DEIM2020 I5-1(day2 p69)

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

京塚　萌々†　　田島　敬史†

† 京都大学大学院情報学研究科　〒 606-8211 京都府京都市左京区吉田本町
E-mail: †kyozuka@dl.soc.i.kyoto-u.ac.jp, ††tajima@i.kyoto-u.ac.jp

あらまし　本研究では，人手でアイテムを複数クラスのうち一つへ分類するタスクの精度向上手法を提案する．このようなタスクでは，一つのアイテムの分類を複数のワーカに依頼し，各ワーカに一つのクラスを選ばせ，これらの多数決を取る手法が広く用いられている．本研究では，各アイテムに対して精度の高いラベルをつけると期待できるワーカからラベルを集めるために，各アイテムに一人ずつワーカを割り当てた上で，次のワーカを割り当てる際にはこれまでに割り当てたワーカの結果に基づいてそのアイテムに最も精度が高いラベルをつけられると期待できるワーカを逐次的に割り当てる．また，タスクやデータによっては，ワーカがアイテムに適したクラスをただ一つ選択することが困難な場合がある．その場合，多数決を用いるのであれば，必ずしも一つに絞らせる必要はなく，むしろ複数の候補を選択させた方がより多くの情報を得られる可能性がある．そこで，本研究の提案手法では，各ワーカにより可能性が高い順に複数の候補クラスを選択することを許容して分類を行わせる．本研究では，これらのアプローチによって，最終的に多数決で推定されるラベルの精度の向上を図る．

キーワード　ワーカ割当て，タスク割当て，マルチクラス分類

1 はじめに

近年，クラウドソーシングサービスを用いて不特定多数のワーカにオンラインで作業を依頼することで安価に大量の成果物を得ることが一般的になった．例えば，大量の画像に写っている被写体の分類をクラウドワーカに依頼することで，専門家に依頼するより短時間で安価に成果物が取得できるようになった．しかし，クラウドソーシングサービスで作業を引き受ける不特定多数のワーカは専門家に比べ能力不足であったり，報酬目当てで意図的に手抜きをしたりすることが想定される．そのため，ワーカから得られたラベルの精度を向上することはクラウドソーシング研究における重要なテーマである．

本研究では，人手でアイテムを複数クラスのうち一つに分類するタスク（マルチクラス分類タスク）の精度を向上するためのアプローチを二種類提案する．

一つ目のアプローチは，各アイテムにワーカを逐次的に割り当てることである．マルチクラス分類タスクでは，一つのアイテムを分類するタスクを複数のワーカに割り当て，各ワーカに適合する一つのクラスを選ばせた後にこれらの多数決を取る手法がよく用いられる．このとき，ワーカをランダムに選択するのでなく，各アイテムに関して質の良いラベルをつけられると思われるワーカを選択すれば，より精度の良いラベルが得られる可能性が高くなると考えられる．そこで，各アイテムに一人ずつワーカを割り当てた上で，二人目以降のワーカを割り当てる際には，それまでに割り当てたワーカが選択したクラスに基づいてそのアイテムを分類する際に最も精度が向上することを期待できるワーカを割り当てる．このアプローチによって，後に割り当てられたワーカほどそのアイテムに対して精度の良いラベルを選択できることが期待できる．

二つ目のアプローチとして，ワーカが適合ラベルをただ一つに絞りきれない場合に，適合すると思われる順に複数のラベルを選択することを許容する．マルチクラス分類タスクにおいて，タスクやデータによっては，ワーカがアイテムに適したクラスをただ一つ選択することが困難な場合があると考えられる．例えば，大量の犬の写真を犬種ごとに分類するタスクの場合，犬に詳しくないワーカにとっては写真に写る犬の犬種をただ一つ決定することは難しいと考えられる．このとき，最終的な正解ラベルの推定に多数決を用いるのであれば，各ワーカがアイテムに適合するクラスを必ずしも一つだけに絞る必要はないと考えられる．むしろ，適合するクラスの候補となりうる複数のクラスを選択させた方がそのアイテムに関して得られる情報は多いと考えられる．

本論文では，これらのアプローチをそれぞれ適用し，最終的に推定されるラベルの精度を向上する手法を提案した．さらに，Amazon Mechanical Turk (MTurk)[1]を利用して収集したデータに提案手法を適用するシミュレーション実験を行い，提案手法の有用性を評価した．

2 関連研究

2.1 品質管理

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ラベル

1：https://www.mturk.com

を推定するための最も単純なアプローチは，同じアイテムに対して複数のワーカーからラベルを得て多数決を取ることであるが，単純な多数決では真面目にタスクをこなすワーカーと手抜きをするワーカーの重みが等しくなってしまうので，正答率が低いワーカーに対して重みを小さくしたり，割り当てるタスクを減らしたりするような手法が提案されている．タスクにどのようにワーカーを割り当てるかという問題に関する研究は数多く，ハンガリアンアルゴリズムはその古典的なものである [1], [2]．ハンガリアンアルゴリズムではワーカーとタスクの間に定義されるコスト行列が与えられたとき，ワーカー 1 人にタスクを 1 件ずつ割り当てる最適な組み合わせを求めることができる．現在では線形計画問題によるアプローチが行われており，各ワーカーがこなせる作業量などの制約も考慮し，ワーカー全体がこなすタスク数を最小化したり，品質を最大化したりした上で真のラベルを推定する問題に帰着する手法が提案されている [3], [4]．

