DEIM Forum 2023 4b-8-5

クラウドワーカの模倣モデルと投票作業の一致率を用いた結果集約方法

太田　奈那† 　鈴木　 優†

† 岐阜大学工学部電気電子・情報工学科　〒 501–1193 岐阜県岐阜市柳戸 1 番 1

E-mail: †{y3033029@edu.gifu-u.ac.jp，ysuzuki@gifu-u.ac.jp}

あらまし　クラウドソーシングは，インターネットを介して人を募集するためスパムワーカが一定数存在する．スパムワーカが行った作業結果も含めて集約を行うと，作業依頼者の求める結果が得られないことがある．そこで，全データを集約した全体の結果と個人の作業結果との一致率が高いほど品質の高いワーカであるとし，品質の低いワーカの影響力を抑える結果集約手法を提案する．本手法では，ワーカの模倣モデルによって得られた結果をもとに算出した一致率を用いて，各ワーカの票に重み付けして多数決を行う．票の重みをワーカごとに変えることによって品質の低いワーカの影響力を抑えることができ，品質の高いワーカの作業結果が反映されやすくなると考えた．それにより，複数人の作業に対する多数決にて付与された評価ラベルと比較して，手法を適用させることによって付与された評価ラベルの方が作業依頼者の求めるものに近い作業結果を得ることができると考えた．そこで，本手法の有効性を確かめるために実験を行った．その結果，本手法を適用することによって集約した結果の *Accuracy* は 0.6685 となり，品質の低いワーカの影響力を抑え，作業依頼者の求める結果に近い作業結果が得られることを確認した．

キーワード　機械学習，クラウドソーシング，一致率

1　はじめに

クラウドソーシング [1] とは，インターネット上で募集した不特定多数の人に作業を委託することである．クラウドソーシングを行うことにより，複雑な作業や一人で行うには困難な作業を複数の人に分担して行ってもらうことができる．そのため，一人ひとりの作業は単純になり，一人にかかる負担を軽減できる．また，インターネット上で作業が行われるため，誰でも簡単に好きな時間に好きな場所で作業を行うことができる．

しかし，インターネットを介してワーカを募集しているため，スパムワーカと呼ばれる品質の低いワーカが一定数存在する．スパムワーカとは，数をこなすためだけに故意に不適切な作業をするワーカである．スパムワーカが存在する状態で結果集約を行うとスパムワーカの影響を受けてしまい，作業依頼者の求める結果が得られないことがある．スパムワーカは他のワーカと違う結果を出すため，スパムワーカを取り除くことができるのではないかと考えられる．そのためには，作業に対する正しい評価が必要になる．そして，その評価は作業依頼者が求める結果に近い作業結果となっていなければならない．

作業依頼者が求める結果を得るためには，より多くのワーカの意見が必要であると考えた．しかし，より多くのワーカの意見を集めるには時間や費用がかかる．そこで，ワーカの作業を模倣する分類器を構築することにより，擬似的にワーカを増やすことができると考えた．擬似的にワーカを増やすことによって，時間や費用を抑えた上で多くのワーカの意見を集めることが可能になる．そして，その意見を集約することにより作業依頼者の求める結果を得ることができると考えられる．

模倣モデルを構築するワーカの中には，先ほど述べたスパムワーカが存在している可能性がある．スパムワーカの作業結果を含めて結果集約を行った場合，作業依頼者の求める結果を得ることは難しい．そこで，ワーカの品質を考慮した結果集約を行うことによって，スパムワーカの影響力を抑えることができるのではないかと考えた．ワーカ個人が付与した評価ラベルと複数人の作業に対する多数決にて付与された評価ラベルとの一致率をもとにワーカの品質を求める．その品質を用いて結果集約を行い，作業依頼者の求める結果を得ることを目指す．

本研究では，クラウドソーシングの質を向上させるための異なる三つの結果集約手法を提案する．一つ目は，一致率をもとにワーカの選定を行って多数決をとる手法である．品質の低いワーカを取り除いて品質の高いワーカの作業のみを使用することにより，作業依頼者の求める結果が得やすくなると考えた．二つ目は，一致率をもとにワーカの品質を求め，その品質を用いて各ワーカの票に重み付けして多数決をとる手法である．票の重みをワーカごとに変えることにより，品質の低いワーカの影響力を抑えることができると考えた．三つ目は，一致率をもとにワーカが一つのデータに対して投票できる確率を算出し，無効票を決定して多数決をとる手法である．集約時にデータを使用するかどうかを確率的に決定することにより，品質の低いワーカが行った作業の中に含まれる有用なデータが使用される可能性を残せると考えた．

これら三つの手法で共通している点は，ワーカの一致率をもとにした多数決を行うことである．そのため，一致率が高いワーカの作業結果は多数決に反映されやすく，低いワーカの作業結果は多数決に反映されにくくなる．これにより，模倣モデルを構築したワーカの中にスパムワーカが含まれていたとしても，その影響は小さくなる．そして，作業依頼者の求める結果に近い作業結果を得ることができるのではないかと考えられる．

提案手法がクラウドソーシングの質の向上に有効であるかどうかを確かめるために評価実験を行った．各手法を用いて付与

される評価ラベルに加えて，複数人の作業に対する多数決にて付与された評価ラベルが，作業依頼者の求める結果とどの程度一致するかを比較することにより手法の有効性を確かめた．

三つの手法を用いてそれぞれ実験を行ったところ，品質を票の重みとする手法を用いた場合に *Accuracy* は 0.6685 となり，作業依頼者が求める結果と一致したデータが最も多いという結果が得られた．ワーカの選定をして多数決をとる手法や無効票を確率的に決定する手法を用いると，品質の低いワーカの作業データを取り除くことは可能であるが，そのワーカの有用な作業データまで取り除かれてしまうことがある．また，無効票を確率的に決定する手法では品質の高いワーカの作業結果であっても取り除かれてしまうことがあり，正しい評価ラベルを得ることが難しいと考えられる．一方，品質を票の重みとする手法では品質の低いワーカであっても作業結果が残るため，有用な作業データが取り除かれることもなく，評価ラベルを決める際に少なからず影響を与えることが可能である．そのため，品質が低いワーカの影響を抑えつつ，正しい評価ラベルを得ることが可能であったのではないかと考えられる．

