# 人間+AI Crowdの相互作用による タスク結果品質の管理手法

小林正樹[1]　若林啓[2]　森嶋厚行[3]

本論文では，群衆によって開発された不特定多数のAIプログラムを，人間のクラウドワーカと同様に扱う"人間+AI Crowd"世界でのクラウドソーシングにおいて，必ずしも正答するとは限らない人間およびAIワーカの相互作用により，低コストで高品質なタスク結果を実現するタスク割り当て問題を扱う．この"人間+AI Crowd"世界では，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，複数のAIワーカの学習と評価に利用される．そして，最初は人間によって行われていた作業のうち，AI化が可能な部分が徐々に自動的にAIによる作業に置き換えられる．したがって，人間ワーカから得られるタスク結果品質は特に重要であるが，現実のクラウドソーシングでは，様々な理由で人間ワーカからのタスク結果が不正確な可能性があり，これにより最終的なタスク結果品質の低下を引き起こす．そのため，多数決などの集約手法を適用することが一般的であるが，このコストをできるだけ削減したい．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，人間ワーカの追加タスクの必要性を判定することで，人間ワーカのタスク数の増加を抑えながら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を用いた実験結果から，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．

## 1　はじめに

クラウドソーシングは不特定多数の人間のワーカに作業を依頼することで，依頼者の課題を解決するための有効な手段である．クラウドソーシングの課題として，実際に人間のワーカが作業することから，得られた作業結果に誤りを含む可能性があることや，作業を依頼可能な人間ワーカは限られているため，作業効率には限度があることが挙げられる．

一方で近年，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，SIGNATE*3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最先端のAI技術の恩恵を受けられる．

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，作成した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できる．しかし，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．

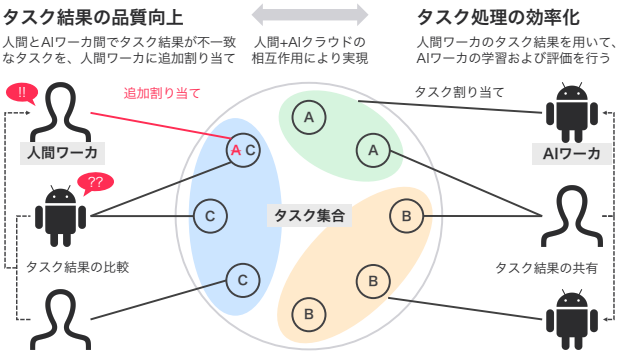


図1　本論文では，人間ワーカとAIワーカの相互作用を設計することで，全体的なタスク結果品質を改善しながらタスクを処理する仕組みを提案する．

本研究はこの問題について，AIプログラムが利用するモデルやアルゴリズムの情報を使わず，ブラックボックスな"AIワーカ"としてモデル化して扱う"人間+AI Crowd"アプローチを提案する．まず，先行研究で「人間ワーカは正しい」という前提で，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 [9, 10]．このアルゴリズムは，人間ワーカからの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ワーカの出力の全体を評価するのではなく，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）に評価し，より多くのタスクをAIワーカに割り当てる．

先行研究 [9, 10]のアルゴリズムでは，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．そのため，ワーカは常に正しいという前提を置いているが，現実には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．クラウドソーシングにおける品質管理の研究では，異なる複数の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して，より正確なタスク結果を得る多数決などの集約に基づく方法が広く用いられている [12]．これにより，全データに対して複数のタスク結果の集約を適用すれば，人間ワー

[1] 学生会員　筑波大学図書館情報メディア研究科
makky@klis.tsukuba.ac.jp
[2] 正会員　筑波大学図書館情報メディア系
kwakaba@slis.tsukuba.ac.jp
[3] 正会員　筑波大学図書館情報メディア系
morishima-office@ml.cc.tsukuba.ac.jp
*1 https://www.kaggle.com
*2 https://www.aicrowd.com
*3 https://signate.jp

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．

そこで本論文では，人間ワーカとAIワーカの結果を相互に比較して，全体的なタスク結果の品質向上を行う仕組みを提案する（図1）．本論文では，正解のあるタスクを対象とし，品質が高い結果とは，その正解率が高いことを意味する．また，品質が高いワーカとは，そのタスクに対して高品質が期待されるワーカである．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の品質（正解率）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と大きく変わらない品質のまま，人間ワーカへの追加割り当て数を大幅に削減できる可能性が示唆された．本手法は理論的に最悪のケースでも既存手法と同等の結果をもたらすため，実用的であると考えられる．

