

ヒューマンコンピューテーションにおける タスク割り当て手法の提案

A Task Allocation Method to Human Computation

巻口 誉宗*¹ 東 正造*¹ 下村 道夫*¹ 金丸 直義*¹
Makiguchi Motohiro Azuma Shozo Shimomura Michio Kanamaru Naoyoshi

*¹NTT サービスエボリューション研究所

Establishing a quality control methodology for obtaining a data set consisting of the results of tasks performed by workers is one of the most important problems to solve in Human Computation (HC). Our research proposes a new quality control methodology based on "efficient task allocation algorithm" which selects suitable workers for a given task. In concrete, the algorithm classify the workers into group by NMF clustering of the workers based on the result of the tasks performed by the workers in the past, aiming to reduce the cost of worker-sectioning process and easily select a large number of workers without sacrificing the quality of the generated data set. We have shown that the newly proposed algorithm successfully reduced the cost of worker-sectioning process for a given task, and can select a large number of suitable workers maintaining quality of the generated data set.

1. はじめに

ヒューマンコンピューテーション (Human Computation, 以下 HC) は, 計算機技術のみでは自動的に処理することが困難な問題を, 人の処理能力によって解決しようとする手法 [1] である. 計算機によって画像や音声, 自然言語等のメディアを認識し, 加工するメディア処理, 知識処理エンジン (以下, エンジン) の中には, 現時点ではサービス提供に必要な認識精度基準を満たしていないものや, 更なる認識精度向上のために学習データを大量に要するものがある.

例えば, ユーザが知りたい対象の写真をサーバに送り, 画像認識エンジンによる認識結果をユーザに通知する画像認識サービスにおいて, エンジンでの認識精度が低い処理対象がサーバに送られた際に, ユーザ以外の人を作業員 (以下, ワーカー) としてエンジンの認識結果の修正や補足, 認識処理自体の代行を実現することで, サービスの品質向上が期待できる. さらに, エンジンの認識精度が低い処理対象に対する学習データをワーカーが生成することで, エンジンの認識精度自体の向上も期待できる.

近年, ワーカーを収集する手段のひとつとして, Amazon Mechanical Turk[2] 等に代表されるクラウドソーシング (Crowdsourcing) が注目されている. クラウドソーシングは, インターネットを通じて大量のワーカーを収集できるため, ワーカーを専属で確保する場合と比較して迅速かつ安価にタスクを発注できる. 一方で, 不特定多数のワーカーにタスクを依頼するため, スパムワーカー等による品質低下や, 専属ワーカーとの比較において専門知識や特定のスキル等を必要とする専門性の高いタスクの品質担保が困難といった問題がある.

本研究では, クラウドソーシングのような多数のワーカーへのタスク依頼を前提とし, 品質の担保を実現する方式を検討する. 具体的には, ワーカーの過去のタスク処理結果を元に, 類似したワーカー同士をクラスタリングし, タスクに適したワーカー群を推定するタスク割り当て手法を提案する.

2章では, クラウドソーシングにおける品質低下を防ぐ品質管理手法について, 先行研究を述べる. 3章では品質管理手法の中でも, ワーカーへのタスク割り当て手法に注目し, 従来手

法の整理と提案手法について述べる. 4章では実データ収集実験, ならびに解析について述べる. 5章では実験結果, および考察を述べる. 最後の6章はまとめである.

2. 品質管理手法

ワーカーのタスク処理結果には様々なノイズが含まれる可能性があり, クラウドソーシングをはじめとした HC に関する研究では, ノイズを除去し, より高品質な処理結果を得る品質管理手法 (Quality Control) が, 重要な要素として注目されている ([3, 4, 5, 6, 7]).

処理結果の品質を低下させるノイズとして, 意図的に低品質な処理結果を返すスパムワーカー, タスク処理に対する能力や知識が不足している適性外ワーカー. 外的・内的要因で本来のパフォーマンスを発揮できないヒューマンエラーワーカーが考えられる.