本稿における貢献は以下のとおりである．

- ワーカの模倣モデルを構築し，モデルの精度を確かめた．
- ワーカの模倣モデルによる評価ラベルと投票作業から得た評価ラベルの一致率を用いてクラウドソーシングの質を向上させることができた．

2 関 連 研 究

クラウドソーシングの質の向上に関する研究はいくつか存在する．西ら [2] の研究では，ソーシャルネットワークを用いたワーカの品質向上を目指している．作業を行うワーカは他一人に作業を委託することができ，報酬は作業に正解したワーカとそのワーカに作業を委託したワーカに支払われる．こうすることによって，能力の低いワーカは能力の高いワーカに作業を委託するようになり，品質の高いワーカの作業結果を得られる．芦川ら [3] [4] の研究では，ワーカに作業の適性があるかどうかのフィルタリングを行うことにより，クラウドソーシングの質の向上を目指している．ワーカの作業前，作業中，作業後に加えて，得られた結果を用いて推測された未知データの結果精度を用いてフィルタリングを行う．フィルタリングを行ってワーカを絞ることにより，クラウドソーシングの精度を向上させている．Halpin ら [5] の研究では，スパムワーカの検出手法を提案している．ワーカごとに作業数や一つの作業を行うのにかかった平均時間などの特徴を用い，機械学習を行うことによってスパムワーカの検出を行っている．松原ら [6] の研究では，ワーカに適した作業の割り当てを行っている．提案されている手法では，複数のタスクをワーカに割り当てる場合において，まず各ワーカに希望するタスクの優先順位をつけさせる．各ワーカがつけたタスクの優先順位をもとに各ワーカに対して作業を割り当て，虚偽の順位をつけたワーカに対して不利益が生じるように設定する．そのため，ワーカは真実の優先順位をつけることとなり，ワーカごとに適した作業を割り当てることにより精度の高い結果を得ている．上記の研究ではワーカの品質に着目することにより，クラウドソーシングの質の向上を目指している．

本研究では，ワーカごとの模倣モデルを作成し，模倣モデルを用いて作業データを増量することによってクラウドソーシングの質の向上を目指す．一つのデータに対する作業結果を増やすことによって多くの意見を得ることができ，作業依頼者の求める結果に近い評価を得ることができるのではないかと考えた．そのため，上記の研究とはクラウドソーシングの精度向上を目指す点では同じであるが，精度を向上させるためのアプローチとして作業データの増量をしているという点で異なっている．

また，結果集約手法に関する研究もいくつか存在する．Dawid ら [7] の研究では，EM アルゴリズムを用いたラベル付与の手法について提案している．ワーカが各ラベルを回答したときの正解率を EM アルゴリズムを用いて推定する．推定して得られた正解率が最も高いラベルを真のラベルとしてデータに付与するという手法である．小山ら [8] の研究では，高精度なラベル統合方法について提案している．ワーカに行った作業の処理結果をどの程度確信しているのかを申告してもらい，その確信度をもとにワーカの評価がラベルを付与する際に必要な情報であるかどうかを確率的に判断している．また，自己申告した確信度は正解率と相関があると考えており，確信度を用いることによって高い精度で適切なラベルを付与している．

本研究では，クラウドワーカの模倣モデルを作成することによって求めた一致率を用いてワーカごとの票の重みを設定したり，一致率をもとにワーカが一つのデータに対して投票できる確率を算出して無効票を決定したりすることによって多数決をとってラベルを付与する．そのため，作業結果の質を高めるために品質の低いワーカの評価が結果に現れにくくするという点で上記の研究と異なる．また，我々が一致率を求めることによりワーカが自己申告する必要がなく，ワーカへの負担が少ない．

3 提 案 手 法

本研究は，クラウドワーカの模倣モデルによる予測結果と複数人の作業に対する多数決にて付与された評価ラベルとの一致率を用いた結果集約手法によって，クラウドソーシングの質を向上させることを目的とする．模倣モデルによる予測結果と複数人の作業に対する多数決にて付与された評価ラベルとの一致率が高いほど品質の高いワーカであると考えられる．そのため，一致率を用いた結果集約手法を用いることによってより良いワーカの結果を反映させることができ，作業依頼者の求める結果に近い作業結果が得られるのではないかと考えた．

そこで本研究では，以下の異なる三つの結果集約手法によってクラウドソーシングの質を向上させることを目指す．

(1) 一致率をもとにワーカの選定をして多数決をとる
(2) 一致率をもとに各ワーカの票に重み付けして多数決をとる
(3) 一致率をもとに各データに対して投票できる確率を算出し，無効票を決定して多数決をとる

各手法について 3.2.1 項から 3.2.3 項で詳しい説明を行う．

3.1 分類器の構築

本研究では，各ワーカの作業を模倣するような分類器を構築する．ワーカの模倣モデルを構築する際，東北大学の乾・鈴木研究室で構築された訓練済み日本語 BERT モデル [1]を使用して，テキストデータに対して評価ラベルを予測するような分類器を構築する．ファインチューニング時には，BERT モデルの最終層のパラメータのみを更新するように設定した．ここで，以下の異なる二種類の方法を用いて BERT モデルのファインチューニングを行う．

