クラウドソーシングにおけるAIを利用したタスク削減手法

山下　裕† 小林　正樹†† 若林　啓††† 森嶋　厚行†††

† 筑波大学 知識情報・図書館学類　〒305-0821 茨城県つくば市春日 1-2
†† 筑波大学 図書館情報メディア研究科　〒305-0821 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　〒305-0821 茨城県つくば市春日 1-2
E-mail: †s1811547@s.tuskuba.ac.jp, ††makky@klis.tsukuba.ac.jp, †††{kwakaba,mori}@slis.tsukuba.ac.jp

あらまし　クラウドソーシングにおいて，品質の高い回答を得るには多くのコストがかかる．そこで本研究では，品質を保ちながら人間ワーカが行うタスク数を削減することを目指す．クラウドソーシングにおける品質管理の手法として，EM アルゴリズムを用い正解と人間ワーカの能力を交互に推定する Dawid と Skene のモデルがある．本研究では，この Dawid と Skene のモデルを応用した手法を提案する．この手法には次の 2 つの特徴がある．1 つ目は Dawid と Skene のモデルの期待値に基づいて，動的な割り当てを行うことである．2 つ目は AI ワーカを導入し，1 人の人間ワーカとして扱うことである．これが有効な手法か確かめるため，様々な種類のデータに対して実験を行い，結果を検証した．結果から特定の種類のタスクにおいて，品質を保ったままタスクを削減できることがわかった．

キーワード　クラウドソーシング，機械学習，タスク割り当て

1 はじめに

クラウドソーシングとは，不特定多数の群衆に仕事を依頼するものである．近年このクラウドソーシングが，教師データのアノテーションなど多くの場面で使用されるようになっている．

クラウドソーシングを実行する際の品質管理手法として最も単純なものが，単純多数決である．図 1 の上の図は，この単純多数決を表したものである．同じタスクに対して複数のワーカに回答を依頼し，多数決をとることで品質を向上させるものである．

単純多数決ではワーカの能力を等しいと仮定するが，実際の状況ではワーカの能力は異なることが多い．そのため，ワーカの能力に基づく重み付き多数決が使用される．図 1 の下の図は，この重み付き多数決を表している．あらかじめ正解のわかっているタスクを使用し，正解率の高かったワーカの票は重く，低かったワーカの票は軽くして多数決をすることで品質を高めることができる．ワーカの能力に基づく重み付け多数決では，あらかじめ正解のわかってるデータを用意することや，ワーカの能力を測るためにその分のタスクやってもらう必要がある．

他にも EM アルゴリズムを用いて，タスクの正解とワーカの能力を同時に推定する Dawid と Skene の提案した手法 [1] がある．図 2 は，Dawid と Skene の提案したモデルを表している．EM アルゴリズムは，タスクの正解を推定する E ステップと，ワーカの能力を推定する M ステップで構成されている．この手法は，ワーカの能力を測るためのタスクを必要とせず，かつ多数決と比べて品質を高めることが知られている．だが依然として，高い品質の回答を集めるにはより多くのワーカに回答をしてもらう必要があり，コストがかかる．図 2 の場合，分類対象タスク 5 つに対して，25 タスクの回答を依頼する必要がある．

