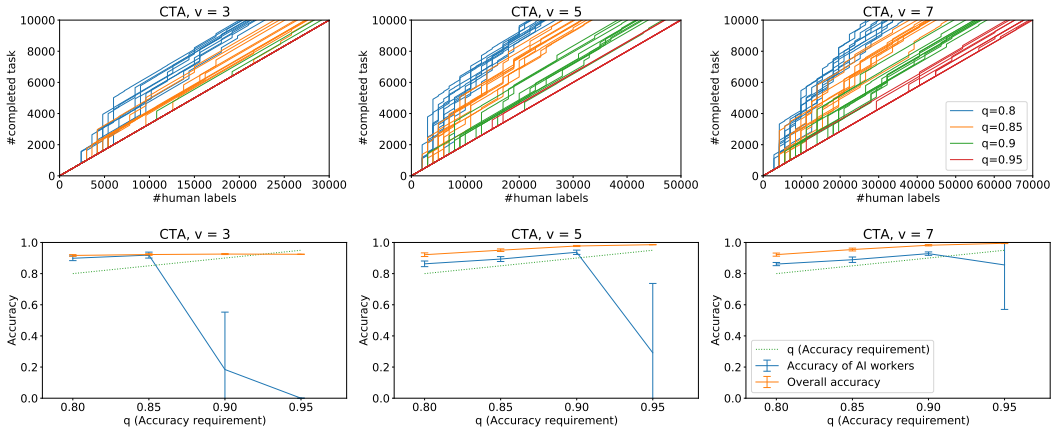

図 5　多数決の人数を変えた実験の結果: 人間ワーカへのタスク割り当て数と完了タスク数の関係 (上)，要求精度と実際のタスク結果の精度の関係 (下)

5.2.4　人間ワーカへの追加割り当ての戦略を変えた実験

人間ワーカと AI ワーカのタスク結果の比較方法が AI ワーカへのタスク割り当て数に与える影響を明らかにするための実験を行った．実験では，人間ワーカの正答率を $h = 0.8$，各タスクの人間ワーカへの最大割り当て数 $v = 5$, $\epsilon = 0.2, epsilon\_linear, epsilon\_sigmoid$ とし，追加割り当ての基準が一致およびランダム選択の条件を比較した．

追加割り当ての基準を人間と AI ワーカの一致とした場合の結果を図 7 に示す．不一致に基づく場合の結果である図 6 では AI ワーカが処理したタスクの精度が q を上回っていた $q = 0.9$ について，一致に基づく設定では q を満たさなかった．このことから，一致に基づく追加割り当てよりも不一致に基づく追加割り当ての方が，より多くのタスクを AI ワーカに割り当てた．

ランダム選択の結果を図 8 に示す．不一致選択の設定では AI ワーカの精度が要求精度を満たすケースが見られた $\epsilon = 0.2, q = 0.95$，および $\epsilon = sigmoid, q = 0.95$ の結果について，ランダムの選択の設定では要求精度を満たせなかった．このことから，要求精度が低い条件では追加割り当てをランダムに選ぶ場合と不一致に基づいて選ぶ場合で違いは無いが，要求精度が高い条件では不一致に基づく戦略が有効であった．

5.3　考　　察

人間ワーカの正答率を変えた実験では，h が低い設定では CTA で要求精度を満たす割り当てを得られないことを示した．これは，CTA が人間ワーカのタスク結果が正しいことを前提とするアルゴリズムだからである．このことから，人間ワーカから得られるタスク結果品質を高めることが重要であると言える．

多数決の人数を変えた実験では，v を高く設定するほど要求精度を満たせる割合が向上した．人間ワーカに割り当てる全てのタスクに多数決を適用することが，前述の課題に有効である．

活用と探索パラメータを変えた実験では，ϵ を適切に設定することで，要求精度を満たしながら人間ワーカへの追加タスク割り当て数を削減できることを示唆する結果が得られた．タスクの進行に伴って ϵ を変動させた設定は，固定値を用いる設定よりも人間ワーカへの追加割り当て数が少なかったことから，タスクの進行状況やその他の情報から活用と探索のパラメータを調節することが有効であると言える．