Model 1　ワーカ個人の作業結果を集めたデータセットを用いて，BERT モデルをファインチューニングする．

Model 2　(i) 複数人の作業に対する多数決にて評価ラベルを付与したデータセットを使用して，BERT モデルをファインチューニングする．

(ii) ワーカ個人の作業結果を集めたデータセットを使用して，(i) で構築したモデルをファインチューニングする．

Model1 では，ワーカのデータ数が少なかった場合に精度の高い分類器を構築することは難しい．そこで，Model2 を用いることによってデータ数の少なさを補うことができ，どのようなワーカも精度の高い分類器を構築することができるのではないかと考えた．そのため，上記に示した二種類の方法を用いて各ワーカの模倣モデルを構築する．

3.2 結果集約手法

本節では，異なる三つの結果集約手法について説明する．三つの結果集約手法に共通する，模倣モデルによる予測結果と投票作業による作業結果の一致率を算出する手順を説明する．

各ワーカを $w_i \in W$，$i = 1, 2, \cdots, n$ とし，各ワーカが実際に作業したデータを集めたデータセットを D_i とする．ここで n はワーカの人数とする．そして，複数人の作業に対する多数決にて評価ラベルを付与したデータセットを D_{all} とする．また，評価ラベルを付与するテキストデータの総数を N とし，各データを d_j，$j = 1, 2, \cdots, N$ とする．複数人の作業に対する多数決にて評価ラベルを付与したデータセット D_{all} の中から，ワーカ w_i が作業していないデータ d_j を抽出してデータセット $D'_i = \{d_j | d_j \in D_{all}, d_j \notin D_i\}$ を作成する．構築した模倣モデルにて予測したデータセット D'_i に含まれるデータ d_j の評価予測を，データ d_j の評価ラベルとして付与する．このとき，ワーカ w_i の模倣モデルによって付与されたデータ d_j の評価ラベルを $l_i(d_j)$ と表す．評価ラベル $l_i(d_j)$ と複数人の作業に対する多数決にて付与された評価ラベル $l_{all}(d_j)$ を比較して，一致率 A_i を求める．ワーカ w_i の一致率 A_i は以下のように算出する．

$$A_i = \frac{\sum_{j=1}^{N} M(l_i(d_j), l_{all}(d_j))}{N} \tag{1}$$

$$M(l_i(d_j), l_{all}(d_j)) = \begin{cases} 1 & (l_i(d_j) = l_{all}(d_j)) \\ 0 & (l_i(d_j) \neq l_{all}(d_j)) \end{cases} \tag{2}$$

1：https://github.com/cl-tohoku/bert-japanese

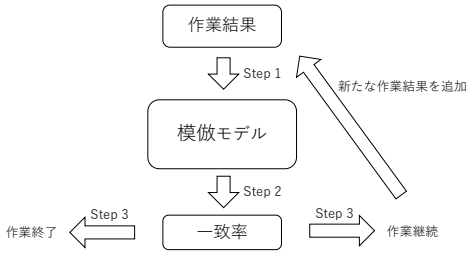

図 1　一致率によるワーカの選定の概要

式 (2) は，ワーカ w_i の模倣モデルによって付与された評価ラベルと複数人の作業に対する多数決にて付与された評価ラベルを比較したとき，一致していたら 1，異なっていたら 0 を出力する関数である．式 (1) を用いて算出された一致率 A_i を使用して三つの結果集約手法で多数決をとる作業を行う．

3.2.1 一致率によるワーカの選定

構築した模倣モデルによって得られた一致率 A_i をもとにワーカを選定し，多数決をとる手法について説明する．この手法では逐次的にワーカの作業の質を確認して，品質の低いワーカを除去する．手法の概要は図 1 に示す通りである．

Step 1　ワーカ w_i が行った作業結果を訓練データとして模倣モデルを作成する．

Step 2　作成した模倣モデルを用いてワーカ w_i が作業を行っていないデータの評価ラベルを予測し，複数人の作業に対する多数決にて付与された評価ラベルとの一致率 A_i を算出する．

Step 3　算出した一致率 A_i が閾値を上回っていたら作業を継続し，下回っていたらその時点で作業を終了する．

Step3 で算出した一致率 A_i が閾値を上回っていた場合は作業を継続し，新たな作業結果を追加して Step1 から Step3 の手順を繰り返し行う．ワーカの作業の質を逐次的に確認する作業を繰り返し行うのは，最初は真面目に作業を行っていても途中から適当にこなしていたり，判定することが簡単なデータが作業序盤に割り振られていることにより作業の質が一時的に良くなっていたりする可能性があるためである．最終的に一致率 A_i が閾値を下回らなかったワーカの作業結果を使用して多数決をとる．ここで，Step1〜3 の手順を繰り返し行っている回数を m，k クラス分類のチャンスレートを $C(C = \frac{1}{k})$，閾値の最大値を θ_{max} として閾値 θ_m を以下の式を用いて定める．このとき，最大の閾値 θ_{max} は $C < \theta_{max} \leqq 1$ の範囲で値をとる．

$$\theta_m = 2\,(\theta_{max} - C)\, sigmoid\left(\frac{m}{5}\right) - \theta_{max} + 2C \tag{3}$$

$$sigmoid(x) = \frac{1}{1 + exp(-x)} \tag{4}$$

式 (4) はどのような入力値も 0.0〜1.0 の範囲の数値に変換して出力することができるシグモイド関数である．式 (3) を用いて閾値を定めることによって，入力値が正の数のときにチャンス

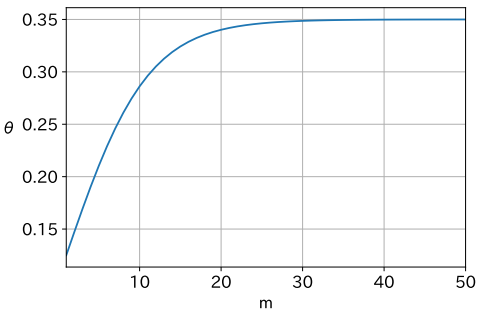

図 2　8 クラス分類で最大の閾値を 0.35 としたときの閾値の推移

レートから閾値の最大値までの範囲の中で数値を出力することができる．閾値の最大値とは，手順を繰り返し行うごとに増加する θ_m が最終的に収束する値のことである．この値を自身で定めることにより，使用するデータセットの難易度に合わせた閾値を定めることができるようになる．また，下限としてチャンスレートを用いることにより，無作為に選んだ場合よりも一致率の低い分類器を取り除き，ワーカの選定をすることができる．図 2 は，8 クラス分類の分類器を構築する際に最大の閾値を 0.35 としたときの閾値の推移の例を表している．

