DEIM Forum 2021 A24-2

クラウドソーシングにおける多数決の精度向上のためのチーム編成手法

松田　浩幸†　　田島　敬史†

† 京都大学大学院情報学研究科　〒 606–8501 京都府京都市左京区吉田本町
E-mail: †matsuda-hiroyuki@dl.soc.i.kyoto-u.ac.jp, ††tajima@i.kyoto-u.ac.jp

あらまし　本研究では，クラウドソーシングにおいて多数決の精度が高くなるようにワーカーのチーム分けを行う手法を提案する．多数決は一つの不正解の選択肢に回答が偏ると失敗する．また，能力の高いワーカーを同じチームに集めると，他のチームの多数決の精度が下がる．我々の以前の研究では，まず，ワーカーの回答データを one-hot ベクトル化したもの同士のユークリッド距離を用いてクラスタリングを行う． 各クラスタは回答の傾向が近いワーカーで構成されるので，異なるクラスタのワーカーを組み合わせることで，チームを編成した．本研究では，新たにワーカー間の距離として，ワーカー同士の親和度という概念を導入し，二人のワーカー A と B に対し，A と平均的なワーカーで形成したチームで多数決をとる場合の正解率と，B も含めたチームで多数決をとる場合の正解率の差の全テスト問題における平均で定義する．また，ハンガリアンアルゴリズムによって距離の遠いワーカー同士が同じチームに属するように，チーム編成を行う．人工データとクラウドソーシングのプラットフォーム上で集めたデータの双方で実験を行なった結果，比較手法及び我々のこれまでの手法よりも正解率の高いチームを編成することができた．
キーワード　クラウドソーシング，多数決，チーム編成，品質管理

1 はじめに

クラウドソーシングとは，インターネットなどを介して不特定多数の人々に作業を依頼する仕組みのことである．例えば，文章の校正をするタスクや，Web サイトの感想を入力するタスクなど様々なタスクを比較的安価なコストで依頼することができる．また，画像などのアノテーションはこれまで少数の専門家によって行われることが多かったが，クラウドソーシングによって専門家ではない不特定多数のワーカーによっても行われるようになった [3]．これにより，機械学習などに用いる大量の正解データなど，計算機では生成するのが困難であったデータを低コストで収集することが可能になった．そして，近年ではクラウドソーシングをサポートする様々なサービスが広く用いられるようになってきており， Amazon によって開始された Amazon Mechanical Turk [1] などのクラウドソーシングのプラットフォームによって市場は急速に拡大している．

しかしながら，クラウドソーシングのワーカーは能力や意欲にばらつきがあるために，得られるデータの品質にばらつきが生じる．よって，品質管理はクラウドソーシングにおける重要なテーマの一つである．安定して高品質な結果を得るための仕組みとして広く採用されているのが，単一のタスクを複数のワーカーに依頼し，その回答を統合して信頼性を向上させるという手法である．クラウドソーシングで依頼される典型的な作業の一つであるデータへのラベル付けなどの多クラス分類問題に対しては，多数決が有用な統合手法の一つとして挙げられる．

実際のクラウドソーシングにおいて募集した複数のワーカーをいくつかのチームに分け，それぞれのチームにタスクを依頼する場合を考える．各チーム内のワーカーには同じタスクを依頼し，それぞれの回答の多数決をとってデータを収集する．このとき，どのようにチーム編成を行うかは重要である．今回は，あらかじめ正解が既知である複数の問題を用意し，全ワーカーに全問題に回答してもらい，その回答データを収集できていると仮定する．このとき，最も正解率が高くなるようなチーム分けを行う最善の方法は，チーム編成の全ての組合せの正解率の期待値を計算し，最も高かったチーム編成を採用することである．しかしながら，この方法には大きな問題点がある．それは計算量が膨大になってしまうことである．例えば 40 人のワーカーを 5 人チームで 8 個に分ける場合は，約 12 京通りとなってしまう．これは，現実的な時間で計算することは不可能である．

多数決の精度が高くなるようなチーム編成を行うために，多数決が失敗してしまう状況について考える．多数決が失敗する状況の一つは，一つの不正解の選択肢にワーカーの回答が偏ってしまうような場面である．できるだけこのような状況を回避するためには，チーム内の各ワーカーの回答が不正解である選択肢のいずれかに偏る確率が低くなるようにすることが重要である．各ワーカーの中には回答の傾向が似通ったワーカーが存在するので，正解率が高かったとしても回答の傾向が近いワーカー同士は同じチームに入れるべきではないと考えられる．また，チーム編成を行う場合の特有の注意点として，能力の高いワーカーを同じチームに集めると，他のチームの多数決の正解率が低くなるという問題がある．そのため，能力の高いワーカーは全チームの正解率を向上させるために，各チームに分散することが望ましいと考えられる．

我々の以前の研究 [18] では，まず，ワーカーの回答データを one-hot ベクトル化したもの同士のユークリッド距離を用いてクラスタリングを行う． 各クラスタは回答の傾向が近いワーカーで構成されるので，異なるクラスタのワーカーを組み合わせることで，チームを編成した．本研究では，新たにワーカー間の距離として，ワーカー同士の親和度という概念を導入する．

二人のワーカー A と B に対し，A と平均的なワーカーで形成したチームで多数決をとる場合の正解率と，B も含めたチームで多数決をとる場合の正解率の差の全テスト問題における平均で定義する．この距離を用いて以前の研究と同様にクラスタリングを行った後，距離の遠いワーカー同士が同じチームに属するように，チーム編成を行う．距離の近いクラスタのペアから順番にハンガリアンアルゴリズムを用いて同じチームに配属させるワーカーのペアを作ることで，チームを形成する．