また，統計学の分野では Dawid ら [5] が EM アルゴリズム [6] を利用し，ワーカーの能力のパラメーターと真のラベルの推定値を交互に更新することで真のラベルを推定する手法を提案しており，近年のクラウドソーシングにおける真のラベル推定の研究の基本となっている．

Oyama ら [7] は，ワーカーから回答を得ると同時に自分の回答が正しいと思うか申告させた値（確信度）を利用して真のラベルを推定する手法を提案している．ここでは， Dawid らのアルゴリズムに確信度を確率変数として追加し，拡張を行っている．

本研究では各ワーカーがうまく判定できるアイテムのラベルの組み合わせを利用してワーカー割り当てを行うことで精度を向上することを意図している．クラウドソーシングでは，精度が低いワーカーが混じっていてもそのノイズを取り除いてタスクの品質を保つために多数決を用いるが，多数決の結果から外れている頻度が高いワーカーはスパマーであるとして除去される．しかし，ワーカーによっては多段階の判定において常に一段階低いまたは高い回答をしていたり，問題の読み違いにより常に反対の回答をしたりというように，不正をしているわけではなく回答にバイアスがかかっていることがありうる．そこで，Ipeirotis [8] らは，精度が低いワーカーをただ排除するのではなく，ワーカーの特性を生かした品質管理の手法を提案した．彼らは，ワーカが与えたラベルをワーカの誤り率を反映した確率分布を表すソフトラベルに変換し，ソフトラベルを利用してワーカが誤分類を行った際のコストの期待値を計算することで，より正確にワーカの能力を推定した．

2.2 マルチクラス分類

クラウドソーシングを利用してマルチクラス分類タスクを行う研究について述べる．

Vempaty [9] らは，マルチクラス分類タスクを多数の二値分類タスクに分解し，その結果を 0/1 の列で表現した上でクラスにデコードする手法を提案した．

また，Duan ら [10] は，ラベリングタスクをいくつかのサブタスクに分割して二層の階層タスクを構成するというアプローチを考案した．さらに，それぞれの階層のサブタスクに適したワーカーを割り当てるグリーディーアルゴリズムを提案し，適用することで精度が向上したと報告している．Duan らの手法は，上位階層のタスクにおいてデータを複数のラベルのグループに割り当てることになるところが本研究と似ていると言える．しかしながら，Duan らの手法では全てのデータが固定された階層に基づいて分類されるため，上位階層での判定誤りを下位階層で修正できないのに対し，本研究ではデータごとにラベルのグループが変わりうるため，以前にデータに割り当てられたワーカの判断が誤っていてもそれ以降のワーカによって修正できる可能性があるという利点がある．

マルチクラス分類タスクに関する以上の研究は，本来のタスクをサブタスクに分解することによって精度の向上を意図したものであり，タスク自体は変えずに回答方法を複数回答可とする本研究のアプローチとは異なる．

2.3 マルチラベルタスク

クラウドソーシングにおいて，データに複数のラベルを付与するタスク（マルチラベルタスク）に関する研究も行われている．

Nowak ら [11] は画像データに複数のラベル付けを行うタスクを Amazon Mechanical Turk(www.mturk.com) に依頼し，非専門家である多数のワーカーから得たラベルを多数決や精度を用いて統合すると専門家によるアノテーションに匹敵する品質になると結論づけた．

また，Kanehira ら [12] は，複数のワーカーにラベル付与を依頼し複数あるいは全てのワーカーに共通する回答を真のラベルとみなして得たマルチラベルのデータセットに関して，付与されたラベルは信頼できるが，データに本来付与されるべきであるのに付与されていないラベルが存在する可能性があるという性質を指摘し，そのような不完全なマルチラベル学習データから識別器を学習する研究を行った．このように，クラウドソーシングにおけるマルチラベル分類タスクに関しての研究はなされているが，本研究のようにマルチクラス分類タスクに関してワーカーからマルチラベルを得るアプローチはなされていない．

2.4 ワーカの逐次選択

データにラベルを付与するワーカを逐次的に選択する研究はいくつか行われているが，いずれもワーカ割り当てやワーカの誤分類にコストを定義したり予算を定義したりすることでラベルを追加取得するかどうかを動的に判定するものであり，あらかじめアイテムごとに割り当てるワーカの人数を決定した上でどのワーカからラベルを取得すべきか判定する本研究とは異なる．

Sheng ら [13] は各アイテムのラベルを繰り返し取得することの有効性について研究し，すべてのアイテムに繰り返しラベリングを行うことが必ずしも精度の向上につながるわけではなく，ワーカから得るラベルの不確実性を考慮して追加でラベルを得

るべきアイテムを選択した上で追加ラベルを得ることが効果的であると結論づけた．

Gao ら [14] は，真のラベルを推定できたときの利益や誤ったときのコスト，ワーカからラベルを得るコストを定義した上で，アイテムに対し新たにラベルを得たときの利益の期待値を計算してラベルを得るかどうか判定する手法を二つ提案した．

Li ら [15] は，予算が限られている場合に，アイテムにワーカを割り当てるコストを予算の範囲内に収めつつ，あらかじめ定めた品質要件を達成するアイテム数が最大になるようにワーカを割り当てるべきアイテムを選択する手法を提案した．

これらはワーカを追加で得るべきアイテムを選択するための研究であり，アイテムにラベルを追加すべきワーカを選択する本研究とは設定自体が異なるものである．

3 提案手法

本研究の提案手法について述べる．

この研究の目的は，アイテムを複数のクラスのうち適合する一つに分類するタスクにおいて，一つのアイテムにつき複数人のワーカを割り当てて得たラベルの多数決を取って推定する正解ラベルの精度を向上することである．本研究では，あらかじめ正解ラベルのわかっているアイテムがあり，アイテムの正解ラベルの分布（事前分布）が事前に分かっているものとする．また，クラス集合を C とする．