3.2.2　一致率による票の重み設定

ワーカの模倣モデルによって得られた一致率をもとに，各ワーカの票に重み付けすることにより多数決をとる手法について説明する．この手法では，品質の低いワーカの作業結果が集約結果に反映されにくくなるように，多数決をとる際に各ワーカの票に重み付けして集約をする．ワーカ w_i の一致率 A_i を算出する流れについては StepA，算出した一致率 A_i をもとに票の重みを設定して多数決をとる流れは StepB の手順で行う．

Step　A-1　データセット D_i から無作為に抽出したデータを訓練データとして学習を行い，ワーカ w_i の模倣モデルを構築する．

A-2　模倣モデルを用いてデータセット D_i' に含まれるテキストデータ d_j の評価ラベルを予測し，複数人の作業に対する多数決にて付与された評価ラベルとの一致率 A_i を算出する．

Step　B-1　A-2 で算出した一致率 A_i をもとに票の重みを設定し，作業結果に対して重み付けを行う．

B-2　同じデータに対して評価を行ったワーカのデータを集約し，結果を出す．

B-1 で行われるワーカ w_i の票の重み設定は以下の手順で行う．ワーカ w_i の品質を q_i として，最大値 1，最小値 0 となるように一致率 A_i を正規化する．正規化する際に k クラス分類のチャンスレート C を含めることによって，全ワーカの一致率 A_i がチャンスレート C を上回っていたとき，全ワーカが投票権を得ることができる．ワーカの一致率 A_i とチャンスレート C を含む集合 $Q = \{x | x = C, A_i (i = 1, 2, \cdots, n)\}$ を用いて，ワーカの品質 q_i を以下のように求め，ワーカ w_i の票の重みとする．

$$q_i = \frac{A_i - x_{min}}{x_{max} - x_{min}} \tag{5}$$

3.2.3　一致率による無効票の決定

ワーカの模倣モデルによって得られた一致率をもとに算出した確率を用いて，無効とする票を決めて多数決をとる手法について説明する．この手法では，品質の低いワーカの作業結果を集約結果に反映されにくくするために，作業集約時に無効票を確率的に決定する．ワーカ w_i の一致率 A_i を算出する流れについては 3.2.2 項の StepA と同じである．また，算出した一致率 A_i をもとに無効票を決める流れは StepC の手順で行う．

Step　C-1　模倣モデルを用いてデータセット D_i' に含まれるデータ d_j の評価ラベルを予測する．

C-2　A-2 で算出した一致率 A_i の値をもとにワーカ w_i の票が有効か無効を決定する．

C-2 では，A-2 で算出した一致率 A_i の値をもとにワーカ w_i のデータ d_j に対する票を有効とする確率 p_i を決定している．各ワーカの一致率のうち最大となる値を A_{max} とし，確率 p_i を以下の式を用いて算出する．

$$p_i = sigmoid\left(10 \times \frac{A_i}{A_{max}} - 5\right) \tag{6}$$

式 (6) はシグモイド関数をもとにしており，$A_i = \frac{A_{max}}{2}$ のとき $p_i = 0.5$ となる．そのため，一致率 A_i が $\frac{A_{max}}{2}$ を上回ると票が有効になる確率 p_i が高くなり，意見が反映されやすくなる．一方で，一致率 A_i が $\frac{A_{max}}{2}$ を下回ると票が有効になる確率 p_i が低くなり，意見が反映されにくくなる．このようにして無効票を決定することによって多数決をとる．

4　評 価 実 験

本研究では，異なる三つの結果集約手法を提案した．そこで，各手法が有効であるかどうかを確かめるための実験を行った．

4.1　データセット

本研究では，岐阜大学鈴木研究室にてクラウドソーシングにより構築されたデータセットを使用する．このデータセットは 605 人のクラウドワーカによって作業されたデータで構築され，全部で 250,354 件のデータが含まれている．このデータはワーカ ID，ツイート ID，ツイート内容，評価ラベルの四つのカラムで構成されている．またこのデータセットは，ツイート内に含まれる「笑」がポジティブな意味を持つのかネガティブな意味を持つのか，ネガティブの中でもどのような場面で使用されているのかを知ることを目的に作成されたものである．評価ラベルは表 1 に示した 8 種類のラベルを使用している．データセットを構築する際，評価ラベルは以下の手順で付与される．

(1)　「笑」を含むテキストデータに対してポジティブ，ネガティブ，ニュートラル，ポジ＋ネガ，その他の中から一つの評価ラベルを付与する．

(2)　(1) でネガティブが付与されたテキストデータに対して攻撃性あり，攻撃性なし，自虐，判断できないの中から一つの評価ラベルを付与する．

このように段階を踏むことによって，ネガティブの評価ラベル

が付与された場合にさらに細かい評価ラベルを付与した．以上で述べた作業は，一つのデータに対して 5 人のワーカによって行われることが想定されている．

上記のデータセットは一つのデータに対して複数のワーカの作業結果が存在しているため，多数決をとり評価ラベルを一つに決定する．データセットの中には作業人数が 5 人に満たないデータが存在するため，5 人の作業結果が残っているデータのみを用いて多数決をとることにする．複数人の作業に対する多数決にて評価ラベルを付与したデータセットの作成方法について説明する．クラウドソーシングによって構築されたデータセットの中から，作業人数が 5 人のテキストデータを抽出する．この抽出したテキストデータに評価ラベルを付与する際，上記で述べた手順 (1)，(2) に倣い 5 人のワーカの作業結果を使用して多数決をとる．多数決にてただ一つの評価ラベルに決まったデータを集め，データセットを作成する．投票作業にてただ一つの評価ラベルが付与されたデータは全部で 29,711 件である．

