タスク結果品質を考慮した人間+AI クラウドへのマイクロタスク割当て

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，人間ワーカと AI ワーカを含むクラウド（群衆）に動的にタスクを割り当てる問題に取り組む．近年，深層学習に代表される AI プログラムの作成をクラウドソースすることが普及し始めている．一方で，作成された AI プログラムを別のソフトウェアや人間ワーカとどのように組み合わせれば，効率的なタスク処理を実現できるかを依頼者が判断するのは容易ではない．本論文では，AI プログラムをクラウドソーシングにおけるワーカとしてモデル化することによりこの問題に取り組む．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，AI ワーカによるタスク結果を部分的に採用し，採用できない部分については他の AI ワーカや人間ワーカからのタスク結果を採用する．実験結果は，提案手法が能動学習に基づく手法と比較して，要求精度を満たしながらより多くのタスクを AI ワーカに割り当てることを示した．この結果は，性質が分からない 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す．

キーワード　クラウドソーシング, マイクロタスク, タスク割り当て

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)” における各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[2]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 を含むソフトウェアの作成をアウトソースして活用する試みが普及しつつある．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 [3]，AI をワーカとして受け入れるための環境整備が進められている．しかし，AI ワーカを入手した後に，それらを評価して問題解決に適用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ラベルを入手したい場合に，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

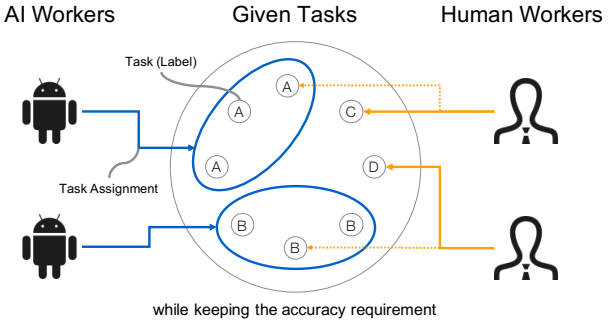

図 1　本論文では，人間ワーカと AI ワーカの適切な分担により，マイクロタスク処理の効率化を目指す．

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．提案手法は，人間は正しい結果を返すと仮定し (実際には，クラウドソーシングの分野で広く用いられているスパム除去やタスク結果の集約などの手法で品質向上を行う必要がある [1])，AI ワーカの性質は未知であるとする．このような状況において，提案手法は精度要件を満たしながら，より早くかつ低いコストでタスク結果を入手す

1：単純なアルゴリズムやルールベース，深層学習なども含む広義の AI プログラム
2：https://www.kaggle.com
3：https://crowd4u.org

るためのタスク割り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 ワーカの全体の性能を見るのではなく，AI ワーカ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り，それについてはパラメータで調整できることである．すなわち，品質は犠牲にしてでも安く・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，を依頼者が選択できる．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 ワーカに得意な種類のタスクのみを割り当てる．この統計的検定は，人間ワーカ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，提案手法が能動学習を用いた手法と比較してより多くのタスクを AI ワーカに割り当てることが確認された．提案手法において，AI ワーカと人間ワーカへの割り当てはコストと品質のトレードオフの関係にあり，要求されたタスク結果品質に応じて調節可能であった．

この結果は，人間ワーカと性質が未知である AI ワーカ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．

本研究の貢献は以下の通りである．

（1） 入力をタスク集合と AI ワーカ集合および全体的なタスク結果の要求精度とし，要求精度を満たすタスク割当を求める問題を，AI+人間クラウドタスク割当問題として提起した．

（2） 人間ワーカによるタスク結果を用いて AI ワーカの得意な部分タスク（タスククラスタ）を発見し，タスククラスタに基づいて AI ワーカにタスクを割り当てる手法を提案した．いくつかの仮定の下で，提案手法により得られるタスク割り当てが要求精度を満たすこと理論的に解析した．

（3） 提案手法により，要求精度を満たすようなタスク割り当てを実現できることを実験的に示した．提案手法は能動学習に基づく手法と比較して，より多くのタスクを AI ワーカに割り当てること，要求精度に応じて AI ワーカに割り当てるタスクを柔軟に変更することを確認した．

2 関 連 研 究

本研究で扱う問題に関連する様々な研究が取り組まれている．能動学習では，AI の性能を効率的に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 [2]．データに対してラベル付与を行うためのワーカを自動的に決定するという点で，本論文で取り組む問題に関連する．能動学習と本論文で取り組む問題に本質的な違いがあり，能動学習では人間ワーカに対する割り当てを決定するのに対し，本論文で取り組む問題では AI ワーカに割り当てるタスクを決定する．本論文の提案手法においても人間ワーカへのタスク割り当てはランダムに決定される．能動学習における AI の学習を効率化するためのクエリ戦略を，提案アルゴリズムと組み合わせることは重要な今後の課題である．

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 [3]．ア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，人間ワーカがタスク処理を担当する．

人間と計算機のコラボレーションに関する様々なパターンの研究 [4] [5] が取り組まれている．典型的なアプローチは，AI プログラムを用いて，人間のワーカによる柔軟な処理が必要なデータを識別することである [6] [7] [8]．一方で，AI の出力を用いて人間ワーカのトレーニングを行う [9] といったアプローチも検討されている．本研究の提案手法は，AI ワーカによって処理可能なタスクを識別するものであり，これらのアプローチと相互に組み合わせることが可能である．