2.1 品質管理手法の先行研究

これらのノイズを除去するための品質管理手法の先行研究は, 「ワーカーの処理結果解析による高品質処理結果の算出 (アプローチ 1)」, 「ワーカーの処理結果自体の品質向上 (アプローチ 2)」の2つのアプローチがあると考えられる.

アプローチ 1 は, ノイズの混入を前提とした処理結果から, 高品質な処理結果を算出するアプローチである. 例として, ワーカーの属性推定に基づき正解を算出する手法 [3] や, 機械学習の識別機を処理結果から直接推定する手法 [4] が提案されている. このアプローチでは, 同じタスクを複数のワーカーが処理する「処理結果の冗長性確保」が前提とされており, ワーカーへ依頼するタスク数が増加するため, 報酬コスト, 時間コストの増加が懸念される.

アプローチ 2 は, ノイズ混入を事前に抑制し, 処理結果自体の品質を向上させるアプローチである. 例として, タスクの簡易化や報酬設計の工夫, 適切なワーカーの選出等が挙げられる. 中でも, これから依頼しようとするタスクに対して適切なワーカーを選出し, タスクを割り当てる手法は, 処理結果の冗長性を減らした上でも高品質な処理結果が得られると考えられ, 報酬コスト, 時間コストの削減が期待できる. タスク割り当て手法に関連した先行研究として, 芦川らは, タスク依頼者がワー

カの情報を管理可能な Private Crowdsourcing System を提案し、ワーカのスキルや経験値に基づきタスクをマッチングすることで品質向上が実現される可能性について述べている [7].

次章ではタスク割り当て手法について既存手法を整理することで、既存手法のデメリットを明確化し、デメリット解消のための要件を定義する。そして要件を満たすと考えられる提案手法を述べる。

3. タスク割り当て手法

タスクにワーカを適切に割り当てるためには、ワーカとタスクの属性推定が不可欠である。属性推定には、人による属性推定と、計算機による属性推定の2パターンがある。以下にそれぞれのパターンのメリットとデメリットをまとめ、デメリットを解消するための要件を定義する。

3.1 人によるワーカの属性推定

本例として、リサーチ会社のスクリーニングに代表される「自己申告」がある。自己申告のメリットは、詳細な属性情報を容易に取得できる点等が挙げられる。一方、スパムワーカや、ヒューマンエラーワーカ、自らの状態を正しく把握していない適性外ワーカからは、信頼できる申告結果が得られない等のデメリットが考えられる。ワーカの自己申告の信頼度について、小山ら [5] は、ワーカが申告する「タスク処理結果への自信度」と実際の正解率の間に差があることを報告している。このデメリットを解消するためにワーカの自己申告情報以外の情報利用を要件として定義する（要件1）。

また、人によるワーカの属性推定の別の例として、タスクを処理したワーカの評価を、リクエストや別のワーカが行う「他者評価」[6]がある。この手法は、タスクの処理結果が大量にある場合、評価するワーカ数が増加し、報酬コスト、時間コストが増加するといったデメリットがある。ワーカ数の増加はアプローチ1においてもデメリットと考えられるため、このデメリットを解消するためにタスク依頼数の削減を要件として定義する（要件2）。

3.2 計算機によるワーカの属性推定

本例として、テストタスクが挙げられる。正解が既知のテストタスクをワーカに処理させ、正解率を算出することで、そのタスクに適したワーカの選定や、スパムワーカ、適正外ワーカを除外できると考えられる。テストタスクのデメリットとして、テストタスクの処理にかかる報酬コストや時間コストの増加、テストタスクを用意するコストの増加等が考えられる（導かれる要件は要件2と等しい）。

計算機による他の属性推定手法の例として、処理時間の解析がある。例えば NTT コム リサーチ [8] では、アンケートの回答時間が短いワーカ5%を解答集計に含めないという品質管理をオプションとして用意している。この手法では、アンケートのように一定以上の想起時間を要するタスクにおいては、処理時間の短さからスパムワーカを判断し、除去することができると考えられるが、画像のカテゴリ分類のようにワーカの得意不得意によって処理時間にばらつきが出るようなタスクにおいては、処理を早く正確に行えるワーカを排除してしまう可能性がある。