4.2 節から 4.4 節の実験では，投票作業によって得られたデータセット D_{all} と，実際に作業を行ったデータが 500 件以上あるワーカのデータセット D_i を使用する．ここで作業データが 500 件以上あるワーカを対象にした理由は，データ数が少ないとワーカの特徴を捉えたモデルを構築することが難しいと考えたためである．このワーカのデータセット D_i は，クラウドソーシングによって得られたデータの中から，ワーカ ID を用いてデータを抽出することによって作成する．

作業データが 500 件以上あるクラウドワーカは 127 人であったため $n = 127$ として，これらのデータセット $D_i(i = 1, 2, \cdots, 127)$ を使用して三つの結果集約手法の有効性を確かめるための実験を行う．また，多数決をとるデータとして，ワーカ $w_i(i = 1, 2, \cdots, 127)$ 以外のワーカの作業データを取り除いた後に残ったデータのうち，5 人のワーカ w_i によって評価ラベルが付与されているデータを使用する．このデータの中から無作為に抽出した 2,000 件のデータを集めたデータセット D_{eva} を使用して多数決をとり，それぞれの手法の評価を行う．

手法の評価を行うにあたって，データセット D_{eva} に含まれるデータ d_j に対してあらかじめ著者が正解ラベルを付与した．この正解ラベルが付与されたデータセットを D_{true} とする．この付与された正解ラベルと，それぞれの実験で付与された評価ラベルや複数人の作業に対する多数決にて付与された評価ラベルをそれぞれ比較して *Accuracy* を算出する．ここで得られた *Accuracy* をもとに実験の結果を比較していく．

表 1　評価ラベルの内容と正解ラベルの数

ラベル番号	ラベル内容	データ数 (件)
0	ポジティブ	8,955
1	ニュートラル	12,801
2	ポジ+ネガ	1,188
3	その他	706
4	攻撃性あり	1,382
5	攻撃性なし	3,116
6	自虐	1,524
7	判断できない	39

4.2　実験 1：一致率によるワーカの選定

本節では，3.2.1 項で説明した，一致率をもとにワーカを選定することによって多数決をとる手法の有効性を確かめるための実験について述べる．

4.2.1　実 験 手 順

本実験では，4.1 節で説明したワーカ w_i のデータセット D_i を使用して模倣モデルを構築する．このとき，データに評価ラベルを付与した手順と同様の手順で評価ラベルの予測を行うため，二つの模倣モデルを構築する．一つ目は，ネガティブかポジティブかなどを判定するモデルである．二つ目は，ネガティブの中でさらに細かい評価ラベルを判定するモデルである．この二つのモデルを組み合わせることによってデータ d_j に対する評価ラベルを予測する．また，3.1 節で説明した Model1 にて模倣モデルを構築する．なお，3.2.1 項で説明した手法は，逐次的にワーカの作業の質を確認する手法であるため，データセットの中から任意のデータ数を一纏まりとして実験を行った．

ワーカ w_i のデータセット D_i から抽出するデータを 200 件とし，このデータを訓練データとして模倣モデルを構築する．訓練データは，各評価ラベルのデータ数が等しくなるように，データ数が最も多い評価ラベルのデータ数に合わせてデータの複製を行う．模倣モデルを構築する際，300 エポックの学習を行う．また，データセット D_i' から抽出した 6,000 件のデータをデータセット D_i'' とする．構築した模倣モデルを用いてデータセット D_i'' の評価ラベルを予測する．予測にて付与された評価ラベルと複数人の作業に対する多数決にて付与された評価ラベルとの一致率を式 (1) を用いて算出する．算出した一致率 A_i が定めた閾値を上回っているかどうかを判定する．このときの閾値は，$k = 8, \theta_{max} = 0.35$ として式 (3) に従って定める．一致率 A_i が閾値を上回っていた場合，訓練データを 200 件増やして，模倣モデルを構築する作業から一致率の判定をする作業までの流れを再び行う．一致率 A_i が閾値を下回る，あるいは，データセット D_i に含まれるすべてのデータを使用して模倣モデルを構築した場合は作業を終える．一致率 A_i が閾値を下回ったワーカの作業データは取り除き，取り除かれなかったワーカの作業データを使用してデータセット D_{eva} に含まれるデータの評価ラベルを多数決にて付与し，*Accuracy* を算出する．

4.2.2　結果・考察

Model1 を用いてワーカの品質を測定することによってワーカの選定を行い，多数決をとった結果を表 2 の実験 1 に示す．また，表中の baseline は，複数人の作業に対する多数決にて集約された結果である．複数人の作業に対する多数決にて集約さ

表 2　各手法を適用した際の *Accuracy*

		Accuracy
baseline		0.6625
実験 1		0.5845
実験 2	Model1	**0.6685**
	Model2	0.6625
実験 3	Model1	0.5710
	Model2	0.5910

れた結果の *Accuracy* と比較すると，実験 1 の手法を用いて算出した *Accuracy* の方が低くなっていることがわかる．

Step1〜3 を繰り返し行うことによって取り除かれたワーカは 127 人中 4 人のみであった．取り除かれたワーカが実際に行った作業結果と複数人の作業に対する多数決にて付与された評価ラベルとの一致率を確認したところ，他のワーカと比べて少し低い値になっていた．しかし，取り除かれたワーカよりも一致率の低いワーカが存在していた．そのため，必ずしもワーカが実際に行った作業結果と複数人の作業に対する多数決にて付与された評価ラベルとの一致率の低いワーカが取り除かれるわけではないと考えられる．また，訓練データとして使用するデータ数が多くなるほど閾値が大きくなるため，作業データ数の多いワーカの方が取り除かれやすい傾向にあると考えられる．