本研究の貢献は次の通りである:

- 先行研究 [9, 10]で提案した人間＋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評価した．
- 既存のタスク割り当てアルゴリズム [9, 10]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，人間ワーカへの追加割り当て数を大幅に削減しながら，多数決を適用する手法と大きく変わらない品質の結果をもたらすケースがあることを示した．本手法での最悪のケースは，最大限人間に割り当てる場合だが，その場合も既存手法と同等の結果になる．

## 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様々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，品質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 [4, 8]. 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．本論文では人間ワーカのタスク結果品質を多数決により改善できることを前提としたアルゴリズムを提案する．一方で，大半のワーカが誤った回答をするような，多数決が機能しない状況に対応するために，各ワーカの能力や対象のタスク以外への結果を考慮した回答統合手法と提案アルゴリズムを組み合わせることは興味深い課題である [5, 13, 20].

人間により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り次に人間によるラベル付けを必要とするデータを決定して，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 [11, 17, 18]. 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対象の機械学習モデルの性能の最大化であるのに対し，本研究ではAIワーカと人間ワーカに適切にタスクを割り当てることで，全体として要求精度を満たすのが目的である．(2) 能動学習は学習ループの実行を開始する前に，対象の機械学習モデルを与える必要があるが，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のAIワーカの増減を許す．

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）し，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 [6]. しかし，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求められる．

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 [1, 2, 7, 14–16, 19]. 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依頼者の要求精度に応じて自動的に決定する．

## 3 問題設定と既存手法

先行研究 [9, 10]で提案された人間ワーカおよびAIワーカに対するタスク割り当て問題と，そのタスク割り当てアルゴリズムについて説明する．

### 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 $w \in W$ は関数 $w : T \to \mathbb{N}$ であり，入力のタスクに対して $A$ に限定されないラベルを返す．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．個々の人間ワーカ毎のスキルや能力を考慮しないことから，人間ワーカを単に ‘h’ と表記する．

本研究が扱うのは，依頼者が設定する全体的なタスク結果の要求精度 $q$ （$0 \le q \le 1$）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．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 $(t, w)$ で構成される．ただし，$w \in W \cup \{\text{‘}h\text{’}\}$ および $t \in T$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（タスク集合の部分集合）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，多くの場合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． $R_w = \{(t_i, t_j) \mid w(t_i) = w(t_j), t_i, t_j \in T, t_i \neq t_j\}$ をあるAIワーカ $w$ 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 $w$ から得られるタスククラスタ集合は商集合 $C_w = T/R_w = \{T_{w,0}, T_{w,1}, ..., T_{w,\mathbb{N}}\}$ 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．全AIワーカから得られる全てのタスククラスタは $C = \bigcup_{w \in W} C_w$ と表記する．

先行研究 [9,10]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 [9,10]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 $\alpha$ 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 $S$ であり，これはアルゴリズム中の *assign* 関数の実行履歴に相当する． 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，$ans : T \to A \times (W \cup \{\text{`}h\text{'}\}) \cup \{(\emptyset, \emptyset)\}$ 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．あるタスクに対するラベルとワーカが未定の場合は$(\emptyset, \emptyset)$を返す．選択されたタスクは，*assign*関数によって人間ワーカに割り当てられ，*task_result*関数によってタスク結果を取得する（2行目）．その後，得られた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，タスククラスタ中の人間ワーカによるラベル済みタスクのうち，最頻出ラベルを $\hat{a}$ とおく． 6行目では *statistical_test* 関数により，タスククラスタの各タスクのラベルを $\hat{a}$ とした場合に，タスククラスタから得られるタスク結果品質が要求精度 $q$ を有意水準 $\alpha$ で上回るかを判定する．先行研究では *statistical_test* 関数として二項検定を用い，タスククラスタから得られるタスク結果品質と要求精度の間に差がないことを帰無仮説として，片側検定を行った．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$(\hat{a}, w)$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 $\alpha$.
**Output:** A sequence of (task, worker) pairs.

