卒業論文

クラウドソーシングにおける
作業精度推定の不確かさを考慮した結果集約手法

城所 祥太

2024 年 3 月 25 日

岐阜大学 工学部 電気電子・情報工学科 情報コース
鈴木研究室

本論文は岐阜大学工学部に
学士（工学）授与の要件として提出した卒業論文である.

城所 祥太

指導教員：
鈴木 優 准教授

クラウドソーシングにおける
作業精度推定の不確かさを考慮した結果集約手法*

城所 祥太

内容梗概

本研究では，作業精度の高いワーカの作業情報を用いてタスク型クラウドソーシングの結果を集約することにより，高品質な作業結果を出力することを目的とする．ワーカの作業精度を推定する際に，総作業数が少ないワーカの正確な作業精度を推定できない問題がある．そこで我々は，ワーカの作業精度を推定する際の不確かさを考慮したワーカの選出に着目した．本手法では，推定の不確かさを考慮する方法として，確率分布に従うサンプリングの結果を用いる．確率分布に従うサンプリング値を元にワーカの選出を行うことによって，総作業数が多く，作業精度の高いワーカを優先的に選出することができると考えた．我々は，手法の有効性を検証するために実験を行った．実験では，異なる三種類の手法による集約結果を比較し，推定の不確かさを考慮したワーカの選出が高品質な作業結果の出力に有効であるか調査した．

キーワード

クラウドソーシング，確率的データ処理

*岐阜大学 工学部 電気電子・情報工学科 情報コース 卒業論文, 学籍番号: 1203033062, 2024 年 3 月 25 日.

目次

図目次

表目次

第 1 章　はじめに

本研究は，クラウドソーシングにおいて作業精度の高いワーカの作業情報を用いた結果集約によって，高品質な作業結果を出力することを目的とする．クラウドソーシング [1] とは，インターネット上で不特定多数の作業者に作業を依頼する仕組みである．タスク型クラウドソーシングでは，複数のワーカに同じタスクを依頼し，それらの作業情報を用いた多数決によって最終的な作業結果を決定する方法がある．複数の作業情報を用いた多数決結果は，多くのワーカの意見を反映させることができる．そのため，一つのタスクに対して一人のワーカに作業を依頼した場合に比べ，複数の作業情報を用いた多数決結果は信頼度の高い作業結果となる．しかし，不適切な作業を行うワーカを含む場合，多数決結果が作業依頼者の求める作業結果と異なることもある．そこで我々は，作業精度の高いワーカの作業情報を集約することによって，作業依頼者の求める作業結果を得る集約手法を提案する．

本手法では，ワーカの作業情報から確率分布を生成し，その確率分布から生成する乱数を用いた結果集約を行う．この集約方法を用いることによって，推定されるワーカの作業精度の不確かさを考慮することができると考えた．個人の作業情報と作業依頼者による正解の作業結果の一致率が高いワーカは，作業精度の高いワーカであると考えられる．作業精度の高いワーカの作業情報を用いた多数決により，高品質な作業結果を得ることができる．しかし，ワーカの作業精度を推定するためには二つの課題がある．一つ目は，ワーカの作業精度を推定するために必要となる，作業依頼者による正解の作業結果を大量に用意することが難しい点である．二つ目は，総作業数が少ないワーカの正確な作業精度を推定することができない点である．

ワーカの作業精度を推定するために，まず，個人の作業情報と作業依頼者による正解の作業結果を比較する必要がある．十分な精度で作業精度を推定するためには，作業依頼者が正解の作業結果を大量に用意する必要がある．しかし，大量の正解の作業結果を用意することは作業依頼者に大きな負担がかかる．そこで我々は，作業依頼者による正解の作業結果の代わりに，複数の作業情報を用いた多数決結果を利用することができるのではないかと考えた．