収集したラベルを後からグループ化し直すというアプローチでタスク結果を集約するアプローチが提案されている [10]．この方法では，実際のクラスの種類やワーカから提案されたラベルの種類に依存するのではなく，データの性質に基づいてより柔軟な単位でクラスタリングを行う．この様なアプローチは，AI ワーカからより多くの，信頼性の高いタスククラスタを選択することに適用できると考えられる．一方で，分類器の学習において入力に対してラベルを返さないという出力を許す手法が提案されている [11] [12]．この手法では分類器自体にラベルを返すかどうかを決定する機能を含めるのに対して，本研究で提案する手法では，AI ワーカからの回答を受け入れるかどうかの決定はシステムが行う．

ネットワークを通じたソフトウェア開発の外部委託 [13] は広く利用されており，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 が挙げられる．

クラウドソーシングにおいて，高い品質の結果を低コストで入手することが重要であり [14]，本研究で提案するアルゴリズムは AI ワーカを活用することで，要求精度を満たしながら人間ワーカへのタスク割り当てを削減する．

3 人間+AI クラウドタスク割当問題

本節ではまず，本論文で定義する人間+AI クラウドタスク割り当て問題について述べ，能動学習に基づく解法および本研究で提案する手法について述べる．

3.1 問題の定義

表 1 に，本論文で用いるタスク，ラベルおよび AI ワーカ等の表記法を示す．タスク集合 T の各要素には真の正解が存在すると仮定する．AI ワーカの集合を W と表す．W には教師あり学習アルゴリズムや教師なしアルゴリズムを含む不特定多数の AI ワーカが含まれる．各 AI ワーカ $w_i \in W$ は，その AI ワーカ振る舞いを表す関数 $C_{w_i} : \mathcal{P}(T \times A) \to \mathcal{P}(\mathcal{P}(T))$ を持

表 1　記号の表記

$T = \{t_1, \cdots, t_M\}$	与えられた多クラス分類タスクの集合.
$A = \{a_1, \ldots, a_N\}$	タスクにおけるクラス（ラベル）の集合.
$W = \{w_1, w_2, \ldots\}$	AI ワーカの集合．教師あり学習アルゴリズムや教師なし学習アルゴリズムなどを含む.
$C_{w_i} : \mathcal{P}(T \times A) \to \mathcal{P}(\mathcal{P}(T))$	w_i が生成するタスククラスタ集合を返す関数.
$C(D) = \bigcup_{w_i \in W} C_{w_i}(D)$	AI ワーカ集合 W の各ワーカから生成されたタスククラスタの集合.

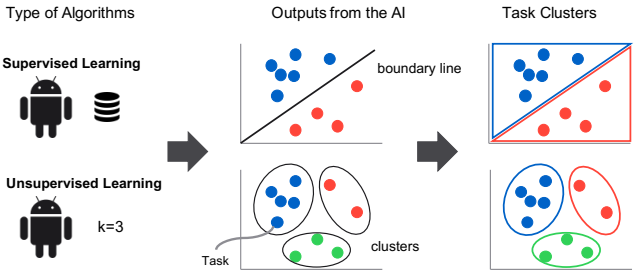

図 2　AI ワーカの出力をタスククラスタとして扱う

つ．関数 C_{w_i} はタスクと人間ワーカによるラベルのペアの集合 D を入力とし，出力はタスククラスタの集合 $\{T_{w_{i,1}}, T_{w_{i,2}}, \ldots\}$ である．ここで，$T_{w_{i,j}}$ の直和集合 $T_{w_{i,1}} \oplus T_{w_{i,2}} \oplus \ldots$ が T となる．D はその AI ワーカの学習のために利用することができ，教師あり学習アルゴリズムではタスクとラベルの両者が使われるが，一方で教師なしアルゴリズムではタスクのみが使われる．

図 2 に示すように，各 AI ワーカの出力をタスククラスタ集合として定式化する．タスククラスタとは，ある AI ワーカによって同じ種類のラベルが付与されるタスクの集合である．w_i が教師あり学習アルゴリズムである場合，$C_{w_i}(D)$ が返すのは AI ワーカが D で学習した後の分類結果である．w_i が教師なし学習アルゴリズムである場合，$C_{w_i}(D)$ が返すのはクラスタリング結果である．この場合，$C_{w_i}(D)$ の出力結果は D には依存せず，各タスクが属するクラスタリング結果に基づく．各タスククラスタのラベルが A の要素と対応付けられるかどうかは AI ワーカのアルゴリズムや実装に依存する．

人間+AI クラウドタスク割当問題で求めるのは，タスクとワーカのペア (t_j, w_i) を要素とするタスク割り当ての列 S である．S にはタスク集合のすべての要素に対するタスク割り当てを含む必要があるため，$|S| = |T|$ である．タスク t_j を担当するワーカ w_i は，AI ワーカの集合の要素または人間ワーカの要素 $w_i \in W \cup \{'h'\}$ である．

定義 1 (人間+AI クラウドタスク割当問題)．多クラス分類タスクの集合 T，AI ワーカの集合 W，および要求精度 $0 \leqq q \leqq 1$ が与えられ，人間ワーカは常に正解を回答すると仮定する．ここで，AI ＋人間ワーカ割当問題とは，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．