本研究では次の二つのアイデアに基づいた手法を提案する．

- アイテムに割り当てるワーカを逐次選択する
- ワーカが複数のラベルを選択することを許容する

（本研究ではラベルを適合していると思われる順にランキングさせ，2位ラベルまでを利用する）

提案手法の大まかな流れは次のようになる．

（1） 全てのワーカが正解が既知のアイテムに与えたラベルから混同行列を作成する

（2） 正解が未知の各アイテムについてあらかじめ決めた人数のワーカを割り当ててラベルを得る

（3） 各アイテムに対して得たラベルから真のラベルを推定する

それぞれの段階に提案したアイデアを適用した手法について述べる．

3.1 混同行列の作成

まず各ワーカが各カテゴリのアイテムに付与するラベルの精度を測定するために，正解が既知のアイテムについてマルチラベル分類を行うタスクを投稿する．この結果を利用して各ワーカの混同行列を作成するが，本研究ではワーカが一つのアイテムに対してラベルを複数選択することを許容する場合があるので，混同行列をラベル複数選択に対応させる場合とさせない場合の両方について述べる．

3.1.1 複数ラベル選択を許容しない場合

複数ラベル選択を許容しない場合，各アイテムに対し一人のワーカがつけられるラベルは最も適合していると思われる一種類のみである．したがって，作成する混同行列は従来のものと同じである．この混同行列を従来の混同行列と呼び，ワーカ j の従来の混同行列を $\pi_O(j)$ と表す．分類するクラス数が M クラスの場合，ワーカ j の混同行列 $\pi_O(j)$ は $M \times M$ 行列であり，$\pi_O(j)_{c,c'}$ 成分は正解ラベル $c \in C$ のアイテムのうちワーカ j が $c' \in C$ のラベルをつけた割合ということになる．

3.1.2 複数ラベル選択を許容する場合

従来の混同行列は正解ラベルとワーカの選択ラベルが一対一対応するが，ワーカが一つのアイテムに対し複数のラベルを選択することを許容する場合，必ずしも正解ラベルとワーカの選択ラベルは一対一対応しない．そのため，本研究では複数ラベル選択に対応した二種類の混同行列を提案する．

- ラベルを複数選択した場合の組み合わせを含めて選択ラベルの列を拡張した混同行列（拡張した混同行列）

混同行列の列はワーカが選択したラベルを表すが，そこに複数選択した際の組み合わせの列を加える．M クラス分類を行うとすると，M クラス中2位までラベルを選択する組み合わせは $M(M-1)$ 通りである．したがって，ワーカが選択するラベルの列数はラベルを一つだけ選択した場合も含めて $M + M(M-1) = M^2$ である．この混同行列はワーカが選択するラベルの組み合わせをすべて網羅しているので，正解ラベルとワーカの選択ラベルの一対一対応がとれる．この混同行列を拡張した混同行列と呼び，ワーカ j の拡張した混同行列を $\pi_E(j)$ と表す．$\pi_E(j)_{c,(c',c'')}$ はワーカ j が正解ラベル $c \in C$ のアイテムに対し最も適合したラベルとして $c' \in C$，2番目に適合したラベルとして $c'' \in C$ を選択した割合ということになる．

- ラベルを複数選択した場合，合計が1になるように各順位の選択ラベルに重みを加える混同行列（重み付き混同行列）

ワーカがラベルを複数選択したとき，一つだけ選択しているときより各ラベルへの確信度合いが低く，また1位ラベルより2位ラベルの方が確信度合いが低いと考えられる．そこで，1位ラベルと2位ラベルの重みの合計がラベルを一つだけ選択したときと同じになるように1位ラベルと2位ラベルにつける重みを定め，M クラス分類に対して $M \times M$ の混同行列を作ることを考える．この混同行列を重み付き混同行列と呼び，ワーカ j の重み付き混同行列を $\pi_W(j)$ と表す．

正解ラベル c のアイテムに対してワーカが1位として c'，2位として c'' を選択した際，混同行列の要素 $\pi_{c,c'}$ に $\delta(0.5 \leqq \delta < 1)$，$\pi_{c,c''}$ に $1-\delta$ を加える．正解が既知のアイテムにワーカ j がつけたすべてのラベルについてこの処理を行ったあと，混同行列の各行の和が1になるように正規化を行う．

この混同行列を用いる場合，正解ラベル c のアイテムに対してワーカ j が1位として c'，2位として c'' を選択する条件付き確率を $P_j(c'|c)\delta + P_j(c''|c)(1-\delta)$ として計算する．

拡張した混同行列にはワーカが複数ラベルを選択したときも一つだけ選択したときと全く同様に扱うことができるという長所があるが，混同行列のサイズが大きくなってしまうため計算時間が長くなる，行列が疎になりやすいという短所がある．

重み付き混同行列は，従来の混同行列と同じサイズで複数ラ

ベル選択に対応することを意図しているが，拡張した混同行列と異なりワーカが学習データに対して選択したラベルの組み合わせの情報が失われる．そこで，重み付き混同行列に関しては，拡張した混同行列で高精度の分類を実現した手法で実験を行うことで，ワーカの選択ラベルの組み合わせが失われても精度が維持できるか確認する．