本論文の構成は以下の通りである．第 2 章において，本研究と関連のあるクラウドソーシングの統計的な品質管理手法に関する研究など，提案手法に関連した研究について述べる．第 3 章では，本研究で扱う問題の設定について説明する．第 4 章では，提案手法の説明を行い，第 5 章において提案手法の性能評価のための計算機実験について述べる．第 6 章では，本論文の結論を述べるとともに，今後の課題について説明する．

2 関 連 研 究

クラウドソーシングにおける品質管理については様々な研究が行われている．不特定多数のワーカーの回答から真の正解を推定するための最も単純な手法は，同じ問題に対して複数のワーカーから回答を得て多数決を取ることである．しかし，単純な多数決では能力の高いワーカーと同じ選択肢を選び続けるような悪質なワーカーの回答の重みが変わらないので，多数決をしても品質が向上しない．そこで，正解率によって各ワーカーの回答の重みを調整したり，割り当てるタスクを増減させるような手法が提案されている．Dawid ら [4] は，信頼度の異なる複数の医師が複数の患者を診断した結果から正しい診断結果を推定する，EM アルゴリズムを利用した手法を提案した．この研究は医療の診断における文脈で行われたものであったが，クラウドソーシングの品質管理の文脈においても様々な研究の基礎となっており，Dawid らの手法を拡張するような研究も盛んに行われている．例えば，各ワーカの性能と各タスクの難易度を考慮した手法 [16] や，ワーカの性能とタスクの難易度を多次元化した手法 [15] などが挙げられる．また，Joglekar ら [5] [6] は，多数決における回答の一致を用いた，ワーカの誤り率を推定する手法を提案した．

Kuncheva ら [10] は，複数の分類器によるアンサンブル学習において，分類器の精度，数，ユールの Q で計算される分類器間の相関関係が多数決の精度にどのような影響を与えるかについて研究した．正の相関がある分類器同士では，ある分類器が正解できる問題は他の分類器も正解できる可能性が高くなり，逆に，ある分類器が正解できない問題は分類器も正解できない可能性が高い．多数決では，正解の票が過半数より多くあっても，多数決の精度改善には無駄な票となってしまうので，アンサンブルによる正解率の向上は期待できない．負の相関がある分類器同士では，ある分類器が正解できない問題に対しても，他の分類器が正解できる可能性があるので，多数決で正解することができる可能性が向上する．つまり，複数の分類器で多数決をとる場合，分類器間の負の相関は有益であるということがわかる．本研究では，この考えに基づき，多数決を行うそれぞれのチームは回答の傾向が異なるワーカーによって構成されるようにチーム編成を行う．

Wu ら [17] は，クラウドソーシングにおいて単一のタスクを複数のワーカーに依頼するときに，多様性を重視したワーカーを選出する手法を提案した．各ワーカーの類似度をプロフィールの一致度や過去の異なるタスクにおける回答データなどから計算し，多様性の度合いとして扱う．そして，ワーカー同士の類似度の平均が低い，つまり多様性の高いようなワーカーの組み合わせを選出する．この手法はワーカーの類似度を計算し，類似度の低いワーカーを選出するという点で提案手法に近い手法であるが，多様な解を収集すること自体を目的としており，提案手法では多数決で正解率が高くなるようなチーム編成を行うことが目的であるという点で異なる．

3 問 題 設 定

本研究で扱う全ワーカーの中から多数決の精度が高くなるようなワーカーのチーム編成を行う問題について定義する．まず，ワーカーが取り組むタスクとして，典型的なクラウドソーシングのタスクである多クラス分類問題を用いる．多クラス分類問題では，ある問題に対して複数の選択肢の中から正解だと思われる選択肢を回答させる．例えば，「画像に写っているイヌ科動物の種類はどれであるか」という問題に対して，「(a)Alaskan Malamute，(b)Siberian Husky，(c)Samoyed，(d)German Shepherd，(e)Gray Wolf，(f)Coyote，(g)Dhole」の選択肢の中から答えるタスクである．

本研究で扱う問題は，次のように表される． チーム編成を行うワーカーを n 人とし，チームの数を k，1 チームの人数は今回は n/k が割り切れる場合のみを考え $v(= n/k)$ 人とする．また，正答が既知であるテスト問題集合を Q，多クラス分類問題の選択肢集合を D とし，ワーカーは Q の各テスト問題に対して選択肢を一つ選んで回答しているとする．また，正解が $d \in D$ のときに各ワーカー $w \in W$ が各選択肢 $e \in D$ を正解と判定する確率 $\pi^w_{d,e}$ が与えられているとする．本問題では，n 人のワーカーから間違い方の傾向が異なるワーカーを組み合わせることで，多数決の精度が高いチーム分けを行うことが目的である．

3.1 相対多数決

本研究では様々な多数決の手法の中でも単純な相対多数決を考える．相対多数決では，選出された各ワーカーの回答の中で最も多く回答された選択肢を正解の選択肢であると推定し，多数決の結果として出力する．最多の回答が過半数を超えていなくても，最多の票数であればその選択肢を採用する．最多の回答となった選択肢が複数存在する場合は，それらの選択肢の中から等確率で一つの選択肢を選び，正解の選択肢であると推定することとする．