表 3 の実験 1 に示すのは，複数人の作業に対する多数決にて付与された評価ラベルが，3.2.1 項にて説明した手法を用いることによって変化したツイートの例である．一つ目のツイートは，複数人の作業に対する多数決にて付与された評価ラベルは不正解であったが，手法を適用して付与されたラベルは正解であった例である．ワーカによって付与されたラベルはネガティブが多数であり，投票作業によって「攻撃性なし」が付与された．しかし，手法を適用して付与されたのは「ニュートラル」であった．そのため，ネガティブと評価したワーカは，一致率が閾値を下回ったために取り除かれた品質の低いワーカであると考えられる．また，複数人の作業に対する多数決にて付与された評価ラベルが不正解であり，手法を適用して付与されたラベルが正解であったデータは 2 件のみであった．このことから，本実験にて取り除かれたワーカ以外にも品質の低いワーカが存在し，その全てを取り除くことができなかったと考えられる．

二つ目のツイートは，複数人の作業に対する多数決にて付与された評価ラベルは正解であったが，手法を適用して付与された評価ラベルは不正解であった例である．手法を用いることによって，評価ラベルが「攻撃性なし」から「ニュートラル」に変化した．このことから，ネガティブの評価ラベルを付与したワーカは，手法を用いたことによって取り除かれた品質の低いワーカであると考えられる．しかし，手法を用いて付与された評価ラベルが正解ラベルと一致しなかったことから，取り除かれたワーカの作業結果の中にも有用なデータがあったと考えられる．また，複数人の作業に対する多数決にて付与された評価ラベルが正解であり，手法を用いて付与された評価ラベルが不正解のデータは 210 件であった．このことから，取り除かれたワーカの作業データを全て取り除いてしまうと，有用なデータまで取り除かれてしまうということが考えられる．

本実験では，品質の低いワーカを全て取り除くことはできず，質の悪い作業データが残った．そのため，正解ラベルと同じ結果を得ることができなかったと考えられる．また，必要なデータが取り除かれてしまったという結果から，取り除かれたワーカの作業データの中にも有用なデータがあったと考えられる．

4.3 実験 2：一致率による票の重み設定

本節では，3.2.2 項で説明した，一致率をもとにワーカごとに票の重みを設定することによって，多数決をとる手法の有効性を確かめるための実験について述べる．

4.3.1 実 験 手 順

4.1 節で説明したワーカ w_i のデータセット D_i を使用して模倣モデルを構築する．このとき，4.2 節と同様にポジティブかネガティブかなどを判定するモデルと，ネガティブの中でさらに細かい評価ラベルを判定するモデルの二つの模倣モデルを構築する．モデルの構築には 3.1 節で説明した二種類の方法を用いて，合計で四つのモデルを構築する．また，模倣モデルを構築するとき 4.2 節と同様に 300 エポックの学習を行う．データセット D_i を訓練データ 6 割，検証データ 2 割，テストデータ 2 割に分けて学習を行う．このとき，訓練データは 4.2 節と同様に最もデータ数が多い評価ラベルのデータ数に合わせてデータの複製を行う．構築した二種類の模倣モデルを用いて評価ラベルの予測を行い，3.2 節の式 (1) を用いて各ワーカの一致率 A_i を算出する．算出した全ワーカの一致率と 8 クラス分類のチャンスレート 0.125 を用いて最大値 1，最小値 0 となるように正規化を行い，ワーカ w_i の品質 q_i を求める．この品質 q_i をワーカ w_i の票の重みとしてデータセット D_{eva} に含まれるデータの評価ラベルを多数決にて付与し，*Accuracy* を算出する．

4.3.2 結果・考察

Model1 と Model2 のそれぞれを用いて算出した一致率をもとにワーカの品質を求め，その品質を用いて各ワーカの票に重み付けして多数決をとった結果を表 2 の実験 2 に示す．複数人の作業に対する多数決にて集約された結果と二種類のモデルを用いて手法を適用した結果を比較すると，Model1 を用いて手法を適用させた場合の *Accuracy* が最も高くなっている．

Model2 では，各ワーカ用にファインチューニングする前のモデルが，ワーカの一致率を出す際の正解のデータを使用して学習している．そのため，データ数が少ないワーカほどファインチューニングによるパラメータの変化量が少なく，各ワーカの特徴が捉えきれない．それにより，ファインチューニング前とほぼ同じ予測をするため，ワーカの質に関係なく一致率が高くなりやすく，全体を通して一致率の差が現れにくい．そのため，ワーカ w_i の品質 q_i を票の重みとして設定したとしても，複数人の作業に対する多数決にて付与された評価ラベルと変わらない評価ラベルが付与されやすくなる．実際に，複数人の作業に対する多数決にて付与された評価ラベルと，Model2 を用いて評価ラベルの予測をした場合に付与された評価ラベルが変化していたデータは 2,000 件のうち 9 件のみであった．

一方，Model1 を用いると，ワーカ w_i の特徴をそのまま反映させたモデルが構築できる．そのため，ワーカ w_i の本来の一致率 A_i を算出することができると考えられ，複数人の作業に対する多数決にて付与された評価ラベルに依存しない品質 q_i を求めることができる．そして，より品質の高いワーカの作業結果が反映されやすくなると考えられる．また，複数人の作業に対する多数決にて付与された評価ラベルと，Model1 を用いて評価ラベルの予測をした場合に付与された評価ラベルが変化していたデータは 2,000 件のうち 85 件であった．