3.1.3 混同行列のスムージング

正解が既知のアイテムに対してワーカから得たラベルのみで混同行列を作成すると，正解が未知のアイテムに対してワーカが学習データにないラベルをつけたときに条件付き確率が 0 になってしまい，精度が下がる原因となる．特に，ワーカが複数のラベルを選択することを許容し，ワーカの選択するラベルの組み合わせも含めて混同行列を拡張した場合，混同行列のサイズが大きいために要素が 0 の成分が多くなる可能性が高い，そこで，スムージングの方法として次の二つを提案する．

- 全要素に 1 を足す（add-one スムージング）

正解ラベルとワーカの選択ラベルのすべての組み合わせが一度は出現するとし，出現回数に 1 回追加した混同行列を作成する．混同行列の列数（選択ラベルの組み合わせの数）が M'，クラス c に属すアイテムの数が n_c，クラス c に属すアイテムのうちワーカ j がクラス c' に分類したものの数を $n^j_{(c,c')}$ とすると，ワーカ j の混同行列の成分 $\pi^j(c,c')$ は次のように求められる．

$$\pi^j(c,c') = \frac{n^j_{(c,c')} + 1}{n_c + M'}$$

- 全ワーカのつけたラベルから作った混同行列と重み付き和をとる（グローバルスムージング）

正解が既知のアイテムに関して，ワーカ全員の選択ラベルからも混同行列（グローバルな混同行列）を作成する．ワーカ個人の混同行列の重みを $\gamma(0.5 < \gamma < 1)$，全員の混同行列の重みを $1-\gamma$ として重み付き和をとったものをワーカ個人の混同行列として用いる．

それぞれのワーカの選択ラベルにはワーカ個人に由来するバイアスがかかっていると考えられるが，同じアイテムに対して選択するラベルの傾向はある程度似通うと考えられる．そこで，ワーカ全員のグローバルな混同行列を作成してワーカ全員の選択ラベルの傾向をつかみ，ワーカ個人の混同行列と重ねることで，正解が既知のアイテムへの付与ラベルからは観測できなかった組み合わせの出現確率を予測することを試みる．ワーカ全員のラベルから作った混同行列を π^G，スムージングを行う前のワーカ j の混同行列を π^j_b とすると，スムージング後のワーカ j の混同行列の成分 $\pi^j(c,c')$ は次のように求められる．

$$\pi^j(c,c') = \gamma \times \pi^j_b(c,c') + (1-\gamma) \times \pi^G(c,c')$$

3.2 ワーカ割り当て

各アイテムに割り当てるワーカをあらかじめ決めた人数ずつ選択して割り当て，ラベルを得る．

3.2.1 アイテムの真のラベルの推定分布

アイテムがクラス $c \in C$ に所属する事前確率を $P(c)$，正解ラベルが c のアイテムにワーカ j がラベル c' をつける確率を $P_j(c'|c)$ とする（各混同行列 $\pi(j)_{c,c'}$ 成分で表される）．アイテム i に対してラベルを n 回得たときアイテム i がクラス c に所属する確率を $P_i(c)_n$ と表す．ラベルを得ていないとき，アイテム i がクラス c に所属する確率 $P_i(c)_0$ は事前確率 $P(c)$ と等しい．

アイテム i に対して j 人目のワーカ j がクラス c_{ij} を選択したとき，アイテム i がクラス c に所属する確率 $P_i(c)_j$ は，$n-1$ 人のワーカがラベルをつけたときにアイテム i がクラス c に所属する確率 $P_i(c)_{j-1}$ を用いてベイズの定理より次のように求められる．

$$P_i(c)_0 = P(c)$$
$$P_i(c)_j = P_i(c)_{j-1} \times \frac{P_j(c_{ij}|c)}{\sum_{c' \in C} P_j(c_{in}|c') P_i(c')_{n-1}}$$

この計算によって，ワーカから新しいラベルを得るたびに各アイテムが所属するクラスの推定分布を更新する．

3.2.2 割り当てるワーカの選択

本研究では各アイテムにすでに付与されたラベルを利用してワーカの逐次選択を行う場合があるので，逐次選択を行う場合と行わない場合の両方についてワーカの割り当て手法を説明する．各アイテムに割り当てるワーカを K 人と定めた上で，ワーカ一人に割り当てられるアイテムの数がほぼ同数となるように割り当てを行う．アイテム数を N，ワーカの人数を W 人とすると，ワーカ一人あたり $N \times K \div W$ 個のアイテムが割り当てられることになる．

ワーカの逐次選択を行わない場合は各アイテムに割り当てるワーカをランダムに選択する（ランダム選択）．

ワーカの逐次選択を行う場合，アイテムの真のラベルの推定分布とワーカの混同行列を利用してワーカを選択する．このときの手法を三つ提案する．

- 各アイテムについて，正解を選択する確率が最大のワーカを選択する（個人確率）

各アイテムに割り当てる一人目のワーカはそれぞれランダムに選択する．二人目以降はワーカが付与するラベルが正解である確率を計算し，最大となるワーカを選択する．ワーカ個人の回答が正解である確率を用いて割り当てるワーカを選択するので，この手法を個人確率手法と呼ぶ．現時点でのアイテム i の真のラベルが $c \in C$ である確率を $P_i(c)$，正解ラベルが c のアイテムをワーカ j がクラス $c' \in C$ と分類する確率を $P_j(c'|c) = \pi(j)_{c,c'}$ とすると，アイテム i に対してワーカ j が選択するクラス c_{ij} が正解である確率 $P(c_{ij} = t_i)$ は次のように計算できる．