3.2 組み合わせ爆発

本問題の解法として考えられる単純な方法とは，全てのチー

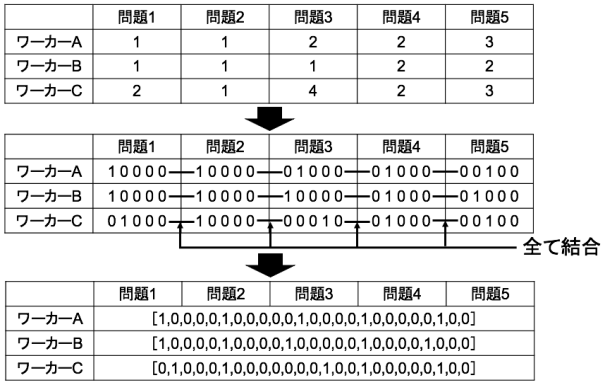

図 1　ワーカーの特徴ベクトルの定義の流れ

ム編成の組合せについて多数決の正解率を計算し，最も高かったチーム編成を採用することである．しかし，この方法には大きな問題点がある．それはワーカーの人数が増えるにつれ，チーム編成の組合せは爆発的に増加していく，つまり組合せ爆発が起こってしまいことである．ワーカーの数を w 人として v 人のチームを w/v 個作る場合，全通りのチーム分けには $\frac{{}_wC_v \times {}_{w-v}C_v \times \cdots}{(w/v)!}$ 通りの分け方が存在する．例えば 40 人のワーカーを 5 人チームで 8 個に分ける場合は，約 12 京通りとなってしまう．これは，現実的な時間で計算することは不可能である．よって，単純に全てのチーム編成の組合せについて多数決の正解率を計算することはできないので，少ない計算時間で多数決の精度の高いチーム編成を行うことができる手法が必要である．本研究で提案する手法は，このような問題から，できるだけ現実的な計算時間でチーム編成を行うことを目指して考案したものである．

4　提 案 手 法

提案手法は，ワーカーをクラスタリングし，異なるクラスタのワーカーを組み合わせることで多数決の精度が高くなるようなワーカーのチームを編成する手法である．多数決が失敗してしまう状況として考えられるのは，一つの間違いの選択肢にワーカーの回答が偏ってしまった結果，間違いの選択肢が多数派となるような場面である．できるだけこのような状況を回避するためには，チーム内の複数のワーカーが間違いの選択肢の一つに偏る確率が低くなるようなワーカーを選出することが重要である．

4.1　ワーカー間の距離の定義

提案手法ではまず，後のクラスタリングで利用するために，ワーカー間の距離を過去の回答データから定義する．今回は 2 種類の定義について説明する．

4.1.1　ワーカーの回答のベクトル化

まずは，我々のこれまでの研究で提案した，テスト問題の回答データからワーカー間の距離を定義する従来手法について説明する．ワーカー間の距離を定義する流れについて図 1 に示す．ワーカーの過去の回答データから，各タスクごとに回答の選択肢に対応した one-hot ベクトルへと変換する．ここで変換される one-hot ベクトルとは，次元数が全選択肢の個数と同じであり，回答の選択肢に対応した次元のみが 1，それ以外が 0 となるようなベクトルのことである．例えば，テスト問題の選択肢が 1 から 5 の 5 個あるときに，ワーカーの回答が 2 であったときは (0,1,0,0,0) というベクトルに変換される．この操作を全てのテスト問題の回答に対して行った後，各回答の one-hot ベクトルを全て連結する．つまり，連結後のベクトルの次元数は選択肢の個数 $|D|$ とテスト問題数の $|Q|$ の積となる．本論文ではこのように定義したベクトルを「ワーカーの特徴ベクトル」と呼ぶことにする．また，提案手法では，ユークリッド距離を用い，ワーカー間の距離として定義する．これによって，あるワーカーともう一方のワーカーが同じ問題に対して異なる選択肢で間違ったときにワーカー間の距離が生まれるので，両方が同じクラスタに属しずらくなる．逆に，間違い方が同じであるようなワーカー同士の距離は近くなる．

4.1.2　ワーカー間の親和度

二つ目のワーカー間の距離として，ワーカー同士の親和度という概念を導入する．二人のワーカー組に対し，一方のワーカーと平均的なワーカーで形成した v 人のチームで多数決をとる場合の正解率と，もう一方のワーカーも含めた v 人のチームで多数決をとる場合の正解率の差で定義される．例えば，ワーカー A とワーカー B に対して 5 人のチームを作る場合を考える．あるテスト問題 q において，A が正解で B が不正解の場合，A と平均的なワーカー 4 人による 5 人のチームで多数決をとった場合の正解率を計算する．平均的なワーカーとは，テスト問題 q の各選択肢を全ワーカーが選んだ割合に応じて確率的に選択するようなワーカーである．次に，平均的なワーカーの一人を B と入れ替え，5 人で多数決をとった場合の正解率を計算する．A が不正解で B が正解の場合は，上記の A と B を入れ替えて計算する．双方ともに正解や不正解だった場合も上記と同様に計算する．以上を全問題に対して計算し，その平均値がワーカー A とワーカー B の「親和度」となる．

4.2　ワーカーのクラスタリング

4.1 において定義された距離によるクラスタリングを行う．提案手法ではクラスタリングには特徴ベクトルのユークリッド距離に対しては Balanced K-means 法を，親和度に対しては Balanced K-medoids 法を用いた．クラスタ数である K は選出するチームの人数である v とする．これによって，最後に異なるクラスタからワーカーを選出するときに，各クラスタから一人ずつ選出すればチームの人数を v 人とすることができる．

4.2.1　最適化による Balanced K-means 法

最適化による Balanced K-means 法 [14] とは，一般的な K-means 法に少し変更を加え，クラスタのサイズが均等になるようにクラスタリングを行う手法である．K-means 法では，データ点を最近傍のセントロイドを持つクラスタに割り当てるステップと，クラスタの重心を再計算することでセントロイドを更新するステップを繰り返していく．最適化による Balanced K-means 法のアルゴリズムは，セントロイドを更新するステップは K-means 法と同様であるが，データ点を各クラスタに割り

当てる方法が異なっている．最適化による Balanced K-means 法では，単純にデータ点から最近傍のセントロイドを持つクラスタを選ぶのではなく，クラスタのサイズが均等になるように制約を加えた最適化問題を解くことで，割り当てるクラスタを計算する．今回，各クラスタのサイズが均等になるようなクラスタリングアルゴリズムを利用する理由は，この後の異なるクラスタから 1 人ずつワーカーを選出するときに，クラスタのサイズが均等でない場合は，いずれかのクラスタで先に選出できるワーカーがいなくなってしまい，均等なチーム編成ができなくなるからである．