AI ＋人間ワーカ割当問題にはすべてのタスクを人間ワーカに割り当てるという自明な解が存在する．ここで，あるタスク割り当て S_1 と，$(t_j, 'h')$ である要素が S_1 よりも少ないタスク割り当て S_2 が存在する時，S_2 は S_1 よりも効率的であると呼ぶ．

タスク割り当て S_2 よりも効率的なタスク割り当てが存在しない場合，S_2 は最適であると呼ぶ．AI ワーカ集合は不特定多数の AI ワーカで構成されるため，最適なタスク割り当てを発見するのは困難である．

定理 1 (非決定性)．AI ワーカの性能が不明である場合，タスク割り当てが AI ＋人間ワーカ割当問題において最適であるかどうかは決定不能である．

証明．最適であるとみなすことが出来るいかなるタスク割り当てにおいても，AI ワーカの性能が不明であることから，そのタスク割り当てよりも効率的なタスク割り当ての存在を否定することは出来ない．□

したがって，適切なタスク割り当てを得るために，AI ワーカの性能を外部から測定する必要がある．さらに，AI ワーカの学習に用いるデータ数が増加することで，AI ワーカの性能が変化する可能性があるため，性能の測定は動的に行う必要がある．次の小節では，AI ワーカ割当問題への回答を計算するための統計的検定に基づくアルゴリズムを提案する．アルゴリズムでは，人間ワーカに割り当てるタスクをランダムに選択するため，常に最適に近いタスク割り当てを出力ことは保証出来ない．しかし，実験結果に示すように，ベースライン手法である能動学習に基づく手法と比較して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．

3.2　能動学習に基づく解法

人間+AI クラウドタスク割当問題に取り組むための解法として，既存手法である能動学習に基づく手法をアルゴリズム 1 に示す．この手法では，能動学習におけるクエリ戦略に基づいて人間ワーカに割り当てるタスクを決定する．クエリ戦略によって人間ワーカへのタスク割り当てを制御することで，AI ワーカの学習に有用なデータに優先的にラベルを付与し，これにより学習した AI ワーカに対して早期にタスク割当を行うことを目指す．この手法を利用するためには，AI ワーカが教師あり学習または半教師あり学習に基づくアルゴリズムであり，クエリ戦略の処理に必要な各タスクの情報量尺度を入手可能であることが求められる．

AI ワーカの学習および評価は，人間ワーカによってラベル付されたタスクを用いて行う．AI ワーカを評価した結果，精度が要求精度を満たした場合に，その時点でラベルが付与されていないタスク全てを AI ワーカに割り当てる．

3.3　AI ワーカからのタスク結果品質を保証する解法

人間+AI クラウドタスク割当問題では出力されるタスク割り当てを行うことでタスク結果品質が少なくとも q を満たす必要がある．ここでは，AI ワーカによるタスク結果の品質が少なくとも q を満たすようなタスク割当を行うアルゴリズムである

Algorithm 1 能動学習に基づく手法

Input: T, A, w, q

Output: Z

let $D = \{(t, a) | \exists t \in T\ ans(t) = (a,' h')\}$
for all $t \in T$ s.t. $ans(t)! = (\phi, \phi)$ **do**
  if $accuracy\_score(w, D) > q$ **then**
    update ans of all $t' \in T$ s.t. $ans(t') = (\phi, \phi)$, $ans(t') = (predict(w, t'), w)$
  end if
  let $t' = query\_strategy(t \in T \text{ s.t. } ans(t) == (\phi, \phi))$
  let $a' \leftarrow humanworker\_result(t', A)$
  update ans so that $ans(t) = (a',' h')$
end for
return $\{(t_1, ans(t_1).result), (t_2, ans(t_2).result), \ldots\}$

OBA(Observation-based Assignment) について説明する．

定義 2 (単純人間+AI クラウドタスク割当問題)．分類タスクの集合 T，AI ワーカの集合 W および AI ワーカによるタスク結果の品質要件 $0 \leqq p \leqq 1$ が与えられている．単純人間+AI クラウドタスク割当問題で求めるのは，AI ワーカによるタスク結果が q を満たすようなタスク割り当て集合である．

Observation-based Assignment (OBA). OBA の背景にある考え方は，AI ワーカの結果と人間ワーカのタスク結果を比較することである．これは，人間ワーカは常に正しくタスクに回答することが出来るという仮定に基づく（この仮定は現実には成立せず，クラウドワーカから得られるタスク結果の品質管理するための手法等を組み合わせることが求められる）．タスククラスタを評価するために，タスククラスタに含まれるタスクのうち，人間ワーカによるラベルが付与されているものに注目する．統計的検定により，タスククラスタに含まれるタスクに正解ラベル（人間ワーカによる最頻の種類のラベル）を付与する確率が要求精度を満たしていることを評価する．

入力. OBA の入力は，タスク集合 T，ラベル候補の集合 A，AI ワーカの集合 W，C_{w_i} の集合および q である．ここで，q は AI ワーカの性能が満たすべき要求精度である．ただし，$0 \leqq q \leqq 1$ である．