1: **for all** $t \in T$ s.t. $ans(t) = (\emptyset, \emptyset)$ in a random order **do**
2: $\quad a' \leftarrow task\_result(assign(t, \text{`}h\text{'}))$
3: $\quad$ update $ans$ so that $ans(t) = (a', \text{`}h\text{'})$
4: $\quad$ **for all** $T_{w,i} \in C$ **do**
5: $\quad\quad \hat{a} = \arg\max_{a \in A} |\{t \mid t \in T_{w,i}, ans(t) = (a, \text{`}h\text{'})\}|$
6: $\quad\quad$ **if** $statistical\_test(T_{w,i}, \hat{a}, q, \alpha)$ **then**
7: $\quad\quad\quad$ for $t' \in T_{w,i}$ s.t. $ans(t') = (\emptyset, \emptyset)$, $assign(t', w)$ and update $ans$ so that $ans(t') = (\hat{a}, w)$
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状況を考慮した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は，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への多数決の適用である．多数決により一部の回答が不正確であっても全体として品質の高いタスク結果を得ることができる．図2は，2値分類における，多数決を行う人数（$v$）と個々の人間ワーカの正答率（$h$）の組み合わせにおいて，多数決結果の正答率のシミュレーション結果を図示したものである．多値分類で間違いパターンが偏らない場合などでは多数決が有効に働く閾値は0.5ではないが，傾向は同様になる．つまり，ワーカの平均正解率が閾値を超えている場合に多数決は有効であり，より多くのワーカが参加すれば正解

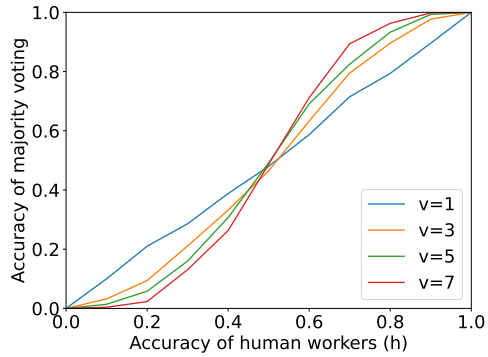


図2　個々のワーカの正答率と多数決する人数ごとの正答率の関係.

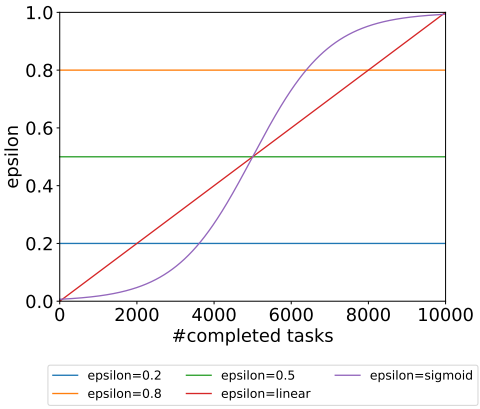


図3　総タスク数が10000の場合の，タスク割り当ての進行と各探索方策における $\epsilon$ の関係.

率が上がる．本論文の提案アルゴリズムは，このような多数決によるタスク結果品質の改善が有効なケースを対象にする．

一方，人間ワーカに割り当てる全タスクへの多数決は，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たタスクに注目する．不一致であるは，人間ワーカもしくはAIワーカのいずれかもしくは両方が誤りである可能性が高いので，(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う．この処理を繰り返すことで，より正確に全タスクを処理することを目指す．

#### 4.2.2　活用と探索

本研究の問題設定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のタスク結果品質が重要である．しかし，多数決による品質管理には十分な票数が必要である．

そこで，人間ワーカへの追加割り当てを探索，AIワーカへのタスク割り当てを活用とし，両者の処理をバランス良く実行することを考える．ここで，活用と探索の処理の使い分けを制御するパラメータ $\epsilon$ を導入する．ただし，$0 \leq \epsilon \leq 1$である． $[0, 1]$ の範囲の一様分布 $U$ から乱数 $u \sim U$ を取得する． $\epsilon < u$ の場合，探索を行う．具体的には，4.2.1節で説明した通り，人間ワーカとAIワーカのタスク結果の不一致に基づき，人間ワーカに対して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を向上させる．一方で $\epsilon \geqq u$ の場合，活用を行う．具体的には，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．