人間ワーカへの追加割り当て戦略を比較した実験では，特に要求精度が高い設定において，人間ワーカと AI ワーカの不一致に基づく戦略が有効であった．本論文では，各 AI ワーカと人間ワーカのタスク結果が不一致であった数が多いタスクから

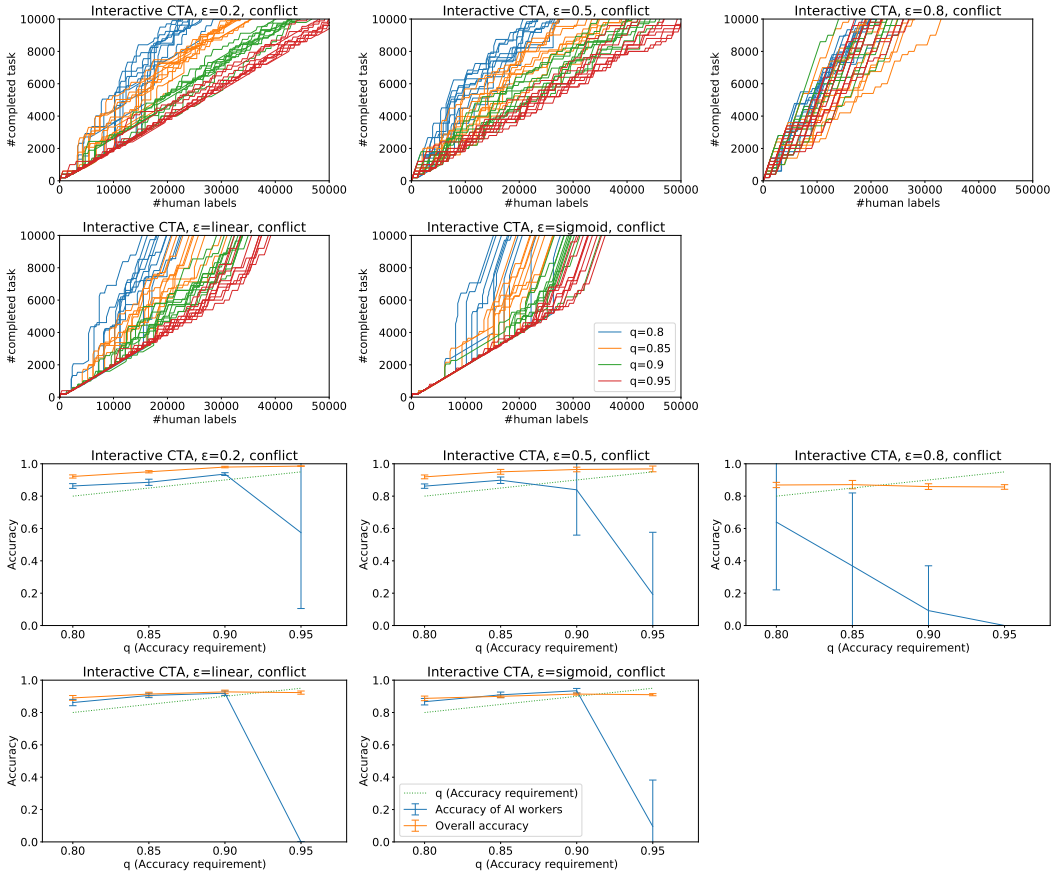

図 6　活用と探索のパラメータを変えた実験の結果: 人間ワーカへのタスク割り当て数と完了タスク数の関係 (上)，要求精度と実際のタスク結果の精度の関係 (下)