出力. OBA の出力は，タスクと回答のペアで構成されるタスク割り当ての集合 $\{(t_1, a_{k1}), \ldots, (t_M, a_{kM})\}$ である．ただし，$a_{ki} \in A$ である．この出力は列ではなくタスクと回答のペアの集合である．

手続き. OBA のアルゴリズム 2 を示す．$ans : T \to A \times (W \cup \{'h'\}) \cup \{(\phi, \phi)\}$ を更新可能な関数である．この関数は，タスク集合 T を入力とし，タスク結果と割り当てられたワーカを出力する．ここで，人間ワーカは $'h'$ と表記する．

初期状態において，$ans(t)$ は $t \in T$ に対して (ϕ, ϕ) を返す．これはタスクがワーカに割り当てられておらず，タスク結果が存在しないことを表す．提案手法は $ans(t) = (\phi, \phi)$ であるタスク t を人間ワーカに割り当てる．タスク割り当ての進行に応じて ans を更新し，すべての t について $ans(t)$ が (ϕ, ϕ) を返さなくなるまで処理を繰り返す．

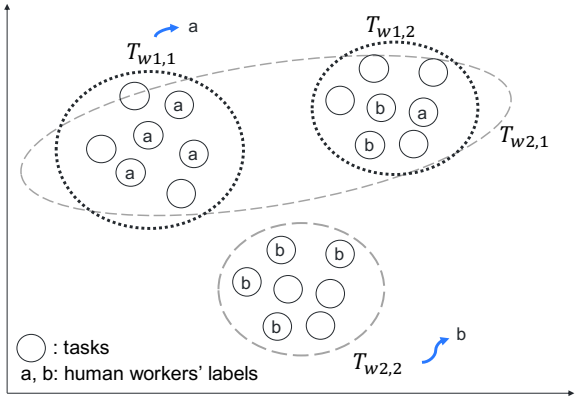

図 3　人間ワーカによるラベルに基づいてタスククラスタが採用できるかどうかを決定する

Algorithm 2 Observation-based Assignment (OBA)

Input: T, A, W, C_{w_i}, q

Output: Z

let $D = \{(t, a) | \exists t \in T\ ans(t) = (a, `h')\}$
for all $t \in T$ s.t. $ans(t)! = (\phi, \phi)$ **do**
  for all $T_{w_{i,j}} \in C(D)$ **do**
    if $statistical\_test(T_{w_{i,j}}, D, q)$ **then**
      let $\hat{a}$ be the label for $T_{w_{i,j}}$
      update ans so that for all $t' \in T_{w_{i,j}}$, $ans(t') = (\hat{a}, w_i)$
    end if
  end for
  let $t' = t \in T$ s.t. $ans(t) == (\phi, \phi)$
  let $a' \leftarrow humanworker\_result(t', A)$
  update ans so that $ans(t') = (a',' h')$
end for
return $\{(t_1, ans(t_1).result), (t_2, ans(t_2).result), \ldots\}$

人間ワーカからあるタスク t に対するラベル（例えば a or b）を受け取るたびに，提案手法は次の処理を行う (図 3)．D はその時点までに人間ワーカによって処理されたタスクとそのラベルのペアの集合である．次に，各タスククラスタ $T_{w_{i,j}} \in C(D)$ について，$T_{w_{i,j}}$ に含まれるタスク全てに人間ワーカによる最頻ラベル $\hat{a}$ (例えば a や b) を付与することが出来るかどうかを確認するための統計的検定を行う．統計的検定では，タスククラスタにおける人間ワーカによるラベルが付与されているタスクに注目する．ここで，人間ワーカは割り当てられたタスクに対して正しく回答できることと仮定する．クラウドソーシングにおける品質管理手法を活用することでこれを実現できると考える．この手順では，タスク割り当ての品質が要求精度を上回ることを保証することができるような，任意の統計的検定手法を用いることが出来る．タスククラスタに含まれるタスクのラベルが，人間ワーカによる最頻の種類のラベルである割合が有意に高いと認められたら，タスククラスタに含まれる各タスク $t \in T_{w_{i,j}}$ の $ans(t)$ のラベルを $\hat{a}$ に更新する．

3.4　全体的なタスク結果品質を保証する解法

全体的なタスク結果品質が要求精度を満たせるようなタスク割り当てを実現する解法について議論する．OBA では評価対

Algorithm 3 モンテカルロシミュレーション

Input: C, N, q, α

Output: is_acceptable

1: let $success = 0$
2: let $X[c][i] \sim Beta(1 + c.r, 1 + c.n - c.r) \forall c \in C, 1 \leqq i \leqq N$
3: **for** i in range(N) **do**
4: let $overall\_accuracy = \frac{\sum_{c \in C} X[c][i] * c.size}{\sum_{c \in C} c.size}$
5: **if** $overall\_accuracy > q$ **then**
6: $success+ = 1$
7: **end if**
8: **end for**
9: let $p\_value = 1 - (success/N)$
10: return $p\_value < \alpha$

象のタスククラスタだけに着目して採用するかどうかを判断したのに対し，ここで扱う解法では，これまでに採用したタスククラスタと次の候補であるタスククラスタを考慮することが求められる．