そこで本研究では，品質を保ちながら人間ワーカの回答するタスク数を抑え，コストを削減することを目指す．

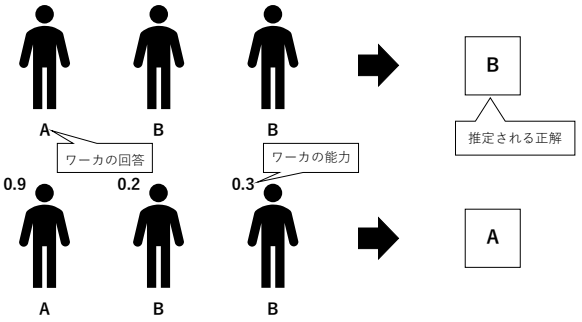

図 1　単純多数決（上）と重み付き多数決（下）

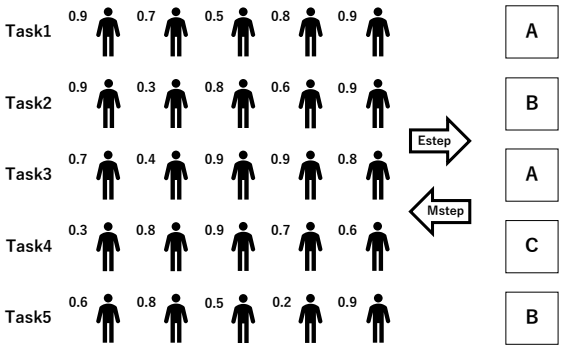

図 2　Dawid と Skene のモデル

今回提案する手法は，Dawid と Skene のモデルを応用したものである．この手法には次の 2 つの特徴がある．1 つ目は Dawid と Skene のモデルの期待値に基づいて，動的な割り当てを行うことである．2 つ目は AI ワーカを導入し，1 人の人間ワーカとして扱うことである．

実験では，AI と人間の両方が優れたタスク，AI と人間の両

方が不得意なタスク，AI が人間よりも優れたタスク，人間が AI よりも優れたタスクの全部で 4 種類のタスクを用意し結果を検証した．加えて，スパムワーカが一定数いるという状況でも実験を行った．

実験の結果，AI と人間が優れたタスクの場合，同等以上の品質でタスク数を削減することが出来た．スパムワーカが一定数いるとき，より効果を発揮した．人間が AI よりも優れたタスクの場合，タスク数を削減することは出来なかったものの，品質を悪化させることはなかった．スパムワーカが一定数いるとき，品質を保ちながらタスク数を削減することができた．これらから，人間ワーカからある程度正しい回答が手に入るタスクであれば，提案手法の有効性があることが示された．また，スパムワーカがいることで，より効果を発揮することが示された．

2 関 連 研 究

クラウドソーシングにおける品質管理手法では，多くの研究がなされている．一番簡単な手法として単純多数決がある．同じタスクに対して複数のワーカに回答を依頼し，多数決をとることで品質を向上させようとするものである．この多数決では，ワーカの正解率がある程度高くなくてはいけない [2].

単純多数決ではワーカの能力を等しいと仮定していたが，実際の状況ではワーカの能力は異なっていることが多い．こういった場合，ワーカの能力に基づく重み付き多数決が使用される．あらかじめ正解のわかっているタスクなどを使用し，ワーカの能力パラメータを推定する．正解率の高かったワーカの票は重く，低かったワーカの票は軽くして多数決をおこなうことで品質を高めることができる [3].

ワーカの能力に基づく重み付け多数決では，あらかじめ正解のわかっているデータを用意することや，ワーカの能力を測るためにその分のタスクやってもらう必要がある．EM アルゴリズムを用いて，タスクの正解とワーカの能力を同時に推定する Dawid と Skene の提案した手法がある．この手法は，ワーカの能力を測るためのタスクを必要とせず，かつ通常の多数決と比べて品質を高めることが知られている [1]．本研究では，これを応用した手法を検討する．

他にもタスクの難易度を考慮するもの [4] や，タスクとワーカの相性を考慮するもの [5]，ワーカの確信度を考慮するものがある [6]．だが，本研究でタスクの正解推定で考慮するものは，AI ワーカを含めたワーカの能力のみである．

多くの品質管理手法では，多数派の意見が採用されることが多いが，多くのワーカにとって難しいタスクでは，多数派が正解であるとは限らない．そこで専門家が少数しかいないような状況においての手法も提案されている [7].

3 提 案 手 法

本研究では，Dawid と Skene のモデルを応用した手法を提案する．提案手法には，2 つの特徴がある．1 つ目は，Dawid と Skene のモデルの期待値に基づいて，動的な割り当てを行う．2 つ目は，AI ワーカを導入し，1 人の人間ワーカとして扱う．

3.1 Dawid と Skene のモデル

Dawid と Skene のモデルでは，EM アルゴリズムを用いて，分類対象タスクの正解と人間ワーカの能力を交互に推定する．具体的には，E ステップと M ステップという以下の 2 つのステップを交互に行う．

記号の表記　分類対象タスク $i \in I$．ワーカ $k \in K$．ラベル $j \in J$，$q \in J$，$l \in J$，$n_{il}^{(k)}$ はワーカ k が分類対象タスク i に対して，ラベル l と回答したか．回答した場合には 1，回答していない場合は 0 が値になる．$\alpha_{jl}^{(k)}$ は，ワーカ k が真のラベル j のとき，l と回答する確率．T_{ij} は，分類対象タスク i の真のラベルが j であるか．真の場合には 1，偽の場合は 0 が値になる．E_{ij} は，分類対象タスク i の真のラベルが j である期待値を表す．