$$P(c_{ij} = t_i) = \sum_{c \in C} P_i(c) P_j(c'|c)$$

アイテム i にまだラベルを付与していないワーカ全員について個人精度の期待値を計算し，最大となるワーカからアイテム i のラベルを得る．

- これまでに付与されたラベルと選択するラベルの多数決ラベルが正解である確率が最大のワーカを選択する（多数決確率）

各アイテムの一人目のワーカはそれぞれランダムに選択する．二人目以降は付与するラベルとこれまでにそのアイテムに付与されたラベルを合わせて多数決を取ったラベルが正解である確率を計算し，最大となるワーカを選択する．多数決ラベルが正解である確率を用いて割り当てるワーカを選択するので，この手法を多数決確率手法と呼ぶ．アイテム i に対してワーカ j がクラス $c_{ij} \in C$ を選択したとき，すでにアイテム i に付与されているラベルと合わせて最も選択された数が多いクラス $\hat{c_{ij}}$ が決まる．多数決クラス $\hat{c_{ij}}$ が正解である確率 $P(\hat{c_{ij}} = t_i)$ は次のように計算できる．

$$E(P(\hat{c_{ij}} = t_i)) = \sum_{c \in C} P_i(\hat{c_{ij}}) P_j(c' | \hat{c_{ij}})$$

アイテム i にまだラベルを付与していないワーカ全員について多数決クラスの精度の期待値を計算し，最大となるワーカからアイテム i のラベルを得る．

- これまでに付与されたラベルと選択するラベルによって最尤推定されるラベルが正解である確率が最大のワーカを選択する（最尤推定確率）

各アイテムの一人目のワーカはそれぞれランダムに選択する．二人目以降は付与するラベルとこれまでにそのアイテムに付与されたラベルを合わせて最尤推定されるラベルの精度の期待値を計算し，最大となるワーカを選択する．ワーカ $1, ..., j-1$ がラベル $c_1, ..., c_{ij-1}$ を付与したアイテム i に対してワーカ j がクラス $c \in C$ を選択したとき（$c_{ij} = c$），最尤推定されるアイテム i の真のクラス $\hat{t_i}(c)$ とその精度の期待値 $P(\hat{t_i}(c_{ij}) = t_i)$ は次のように計算できる．

$$\hat{t_i}(c_{ij}) = \arg\max_{c \in C} \prod_{k=1}^{j} P(c_{ik} | t_i = c)$$

$$P(\hat{t_i}(c_{ij}) = t_i) = \sum_{c \in C} P_i(\hat{t_i}(c)) P_j(c' | \hat{t_i}(c))$$

アイテム i にまだラベルを付与していないワーカ全員について最尤推定ラベルが正解である確率を計算し，最大となるワーカからアイテム i のラベルを得る．

各アイテムについて一人目のワーカをランダムに選択したあと，あらかじめ決めた人数になるまで上述の尺度を用いてワーカを順に選択してラベルを得る．

3.3 ラベルの推定

正解が未知のアイテムのラベルを集め終えたら，各アイテムに関して集まっているラベルを使用して最尤推定を行う．アイテム i について，ワーカ $1, ..., j$ がラベル $c_{i1}, ..., c_{ij}$ をつけたときアイテム i の真のラベル $\hat{t_i}$ は次のように推定される．

$$\hat{t_i} = \arg\max_{c \in C} \prod_{k=1}^{j} P_k(c_{ik} | c)$$

3.4 比 較 手 法

本研究では二つのアイデアに基づいた手法を提案している．

- アイテムに割り当てるワーカを逐次選択する
- ワーカが複数のラベルを選択することを許容する

これらのアイデアを組み合わせると，大きく分けて次の三種類になる．

- ワーカ逐次選択のみ用いる手法
- 複数ラベル選択のみ用いる手法
- ワーカ逐次選択と複数ラベル選択を組み合わせた手法

混同行列の作成，ワーカの選択，ラベル推定の三段階でそれぞれ提案した手法の組み合わせのうち，ラベルの複数選択を許容せずに従来の混同行列を用い，ランダムに選択したワーカのラベルから最尤推定を行う手法は，本研究で提案しているアイデアを用いていない手法と言える．本研究ではこの手法を比較手法として扱う．

4 実　　験

Amazon Mechanical Turk で画像データの分類を依頼する実験を行った．複数の提案手法の比較を行うために，ワーカにすべての画像データにラベルをつけてもらったデータを収集し，そのデータを用いて各手法のシミュレーションを行った．

まず Amazon Mechanical Turk (MTurk) にマルチクラス分類タスクを投稿した．イヌ属の動物 7 種類のうち 1 種が写った写真 800 枚について，写っている動物の種類をラベリングするタスクを投稿し，ラベルを集めた．このとき，最も適合していると思われるクラスを必須回答とした上で，適合していると思われる順に複数のクラスを回答することを許容した（七択問題であるので，最大 7 位まで回答することができる）．

提案手法のうちいくつかは複数回答された際の 1 位の重み δ や全ワーカの混同行列を利用してスムージングした際の個人の混同行列の重み γ を決定する必要がある．そこで，比較的少人数のワーカが 600 枚の画像にラベルをつけたデータを用いて割り当てのシミュレーションを行い，それぞれの手法で精度が最も高くなる δ, γ を決定した．次に，画像 800 枚により多くのワーカがラベルをつけたデータを用いてシミュレーションを行った．