ここで，全体的なタスク結果品質が要求精度を満たすようなタスククラスタの正解確率は要求精度と等しいとは限らない．教示あり学習アルゴリズムに基づく AI ワーカの性能は人間ワーカによるラベルの増加に応じて変化すると予想される．一方で，人間ワーカによるラベルが増えるにつれて，その時点で未処理なタスク集合に対して求められる実質的な要求精度は問題の入力である要求精度を下回ると考えられる．このような，AI ワーカの性能および未処理のタスクに対して求められる要求精度が動的に変化する状況において，全体的なタスク結果品質を考慮したタスク割り当てを行うことが求められる．

全体的なタスク結果品質が要求精度を満たせるかを計算するにあたり，まず各タスククラスタが正解（タスククラスタに含まれるタスクのうち，人間ワーカによって付与された最頻ラベルを正解ラベルとする）を出力する確率分布を考える．

本研究では，この確率分布が事前確率分布 $p(\theta) = Beta(1, 1)$，尤度関数 $p(X \mid \theta)$ から算出した事後確率分布 $p(\theta \mid X) = Beta(1 + x, 1 + n - x)$ に従うと仮定する．ただし，$Beta(\alpha, \beta)$ はベータ分布である．尤度関数は，n 回の試行のうち r 回で正解ラベルを返す二項分布を仮定する．

全体的なタスク結果品質が要求精度を満たす確率を求めるアルゴリズム 3 を示す．このアルゴリズムはタスククラスタ集合 C, 要求精度 q, シミュレーション回数 N, 有意水準 α を入力とし，全体的なタスク結果品質が要求精度を満たす確率が $1 - \alpha$ を上回る場合に真を，それ以外の場合には偽を返す．2 行目において，各タスククラスタにおける事後分布に従う乱数を N 回生成している．ただし，$c.n$ は人間ワーカによるラベルの数を，$c.r$ は人間ワーカによるラベルのうち最頻ラベルの数を，$c.size$ はそのタスククラスタを採用した場合にあたらにラベルを付与できるタスクの数を表している．

Dynamic Observation-based Assignment (DOBA). ここで，DOBA(Dynamic Observation-based Assignment) と呼ばれるアルゴリズムを提案する．DOBA のアルゴリズムは OBA のアルゴリズム 2 にいくつかの変更を加えることで実現できる．まず，行 3 の前の手順として採択済みタスククラスタを保持するための配列を定義する．採択済みタスククラスタは，アルゴリズム 3 の実行するための入力として与える必要がある．4 行目で行う統計的検定は，任意の検定方法ではなくアルゴリズム 3 を用いる．

3.5 理論解析

ここでは，提案アルゴリズムに関連するいくつかの理論的特性を解析する．

正当性. まず，提案アルゴリズムの出力が品質要件を満たしていることを証明する．

補題 1 (OBA の正当性)**.** OBA は，タスク結果の品質が有意水準 l で q を満たすような AI ワーカにタスクを割り当てる．

OBA のアルゴリズムでは，ラベルの品質が有意水準 l において要求精度 q を満たすことを保証するための統計的検定を行う．このことから提案手法はこの補題に該当する．

効率性. OBA の効率は，各 AI ワーカの実装やアルゴリズムに大きく依存するが，問題設定で述べたとおりこれらはブラックボックスである．従って，最も単純なケースを想定し，効率について議論する．

T をタスクの列，q を要求精度とする．この設定で最も単純な例を検討する．

- 受け入れる候補となるタスククラスタを 1 つもつ AI ワーカを 1 つ用いる．
- $L^* \subseteq T$ は C における AI ワーカの性能が q を上回るようなタスク集合である．ただし，L^* 中の要素 L_i の N タスクに正しいラベルが与えられていると仮定する．
- $C \cap L^* = \phi$ は，AI ワーカの C を採択するために，C 中の十分なタスクを人間ワーカに割り当てることなく，AI ワーカの性能を向上させることが可能であることを意味する．

C が採用されるまでに人間ワーカによって C に含まれるタスクが処理されないことが理想的であるが，一般には考えにくい．C を受け入れるときに，人間ワーカによって処理されていないタスクの部分集合 $C' \subset C$ は次のように表される．

$$C' = C - \Delta C - L',$$

ただし，ΔC は C が OBA で採用されることを統計的に決定するために必要なタスクの集合であり，L' は AI ワーカが L_i 中の N に回答する前に，人間ワーカによって回答されたタスクである．

ΔC の大きさは，使用する統計的検定手法，有意水準および要求精度によって決定される．ここでは，成功確率が p，有意水準が 0.05 の二項検定を使用すると仮定する．ここで，C に含まれるすべてのタスクが人間ワーカによるラベル l であるとする．次に，C を受け入れるための条件は次のとおりである．