E ステップ　E ステップでは，正解のクラス分布パラメータ S_j とワーカの能力パラメータ $\alpha_{jl}^{(k)}$ を固定し，分類対象タスクの正解の期待値を推定する (1).

$$E(T_{ij}=1 \,|\, data) = \frac{\prod_{k\in K}\prod_{l\in J}(\alpha_{jl}^{(k)})^{n_{il}^{(k)}} S_j}{\sum_{q\in J}\prod_{k\in K}\prod_{l\in J}(\alpha_{jl}^{(k)})^{n_{il}^{(k)}} S_q} \tag{1}$$

M ステップ　M ステップでは，分類対象タスクの正解の期待値を固定して，正解のクラス分布パラメータ (2) とワーカの能力パラメータ (3) を推定する．

$$S_j = \frac{\sum_{i\in I} E_{ij}}{I} \tag{2}$$

$$\alpha_{jl}^{(k)} = \frac{\sum_{i\in I} E_{ij} n_{il}^{(k)}}{\sum_{l\in J}\sum_{i\in I} E_{ij} n_{il}^{(k)}} \tag{3}$$

3.2 動的な割り当て

通常 Dawid と Skene のモデルでは，全ての分類対象タスクに等しい数の回答を集め，すべて回答がでそろってから EM アルゴリズムを適応する．本研究では，これを逐次的に行う．具体的には，新しく人間ワーカの回答を得る度に EM アルゴリズムを適応し，期待値の最大値 E_{maxi} が最小のタスクを次のワーカに割り当てる．

図 3 は，この動的な割り当てを表したものである．EM アルゴリズムを適応した結果，右にあるように期待値の最大値 E_{maxi} が得られたとする．この場合，全タスクの期待値の最大値を比較すると，Task4 の期待値の最大値 0.4 が最小になる．そのため，Task4 を次のワーカに割り当てる．この動作を，新しく人間ワーカの回答を得る度に行う．この図からも分かるように，従来の Dawid と Skene のモデルのように 1 つのタスクに対して，等しい数の回答を集めるわけではない．推定結果が怪しいタスクに，回答を集める．