4.1 実験に用いたデータ

800 枚の画像のうちランダムに選択した 200 枚を正解が既知のアイテムとしてワーカの混同行列の作成に用い，残りの 600 枚の画像を正解が未知のアイテムとしてラベル推定を行った．42 人のワーカからデータを収集し，ラベル精度が 25%を下回る 4 人をスパマーとみなし除去した．ワーカのラベル精度の分布は図 1 のようになった．

また，表 1 は，ワーカ 38 人が 800 枚の写真につけたラベルのうち 1 位ラベルから混同行列を作成したものである．ワーカ全員の 1 位ラベルの正答率は 0.718 であった．

この混同行列より，ワーカ全体としては German Shepherd（shepherd）や Samoyed（samoyed）は 9 割近い高精度で判別

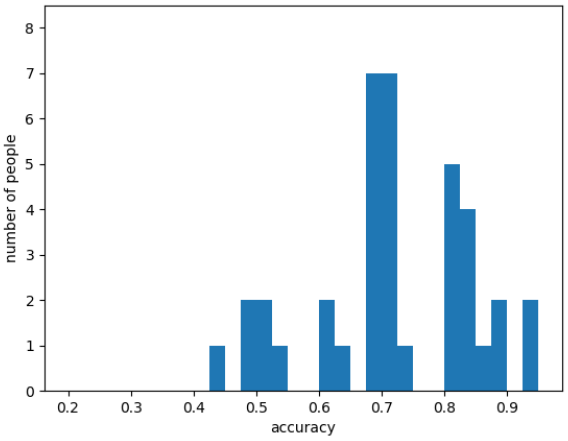

図 1　ワーカのラベル精度の分布（本実験データ）

表 1　ワーカ全員の 1 位ラベルの混同行列

	Malamute	Coyote	Dhole	Wolf	Shepherd	Samoyed	Husky
Malamute	0.575	0.009	0.007	0.039	0.033	0.037	0.301
Coyote	0.015	0.605	0.147	0.170	0.024	0.012	0.027
Dhole	0.008	0.107	0.743	0.099	0.029	0.006	0.007
Wolf	0.0811	0.149	0.039	0.567	0.047	0.044	0.074
Shepherd	0.028	0.014	0.015	0.016	0.890	0.021	0.015
Samoyed	0.013	0.007	0.008	0.005	0.018	0.936	0.013
Husky	0.214	0.009	0.006	0.049	0.031	0.028	0.663

可能であるが，Alaskan Malamute（malamute）と Siberian Husky（husky）の組や Coyote（coyote），Dhole（dhole），Gray Wolf（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．

4.2　実験手順

提案手法は混同行列の作成・アイテムに割り当てるワーカの選択・真のラベル推定の三段階に分けられる．それぞれ説明する．なお，3.4 で述べたように，1 位ラベルのみを用いて従来の混同行列を作成し，ワーカをランダムに選択して最尤推定を行う手法を比較手法として扱う．

4.2.1　混同行列の作成

4.1 で収集したデータのうち正解が既知のアイテムにつけられたラベルのデータをワーカごとに分割し，3.1 で述べた三種類の混同行列を作成する処理を行った．

- 従来の混同行列

ワーカがつけた 1 位ラベルのみを用いて，正解ラベルを行，選択ラベルを列とした 7×7 の混同行列を作成した．

- 拡張した混同行列

ワーカがつけた 1 位ラベルと 2 位ラベルを用いて拡張した混同行列を作成した．

- 重み付き混同行列

ワーカがつけた 1 位ラベルと 2 位ラベルを用いて重み付き混同行列を作成した．

それぞれの混同行列について，3.1.3 で述べた二種類のスムージング処理をそれぞれ行った．三種類の混同行列に二種類のスムージングを行ったので六種類の混同行列ができたことになる．

4.2.2　ワーカの選択

4.2.1 で作成した混同行列を利用して，3.2 で述べたように各アイテムにワーカを割り当てた．本実験では各アイテムに 3 人ずつワーカを割り当ててラベル推定を行うこととした．特定のワーカばかりにタスクが割り当てられることのないよう，ワーカ一人当たり高々$800 \times 3 \div 38 \simeq 63.1 < 64$ 件のアイテムが割り当てられるように設定した上で，3.2 で述べたように次の割り当て処理をそれぞれ行った．

- ランダム選択

各アイテムに 3 人のワーカをランダムに選択し，それぞれラベルを得た．

- 正解を選ぶ確率が最大となるワーカを選択（個人確率）
- 多数決ラベルが正解である確率が最大となるワーカを選択（多数決確率）
- 最尤推定ラベルが正解である確率が最大となるワーカを選択（最尤推定確率）

4.2.3　ラベル推定

得たラベルから最尤推定によって真のラベルを推定した．

4.3　実験の評価

提案手法・比較手法ともにランダムにワーカを選択する要素があるため，各手法についてそれぞれ 10 回ずつ実験を行い，多数決によって推定されたラベルの精度の平均で評価を行った．

4.4　実験結果

各手法 10 回ずつ実験を行った際の平均分類精度は表 2 のようになった．最大値を太字で示した．

5　考　　察

本研究で提案した複数のアイデアの効果について考察を行う．

5.1　混同行列のスムージング

混同行列の要素が 0 になることにより精度が下がることを防ぐために，本研究では add-one スムージングとグローバルスムージングの 2 種類のスムージング手法を適用した．