全体としてより多くのタスク結果をAIワーカから採用するために，タスク割り当ての進行に応じて $\epsilon$ を動的に変動させることを考える．人間ワーカへの割り当てタスク数が少ない状態では，AIワーカからのタスク結果を採用できる可能性は低いので，優先的に探索を実行することが有効である．しかし，人間ワーカへの割り当て済みタスク数が増えるにつれ，AIワーカの性能が十分となる可能性が高くなるので，活用を優先して行うことが有効である．式1と式2に，タスク割り当ての進行に応じて値が変動する $\epsilon_{linear}$ と $\epsilon_{sigmoid}$ を示す．

$$\epsilon_{linear} = \frac{|t \mid t \in T, ans(t) = (\emptyset, \emptyset)|}{|T|} \tag{1}$$

$$\epsilon_{sigmoid} = \frac{1}{2}(1 + \tanh(\frac{1}{2}\alpha\epsilon_{linear})) \tag{2}$$

ただし，実験では $\alpha = 10$ を用いた．図3に，完了タスク数の増加に伴う $\epsilon$ の変化と，実験で比較する $\epsilon = 0.2, 0.5, 0.8$を示す．

### 4.3　提案アルゴリズム

提案アルゴリズムであるInteractive Clusterwise Test-based Assignment（アルゴリズム2）について説明する． ICTA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，高品質な訓練およびテストデータによってAIワーカを活用する．これにより，人間ワーカとAIワーカのタスク結果品質を相互に改善しながらタスク割り当てを行う． 4行目で $\epsilon$ と乱数を比較し，探索（5 - 12行目）を行うか，活用（14 - 19行目）を行うかを決定する． 6 - 9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する．10行目で最も不一致の頻度が高いタスクを選択し，11行目でそのタスクを $v - 1$ 名の追加の人間ワーカに割り当てた上で$majority_vote$関数により多数決を行う．12行目では選択したタスクの結果を多数決の結果で更新する．

## 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 [3]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モデルを初期

**Algorithm 2:** Intaractive Clusterwise Test-based Assignment（ICTA）

**Input:** A set $T$ of tasks, a set $W$ of AI workers, the accuracy requirement $q$, the significance level $\alpha$, and the intaractive threshold $\epsilon$.
**Output:** A sequence of (task, worker) pairs.

**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** $\epsilon < random()$ **then**
    let $Q$ be a multiset
    **for all** $T_k \in C$ **do**
      $\hat{a} = \arg\max_{a \in A} |\{t \mid t \in T_k, ans(t) = (a, `h')\}|$
      $Q \leftarrow Q \cup \{t \mid t \in T_k, a \in A, ans(t) = (a, `h'), a \neq \hat{a}\}$
    **end for**
    let $t'$ be the most frequent element in $Q$
    $a' \leftarrow majority\_vote([ans(t'), task\_result(assign(t', `h'))])$
    update $ans$ so that $ans(t') = (a', `h`)$
  **else**
    **for all** $T_{w,j} \in C$ **do**
      $\hat{a} = \arg\max_{a \in A} |\{t \mid t \in T_{w,j}, ans(t) = (a, `h')\}|$
      **if** $statistical\_test(T_{w,j}, \hat{a}, q, \alpha)$ **then**
        for $t' \in T_{w,j}$ s.t. $ans(t') = (\emptyset, \emptyset)$, $assign(t', w)$ and update $ans$ so that $ans(t') = (\hat{a}, w)$
      **end if**
    **end for**
  **end if**
**end for**

パラメータの設定で用いた: MLPClassifier，ExtraTreeClassifier，LogisticRegression，KMeans，DecisionTreeClassifier，SVC.

実験では，既存手法のCTAと提案手法のICTAを比較する．統計的検定には二項検定を利用し，有意水準は5%とした．実験で使用したソースコードはGitHubリポジトリに公開されている[*4].

### 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 $v \in \{1, 3, 5, 7\}$
- 人間ワーカの正答率 $h \in \{0.8, 0.9, 1.0\}$
- 要求精度 $q \in \{0.8, 0.85, 0.9, 0.95\}$
- $\epsilon \in \{0.2, 0.5, 0.8, \epsilon_{linear}, \epsilon_{sigmoid}\}$
- 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 ($h \in \{0.8, 0.9, 1.0\}$) におけるCTA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．エラーバーは標準偏差を表す．上段の図において，横軸に比例した総タスク数の増加は人間ワーカによるタスク処理を意味し，一方で横軸の値の変動を伴わない総タスク数の増加はAIワーカによるタスク処理を意味する．下段の図において，全体的なタスク結果が，緑色の点線で示される要求精度を下回っている場合，そのタスク割り当ては依頼者の要求を満たしていない．青色で示されるAIワーカが処理した部分のタスク結果は要求精度を上回っている必要はないが，アルゴリズムの評価のため採用したAIワーカの実際の品質を示したものである．