3.3 AI ワーカの導入

上記で説明した動的な割り当てに，AI ワーカを導入し，1 人の人間ワーカとして扱う．ここで AI ワーカとは，人間がやる

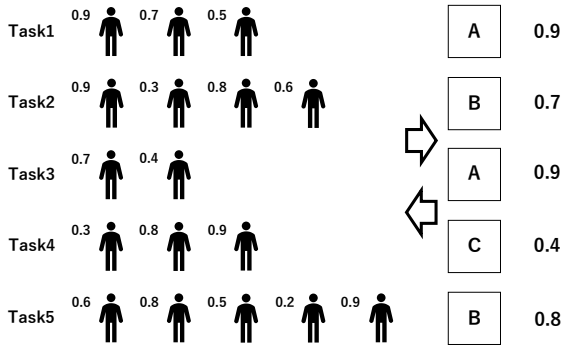

図 3　動的な割り当て（右がそれぞれのタスクにおける期待値の最大値）

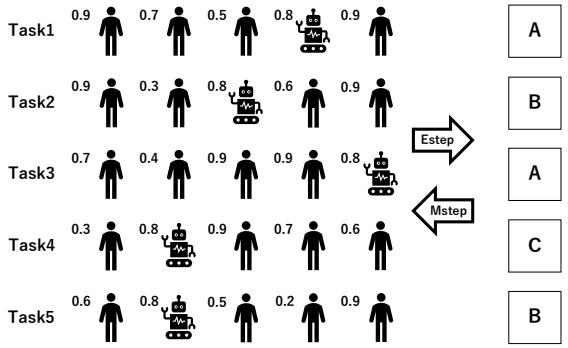

図 4　AI ワーカの導入

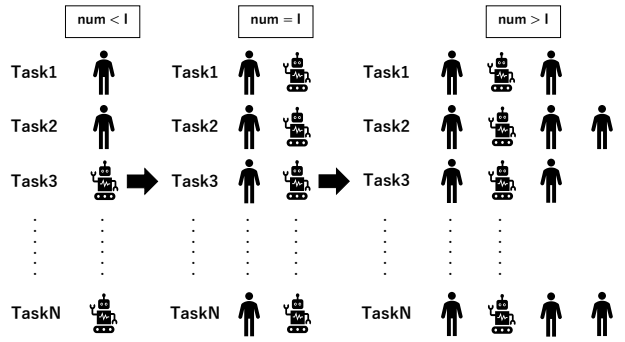

図 5　提 案 手 法

タスクを，人間と同じように処理するアルゴリズムのことである．図 4 は，この AI ワーカの導入を表したものである．EM アルゴリズムの M ステップによって，AI ワーカの能力を含めた，ワーカ全員の能力パラメータを推定する．E ステップでは，AI ワーカの能力を含めた，ワーカ全員の回答に基づいて，分類対象タスクの正解を推定する．

3.4　具体的なアルゴリズム

入力　与えられた分類対象タスクの集合を $T = \{t_1, \cdots, t_I\}$，最大限依頼できるタスク数を M とする．

出力　各分類対象タスクに対して推定されたラベルの集合を $A = \{a_1, \cdots, a_I\}$ とする．

手続き　Algorithm1 は，提案手法のアルゴリズムを示している．それに用いられている関数を以下で示す．

関数 $task\_selection$ は，次にどのタスクを割り当てるか決定する関数である．各分類対象タスクの期待値の集合 $X = \{\{e_{11}, \cdots, e_{1J}\}, \cdots, \{e_{I1}, \cdots, e_{IJ}\}\}$，イテレーション回数を表す変数 num，分類対象タスクの集合 T を入力とする．各分類対象タスクに何人のワーカに回答を依頼するか格納した $Y = \{b_1, \cdots, b_I\}$ を出力する．$num \leqq I$ では，各分類対象タスクに 1 人が回答するように順番に割り当てる．$num > I$ では，各分類対象タスクの期待値の最大値が一番低いものを，次のワーカに割り当てる．

関数 $generate\_data$ は，ワーカの回答データを EM アルゴリズムで処理できるように整形する関数である．各分類対象タスクに何人のワーカに回答を依頼するか格納した Y，変数 num，AI ワーカの回答 $W = \{d_1, \cdots, d_I\}(num \geqq I)$ を入力とする．整形したデータ Z を出力する．

関数 em は，EM アルゴリズムを適応する関数である．整形したデータ Z を入力とする．各分類対象タスクの正解の期待値の集合 X を出力する．

関数 $answer$ は，各分類対象タスクの正解の期待値が最大となっているラベルを取得し，全分類対象タスクの回答の集合をつくる．各分類対象タスクの正解の期待値の集合 X を入力とする．各分類対象タスクに対して推定されたラベルの集合 A を出力する．

関数 $AI\_train\_predict$ は，AI ワーカに EM アルゴリズムによって得た分類対象タスクの回答を学習させ，それをもとに全分類対象タスクの回答を推定させる．各分類対象タスクの推定した正解の集合 A，分類対象タスクの集合 T を入力とする．AI ワーカの回答 W を出力する．

関数 $merge$ は，$num < I$ のとき，各分類対象タスクの推定した正解の集合 A の足りない分を，AI ワーカの回答で補う．各分類対象タスクの推定した正解の集合 A，AI ワーカの回答 W を入力とする．各分類対象タスクの推定されたラベルの集合 A を出力とする．

以上の流れを，M まで繰り返す．

図 5 は，全体の提案手法の全体の流れを表している．$num < I$ の時，各分類対象タスクに 1 つずつ回答を集める．まだラベルの手に入っていない分類対象タスクに関しては，AI ワーカが現状のラベルを学習し，予測することで補う．$num = I$ から AI ワーカは回答を初める．$num > I$ では，AI ワーカの導入に加えて，動的な割り当てを行う．

表 1　比較する手法

	動的に割り当て	順番に割り当て
AI ワーカ有り	AI-Human(dynamic)	AI-Human
AI ワーカ無し	Human(dynamic)	Human

4　実　験　1

提案手法が，品質を保ったままタスクを削減できるか検証する．実験では 4 つの手法を比較した．表 1 は，この 4 つの手法を示している．1 つ目は，AI ワーカを導入し，動的な割り当て