本研究で実験を行った手法のうち，2 種類のスムージング手法それぞれを行った組み合わせは 8 組であった．うち 7 組は add-one スムージングの精度が高く，残りの 1 組は同率であった．混同行列内のゼロ要素が確実になくなる add-one スムージングと比較して，グローバルスムージングではワーカ全体の混同行列でもゼロの要素があればスムージングしても混同行列にゼロ要素が残る．そのため，学習データで登場しない組み合わせの出現時の精度が下がり，add-one スムージングより精度が低くなったと考えられる．ただし，グローバルスムージングと add-one スムージングは独立したスムージング手法であるので併用も可能である．併用した場合に精度がどう変わるか検討の余地があると考えられる．

表 2　10 回分類を行った際の平均精度

ラベル数	混同行列	スムージング	ワーカ選択	平均精度
1	従来	add-one	ランダム	0.821
1	従来	グローバル ($\gamma = 0.7$)	ランダム	0.814
1	従来	add-one	個人確率	0.873
1	従来	グローバル ($\gamma = 0.8$)	個人確率	0.861
1	従来	add-one	多数決確率	0.838
1	従来	グローバル ($\gamma = 0.8$)	多数決確率	0.833
1	従来	add-one	最尤推定確率	0.862
1	従来	グローバル ($\gamma = 0.7$)	最尤推定確率	0.854
2	拡張	add-one	ランダム	0.831
2	拡張	グローバル ($\gamma = 0.7$)	ランダム	0.813
2	拡張	add-one	個人確率	0.873
2	拡張	グローバル ($\gamma = 0.8$)	個人確率	0.870
2	拡張	add-one	多数決確率	0.845
2	拡張	グローバル ($\gamma = 0.8$)	多数決確率	0.845
2	拡張	add-one	最尤推定確率	**0.884**
2	拡張	グローバル ($\gamma = 0.7$)	最尤推定確率	0.837
2	重み付き ($\delta = 0.6$)	add-one	個人確率	0.878
2	重み付き ($\delta = 0.6$)	add-one	最尤推定確率	0.872

以下の考察では add-one スムージングを用いた場合の結果（表 3）を用いる．

表 3　add-one スムージングラベル複数選択の効果

ラベル数	混同行列	ワーカ選択	平均精度
1	従来	ランダム	0.821
1	従来	個人確率	0.873
1	従来	多数決確率	0.838
1	従来	最尤推定確率	0.862
2	拡張	ランダム	0.831
2	拡張	個人確率	0.873
2	拡張	多数決確率	0.845
2	拡張	最尤推定確率	0.884
2	重み	個人確率	0.878
2	重み	最尤推定確率	0.872

5.2　逐次選択の効果

1 位ラベルのみを用いた場合，各アイテムに適したワーカを逐次的に選択して割り当てることで，ランダムにワーカを割り当てたベースラインより精度が高くなった．特に，ワーカ個人が正解を選択する確率や最尤推定ラベルが正解である確率を用いた割り当てによって大幅に精度が高くなった．10 回の試行における精度の分散は，最尤推定ラベルが正解である確率を利用してワーカを割り当てたときに最も小さくなった．

分類結果の混同行列によると，いずれの逐次割り当て手法も，ランダムに割り当てた場合よりほとんどのクラスの精度が高くなっていた．クラスごとの精度の分散はワーカ個人が正解を選択する確率による割り当てにおいて最も小さくなった．ワーカ個人が正解を選択する確率を利用してワーカを割り当てると，全クラスで安定して高精度の分類を行えるとわかった．

最尤推定ラベルが正解である確率を利用した割り当ては，クラスごとの精度の分散が大きく，ランダムに割り当てた場合と近かった．また，区別しづらいクラスの組み合わせがある場合は，片方のクラスの分類精度は高くもう片方のクラスの分類精度は低くなっていた．例えば，区別しづらいクラスの組み合わせである Alaskan Malamute と Siberian Husky については，Alaskan Malamute の精度はランダムにワーカを割り当てたときを下回った一方，Siberian Husky の精度は他の割り当て手法での Siberian Husky の分類精度を大きく上回った．

ワーカがラベルを複数選択することを許容しない場合，各アイテムにラベルを与えるワーカを逐次的に割り当てることで精度の向上が見込めることがわかった．特に，そのアイテムに正解を与える確率が高いワーカを逐次的に割り当てることで，クラス間での分類精度のばらつきが少なくなった．

5.3　ラベル複数選択の許容

複数ラベルを用いてラベル推定を行った場合と 1 位ラベルのみを用いて推定を行った場合の比較を行い，ラベル複数選択を許容することの効果について述べる．

ワーカをランダムに割り当てた場合，2 位までのラベルを用いることで精度が微増した．1 位ラベルのみを用いて分類を行った場合と比較すると，取り違えやすいクラスのペアである malamute と husky の精度がともに約 8%高くなったことから，ワーカが絞り込めなかったカテゴリの情報を得ることで精度を高めることができると考えられる．

その一方，複数選択を行わなかったときには起こらなかったパターンの誤分類が起こった．例えば，誤分類されにくいクラスであった shepherd のアイテムが複数選択の許容により malamute に分類されることがあった．複数ラベルの利用により 1 位ラベルのみ用いた際には存在しなかったノイズを新たに含めて推定を行ってしまうリスクがあると言える．

ワーカ逐次選択を行った場合でも，2 位ラベルまで用いて推定を行うことでラベル全体の分類精度を高めることができた．しかし，ワーカが正解を選択する確率や多数決ラベルが正解である確率による逐次選択では，複数のラベルを利用することでクラスごとの精度の分散が大きくなった．特に取り違えやすいクラスの組において，片方のクラスの精度は低く，もう片方は高くなることで，全体としての精度は変わらないまたは高くなっているが，クラスによっては精度が下がることがあった．