表 3 の実験 2 に示すのは，複数人の作業に対する多数決にて

表 3　各手法によって付与された評価ラベルが変化したツイート例

	「笑」を含むツイート	vote	true	baseline	Model	Model2
実験 1	最寄りのゲーセン 10km 以上離れてんの笑 !?	ニュートラル 2 攻撃性なし 2 自虐 1	ニュートラル	攻撃性なし	ニュートラル	-
	え、ちょっと確認！今週の金曜日に虹会 20 時から ?? えっ、待てよ、!! 部活とか、帰りの時間で、見れんかもじゃん、やばいやばい !! え、終わった…笑	ニュートラル 2 攻撃性なし 2 自虐 1	攻撃性なし	攻撃性なし	ニュートラル	-
実験 2	今週末の楽しみ無くなりそうーさすがにやさぐれそうだわ笑	攻撃性なし 2 自虐 3	攻撃性なし	自虐	攻撃性なし	自虐
	そやゆとって何でこんなにいいの？ふたりの全部がいい笑	ポジティブ 3 ニュートラル 2	ポジティブ	ポジティブ	ニュートラル	ポジティブ
実験 3	おはよう夜勤に向けて寝溜め中。笑	ポジティブ 3 ニュートラル 2	ニュートラル	ポジティブ	ニュートラル	ポジティブ
	リズム感がないのでシェーカーは難易度が高いし、前々から思ってたけど私時々変なところで拍を入れてしまうんだけどつまりは耳が追いかけてるのがどの音なのかがわかるよね…。笑	ニュートラル 2 攻撃性なし 1 自虐 2	自虐	自虐	ニュートラル	自虐

付与された評価ラベルが，3.2.2 項にて説明した手法を用いることによって変化したツイートの例である．

一つ目のツイートは，複数人の作業に対する多数決にて付与された評価ラベルは不正解であったが，Model1 と手法を適用して付与された評価ラベルは正解であった例である．ワーカによって付与された評価ラベルは「攻撃性なし」が 2 票,「自虐」が 3 票であったため，複数人の作業に対する多数決にて付与された評価ラベルは「自虐」となっている．しかし，Model1 と手法を適用して付与された評価ラベルは「攻撃性なし」となっており，正解ラベルと一致している．このことから,「攻撃性なし」の評価ラベルを付与したワーカの品質が高いということが予想できる．また，品質の低いワーカの影響力を抑えることができたため，正しい結果を得ることができたと考えられる．

二つ目のツイートは，複数人の作業に対する多数決にて付与された評価ラベルは正解であったが，Model1 と手法を適用して付与された評価ラベルは不正解であった例である．一つ目のツイートと同様に，ワーカによって付与された評価ラベルが 2 票と 3 票に割れている．しかし，Model1 と手法を適用して付与された評価ラベルは 2 票しか付与されなかった評価ラベルである．品質の低いワーカによって付与された多数派の評価ラベルではなく，品質の高いワーカによって付与された少数派の評価ラベルが集約結果として得られたと考えられる．このことから，提案手法を用いることによって，品質の低いワーカの影響力を抑えることができたと考えられる．しかし，その評価ラベルは正解ラベルと一致していないことから，品質の高いワーカの評価が常に正しいとは限らないということが考えられる．また，品質の低いワーカが行った作業データの中にも有用なデータが存在すると考えることができ，全ての作業データに対して影響力を抑えることが良いわけではないと考えられる．

4.4　実験 3：一致率による無効票の決定

本節では，3.2.3 項で説明した，一致率をもとにワーカがデータに対して投票できる確率を算出して，無効票を決めて多数決をとる手法の有効性を確かめるための実験について述べる．

4.4.1　実 験 手 順

4.3 節の実験と同じようにワーカ w_i の模倣モデルを構築し，一致率 A_i を算出する．本節でも模倣モデルは 4.2 節と同様に二つ構築する．モデルの構築には 3.1 節で説明した二種類の方法を用いて行い，合計で四つのモデルを構築する．また，模倣モデルの構築については 4.3 節と同様の手順で行う．一致率 A_i の値をもとにワーカ w_i のデータ d_j に対する票が有効になる確率 p_i を算出する．例えば，確率 p_i が 0.6 のワーカ w_i の票は 6 割の確率で有効になり，4 割の確率で無効となる．3.2.3 項で説明した手法をデータセット D_{eva} に含まれるデータに対して行い多数決をとって評価ラベルを付与し，*Accuracy* を算出する．

4.4.2　結果・考察

Model1 と Model2 のそれぞれを用いて算出した一致率をもとに，ワーカが一つのデータに対して投票できる確率を算出して，無効票を決定して多数決をとった結果を表 2 の実験 3 に示す．Model1 と手法を適用した場合と Model2 と手法を適用した場合を比較すると，Model2 と手法を適用した場合の *Accuracy* が高くなっている．しかし，どちらの方法を用いた場合でも，複数人の作業に対する多数決にて集約した結果の *Accuracy* より低くなっている．4.3.2 項にて述べたように，Model2 はワーカの品質に関わらず一致率が高くなりやすくなっている．そのため，$\frac{A_{max}}{2}$ を上回る一致率が多く，ほぼ全てのワーカの確率が 1 に近くなる．一方，Model1 はワーカ本来の品質を求めることができ，それに伴い確率の値にも差がみられた．しかし，票が有効になるか無効になるかの 2 択となっているため，確率の値が低いワーカの作業データに含まれる有用なデータまで取り除かれてしまう．そのため，Model2 を用いた場合よりも *Accuracy* が低くなったと考えられる．

表 3 の実験 3 に示すのは，複数人の作業に対する多数決にて付与された評価ラベルが，3.2.3 項にて説明した手法を用いることによって変化したツイートの例である．一つ目のツイートは，複数人の作業に対する多数決にて付与された評価ラベルは不正解であったが，Model1 と手法を適用して付与されたラベルは正解であった例である．複数人の作業に対する多数決にて付与された評価ラベルは多数派のラベルであったが，Model1 と手法を適用して付与されたのは少数派のラベルであった．このことから，少数派のラベルを付与したワーカの品質が高く，票が有効となる確率が高くなったのではないかと考えられる．