Algorithm 1 アルゴリズム

Input: M, T

Output: A

$num \leftarrow 0$
while $num < M$ **do**
  $Y \leftarrow task\_selection(X, num, T)$
  $Z \leftarrow generate\_data(Y, W, num)$
  $X \leftarrow em(Z)$
  $A \leftarrow answer(X)$
  $W \leftarrow AI\_train\_predict(A, T)$
  if $num < I$ **then**
    $A \leftarrow merge(A, W)$
  end if
  $num \leftarrow num + 1$
end while

を行うものである．これが本研究の提案手法に当たる．これを AI-Human(dynamic) と表す．2 つ目は，AI ワーカを導入し，順番に割り当てを行う（動的な割り当てを行わない）ものである．これを AI-Human と表す．3 つ目は，AI ワーカを導入せず，動的な割り当てを行うものである．これを Human(dynamic) と表す．4 つ目は，AI ワーカを導入せず，順番に割り当てを行う（動的な割り当てを行わない）ものである．これを Human と表す．

AI-Human(dynamic) は，全てのタスクに対して 1 つずつ回答が集まったら，動的な割り当てを開始する．Human(dynamic) は，全てのタスクに対して 2 つずつ回答が集まったら，動的な割り当てを開始する．これは，AI-Human(dynamic) は，全てのタスクに対して 1 つずつ回答が集まった時点で AI ワーカが回答を開始し，全てのタスクに対して 2 つずつ回答が集まるからである．

4.1 設　　定

タスク　AI と人間の両方が優れたタスク，AI と人間の両方が不得意なタスク，AI が人間よりも優れたタスク，人間が AI よりも優れたタスクの全部で 4 種類のタスクを用意した．AI と人間の両方が優れたタスクは，図 6 のように空撮画像を水害被害を受けたもの，受けていないもの，雲に隠れていてわからないものの 3 種類に分類するタスクを使用した．これを水害被害判定タスクと表す．AI と人間の両方が不得意なタスクは，図 7 のように絵画 [8] を Alfred Sisley，Camille Corot，Camille Pissarro，Claude Monet の 4 人の画家に分類するタスクを使用した．これを絵画タスクと表す．AI が人間よりも優れたタスクは，図 8 のように肺の X 線画像 [9] を異常のある肺，通常の肺の 2 種類に分類するタスクを使用した．これを肺 X 線画像タスクと表す．人間が AI よりも優れているタスクは，図 9 のように人が写っている画像を，走っているもの，歩いているものの 2 種類に分類するタスクを使用した．これを歩いているか走っているか判定タスクと表す．今回の実験では，それぞれのタスクに 500 枚の画像を用意し，最大で 2500 タスクまで作業を依頼できるというシチュエーションで実験を行っている．

図 6　水害被害タスク（左から水害被害を受けていないもの，受けたもの，雲に隠れていてわからないもの）

図 7　絵画タスク（左から Alfred Sisley，Camille Corot，Camille Pissarro，Claude Monet）

図 8　肺 X 線画像タスク（左から通常の肺，異常のある肺）

人間ワーカ　事前に Yahoo!クラウドソーシング [10] を利用して，1 枚の画像に対して 5 人のワーカに回答を依頼した．水害被害判定タスクでは 118 人，絵画タスクでは 135 人，肺 X 線タスクでは 136 人，歩いているか走っているか判定タスクでは 96 人のワーカに回答を行ってもらった．また全ての人間ワーカの正解率，能力パラメータは未知であるとする．

AI ワーカ　Keras を使用し，アルゴリズムには CNN（畳み込みニューラルネットワーク）を用いた．また，導入する AI ワーカは 1 体とした．AI ワーカのモデルは以下の通りである．

（1）　3×3 のフィルターを用いて 32 回の畳み込みを行う．
（2）　2×2 ごとに，Max-pooling を行う．
（3）　全体の 25% のノードをドロップアウトする．
（4）　3×3 のフィルターを用いて 64 回の畳み込みを行う．