3.3 提案手法

前述の属性推定手法は、いずれもワーカ個人を対象とした手法である。しかし、例えばテストタスクによる属性推定を個人単位で行う際には、推定したい人数分、テストタスクを依頼する必要があるため、タスク依頼数が膨大となり、要件2を満

たせない。また、アンケートのような正解がないタスクにおいては、ワーカー一人ひとりの処理結果よりも、ワーカ集団としての処理結果が重要視されることから、個人単位の属性推定は必ずしも効果的ではないと考えられる。

そこで本研究ではワーカの集団に対する属性推定に注目し、ワーカのクラスタリングによるタスク割り当て手法を提案する。この手法は、ワーカの過去のタスク処理結果から類似した回答傾向を持つワーカ群を作成し、これから依頼したいタスク（新規タスク）に適したエキスパート群の選定を目的とする。

提案手法は、クラスタの作成にワーカの過去のタスク処理結果を用いる点で要件1, 2の双方を満たすことができると考えられる。クラスタリングによるワーカ群の属性推定の先行研究として、Gomes ら [9] は、画像のカテゴリ分類をワーカの主観によって行わせ、その処理結果を元にワーカをクラスタリングする手法を提案している。我々の研究は Gomes らの手法と比較して、画像認識以外のタスクでも利用可能な手法を検討する点、エキスパート群の選定手法を検討する点に差分がある。

以前の我々の検討 [10] では、画像ラベリングタスク、固有表現ラベリングタスクの処理結果を社内クラウドソーシングで収集し、そのタスク処理結果を NMF でクラスタリングすることで、正解率の高いワーカのクラスタを分類できることを示した。本稿では、リサーチ会社のモニターから収集したタスク処理結果を解析する。タスク処理結果を学習用データと評価用データに分け、学習用データで得られたクラスタから推定されるエキスパート群について、評価用データ（新規タスクを想定）における正解率を検証することで、タスク割り当て手法としての提案手法の有効性を検証する。

4. タスク処理結果の実データ解析

提案手法の有効性を検証するため、タスク処理結果の実データ収集実験、ならびに解析を行った。

4.1 タスク設計

今回の検証では、客観的な評価を容易にするため、正解があるタスクを対象とし、「タスクに適したワーカ」を「正解率の高いワーカ」と定義した。専属ワーカが固有表現 (Named Entity) に対してラベリングしたデータを用いてタスクを作成し、専属ワーカの付与したラベルをそのタスクの正解と定義する。固有表現とは、人名や地名等の固有名詞に加え、日付や数量表現を合わせた表現である。今回は関根ら [11] の拡張固有表現階層を標準としたラベリングデータを用いた。

ワーカには、文章とその中に含まれる単語を提示し、提示された単語について、固有表現の該当するクラスを一覧から選択するよう依頼した。クラスの選択肢は、拡張固有表現階層における最上位階層14クラス（例：人名、神名、組織名等）を設定し、ワーカには拡張固有表現階層のリストと例を提示し、標準に従うよう指示した（図1）。

ワーカはリサーチ会社のモニター331名で、固有表現ラベル15問の処理結果を収集した。

4.2 クラスタリング手法

クラスタリング手法には、非負値行列因子分解 (Nonnegative Matrix Factorization: 以下 NMF) を用いた。NMF は I 行 J 列の行列 \mathbf{X} を、 I 行 K 列の特徴行列 \mathbf{T} 、 K 行 J 列の特徴行列 \mathbf{V} に分解し、 \mathbf{T} と \mathbf{V} の積 \mathbf{TV} と、分解前の行列 \mathbf{X} の距離 D が最小となるよう、 \mathbf{T} と \mathbf{V} を最適化する手法である。NMF は文章データや購買データにおける頻出パターンを算出する用途で広く利用されている [12]。NMF の処理結果が

生物分類上はイルカとクジラには差はない。むしろ、ハクジラとヒゲクジラの差の方が生物的にも形態的にも違いが顕著である。しかし世界的にも日常語レベルでは両者は別のカテゴリーとして認識され、別名で呼ばれることが多い。日本語では、成体の体長がおよそ4mをクジラとイルカの境界と考えることが多い。