実験結果から，$h$ が低いほど要求精度を達成できる割合が低下する．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 $q = 0.8$ 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，$h$ が高いと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

CTAで人間ワーカにタスク割り当てる際に多数決を行う設定で，アルゴリズムの挙動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 $h = 0.8$ とし，多数決の人数 $v$ を変化させた． $v \in \{3, 5, 7\}$ の設定における実験結果を図5に示す．

実験結果から，$v$ が高いほど，要求精度を満たせる頻度が増加した． $v = 3$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 $v = 1$ の設定と同等であり，一連の結果から要求精度を満たせる頻度は $v$ 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

$\epsilon$ がAIワーカへのタスク割り当て数および全体的なタスク結果品質に与える影響を明らかにする．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 ($\epsilon \in \{0.2, 0.5, 0.8, \epsilon_{linear}, \epsilon_{sigmoid}\}$) の結果を比較した（図6）．

$\epsilon = 0.2, 0.5$ 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ワーカに割り当てていることが分かる．一方で，$\epsilon = 0.8$ の設定では $q = 0.8, 0.85$ の場合のみ，要求精度を満たすことができた．この結果から，$\epsilon$ の適切な調整により，全タスクに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が示唆された．

要求精度 $q = 0.8, 0.85$ の設定において，$\epsilon = 0.8$ではAIワーカから得られたタスク結果の品質が要求精度を下回ったが，$\epsilon$ をタスク割り当ての進行に伴って変更する $\epsilon_{linear}, \epsilon_{sigmoid}$では上回った．さらに，$q = 0.9$ において全体のタスク結果品質が要求精度を上回った．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ればより多くのタスクをAIワーカに割り当てられることを示している．しかし，$q = 0.95$ の設定では要求精度を満たせず，追加割り当ての許容回数や$\epsilon$の調整手法に改良の余地がある．

#### 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

[*4] https://github.com/crowd4u/HACTAP-Framework

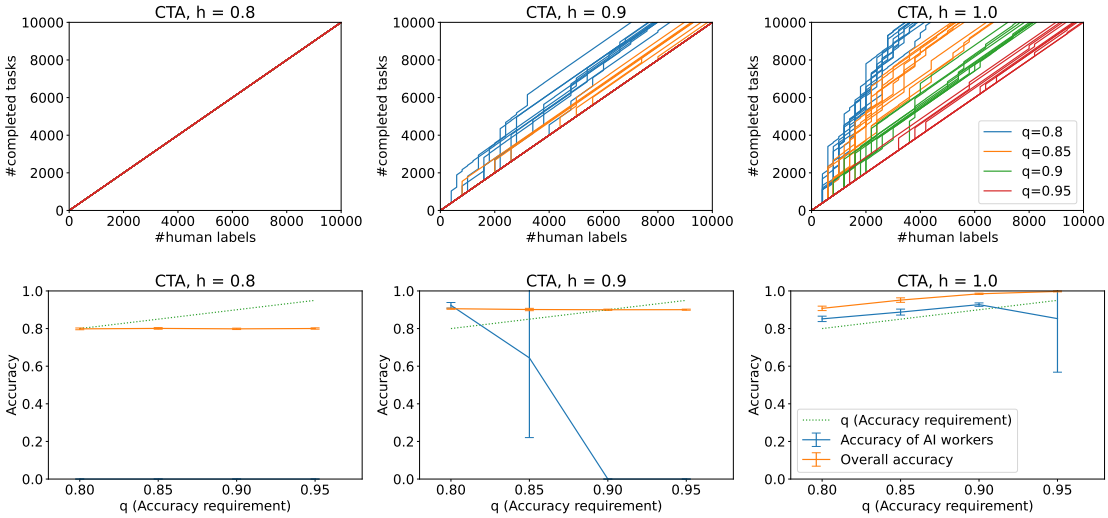


図4　異なる人間ワーカの正答率での実験結果: 人間ワーカへのタスク割り当て数と完了タスク数の関係 (上)，要求精度と実際のタスク結果の精度の関係 (下)