最適化による Balanced K-means 法のアルゴリズムの詳細について説明する．まず，あらかじめ決定しておいたクラスタ数 k 個の最初のセントロイドを K-means++法 [2] によって決定する．次に，全てのデータ点を各クラスタに割り当てる．このとき，本来の K-means 法とは異なり，各クラスタのサイズが均等になるように割り当てを行う．そのために，今回の最適化による Balanced K-means 法の最適化問題を次のように定式化する．

$$\text{minimize } E(c,x) = \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n} x_{i,j}||o_i - c_j||^2 \qquad (1)$$

$$\text{s.t. } x_{i,j} \in \{0,1\} \qquad i = 1,2,...,n \quad j = 1,2,...,k$$

$$\sum_{j=1}^{k} x_{i,j} = 1 \qquad i = 1,2,...,n$$

$$\lfloor \frac{n}{k} \rfloor \leqq \sum_{i=1}^{n} x_{i,j} \leqq \lceil \frac{n}{k} \rceil \qquad j = 1,2,...,k$$

目的関数は各データ点と所属するクラスタのセントロイドとの二乗誤差である．変数 c_j は j 番目のクラスタのセントロイドを表すベクトルであり，$x_{i,j}$ は i 番目のワーカーが j 番目に所属している場合に 1 となり，それ以外は 0 となるような変数である．また，o_i は i 番目のデータ点のベクトルを表す．二番目の制約式は，各ワーカーはいずれかのチームにのみ所属することを表し，三番目の制約式は各チームの所属ワーカー数は $\lfloor n/k \rfloor$ 以上 $\lceil n/k \rceil$ 以下であることを表している．K-means 法と同様に割り当てのステップでは，変数 c を固定し変数 x に関して目的関数 $E(c,x)$ の最小化を行う．この最適化問題は整数線形計画問題であるプので，通常は分岐限定法 [11] などのアルゴリズムを用いて解くことになるが，計算時間が大きくなるという問題点がある．しかし，今回の最適化問題の場合は，変数 $x_{i,j}$ の整数条件を除去することが可能であり，線形計画問題としてシンプレックス法 [8] によって効率的に解くことができる．(1) の最適化問題にスラック変数を加え，制約条件の不等式を除去すると，次式のような整数計画問題の標準形で表すことができる．

$$\text{minimize } E(l,x) = lx^T \qquad (2)$$

$$\text{s.t. } Ax^T = b$$

$$x \geqq 0$$

$$x \in Z$$

整数計画問題の標準形における制約行列 A が完全ユニモジュラ行列であり，ベクトル b が整数ベクトルであるとき，整数計画問題の整数条件を緩和した線形計画問題の解は整数となることが保証されている [12]．完全ユニモジュラ行列とは，行列の任意の正方部分行列の行列式が 0，1 または -1 であるような行列のことである．今回の制約行列 A は [13] により完全ユニモジュラ行列となっている．つまり，元の (1) 式の整数計画問題とそこから整数条件を除去した線形計画問題の解は一致する．よって，整数条件を除去した線形計画問題の解をシンプレックス法によって高速に求めることが可能である．

全データ点の各クラスタへの割り当てを行った後は，K-means 法と同様にセントロイドの更新を行う．(1) 式において変数 x を固定し変数 c に関して目的関数 $E(c,x)$ の最小化を行うが，この最適化問題は K-means 法の場合と同様であるので，クラスタ毎に割り当てられたデータ点から重心を計算する．これらの割り当てのステップとセントロイド更新のステップを繰り返していき，データ点のクラスタへの割り当てが変わらなくなるまで，計算を続ける．

4.2.2 K-medoids 法 (最適化による Balanced K-medoids 法)

K-medoids 法 [7] とは K-means 法と同様にクラスタリングを行う手法であるが，クラスタの代表点の計算方法が異なっている．K-medoids 法ではクラスタの代表点としてメドイドを用いる．メドイドとは、クラスタ内のデータ点の中でそれ以外のデータ点との距離の総和が最小となるようなデータ点のことである．クラスタ C のメドイドは以下のように計算される．

$$\text{medoid}(C) = \arg\min_{x \in C} \sum_{y \in (C - \{x\})} dist(x,y)$$

割り当てのステップや更新のステップはクラスタの代表点の計算方法以外は同様に収束するまで計算する．また，今回は 4.2.1 で説明したように，各クラスタのサイズが均等になる必要がある．最適化による Balanced K-medoids 法では，最適化による Balanced K-means 法と同様に最適化問題を解くことでクラスタへの割り当てを行う．

K-medoids 法（最適化による Balanced K-medoids 法）の特徴は，データ間の距離が定義されていれば，それのみでクラスタリングを行うことができることである．K-means 法におけるセントロイドと異なり，メドイドの計算にはデータ間の距離しか用いないので，各データがベクトルとして表現されている必要がない．提案手法の 2 つ目のワーカー間の距離の定義として「ワーカー間の親和度」を用いるが，この場合ワーカーをベクトルとして表現することができない．そこで，Balanced K-medoids 法を用いることとする．

4.3 異なるクラスタのワーカーの組合せによるチーム編成

最後に異なるクラスタのワーカーによるチーム編成について説明する．4.2 でのクラスタリングによって各クラスタは回答の傾向が近いワーカーで構成される．回答の傾向が異なるワーカーが同じチームに所属するようにチームの編成を行いたいので，異なるクラスタのワーカーを組み合わせてチームを構成する．