回答方向	イルカ	クジラ	ハクジラ	日本語	4m
人名(Person)	○	○	○	○	○
神名(God)	○	○	○	○	○
組織名(Organization)	○	○	○	○	○
地名(Location)	○	○	○	○	○
施設名(Facility)	○	○	○	○	○
製品名(Product)	○	○	○	○	○
イベント名(Event)	○	○	○	○	○
自然物名(Natural_Object)	○	○	○	○	○
病名(Disease)	○	○	○	○	○
色名(Color)	○	○	○	○	○
時間(Time)	○	○	○	○	○
期間(Period)	○	○	○	○	○
数値表現(Numerical)	○	○	○	○	○
該当なし(Other)	○	○	○	○	○

図 1: 固有表現ラベリングタスクの処理画面

ら得られる特徴行列 \mathbf{T} , \mathbf{V} は、それぞれタスク、ワーカがクラス $k \in \{1, \dots, K\}$ に属する強さを表す。

\mathbf{T} と \mathbf{V} の積 \mathbf{TV} と、 \mathbf{X} の距離 D の定義は、一般化 Kullback-Leibler divergence (以下, KL divergence) を用いた [12].

$$D * (X, TV) = \sum_{i=1}^I \sum_{j=1}^J d * (x_{ij}, \widehat{x}_{ij}) \quad (1)$$

$$d * (x_{ij}, t_i^T, v_j) = x_{ij} \log \frac{x_{ij}}{t_i^T v_j} - x_{ij} + \widehat{x}_{ij} \quad (2)$$

ここで $\widehat{x}_{ij} = t_i^T v_j$ であり, KL divergence を用いた \mathbf{T} , \mathbf{V} の各要素 t_{ik} , v_{kj} の最適化は, 以下の更新式で行われる。

$$t_{ik} \leftarrow t_{ik} \frac{\sum_j \frac{x_{ij}}{x_{ij}} v_{kj}}{\sum_j v_{kj}}, v_{kj} \leftarrow v_{kj} \frac{\sum_i \frac{x_{ij}}{x_{ij}} t_{ik}}{\sum_i t_{ik}} \quad (3)$$

代表的なクラスタリング手法である K-means は各ワーカがどれかのクラスに強制的に分類されるハードクラスタリングであるのに対し, NMF はワーカごとにそれぞれのクラスに属する強さ (帰属度) が与えられるソフトクラスタリングである。今回の検討ではそれぞれのクラスへの帰属度が高い順から任意の数のワーカ抽出が可能である点を重視し, ソフトクラスタリングである NMF を選択した。

4.3 処理結果行列の作成

NMF の対象とする行列 \mathbf{X} は, タスク処理結果から以下の手順で作成した。それぞれのタスク番号を $i \in \{0, \dots, I\}$, ワーカの番号を $j \in \{0, \dots, J\}$ とし, 各タスクの処理結果を $[0, 1]$ の 2 値で表現する。すなわちワーカ j の i 番目のタスクへの処理結果ベクトルを \mathbf{y}_i^j とする。例えば固有表現ラベリングタスク i において, クラスの選択肢が [人名, 地名, 施設名] であり, ワーカ j が人名を選択した場合, $\mathbf{y}_i^j = [1, 0, 0]$ と表現される。この処理結果ベクトルを結合し, 処理結果行列 \mathbf{X} を作成した。

4.4 解析と比較手法

解析は下記の手順で行った。

手順 1. 評価対象ワーカの抽出

収集した 331 名分のデータから 66 名のワーカをランダムに抽出し, 評価対象とする。 ($N = 66$)

手順 2. タスクの学習データ, 評価データへの分類

評価対象ワーカのタスク処理結果 75 問のうち, ランダムに抽出した 15 問を学習データ, 残り 60 問を評価データとする。

手順 3. 学習データを用いたワーカ群の抽出

学習データの処理結果を元に, ワーカを $n (< N)$ 人抽出する。抽出方法は後述する。

手順 4. 評価データの正解率算出

手順 3 で抽出した n 人のワーカの評価データから正解率を算出する。

評価対象のワーカから n 人抽出する手法は以下の 5 つを用い, 評価データにおける正解率を比較した。

提案手法 (NMF_Max)