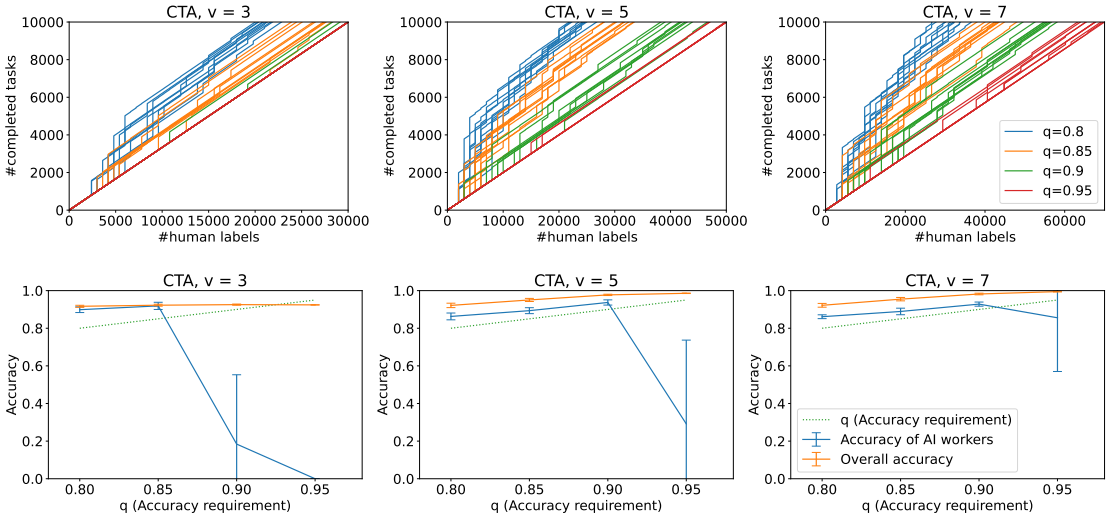


図5　多数決の人数を変えた実験の結果: 人間ワーカへのタスク割り当て数と完了タスク数の関係 (上)，要求精度と実際のタスク結果の精度の関係 (下)

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 $v = 5, \epsilon \in \{0.2, \epsilon_{linear}, \epsilon_{sigmoid}\}$ とし，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左に示す．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．ランダム選択の結果を図7右に示す．不一致選択の設定ではAIワーカの精度が要求精度を上回った $\epsilon_{sigmoid}, q = 0.9$ の設定において，ランダム選択では割り当てられたAIワーカの精度の平均値が要求精度を下回った．このことから，人間ワーカへの追加割り当てを行うタスクを選ぶために，人間ワーカとAIワーカの不一致に基づく手法が有効であった．この結果は，不一致に基づいて追加割り当てを行うことが，AIワーカの訓練および評価のためのデータ品質改善に有効であり，その結果，AIワーカの各タスククラスタが採用出来たからであると考えられる．

### 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かった．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タスクへの多数決が，前述の課題に有効である．