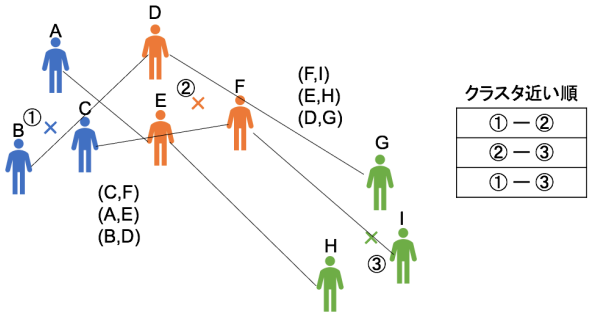

図 2　異なるクラスタのワーカーによるチーム編成の例

例として図 2 のような状況を考える．まず，クラスタ間の距離の小さいクラスタの組からチーム編成を行うために，各クラスタのセントロイド間の距離を計算し，昇順にソートする．図 2 ではクラスタ 1 とクラスタ 2 の距離が最も小さいので，この 2 つのクラスタから処理を開始する．異なるクラスタのワーカーを同じチームに配属するために，クラスタ 1 とクラスタ 2 に所属する各ワーカーの中から同じチームに配属する 1 対 1 のワーカーの組を作る．1 対 1 のワーカーの組を作ることは，クラスタ 1 の各ワーカーをクラスタ 2 の各ワーカーに割り当てる問題と捉えることができるので，提案手法では割り当て問題に対する効率的な解法であるハンガリアンアルゴリズム [9] を用いる．このとき，そのままハンガリアンアルゴリズムを適用すると，距離の総和が最小となるような割り当てを計算してしまう．よって，距離にはあらかじめ −1 をかけておくことで，距離の総和が最大となる割り当てを計算できるようにする必要がある．同様にして，次にクラスタ間の距離が小さいクラスタ 2 とクラスタ 3 に対して，ワー カーのペアを計算する．このとき，クラスタ 1 とクラスタ 2 の割り当て問題で C と F のペアができ，クラスタ 2 とクラスタ 3 の割り当て問題で F と I のペアができるので，(C,F,I) というチームを作ることができる．

5　評 価 実 験

4 で述べた提案手法の性能を評価するために行った実験について説明する．今回は人工データと Amazon Mechanical Turk で収集した爬虫類の画像分類タスクの回答データ，京塚ら [19] が行った犬種の画像分類タスクの実験データを利用した．

5.1　実験データ

5.1.1　人工データ

一つ目の実験データは人工的に生成した回答データである．ワーカーにはある程度の回答の傾向が存在すると仮定し，表 1 の 4 つの混同行列に従って回答するワーカーを複数人用意し，その回答データを生成した．いずれの混同行列も正解の選択肢を選ぶ確率は 0.6 となっているが，不正解の選択肢の中の一つを選ぶ確率も高くなっている．例えば一つ目の混同行列では，正解が 1 のときに不正解の選択肢である 2 を選ぶ確率は 0.25 となっており，他の不正解の選択肢よりも 2 を選ぶ傾向があるということになる．このような確率の偏りを持つ混同行列を 4 種類作成し，生成した回答データを用いて実験を行った．

5.1.2　爬虫類の画像分類タスク

二つ目の実験データは爬虫類の画像分類タスクの回答データである．クラウドソーシングのプラットフォームである Amazon Mechanical Turk を用いて実際にワーカーを募集し，ワーカーには全ての画像に対して回答してもらった．画像には Frilled Agama，Chameleon，Iguana，Komodo Dragon，Gecko，Tuatara，Basilisks，Giant Salamander，Skinks，Newt の 10 種類の中のいずれかの爬虫類の動物が写っており，どれが正しい種類であるかを選択してもらうタスクである．今回の実験では，36 人のワーカーに全 300 枚の画像に対して回答してもらい，各手法のシミュレーションを行った．

36 人のワーカーの正解率の平均は 75.0%となった．また，表 2 は，正解の選択肢に対して，ワーカ 36 人が 300 枚の写真に対して選んだ爬虫類の選択肢の混同行列の平均を表したものである．この混同行列より，ワーカ全体としては Komodo Dragon や Frilled Agama，Chameleon，Giant Salamander は 8 割以上の高い精度で判別可能であるが，Basilisks や Newt の正解率は低い．また，Gecko と Newt の組や Iguana と Tuatara の組は比較的取り違えやすい傾向にあることが言える．また，実験では，300 枚の画像に対する回答データを 200 枚と 100 枚に無作為に分割し，一方をチーム編成を行うための学習データとして扱い，もう一方を実際にそのチーム編成で回答した時の多数決の正解率を測るためのテストデータとして扱う．

5.1.3　犬種の画像分類タスク

Amazon Mechanical Turk において画像データの分類を依頼する実験を行った．画像には Alaskan Malamute，Siberian Husky，Samoyed，German Shepherd，Gray Wolf，Coyote，Dhole の 7 種類の犬種またはイヌ科の動物の犬が写っており，どれが正しいのかを選択肢の中から選んでもらうタスクである．ワーカーには今回は用意した画像データの全てに回答してもらい，その回答データを収集し，各手法のシミュレーションを行った．

56 人のワーカーから回答データを収集し，回答の正解率が 50%を下回る 11 人をスパマーとみなして除去することとした．除去後のワーカーの正解率の平均は 76.6%であった．また，表 3 は，正解の選択肢に対して，ワーカ 56 人が 800 枚の写真に対して選んだ犬種の混同行列を作成したものである．この混同行列より，ワーカ全体としては German Shepherd や Samoyed は 9 割近い高い精度で判別可能であるが，Alaskan Malamute と SiberianHusky の組や Coyote，Dhole，Gray Wolf の組は比較的取り違えやすい傾向にあることが言える．また，今回は 800 枚の画像に対する回答データを半分ずつに分割し，一方をチーム編成を行うための学習データとして扱い，もう一方を実際にそのチーム編成で回答した時の多数決の正解率を測るためのテストデータとして扱う．