学習データの処理結果から NMF によってクラスタを作成する。得られたクラス $k \in \{1 \dots K\}$ について, 帰属度が高い上位 $n' (< N)$ 人のワーカを選定し, 学習データの正解率の平均値を算出する。この平均値が最も高いクラスにおける帰属度の上位 n 人を抽出する。

ランダム抽出 (Random)

ランダムに n 人抽出する。

多数派抽出 (Majority)

学習データの処理結果の多数決をとり, 多数決結果との一致率が高いワーカを上位 n 人抽出する。

エキスパート抽出 1 (Expert_Random)

ランダムに ($n' \times K$) 人のワーカを抽出し, 学習データの正解率が高い順に上位 n 人を抽出する。

エキスパート抽出 2 (Expert_All)

評価対象の全ワーカを対象に学習データの正解率を算出し, 正解率が高い順に上位 n 人を抽出する。

それぞれの抽出手法において, 学習データの正解率が必要となるワーカ数 (テストタスクを依頼するワーカ数) は以下の通りである。今回の解析では, クラス数 $K = 3$, クラス評価用ワーカ数 $n' = 3$ とした。

ワーカ抽出手法	学習データの正解率が必要となるワーカ数
NMF_Max	$n' \times K$
Random	0
Majority	0
Expert_Random	$n' \times K$
Expert_All	N

初期値によるクラスタリング結果の変動, ワーカセット・タスクセットによる変動を排除するため, 乱数の初期値を変えて手順 1 ~ 4 の試行を 100 回繰り返す, 評価データの正解率の平均値を取得した。

5. 結果と考察

5.1 タスク依頼数の削減効果

図 2 に各手法で抽出したワーカ 3 人 ($n = 3$) の平均正解率を示す。評価データの正解率が高い順に, 全ワーカの学習データ正解率に基づくエキスパート抽出 2 (Expert_All), 提

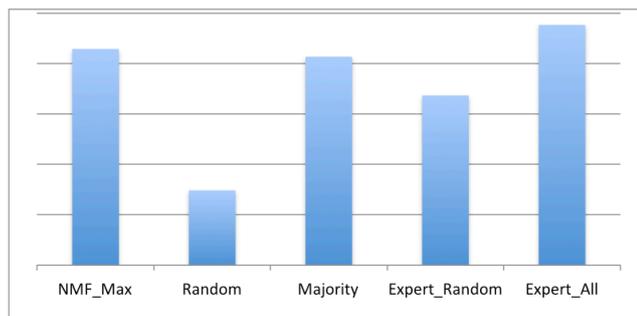


図 2: $n = 3$ における各手法の平均正解率。縦軸は正解率。

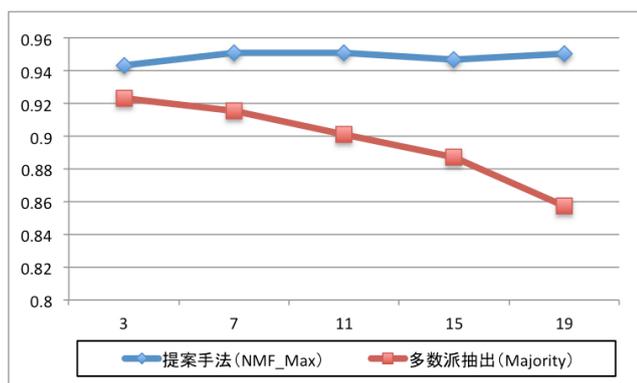


図 3: $n = (3 \sim 19)$ における提案手法 (NMF_Max), 多数派抽出 (Majority) の正解率の推移 (エキスパート抽出 2 の正解率で正規化)。

案手法 (NMF_Max), 一部のワーカの学習データ正解率に基づくエキスパート抽出 2 (Expert_Random), ランダム抽出 (Random) である。