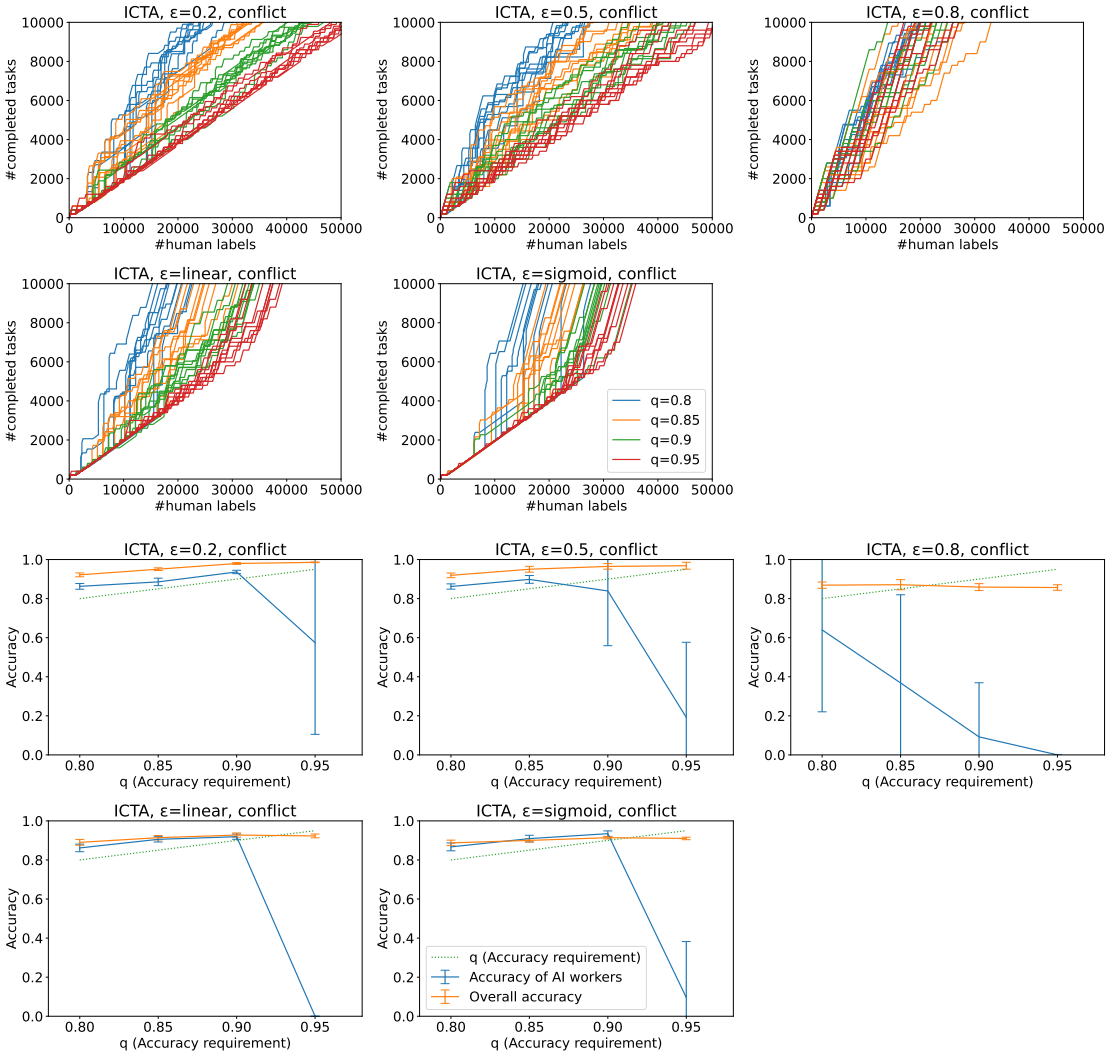


図6　活用と探索のパラメータを変えた実験の結果: 人間ワーカへのタスク割り当て数と完了タスク数の関係 (上)，要求精度と実際のタスク結果の精度の関係 (下)

提案手法において，活用と探索パラメータを変えた実験では，$\epsilon$ の適切な設定により，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 $\epsilon$ 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に基づく$\epsilon$の調節が有効であると言える．一方で，タスクの進行に伴って $\epsilon$ を変動させた設定は，固定値を用いる設定よりも，最初にAIワーカにタスクを割り当てるまでに多くの人間ワーカ割り当てを必要とする傾向が見られた．これは，タスク割り当ての序盤にAIワーカの評価が行われる確率が低いからである．特に要求精度が低い場合に，要求精度を満たすAIワーカが存在するにも関わらず，探索により結果として人間ワーカへのタスク割り当て数が増大すると予想される．このような観点からも $\epsilon$ の動的な変更には改善の余地がある．

提案手法において，人間ワーカへの追加割り当て戦略を比較した実験では，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，不一致であるタスクに基づいてのみ追加割り当てを行ったが，タスククラスタの評価結果に基づいて重みづけするなど，追加割り当てを行うタスクの選択手法には改善の余地がある．さらに，観測された人間ワーカとAIワーカのタスク結果の不一致を，$\epsilon$の動的な調整の根拠として用いることも提案手法の改良の方針として挙げられる．

## 6　まとめ

本研究では，人間ワーカとAIワーカのタスク結果を相互に比較することで，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．実験では，提案手法が人間ワーカとAIワーカの不一致に基づいて人間ワーカへの追加割り当てを行うことで，人間ワーカに割り当てる全タスクに多数決を適用するよりも人間ワーカへの割り当て数を大幅に削減しながら，多数決を適用する場合と同程度のタスク結果品質を得られるケースがあること示した．この結果か