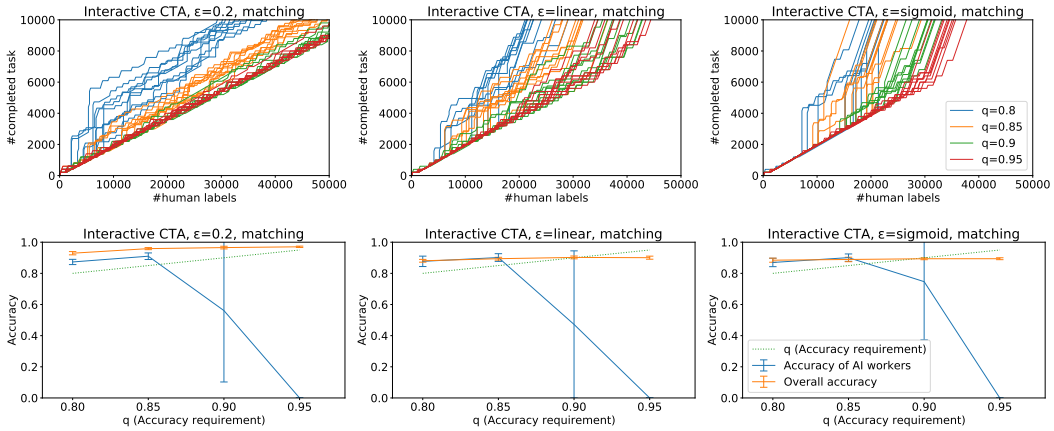

図 7　人間と AI のタスク結果の一致に基づく追加割り当ての実験結果：人間ワーカへのタスク割り当て数と完了タスク数の関係 (上)，要求精度と実際のタスク結果の精度の関係 (下)

追加割り当てを行ったが，各 AI ワーカやタスククラスタの評価結果を用いて重みを調整するなど，追加割り当てを行うタスクの選択方法には改善の余地があると考えられる．

6　ま　と　め

本研究では，人間ワーカと AI ワーカの相互作用を設計することで，人間ワーカからのタスク結果品質を改善しながら，より多くのタスクを AI ワーカに割り当てる手法について議論した．提案手法は，全てのタスクに人間ワーカを冗長に割り当てるよりも少ないタスク割り当て数で，同等の全体的なタスク結果品質をもたらすことが明らかになった．

この結果から，人間ワーカのタスク結果が不正確な状況であり，AI ワーカの品質を統計的に評価することは難しい状況において，人間ワーカのタスク結果の信頼性を高めることで AI ワーカを活用することができるようになることを示した．

今後の課題として，ワーカから得られるタスク結果品質が一定でない場合を考慮したアルゴリズムの開発が挙げられる．

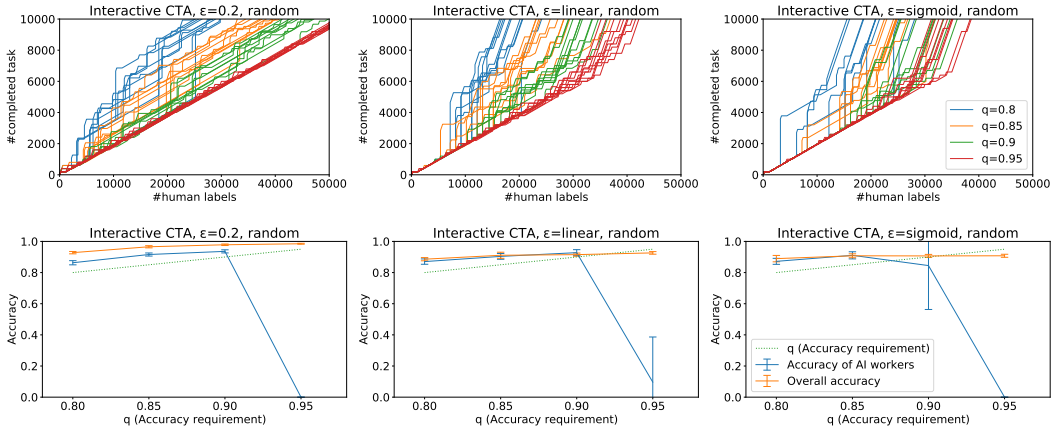

図 8　追加割り当てを行うタスクをランダムに選択した実験の結果：人間ワーカへのタスク割り当て数と完了タスク数の関係 (上)，要求精度と実際のタスク結果の精度の関係 (下)

謝　　辞

本研究の一部は JST CREST (#JPMJCR16E3)，AIP チャレンジの支援による．

文　　献

[1] Masaki Kobayashi, Kei Wakabayashi, and Atsuyuki Morishima. Quality-aware dynamic task assignment in human+ai crowd. In *Companion of The 2020 Web Conference 2020*, pp. 118–119, 2020.