図 9　歩いているか走っているか判定タスク（左から走っているもの，歩いているもの）

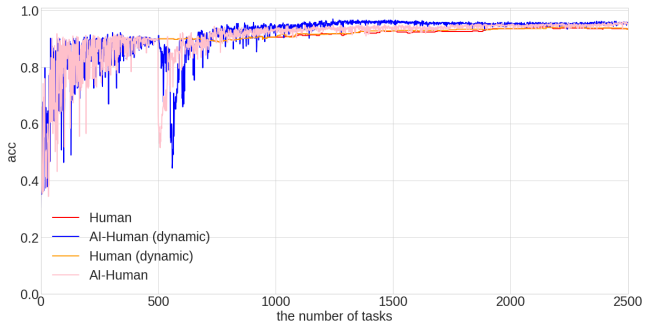

図 10　水害被害判定タスク

（5） 2×2 ごとに，Max-pooling を行う．
（6） 特徴マップをベクトル化する．
（7） 分類クラス数の次元のベクトルの出力．
（8） 活性化関数 softmax を用いる．
（9） 損失関数にはクロスエントロピー誤差を用い，確率的勾配降下法を行う．

4.2 結　　果

図 10 は水害被害判定タスク，図 11 は絵画タスク，図 12 は肺 X 線タスク，図 13 は歩いているか走っているか判定するタスクの結果を示している．横軸は依頼した人間ワーカに依頼したタスク数を，縦軸は正解率であり，グラフはタスク数における正解率の推移を表している．また，横軸は 2500 タスクまであるが，任意の横軸のタスク数 max_tasks と設定したときに得られる正解率を表したグラフと見ることもできる．Human と Human(dynamic) では，500 タスク依頼して初めて全ての画像にラベル付けが行われるので，500 タスクからグラフが始まっている．

4.3 考　　察

まず水害被害タスクを比較する．Human(dynamic) は，AI-Human(dynamic) は，Human の 500 から 700 タスク付近での正解率を下回ってしまっている．しかしながら，Human での 2000 タスク付近の正解率を，750 タスク付近で同程度のものを出せている．これは，約 1250 タスク削減できることを示している．また，最終的な正解率も高くなっている．加えて，AI-Human(dynamic) は，AI-Human のグラフと類似しており，

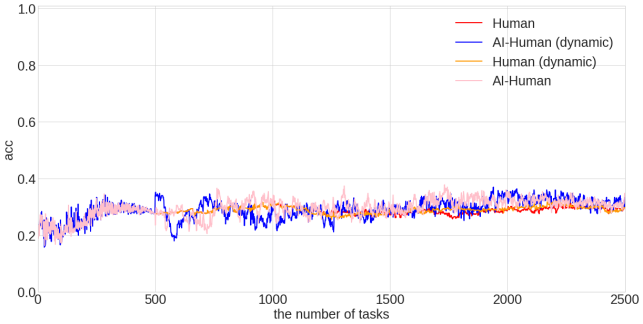

図 11　絵画タスク

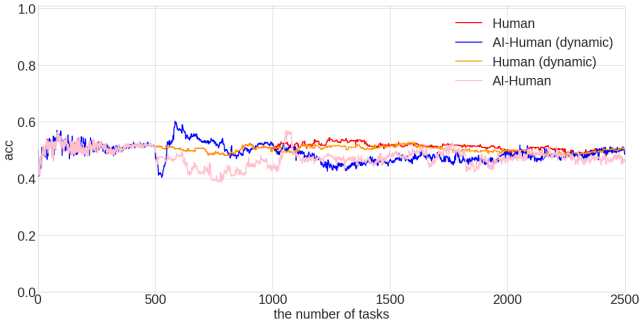

図 12　肺 X 線タスク

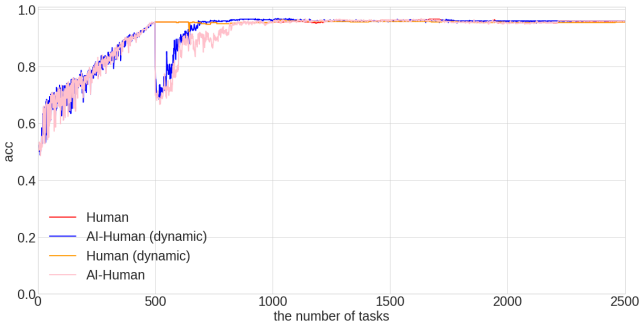

図 13　歩いてるか走っているか判定タスク

2 つの特徴のうち，AI ワーカの導入による効果がでていることが分かる．次に絵画タスク，肺 X 線画像タスクを比較する．これらのタスクでは，AI-Human(dynamic) が Human と比べて，正解率が高くなったり低くなったりしており，効果がない．これは，人間ワーカの正解率が非常に低いことが原因であると考えられる．最後に，歩いているか走っているか判定するタスクを比較する．Human は，Human(dynamic) とほとんど正解率の推移が変わらない．また，AI-Human(dynamic) においても，500 から 700 タスク付近で正解率が下回ってしまっているものの，その後の推移はほとんど変わらない．