エキスパート抽出 2 (Expert_All) は今回の手法の中で最も正解率が高かったものの, 評価対象の全ワーカの学習データ正解率を必要とする (今回の解析では $N = 66$ 人)。一方で, 提案手法 (NMF_Max) においては, 今回の解析ではクラスタ評価人数 $n' = 3$, クラスタ数 $K = 3$ であるため, 9 人分の学習データ正解率で, 正解率の高いワーカ群を推定可能である。

エキスパート抽出 1 は, 提案手法と同じ人数の学習データ正解率を用いた抽出手法であるが, 提案手法よりも正解率が低くなる。このことから, 提案手法は要件 2 のタスク依頼数削減を実現する手法として有効であると考えられる。

5.2 エキスパート群の選定

図 3 に, 提案手法 (NMF_Max), 多数派抽出 (Majority) において, n を 3 ~ 19 まで 4 人ずつ増加させた際の正解率の推移を示す。各手法の正解率は, エキスパート抽出 2 (Expert.all) の $n = (3 \sim 19)$ における正解率によってそれぞれ正規化した。

この図から, 多数派抽出では抽出するワーカ数の増加に伴って正解率が減少していくが, 提案手法は比較的正解率を維持していることがわかる。これは提案手法が正解率の高いワーカをより多く抽出しやすいことを示しており, ワーカが大量に必要な場合等, エキスパート群の選定における提案手法の有効性が示される。

6. まとめ

本稿では HC における品質管理手法としてワーカのタスク処理結果によるクラスタリングを用いたタスク割り当て手法を提案し, 固有表現ラベリングのタスクに対する処理結果の実データを用いてタスク依頼数削減, エキスパート群選定の有効性を示した。今後は固有表現ラベリング以外のタスクについても提案手法の適用可能性の検討を行う。

参考文献

- [1] von Ahn, L.: *Human Computation*, Doctoral Thesis. UMI Order Number: AAI3205378, CMU, (2005).
- [2] Amazon Mechanical Turk, <http://aws.amazon.com/jp/mturk> (2014.02.13).
- [3] Dawid, A.P. and Skene, A. M., *Maximum Likelihood Estimation of Observer Wrorr-Rates Using the EM Algorithm*, Journal of the Royal Statistical Society. Series C, Vol. 28, No.1, pp. 20-28,(1979).
- [4] 梶野 洸, 坪井 祐太, 佐藤 一誠, 鹿島 久嗣, エキスパートによる訓練データとクラウドソーシングで作成した訓練データからの教師付き学習, 人工知能学会論文誌, Vol.28, No.3, pp.243-248, (2013).
- [5] 小山 聡, 馬場 雪乃, 櫻井 祐子, 鹿島 久嗣, クラウドソーシングにおけるワーカの確信度を用いた高精度なラベル統合, 人工知能学会全国大会, 2M5-OS-07b-2, (2013).
- [6] 馬場 雪乃, 鹿島 久嗣, 非定型出力をもつクラウドソーシングタスクにおける成果物の統計的品質推定, 人工知能学会全国大会, 2M5-OS-07b-1, (2013).
- [7] 芦川 将之, 有賀 康顕, 宮村 祐一, *PrivateCrowdSourcing* を用いた言語, 音声資源の収集, 人工知能学会全国大会, 3M3-OS-07d-2, (2013).
- [8] NTT コム リサーチ, <http://research.nttcoms.com/service/qpolicy4.html/> (2014.02.13).
- [9] R. Gomes, P. Welinder, A. Krause, and P. Perona, *Crowdclustering*, Advances in Neural Information Processing Systems, (2011).
- [10] 巻口 啓宗, 並河 大地, 東 正造, 下村 道夫, 金丸 直義, ヒューマンコンピューテーションにおける非負値行列因子分解を用いたタスク割り当て手法の提案, 情報処理学会 第 91 回 グループウェアとネットワークサービス研究会 (2014).
- [11] 関根の拡張固有表現階層 -7.1.0-, <https://sites.google.com/site/extendednamedentityhierarchy/> (2014.02.13).
- [12] 澤田 宏, 非負値行列因子分解 NMF の基礎とデータ / 信号解析への応用, http://www.kecl.ntt.co.jp/icl/signal/sawada/mypaper/829-833_9_02.pdf/ (2014.02.13).