[2] E.F. Moore and C.E. Shannon. Reliable circuits using less reliable relays. *Journal of the Franklin Institute*, Vol. 262, No. 3, pp. 191 – 208, 1956.

[3] Florian Daniel, Pavel Kucherbaev, Cinzia Cappiello, Boualem Benatallah, and Mohammad Allahbakhsh. 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, January 2018.

[4] Burr Settles. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2009.

[5] Natsumaro Kutsuna, Takumi Higaki, Sachihiro Matsunaga, Tomoshi Otsuki, Masayuki Yamaguchi, Hirofumi Fujii, and Seiichiro Hasezawa. Active learning framework with iterative clustering for bioimage classification. *Nature communications*, Vol. 3, p. 1032, 2012.

[6] Yan Yan, Romer Rosales, Glenn Fung, and Jennifer G. Dy. Active learning from crowds. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pp. 1161–1168, USA, 2011. Omnipress.

[7] Jie Yang, Thomas Drake, Andreas Damianou, and Yoelle Maarek. Leveraging crowdsourcing data for deep active learning an application: Learning intents in alexa. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, pp. 23–32, Republic and Canton of Geneva, Switzerland, 2018. International World Wide Web Conferences Steering Committee.

[8] Thomas G. Dietterich. Ensemble methods in machine learning. In *Multiple Classifier Systems*, pp. 1–15, Berlin, Heidelberg, 2000. Springer Berlin Heidelberg.

[9] W. H. Beluch, T. Genewein, A. Nurnberger, and J. M. Kohler. The power of ensembles for active learning in image classification. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9368–9377, Los Alamitos, CA, USA, jun 2018. IEEE Computer Society.

[10] Xin He, Kaiyong Zhao, and Xiaowen Chu. Automl: A survey of the state-of-the-art. *Knowledge-Based Systems*, p. 106622, 2020.

[11] Jie Yang, Alisa Smirnova, Dingqi Yang, Gianluca Demartini, Yuan Lu, and Philippe Cudre-Mauroux. Scalpel-cd: Leveraging crowdsourcing and deep probabilistic modeling for debugging noisy training data. In *The World Wide Web Conference*, WWW '19, pp. 2158–2168, New York, NY, USA, 2019. ACM.

[12] Q. Vera Liao and Michael Muller. Human-ai collaboration : Towards socially-guided machine learning. 2019.

[13] O. Russakovsky, L. Li, and L. Fei-Fei. Best of both worlds: Human-machine collaboration for object annotation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2121–2131, June 2015.

[14] Joseph Chee Chang, Saleema Amershi, and Ece Kamar. Revolt: Collaborative crowdsourcing for labeling machine learning datasets. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI 2017)*. ACM - Association for Computing Machinery, May 2017.

[15] An Thanh Nguyen, Byron C Wallace, and Matthew Lease. Combining crowd and expert labels using decision theoretic active learning. In *Proceedings of the 3rd AAAI Conference on Human Computation and Crowdsourcing*. aaai.org, September 2015.

[16] Azad Abad, Moin Nabi, and Alessandro Moschitti. Autonomous crowdsourcing through human-machine collaborative learning. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, pp. 873–876, New York, NY, USA, 2017. ACM.

[17] Besmira Nushi, Ece Kamar, and Eric Horvitz. Towards accountable ai: Hybrid human-machine analyses for characterizing system failure. In *Proceedings of the 6th AAAI Conference on Human Computation and Crowdsourcing*, 2018.

[18] Yolanda Gil, James Honaker, Shikhar Gupta, Yibo Ma, Vito D'Orazio, Daniel Garijo, Shruti Gadewar, Qifan Yang, and Neda Jahanshad. Towards human-guided machine learning. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, IUI '19, pp. 614–624, New York, NY, USA, 2019. ACM.

[19] Tarin Clanuwat, Mikel Bober-Irizar, Asanobu Kitamoto, Alex Lamb, Kazuaki Yamamoto, and David Ha. Deep learning for classical japanese literature. *CoRR*, Vol. abs/1812.01718, , 2018.