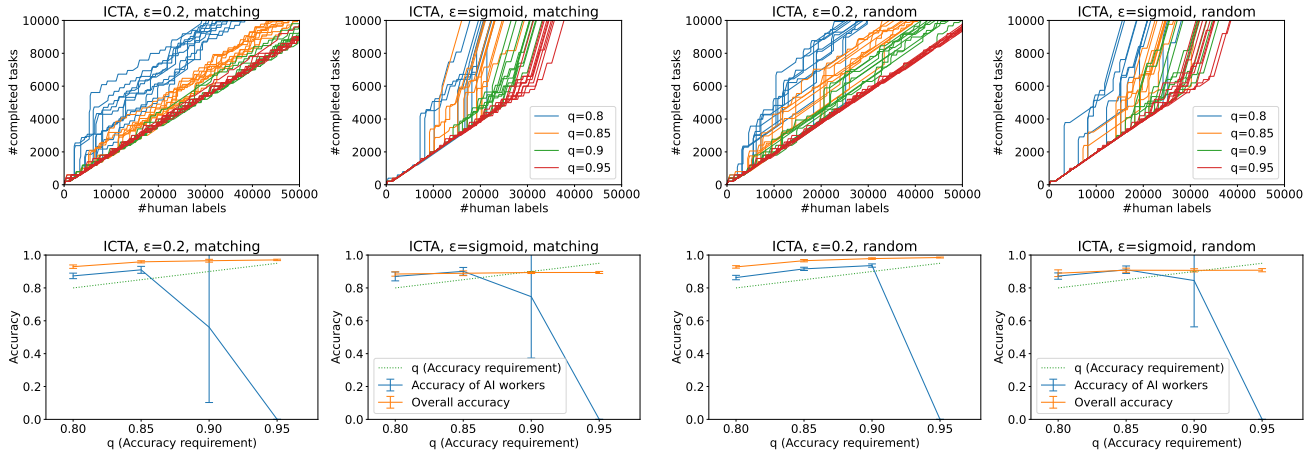


図7　追加割り当てを行うタスクを一致およびランダムに選択した実験の結果：人間ワーカへのタスク割り当て数と完了タスク数の関係 (上)，要求精度と実際のタスク結果の精度の関係 (下)

ら，人間ワーカのタスク結果が不正確であり，AIワーカの学習及びそ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品質を高めることでAIワーカをより活用できることが示唆された．

本論文では，人間ワーカが正答する確率が一定であると仮定し，確率を変化させた実験を行なった．今後の課題として，クラウドソーシング実験の経験則等に基づく合理的な確率分布を仮定したり，個々の人間ワーカの正答率が異なる状況を考慮した実験およびアルゴリズムの改良が挙げられる．

## 謝辞

本研究の一部はJST CREST (#JPMJCR16E3)，AIPチャレンジの支援による．本研究を進めるにあたり，ご助言を頂いた田中和世先生，伊藤寛祥先生に感謝致します．

## 参考文献

[1] 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, pages 873–876, New York, NY, USA, 2017. ACM.

[2] 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)*, May 2017.

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

[4] 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.*, 51(1):7:1–7:40, January 2018.

[5] Alexander Philip Dawid and Allan M Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28, 1979.

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

[7] 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, pages 614–624, New York, NY, USA, 2019. ACM.

[8] Hisashi KASHIMA, Yukino BABA, Kashima Hisashi, and Baba Yukino. Human computation(<special issue>human computation and crowdsourcing). *Journal of the Japanese Society for Artificial Intelligence*, 29(1):4–11, jan 2014.

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

[10] Masaki Kobayashi, Kei Wakabayashi, and Atsuyuki Morishima. Human+ai crowd task assignment considering result quality requirements. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 9(1), 2021.

[11] 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*, 3:1032, 2012.

[12] E.F. Moore and C.E. Shannon. Reliable circuits using less reliable relays. *Journal of the Franklin Institute*, 262(3):191 – 208, 1956.

[13] Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

[14] 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.

[15] 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.

[16] 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)*, pages 2121–2131, June 2015.

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

[18] 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, pages 1161–1168, USA, 2011. Omnipress.

[19] 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, pages 2158–2168, New York, NY, USA, 2019. ACM.

[20] 吉屹李, 雪乃馬場, and 久嗣鹿島. 超問題：専門知識を要するクラウドソーシングタスクの回答統合法. 日本データベース学会和文論文誌, 17-J(8), 3 2019.