5.2　比 較 手 法

提案手法と比較した手法について説明する．一つ目の比較手法はワーカーを無作為に選び出す手法である．全ワーカーの中

表 1　人工データのワーカーの混同行列

		ワーカーの回答				
		1	2	3	4	5
正解	1	0.6	0.25	0.05	0.05	0.05
	2	0.05	0.6	0.25	0.05	0.05
	3	0.05	0.05	0.6	0.25	0.05
	4	0.05	0.05	0.05	0.6	0.25
	5	0.25	0.05	0.05	0.05	0.6

		ワーカーの回答				
		1	2	3	4	5
正解	1	0.6	0.05	0.25	0.05	0.05
	2	0.05	0.6	0.05	0.25	0.05
	3	0.05	0.05	0.6	0.05	0.25
	4	0.25	0.05	0.05	0.6	0.05
	5	0.05	0.25	0.05	0.05	0.6

		ワーカーの回答				
		1	2	3	4	5
正解	1	0.6	0.05	0.05	0.25	0.05
	2	0.05	0.6	0.05	0.05	0.25
	3	0.25	0.05	0.6	0.05	0.05
	4	0.05	0.25	0.05	0.6	0.05
	5	0.05	0.05	0.25	0.05	0.6

		ワーカーの回答				
		1	2	3	4	5
正解	1	0.6	0.05	0.05	0.05	0.25
	2	0.25	0.6	0.05	0.05	0.05
	3	0.05	0.25	0.6	0.05	0.05
	4	0.05	0.05	0.25	0.6	0.05
	5	0.05	0.05	0.05	0.25	0.6

表 2　爬虫類の画像分類タスクにおける全ワーカーの混同行列

	Agama	Basilisks	Chameleon	Gecko	Iguana	Komodo	Newt	Salamander	Skinks	Tuatara
Agama	0.819	0.017	0.029	0.035	0.027	0.006	0.017	0.004	0.009	0.037
Basilisks	0.097	0.583	0.084	0.061	0.032	0.007	0.052	0.003	0.030	0.050
Chameleon	0.033	0.007	0.882	0.006	0.043	0.005	0.007	0.0	0.001	0.016
Gecko	0.005	0.008	0.010	0.664	0.008	0.012	0.161	0.029	0.084	0.018
Iguana	0.015	0.017	0.042	0.015	0.703	0.056	0.006	0.002	0.019	0.125
Komodo	0.001	0.006	0.009	0.003	0.001	0.966	0.003	0.006	0.004	0.001
Newt	0.005	0.016	0.016	0.130	0.018	0.018	0.493	0.195	0.089	0.020
Salamander	0.006	0.002	0.018	0.014	0.004	0.016	0.053	0.857	0.020	0.008
Skinks	0.006	0.009	0.011	0.076	0.017	0.009	0.067	0.029	0.764	0.013
Tuatara	0.039	0.010	0.022	0.022	0.158	0.033	0.008	0.005	0.009	0.693

表 3　犬種の画像分類タスクにおける全ワーカーの混同行列

	Maramute	Husky	Samoyed	Shepherd	Wolf	Coyote	Dhole
Maramute	0.575	0.301	0.037	0.033	0.039	0.009	0.007
Husky	0.214	0.663	0.028	0.031	0.049	0.009	0.006
Samoyed	0.013	0.013	0.936	0.018	0.005	0.007	0.008
Shepherd	0.028	0.015	0.021	0.890	0.016	0.014	0.015
Wolf	0.081	0.074	0.044	0.047	0.567	0.149	0.039
Coyote	0.015	0.027	0.012	0.024	0.170	0.605	0.147
Dhole	0.008	0.007	0.006	0.029	0.099	0.107	0.743

から v 人ずつ無作為に選出し，n/v 個のチームを構成する．この手法を本論文では「無作為編成」手法と呼ぶこととする．二つ目の比較手法は個々のワーカーの正解率が分散するようにチーム編成を行う手法である．まず，全ワーカーを回答データにおける正解率で降順にソートする．そして，正解率が 1 位であるワーカーは 1 つ目のチームに入れ，正解率が 2 位であるワーカーを 2 つ目のチームに入れる．これを繰り返していき，n/v 位のワーカーを n/v 個目のチームに入れた後，次の $n/v+1$ 位のワーカーは 1 位のワーカーが入っている 1 つ目のチームに入れ，$n/v+2$ 位のワーカーは 2 つ目のチームに入れる．これを n 位のワーカーまで繰り返していく．つまり，正解率上位から 1 人ずつ各チームに配分していくことで，個々のワーカーの正解率を分散させ，n/v 個のチームを構成する．この手法を本論文では「分散編成」手法と呼ぶこととする．以上の比較手法の性能と，われわれのこれまでの研究で示した従来手法，提案手法の性能を上記の 3 つの実験データを用いて比較する．

表 4　人工データの実験結果（5 人 × 9 チーム）

手法	正解率	計算時間
提案手法	0.894	0.147s
従来手法	0.890	0.120s
比較手法（無作為編成）	0.870	0.000075s
比較手法（分散編成）	0.872	0.00020s

表 5　人工データの実験結果（3 人 × 15 チーム）

手法	正解率	計算時間
提案手法	0.810	0.132s
従来手法	0.806	0.120s
比較手法（無作為編成）	0.786	0.000055s
比較手法（分散編成）	0.789	0.00019s

表 6　爬虫類画像の実験結果（6 人 × 6 チーム）

手法	正解率	計算時間
提案手法	0.910	0.132s
従来手法	0.905	0.102s
比較手法（無作為編成）	0.883	0.000055s
比較手法（分散編成）	0.885	0.00020s

5.3　実 験 結 果

人工データと犬種の画像分類データについては，45 人のワーカーの回答データを用いて，5 人のチームを 9 個作る場合と 3 人のチームを 15 個作る場合の 2 通りの実験を行なった．爬虫類の画像分類データに関しては，36 人のワーカーの回答データを用いて，6 人のチームを 6 個作る場合と 3 人のチームを 12 個作る場合の 2 通りの実験を行なった．これらの実験データから，特徴ベクトルのユークリッド距離を用いた提案手法とワーカー間の親和度による距離を用いた提案手法，2 種類の比較手法によってチーム編成を行った．各手法によって編成されたチームの多数決の正解率とチーム編成を求める計算時間は表 4，5，6，7，8，9 のようになった．実験結果の正解率と計算時間の数値はいずれも 100 回チーム編成を繰り返したときのそれぞれの平均をとったものである．