$$\frac{n - np}{\sqrt{np(1-p)}} > 1.64.$$

したがって，$n > \frac{1.64^2 p}{1-p}$ となり，$p = 0.9$ のときに $n > 24.2$ で

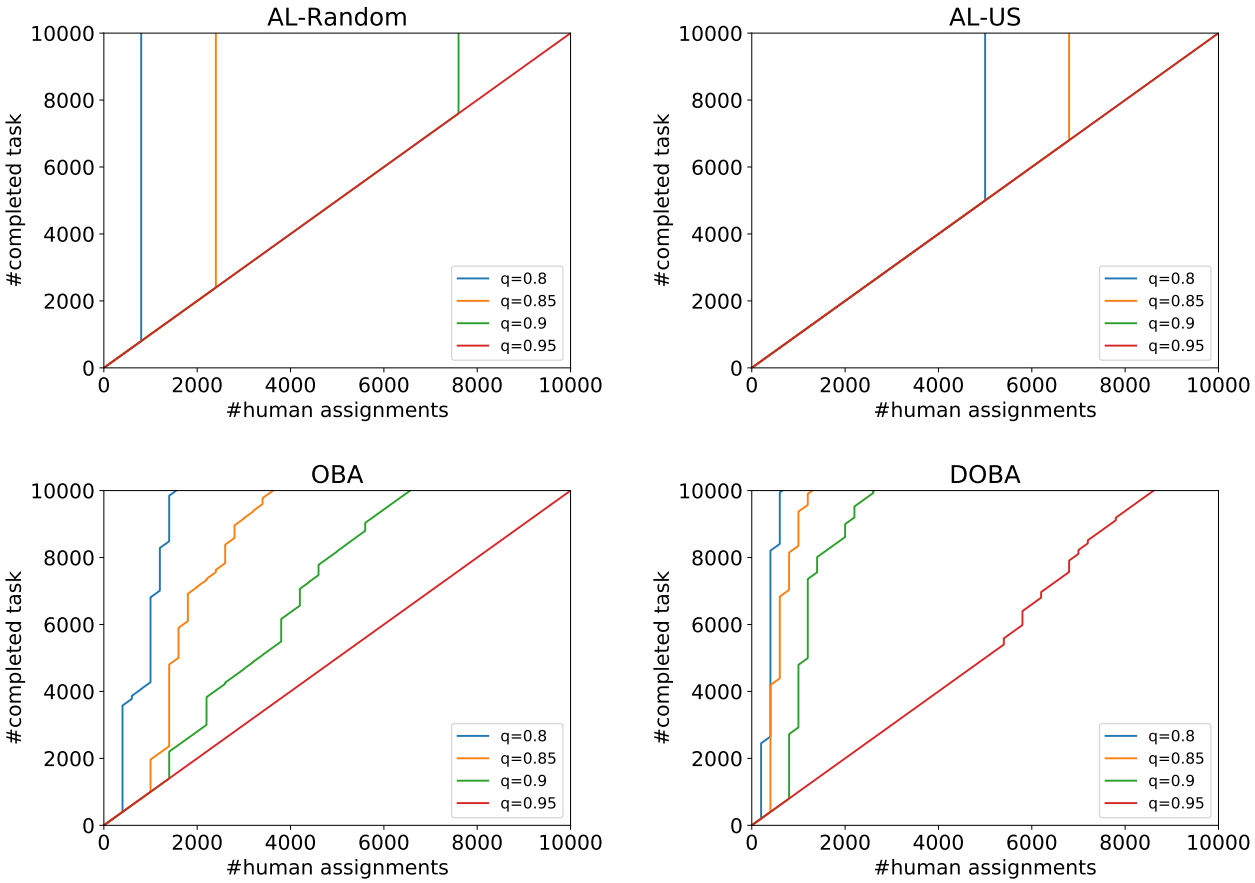

図 4　人間ワーカに割り当てられたタスク数と完了タスク数の関係

あるとする．つまり，$|C| = 25$ であることが求められる．人間ワーカに割り当てるタスクを T からランダムに選択する時，$|L'|$ の期待値は次の式で表すことが出来る．

$$E[|L'|] = \sum_{m=0}^{|C|} mp(|L'| = m \mid |L_i| = N)$$

ここで，$p(|L'| = m||L_i| = N)$ は次の同時確率分布を正規化することで求めることが出来る．

$$HGeom(m; |C| + |L^*|, |C|, N - 1 + m) \times \frac{|L^*| - (N-1)}{|C| + |L^*| - (N - 1 + m)}$$

ここで，$HGeom(k; M, K, L)$ は K 個も成功状態を含む M 個ので構成される M 個の母集団から L この要素を抽出したときに k 個の成功状態が含まれる確率を表す超幾何分布である．例えば，$|C| = 2,000$, $|L^*| = 5,000$, $N = 300$ と仮定すると，AI ワーカの学習のために C に含まれるべき人間ワーカのタスク結果数は $E[|L'|] = 119.98$ である．これは，AI ワーカに割り当てることが出来るタスク数 $|C'|$ がこれらの設定において 1,800 になることを意味する．

上記で議論したのは最も単純なケースである．一方で実際の設定は非常に複雑であり，例えば AI ワーカの性能は人間ワーカによって付与されるラベルの順番などに左右されることが考えられる．また，複数の潜在的なタスククラスタが存在する場合，それらが重複する場合も想定される．本論文の実験結果は提案アルゴリズムが C が十分な大きさである場合に，C' が空となる最悪のケースになる可能性は低いことを示唆している．

4　実　　験

4.1　設　　定

タスク． 手書き文字画像データセットである MNIST 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タスクとした．実験では，各手法により新たに付与したラベルと正解ラベルを照合することで，各手法で得られるラベルの品質を比較した．