これらのことから，提案手法によって，AI と人間が優れたタスク（水害被害タスク）においてタスクを削減できることがわかった．また，最終的な品質を向上させることも可能である．

5 実　験　2

スパムワーカが一定数いる状況で，提案手法が品質を保ったまま，タスクを削減できるか検証する．

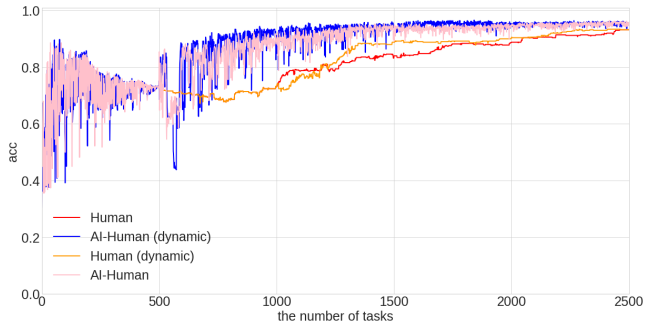

図 14　水害被害判定タスク

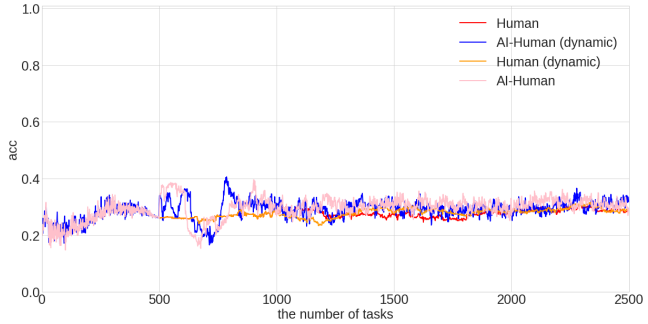

図 15　絵画タスク

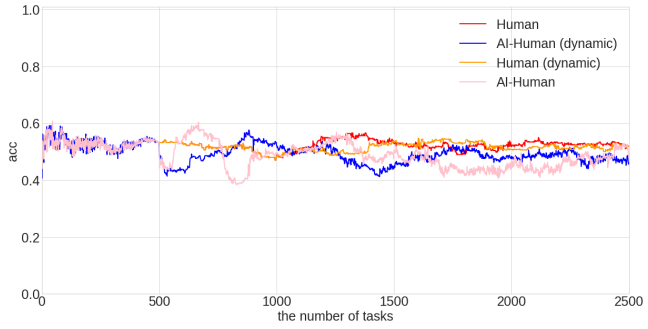

図 16　肺の X 線タスク

5.1　設　　定

実験 1 で使用したデータを，全ワーカの 1/3 をスパムワーカに置き換えて同様の実験を行った．スパムワーカとは，分類対象タスクによらず，選択肢からランダムに選んで回答するワーカのことである．

5.2　結　　果

図 14 は水害被害判定タスク，図 15 は絵画タスク，図 16 は肺 X 線タスク，図 17 は歩いているか走っているか判定するタスクの，スパムワーカが一定数いるという状況下での結果を示している．

5.3　考　　察

まず水害被害タスクを比較する．AI-Human(dynamic) は，Human の 600 タスク付近での正解率を下回ってしまっている．しかしながら，Human での 2500 タスク付近の正解率を，750 タスク付近で同程度のものを出せている．これは，約 1750 タスク削減できることを示している．また，最終的な正解率も高く