表 7　爬虫類画像の実験結果（3 人 × 12 チーム）

手法	正解率	計算時間
提案手法	0.836	0.130s
従来手法	0.830	0.098s
比較手法（無作為編成）	0.820	0.000052s
比較手法（分散編成）	0.822	0.00023s

表 8　犬種画像の実験結果（5 人 × 9 チーム）

手法	正解率	計算時間
提案手法	0.912	0.145s
従来手法	0.905	0.123s
比較手法（無作為編成）	0.895	0.000082s
比較手法（分散編成）	0.889	0.00045s

表 9　犬種画像の実験結果（3 人 × 15 チーム）

手法	正解率	計算時間
提案手法	0.865	0.152s
従来手法	0.859	0.125s
比較手法（無作為編成）	0.852	0.000081s
比較手法（分散編成）	0.850	0.00043s

5.4　実験結果の考察

5.4.1　多数決の精度と計算時間

実験の結果より，いずれのいずれのデータにおいても，提案手法が比較手法や従来手法よりも正解率の高いチームを編成することができた．提案手法と無作為編成手法の 100 回分の正解率のデータについて有意水準 5%で検定を行ったところ，3 種類のいずれの実験データのにおいても，平均値の差は有意であるということが分かった．また，無作為選出の手法ではチーム編成を行う毎に多数決の精度が大きく変動し，犬種画像分類の 100 回分の回答データでの正解率の分散が 0.0042 であった．しかし，提案手法では，100 回分の実験データの分散は 0.00028 であり，安定して多数決の精度が高いチーム編成を行うことができていた．以上の結果から，回答の傾向が異なるようなワーカーを同じチームに配属することは，チーム編成を行う上で有用であるということがわかった．また，本研究で導入した親和度が大きいワーカーを同じチームに配属することは，チーム編成を行う上で一定程度有用であるということがわかった．計算時間についても，実験の結果より，提案手法はいずれも現実的な計算時間でチーム編成が可能であるということがわかった．

5.4.2　手法の問題点

提案手法を最善のチーム編成と比較して分析する．20 人以上の規模ではチーム編成の全組み合わせの多数決の正解率の期待値を計算できないので，今回は 16 人のワーカーを 4 個の 4 人チームに分ける問題を考える．犬種分類タスクの回答データを用いて実験を行なった．無策編成手法でチーム編成を行なった場合，多数決の正解率の平均は 78.6%となり，ユークリッド距離を用いた提案手法でチーム編成を行なった場合，多数決の正解率の平均は 79.8%となった．しかし，最善のチーム編成では多数決の正解率の平均は 80.7%となった．つまり，提案手法にはまだ伸びしろがあるということである．なぜ，提案手法は最善手法に比べて多数決の正解率が下がっているのかを分析した．

ある 16 人のワーカー A〜P を最適化による Balanced K-means 法によってクラスタリングした結果の一例はこのようになった．

- クラスタ 1：[D, G, K, M]
- クラスタ 2：[B,E,H,O]
- クラスタ 3：[F,I,L,N]
- クラスタ 4：[A,C,J,P]

また，最善のチーム編成は次のようになっている．

- 最善チーム:(D, G, L, P),(A,B,M,N),(C,F,H,I),(E,J,K,O)

このとき，最善チームで同じチームに入っているワーカーは異なるクラスタに入っていることが望ましい．なぜなら，クラスタリングの後に，異なるチームのワーカーを組み合わせることでチームを編成するからである．例えば，チーム (A,B,M,N) の各ワーカーは異なるクラスタに分散しているが，ワーカー D と G は同じクラスタ 1 に入ってしまっている．つまり，ある程度はクラスタリングの効果が発揮されているが，一部分ではうまく機能していないところもある．ワーカー D と G は同じクラスタ 1 に入る状況になってしまうと，多数決の精度の高いチームを編成することができなくなってしまう可能性がある．ワーカー D と G が同じクラスタに入ってしまった原因として考えられるものは二つある．一つ目は最適化による Balanced K-means 法の制約条件である．一般的なクラスタリングとは異なり，最適化による Balanced K-means 法はクラスタのサイズを均等にすることを強制するので，場合によっては本来は入るべきでないワーカーまで同じクラスタに属してしまう可能性がある．これはサイズ均等クラスタリング以外のクラスタリング手法でもチーム編成ができるように提案手法を改善する必要がある．二つ目の原因はワーカー間のユークリッド距離や親和度という指標が完全なものではないということである．今後はワーカー間の距離の指標としてどのような点が重要であるのかを分析し，ユークリッド距離や親和度よりもより良い指標を考える必要がある．

また，ユークリッド距離を用いた従来手法も今回の提案手法も，ともに多数決の正解率が高いワーカー同士を異なるチームに分散し，一方のワーカーが不正解のときにもう一方のワーカーが正解するというような互いを助け合えるようなワーカー同士は同じチームに配属できるような距離の指標になっている．しかし，注意しなければならないのは，正解率が 0.5 を超えるような問題に対しては，多数決の正解率が高いワーカーを分散させることは多数決にとって有益であるが，正解率が極端に低いような問題に対しては，帰ってどのチームも正解できなくなってしまう．難易度が高い問題が混在しているようなタスクについては，今と同じ手法を用いても多数決の精度は上がらないと考えられる．よって，難易度が高い問題が存在しているタスクのために，問題の正解率ごとに指標に重みをつけるなど今後の改善が必要である．