能動学習のクエリ戦略． 能動学習に基づく手法では，AI ワーカが未割り当てのタスクに回答することで得られた情報量尺度に基づいて Uncertainty Sampling を行うことで人間ワーカに割り当てるタスクを決定する手法 (AL-US)，未割り当てのタスクからランダムに選出したタスクを人間ワーカに割り当てる手法 (AL-Random) の二種類を実験で用いた．

人間ワーカ． 実験では，人間ワーカ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，人間ワーカは一度に 200 タスクにラベルを付与した．

AI ワーカ． AL-US および AL-Random では，多層パーセプトロンの AI ワーカのみを用いた．OBA と DOBA では，

K-means (k=20), ロジスティック回帰，多層パーセプトロンの３つの AI ワーカを用いた．実験では，機械学習ライブラリである scikit-learn の実装を利用した．

タスククラスタ（AI ワーカ）の評価． AL-US および AL-Random では，統計的検定ではなく AI ワーカによるラベルと人間ワーカによるラベルを用いて 5 分割交差検証により精度を評価した．OBA では，タスククラスタにタスクを割り当てるかどうかを決定するために，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．DOBA のシミュレーションの回数は 100000 とし，$\alpha = 0.05$ とした．

4.2 実験結果

図 4 に各手法における実験結果を示す．各グラフの横軸は人間ワーカに割り当てられたタスク数を，縦軸は人間ワーカまたは AI ワーカに割り当てられたタスク数を表している．縦軸が 10000 に到達したことは，すべてのタスクを人間ワーカまたは AI ワーカのいずれかに割り当てたことを意味する．グラフの各線はそれぞれの要求精度の設定における結果を表している．少ない人間ワーカへのタスク割り当てにおいて，全てのタスクを割り当てることは，より多くのタスクを AI ワーカに割り当てたことを意味する．

AL-Random では，要求精度が 0.8〜0.9 のときに，AI ワーカへの割り当てが行われたが，0.95 の設定では AI ワーカへの割り当ては行われず，全てのタスクを人間ワーカに割り当てた．AL-US では，要求精度が 0.8 〜 0.85 のときに AI ワーカにタスクを割り当てたが，要求精度が 0.9〜0.95 の設定では AI ワーカへのタスク割り当ては行われなかった．AL-Random と AL-US を比較すると，AL-US のほうが AI ワーカへのタスク割当までにより多くの人間ワーカへの割り当てを必要とした．OBA では要求精度が 0.8〜0.9 のときに，AI ワーカにタスクを割り当てた．要求精度が 0.8 および 0.85 の設定では AL-Random の同程度の人間ワーカ割り当てを必要としたが，要求精度が 0.9 の設定ではより多くのタスクを早期に AI ワーカに割り当てることができた．DOBA では各要求精度の設定において，AI ワーカにタスクを割り当てた．全体的な傾向として，OBA よりも多くのタスクを AI ワーカに割り当てた．

図 5 は各手法でタスク割当を行った場合の最終的なタスク結果品質を示している．それぞれのグラフでは各要求精度における全体としての正答率および AI ワーカに割り当てられた部分の正答率を示している．

AL-Random と AL-US を比較すると，AL-US のほうが AI ワーカが担当したタスクの結果品質が明らかに高く，結果として全体としてのタスク結果も高くなった．OBA と DOBA を比較すると，DOBA は OBA と比較して全体的なタスク結果品質がより要求精度に近いことが分かる．

5 考　　察

AL-US では，AI ワーカの学習の効果的なデータに優先的にラベルを付与したが，AI ワーカが要求精度に到達するまでには AL-Random よりも多くの人間ワーカラベルを必要とした．本研究の問題設定では人間ワーカによるラベルと AI ワーカによるラベルの一致に基づいて AI ワーカの性能を評価するため，AL-US では分類が難しいタスク集合で精度の評価が行われる．これにより AI ワーカの性能が低く評価されたと考えられる．

AL-Random では AI ワーカの性能が要求精度に到達しない場合にはすべてのタスクを人間ワーカに割り当てたが，OBA ではタスククラスタに基づく評価により，AI ワーカの全体的な性能が要求精度に到達しない場合においてもタスクを AI ワーカに割り当てることができた．これによりより多くのタスクを AI ワーカに割り当てた．

DOBA では AI ワーカによるタスク結果品質は要求精度を下回るような AI ワーカを採用しながら，全体的な要求精度を満たすようなタスク割り当てが得られた．これは DOBA が全体的なタスク結果品質が要求精度を満たすかどうかに基づいて，タスククラスタの評価を行うためであると考えられる．

提案手法は，人間ワーカが付与するラベルが信頼できると仮定して，人間ワーカによるラベルを用いて AI ワーカにタスクを割り当てるかどうかを判断する．このため，人間ワーカにより正しいラベルを付与することが難しいタスクにおいて，AI ワーカに対してタスクを割り当てることが困難である．多数決に代表されるタスク結果の品質管理手法の導入や，タスク設計の工夫によって人間ワーカから得られるラベルの信頼性を高く保つことが，適切に AI ワーカを採用するために重要である．

6 まとめと今後の課題

本稿では，AI ワーカと人間ワーカ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 (人間+AI クラウドタスク割当問題) に対して，人間ワーカの解答に基づいて統計的検定を行い AI ワーカの得意分野を発見することで，AI ワーカ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無く，人間ワーカ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 ワーカと人間ワーカ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)，AIP チャレンジの支援による．

文　　献

[1] 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.