しかし，Model2 と手法を適用して付与された評価ラベルを見ると，複数人の作業に対する多数決にて付与されたラベルと同じであることがわかる．Model2 を用いた場合の方が一致率 A_i が高くなりやすく差が出にくいため，票が有効となる確率 p_i が高いワーカが多くなる．このことから，Model1 を用いた際に正解ラベルと同じラベルが付与されたことは，必ずしも結果が改善されたとは言えないと考えられる．

二つ目のツイートは，複数人の作業に対する多数決にて付与された評価ラベルは正解であったが，Model1 と手法を適用して付与されたラベルは不正解であった例である．元々は多数派であったネガティブの投票数が，少数派であった「ニュートラル」の投票数を下回ったために少数派である評価ラベルが付与された．このことから，有用なデータが無効票となり取り除かれてしまったと考えられる．

5 おわりに

本研究では，クラウドワーカの品質を考慮した結果集約方法を行うことにより，クラウドソーシングの質を向上させることを目的としている．クラウドワーカの中にはスパムワーカと呼ばれるワーカが一定数存在しており，その影響を受けることによって作業依頼者の求める結果を得られないことがある．しかし，スパムワーカは他のワーカとは違う作業を行うため，正しい作業結果と比較することによってスパムワーカを見つけ出し取り除くことが可能である．そこで本稿では，ワーカの模倣モデルを構築して算出する一致率を用いた異なる三つの結果集約方法を提案した．構築した模倣モデルを使用して得られた予測結果をもとに算出した一致率を用いて結果集約を行うことによって，スパムワーカによる影響を抑えられ，作業依頼者の求める結果に近い作業結果を得られることを考えた．提案手法の有効性を確かめるために実験を行った．

一つ目の実験は，逐次的にワーカの品質を測定することによってワーカの選定を行い，取り除かれなかったワーカの作業結果のみを用いて多数決をとる手法を用いて行った．その結果，手法を適用して集約した結果の *Accuracy* は 0.5845 であった．一方，複数人の作業に対する多数決にて集約した結果の *Accuracy* は 0.6625 であり，手法を適用して付与された結果の方が低くなったことがわかった．この結果から，品質の低いワーカとして取り除かれたワーカの作業データの中に含まれる有用な作業データまで取り除かれてしまったと考えられる．

二つ目の実験は，模倣モデルを用いて算出した一致率をもとにワーカの品質を求め，その品質を各ワーカの票の重みとして多数決をとる手法を用いて行った．その結果，Model1 と手法を適用して集約した結果の *Accuracy* は 0.6685 であり，Model2 と手法を適用して集約した結果の *Accuracy* は 0.6625 であった．Model1 を用いた場合の *Accuracy* が高くなった理由は，より各ワーカの特徴を捉えたモデルを構築できたことによって，ワーカ本来の品質を求めることができたためであると考えられる．そのため，品質の低いワーカの作業結果の重みが小さくなり，評価ラベルを付与する際に及ぼす影響力が抑えられた．

三つ目の実験は，一致率をもとにワーカが一つのデータに対して投票できる確率を算出して無効票を決定して多数決をとる手法を用いて行った．その結果，Model1 と手法を適用して集約した結果の *Accuracy* は 0.5020 であり，Model2 と手法を適用して集約した結果の *Accuracy* は 0.5590 であった．どちらのモデルにおいてもワーカごとの一致率が低く，その値をそのまま確率として用いているためほとんどの票が無効票となってしまっている．そのため，有用な作業データも取り除かれてしまったと考えられる．

一つ目と三つ目の手法では，作業データを取り除くというアプローチを行ったため，有用な作業データまで取り除かれてしまうという結果となった．しかし，二つ目の手法では，品質の低いワーカの影響力を抑えるだけで，そのワーカが行った有用な作業データを活用することができる．そのため，他の手法と比べて良い結果を得ることができた．これらの手法は作業の依頼が終了してデータが揃った状態に適用しているため，依頼終了後にしか品質の低いワーカの特定ができない．そのため，品質の低いワーカに割り振られる作業数が増えてしまう．

今後の展望として作業依頼中に適用できるように拡張することを考えている．それによって，品質の低いワーカの特定が早まり，途中でそのワーカを取り除くことができるようになる．そして，品質の低いワーカに割り振られる予定だった作業を品質の高いワーカに割り振る機会が増え，クラウドソーシングの質が向上すると考えられる．

謝辞　本研究の一部は JSPS 科研費 19H04218 および越山科学技術振興財団の助成を受けたものです．

文　献

[1] Jeff Howe, et al. The rise of crowdsourcing. *Wired magazine*, Vol. 14, No. 6, pp. 1–4, 2006.

[2] 西智樹, 小出智士, 大野宏司, 長屋隆之. ソーシャルネットワークを用いたクラウドソーシングの品質向上. 人工知能学会全国大会論文集 第 27 回 (2013), pp. 3M3OS07d4–3M3OS07d4. 一般社団法人 人工知能学会, 2013.

[3] 芦川将之, 川村隆浩, 大須賀昭彦. プライベートクラウドソーシングにおける精度向上手法. 人工知能学会全国大会論文集 第 28 回 (2014), pp. 1J5OS18b4–1J5OS18b4. 一般社団法人 人工知能学会, 2014.

[4] 芦川将之, 川村隆浩, 大須賀昭彦. マイクロタスク型クラウドソーシングプラットフォーム環境における精度向上手法の導入と評価. 人工知能学会論文誌, Vol. 29, No. 6, pp. 503–515, 2014.

[5] Harry Halpin and Roi Blanco. Machine-learning for spammer detection in crowd-sourcing. In *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

[6] 松原繁夫, 水島拓也. クラウドソーシングにおける複数タスク割当て. 人工知能学会全国大会論文集 第 27 回 (2013), pp. 3M4OS07e3–3M4OS07e3. 一般社団法人 人工知能学会, 2013.

[7] 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.

[8] 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ー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論文集 第 27 回 (2013), pp. 2M5OS07b2–2M5OS07b2. 一般社団法人 人工知能学会, 2013.