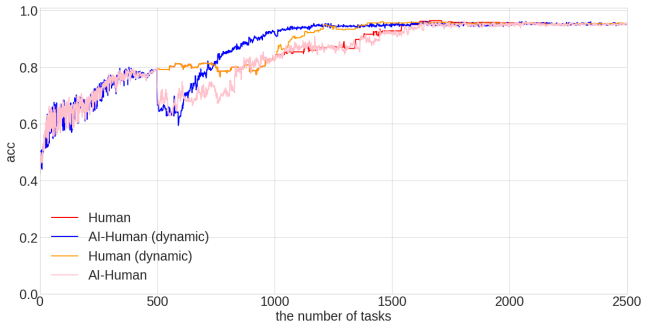

図 17　歩いてるか走っているか判定タスク

なっている．加えて，AI-Human(dynamic) は，AI-Human のグラフと類似しており，2 つの特徴のうち，AI ワーカの導入による効果がでていることが分かる．次に絵画タスク，肺 X 線画像タスクを比較する．これらのタスクでは，AI-Human(dynamic) が Human と比べて，正解率が高くなったり低くなったりしており，効果がない．これは，人間ワーカの正解率が非常に低いことが原因であると考えられる．最後に，歩いているか走っているか判定するタスクを比較する．AI-Human(dynamic) は，Human での 1600 タスク付近での正解率を，1250 タスク付近で同程度のものを出せている．これは，約 350 タスク削減できることを示している．また，Human(dynamic) でも同等のタスク数を削減することが出来ており，2 つの特徴のうち，動的な割り当てによる効果が出ていることが分かる．

これらのことから，提案手法によってスパムワーカが一定数いる状況下では，AI と人間が優れたタスク（水害被害タスク）と人間が AI よりも優れたタスク（歩いてるか走っているか判定タスク）においてタスクを削減できることがわかった．また，AI と人間が優れたタスクでは最終的な品質を向上させることも可能である．

5.4　実験 1 と実験 2 の考察

人間からある程度正しい回答が手に入るタスク（水害被害タスク，歩いているか走っているか判定タスク）場合，500 タスク付近で，人間のみのものに比べて提案手法の正解率がさがってしまう．しかし 1000 タスク以降は，正解率が著しく低くなることはなく，等しいもしくは高くなる．この 1000 タスクというのは，全画像に 2 人の回答があつまり EM アルゴリズムを行えるようになるタスク数である．そのため，こういったタスクの場合，提案手法の有効性がある．また，実験 1 と実験 2 を比較すると，タスクが削減できる場合，実験 2 のほうがより削減することができている．このことから提案手法は，スパムワーカが一定数いる状況でより効果を発揮することも分かる．

6　今後の課題

500 から 700 タスク付近で，正解率が下がってしまうことがあるため，どうするか考える必要がある．

7 終わりに

本論文では，Dawid と Skene を応用した手法を提案し，その結果を検証した．実験の結果，AI と人間が優れたタスクの場合，同等以上の品質でタスク数を削減することが出来た．スパムワーカが一定数いるとき，より効果を発揮した．人間が AI よりも優れたタスクの場合，タスク数を削減することは出来なかったものの，品質を悪化させることはなかった．スパムワーカが一定数いるとき，品質を保ちながらタスク数を削減することができた．これらから，人間ワーカからある程度正しい回答が手に入るタスクであれば，提案手法の有効性があることが示された．また，スパムワーカがいる場合には，より効果を発揮することが示された．

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] 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)*, Vol. 28, No. 1, pp. 20–28, 1979.

[2] Victor S Sheng, Foster Provost, and Panagiotis G Ipeirotis. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 614–622, 2008.

[3] Rion Snow, BrendanO ' connor, Dan Jurafsky, Andrew Y Ng. Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, pp. 254–263, 2008.

[4] Jacob Whitehill, Ting-fan Wu, Jacob Bergsma, Javier R Movellan, and Paul L Ruvolo. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in neural information processing systems*, pp. 2035–2043, 2009.

[5] Peter Welinder, Steve Branson, Pietro Perona, and Serge J Belongie. The multidimensional wisdom of crowds. In *Advances in neural information processing systems*, pp. 2424–2432, 2010.

[6] Satoshi Oyama, Yukino Baba, Yuko Sakurai, and Hisashi Kashima. Accurate integration of crowdsourced labels using workers' self-reported confidence scores. In *Twenty-third International Joint Conference on Artificial Intelligence*, 2013.

[7] Hisashi KASHIMA Jiyi LI, Yukino BABA. Hyper questions: Crowdsourcing answer aggregation method for questions requiring expert knowledge. In *DBSJ Japanese Journal*, pp. 17–J.

[8] Wikiart. `https://www.wikiart.org/`.

[9] `https://lhncbc.nlm.nih.gov/publication/pub9931`.

[10] Yahoo!クラウドソーシング. `https://crowdsourcing.yahoo.co.jp/`.