[2] Yan Yan, RómerRosales, Glenn Fung, Jennifer G. Dy. Ac-

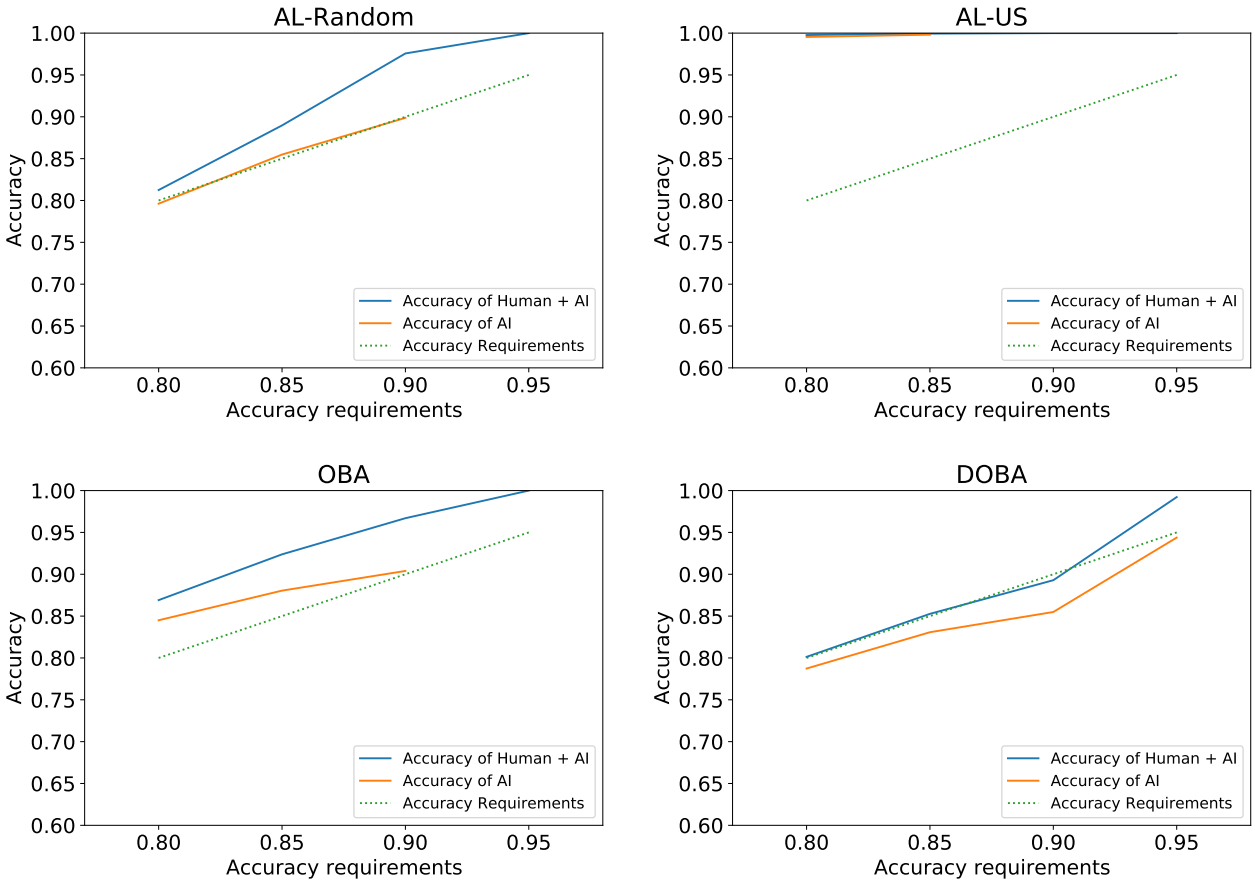

図 5　各要求精度の実験で得られたタスク割り当てにおける，全体でのタスク結果品質および AI ワーカが処理したタスクのタスク結果品質

tive learning from crowds. In *ICML*, pp. 1161–1168, 2011.

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

[4] 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.

[5] 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.

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

[7] 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.

[8] William Callaghan, Joslin Goh, Michael Mohareb, Andrew Lim, and Edith Law. Mechanicalheart: A human-machine framework for the classification of phonocardiograms. *Proc. ACM Hum.-Comput. Interact.*, Vol. 2, No. CSCW, pp. 28:1–28:17, November 2018.

[9] 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.

[10] 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.

[11] Radu Herbei and Marten H. Wegkamp. Classification with reject option. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, Vol. 34, No. 4, pp. 709–721, 2006.

[12] Ignazio Pillai, Giorgio Fumera, and Fabio Roli. Multi-label classification with a reject option. *Pattern Recognition*, Vol. 46, No. 8, pp. 2256 – 2266, 2013.

[13] Alena Naiakshina, Anastasia Danilova, Christian Tiefenau, and Matthew Smith. Deception task design in developer password studies: Exploring a student sample. In *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)*, pp. 297–313, Baltimore, MD, August 2018. USENIX Association.

[14] Karan Goel, Shreya Rajpal, and Mausam Mausam. Octopus: A framework for cost-quality-time optimization in crowdsourcing. In *Fifth AAAI Conference on Human Computation and Crowdsourcing*, 2017.