最尤推定ラベルが正解である確率を用いてワーカを割り当てた場合には，複数のラベルを用いることで 10 回の試行における精度の分散とクラスごとの精度の分散がともに小さくなり，全クラスで高精度の分類を行うことができた．

5.3.1 拡張した混同行列と重み付き混同行列の比較

本研究では複数選択を許容する場合の混同行列として，拡張した混同行列と重み付き混同行列の２種類を提案した．

拡張した混同行列はワーカが複数ラベルを選択した場合でも一つだけ選択した場合と同様に扱うことができるが，選択ラベルの組み合わせの数だけ行列のサイズが大きくなるため，行列が疎になりやすいという問題があった．そこで，混同行列のサイズを大きくせずに複数ラベル選択に対応する手法として重み付き混同行列を提案したところ，拡張した混同行列を用いた場合と平均精度は大きく変わらなかった．適切な重みを定めた上で重み付き混同行列を用いることで，行列のサイズは従来の混同行列と同じまま，拡張した混同行列による分類と同等な精度の分類を行うことができた．ただし，クラスごとの精度の分散は拡張した混同行列を使った場合より大きく，全体としての分類精度は同等でもクラスごとの分類精度にばらつきがあった．

6 結　　論

本論文では，クラウドソーシングで複数のクラスにアイテムを分類するタスクを依頼した際の分類精度を向上するための二つのアイデアを提案した．

一つはアイテムに割り当てるワーカを逐次選択することであった．すでにアイテムに割り当てられたラベルを利用してそのアイテムの真のラベルの推定分布を計算し，ワーカのクラス別分類精度を示す混同行列を利用して各アイテムにより適したワーカを割り当てた．

もう一つはワーカがアイテムに適合するラベルを絞りきれない場合に複数のラベルを選択することを許容することであった．ワーカが複数のラベルを選択した場合に対応した混同行列を提案した．

提案した各手法の有効性を確認するために，Amazon Mechanical Turk にマルチクラス分類タスクを投稿して得た分類データを利用してシミュレーション実験を行い，平均精度の比較を行った．各アイテムに割り当てるワーカを逐次選択することにより，分類精度の向上が見られた．複数ラベル選択の許容に関しては，クラスによって分類精度が向上されるものもあればかえって下がってしまうものもあり，全体としての精度はほぼ変わらなかったが，クラスごとの精度の分散が大きくなった．複数ラベル選択の許容とアイテムに割り当てるワーカの逐次選択を併用することで，クラスごとの精度のばらつきを少なくし，全クラスに関して安定した高精度の分類を行うことができた．

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] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, Vol. 2, No. 1-2, pp. 83–97, 1955.

[2] Harold W Kuhn. Variants of the hungarian method for assignment problems. *Naval Research Logistics Quarterly*, Vol. 3, No. 4, pp. 253–258, 1956.

[3] Chien-Ju Ho, Shahin Jabbari, and Jennifer Wortman Vaughan. Adaptive task assignment for crowdsourced classification. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, pp. I–534–I–542. JMLR.org, 2013.

[4] Chien-Ju Ho and Jennifer Vaughan. Online task assignment in crowdsourcing markets, 2012.

[5] A. P. Dawid and A. 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.

[6] A. DEMPSTER. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, Vol. 39, pp. 1–38, 1977.

[7] Satoshi Oyama, Yukino Baba, Yuko Sakurai, and Hisashi Kashima. Accurate integration of crowdsourced labels using workers' self-reported confidence scores. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, IJCAI '13, pp. 2554–2560. AAAI Press, 2013.

[8] Panagiotis G. Ipeirotis, Foster Provost, and Jing Wang. Quality management on amazon mechanical turk. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, HCOMP '10, pp. 64–67, New York, NY, USA, 2010. ACM.

[9] A. Vempaty, L. R. Varshney, and P. K. Varshney. Reliable crowdsourcing for multi-class labeling using coding theory. *IEEE Journal of Selected Topics in Signal Processing*, Vol. 8, No. 4, pp. 667–679, Aug 2014.

[10] Xiaoni Duan and Keishi Tajima. Improving multiclass classification in crowdsourcing by using hierarchical schemes. In Ling Liu, Ryen W. White, Amin Mantrach, Fabrizio Silvestri, Julian J. McAuley, Ricardo Baeza-Yates, and Leila Zia, editors, *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pp. 2694–2700. ACM, 2019.

[11] Stefanie Nowak and Stefan Rüger. How reliable are annotations via crowdsourcing: A study about inter-annotator agreement for multi-label image annotation. In *Proceedings of the International Conference on Multimedia Information Retrieval*, MIR '10, pp. 557–566, New York, NY, USA, 2010. ACM.

[12] A. Kanehira and T. Harada. Multi-label ranking from positive and unlabeled data. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5138–5146, June 2016.

[13] 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*, KDD '08, pp. 614–622, New York, NY, USA, 2008. ACM.

[14] Jinyang Gao, Xuan Liu, Beng Chin Ooi, Haixun Wang, and Gang Chen. An online cost sensitive decision-making method in crowdsourcing systems. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, SIGMOD '13, pp. 217–228, New York, NY, USA, 2013. ACM.

[15] Qi Li, Fenglong Ma, Jing Gao, Lu Su, and Christopher J. Quinn. Crowdsourcing high quality labels with a tight budget. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, WSDM '16, pp. 237–246, New York, NY, USA, 2016. ACM.