6 おわりに

本研究は，全ワーカーの中から多数決の精度が高くなるようなワーカーのチーム分けを行うことを目的に取り組んだ．提案手法では，ワーカー同士の親和度という概念を導入する．二人のワーカー組に対し，一方のワーカーと平均的なワーカーで形成したチームで多数決をとる場合の正解率と，もう一方のワーカーも含めたチームで多数決をとる場合の正解率の差で定義される．次に，クラスタのサイズが均等となるようなクラスタリングアルゴリズムを用いて，ワーカーをクラスタリングする．その結果，各クラスタは親和度の低いワーカーによって構成される．親和度の大きいワーカー同士を同じチームに入れることで多数決の正解率の向上を図りたいので，異なるクラスタのワーカーを組み合わせることでチームを編成する．距離の近いクラスタのペアから順番にハンガリアンアルゴリズムを用いて同じチームに配属させるワーカーのペアを作り，チームを形成する．実験では，人工データと実際のクラウドソーシングのプラットフォーム上で集めたデータの双方で評価を行なった．比較手法に対して，提案手法ではより良いチーム編成を行うことができた．また，現実的な計算時間でチーム編成を行うことにも成功した．

今後の課題としては，ワーカー間の距離定義の方法を改善することが考えられる．まず，本研究ではワーカーの特徴ベクトルはワーカーのテスト問題の回答データから定義したが，クラウドソーソングにおいてはワーカーが必ずしも全てのテスト問題に取り組むとは限らないので，完全な回答データを得られない可能性がある．そこで，ワーカーの混同行列を利用するなど現実的な問題に対応できる手法に改善する必要がある．また，今回はサイズ均等クラスタリングの手法を用いてクラスタリングを行なっているが，サイズが均等にならない一般的なクラスタリングの手法を利用できるような手法も考える必要がある．最適化問題の手法に関しても，今回は計算時間が膨大となるような定式化になってしまっていたが，目的関数や制約条件を工夫するなどして，現実的な計算時間で解くことができる問題を考案することが今後の課題である．実験については，画像分類の回答データによる実験しかまだ行うことができていないので，他のタスクでは正解率がどのように変化するかなどを調べることで，提案手法と比較手法の差はタスク固有のものかどうかを検証する必要がある．

7 謝　辞

本研究は，JST CREST (JPMJCR16E3)，JSPS 科研費 18H03245 の支援を受けたものである．

文　献

[1] Amazon: Amazon Mechanical Turk, https://www.mturk.com/.
[2] Arthur, D. and Vassilvitskii, S.: K-Means++: The Advantages of Careful Seeding, Vol. 8, pp. 1027–1035 (2007).
[3] Chris, C.-B. and Mark, D.: Creating Speech and Language Data with Amazon's Mechanical Turk，Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk，pp. 1–12 (2010).
[4] Dawid, A. P. and Skene, A. M.: 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).
[5] Joglekar, M., Garcia-Molina, H. and Parameswaran, A.: Evaluating the crowd with confidence, *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, pp. 686–694 (2013).
[6] Joglekar, M., Garcia-Molina, H. and Parameswaran, A.: Comprehensive and reliable crowd assessment algorithms, *2015 IEEE 31st International Conference on Data Engineering*, IEEE, pp. 195–206 (2015).
[7] Kaufman, L. and Rousseeuw, P. J.: *Finding groups in data: an introduction to cluster analysis*, Vol. 344, John Wiley & Sons (2009).
[8] Koberstein, A.: Progress in the dual simplex algorithm for solving large scale LP problems: Techniques for a fast and stable implementation, *Computational Optimization and Applications*, Vol. 41, pp. 185–204 (2008).
[9] Kuhn, H. W.: The Hungarian method for the assignment problem, *Naval research logistics quarterly*, Vol. 2, No. 1-2, pp. 83–97 (1955).
[10] Kuncheva, L., Whitaker, C., Shipp, C. and Duin, R.: Limits on the majority vote accuracy in classier fusion, *Formal Pattern Analysis & Applications*, Vol. 6, pp. 22–31 (2003).
[11] Land, A. and Doig, A.: *An Automatic Method for Solving Discrete Programming Problems*, Vol. 28, pp. 105–132 (2010).
[12] Papadimitriou, C. and Steiglitz, K.: *Combinatorial Optimization: Algorithms and Complexity*, Vol. 32 (1982).
[13] Schrijver, A.: *Theory of linear and integer programming*, John Wiley & Sons (1998).
[14] Tang, W., Yang, Y., Zeng, L. and Zhan, Y.: Optimizing MSE for Clustering with Balanced Size Constraints, *Symmetry*, Vol. 11, p. 338 (2019).
[15] Welinder, P., Branson, S., Perona, P. and Belongie, S. J.: The multidimensional wisdom of crowds, *Advances in neural information processing systems*, pp. 2424–2432 (2010).
[16] Whitehill, J., fan Wu, T., Bergsma, J., Movellan, J. R. and Ruvolo, P. L.: Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise, *Advances in Neural Information Processing Systems 22* (Bengio, Y., Schuurmans, D., Lafferty, J. D., Williams, C. K. I. and Culotta, A., eds.), Curran Associates, Inc., pp. 2035–2043 (2009).
[17] Wu, T., Chen, L., Hui, P., Zhang, C. J. and Li, W.: Hear the whole story: Towards the diversity of opinion in crowdsourcing markets, *Proceedings of the VLDB Endowment*, Vol. 8, No. 5, pp. 485–496 (2015).
[18] 松田浩幸，田島敬史：クラウドソーシングにおける多数決の精度向上のためのワーカー組合せの選出手法，第 12 回データ工学と情報マネジメントに関するフォーラム (2020).
[19] 京塚萌々，田島敬史：アイテムへのワーカ逐次割当てと各ワーカの複数ラベル付与によるマルチクラス分類タスク精度向上手法，第 12 回データ工学と情報マネジメントに関するフォーラム (2020).