我々は，推定される作業精度と本来の作業精度の誤差を考慮したワーカの選出を目指す．以後，推定される作業精度と本来の作業精度の誤差を推定誤差と呼ぶ．

ワーカの総作業数が少ないほど，推定誤差は大きくなるため，推定される作業精度は信頼度が低いといえる．Accuracy や F 値を用いた作業精度推定では，総作業数の違いによる推定誤差を考慮していないため，推定される作業精度の信頼度を評価することができない．そこで，我々は推定誤差を考慮したワーカの作業精度推定を行うために，ワーカの作業情報を元に生成する確率分布に着目した．確率分布とは，ある試行で起こり得るすべての事象（確率変数）の確率を表現したものである．確率分布には，離散型確率分布と連続型確率分布の二種類がある．離散型確率分布とは，確率変数が連続ではなく独立している場合の確率分布である．対して，連続型確率分布とは，確率変数が連続値である確率分布である．

本手法では，推定されるワーカの作業精度を確率変数とする連続型確率分布を用いる．そして，確率分布に従う 0 から 1 の乱数を生成し，生成した乱数をワーカの作業精度と仮定することにより，ワーカ同士の作業精度を比較する．ワーカの作業数が多いほど，生成する確率分布は本来の作業精度に収束する．そのため，この確率分布から生成される乱数は本来の作業精度に近い値となる．一方，ワーカの作業数が少ないほど，生成する確率分布のばらつきは大きくなる．そのため，この確率分布から生成する乱数は本来の作業精度と離れた値を出力しやすくなる．我々は，確率分布から生成する乱数を用いたワーカの選出によって，総作業数が多く，作業精度の高いワーカを優先的に選出することができると考えた．本手法では，各ワーカに対して確率分布を生成し，その確率分布から生成する乱数の値が高いワーカの作業情報を用いた多数決によって最終的な作業結果を決定する．

手法の有効性を検証するために，三種類の集約手法の比較実験を行った．ベースラインとして単純な多数決による方法と，作業情報から求められる F 値が高いワーカの作業結果を用いた多数決により作業結果を決定する方法を用いた．実験の結果，本手法を用いることによる作業精度の高いワーカの作業情報を用いた集約が，ベースラインと比較して高品質な作業結果の出力に有効であることが示された．また，確率分布から生成する乱数を用いることによって，総作業数が多く，作業精度の高いワーカを優先的に選出することができることが分かった．

本論文における貢献は以下のとおりである．

- 確率分布に従うサンプリング値を用いた集約が高品質な作業結果の出力に有効である．

- 確率分布に従うサンプリング値を用いたワーカの選出によって，作業精度が高いワーカのうち，操作行数が多く，推定される作業精度の信頼度が高いワーカを優先的に選出できた．

本論文の構成は以下の通りである．2 章では本論文にて用いた技術や手法の基本的事項について述べる．3 章では関連研究について述べる．4 章では本論文の提案手法について述べる．5 章では提案手法を用いた評価実験の目的や手順，結果・考察などについて述べる．最後に 6 章では本論文のまとめと今後の課題について述べる．

第 2 章　基本的事項

2.1　クラウドソーシング

クラウドソーシングとは，インターネット上で不特定多数のワーカに作業を依頼するシステムである．クラウドソーシングには，プログラミングや WEB デザイン，動画編集など，様々な仕事がある．また，教師あり学習に用いられる教師データの生成の際にクラウドソーシングが用いられることがある．クラウドソーシングを用いる作業依頼者側のメリットは，コストが抑えられる点や，短期的に必要なスキルを調達しやく，業務の効率化を図ることができる点である．また，作業者側のメリットは，好きな時間に作業ができる点や，仕事を選ぶことができる点である．

2.2　確率分布

確率分布とは，ある試行において起こり得る全ての事象の発生確率を表現したものである．この時，ある試行において起こり得る事象を確率変数という．確率分布は，確率変数が連続か不連続かによって離散型確率分布と連続型確率分布の二種類に分けられる．以下に二種類の確率分布の説明を示す．

2.2.1　離散型確率分布

離散型確率分布とは，確率変数が離散値である確率分布である．離散型確率分布における確率変数 X がある値 x を取るときの確率 $f(x)$ は確率質量関数と呼ばれる．確率質量関数は，出力がある事象の発生確率であることに注意が必要である．図 2.1 に，離散型確率分布を用いた例を示す．図 2.1 は，さいころを一回投げる場合に，出る目を $x_i(i = 1, 2, 3, 4, 5, 6)$ とするとき，全ての x_i の確率を示したものである．

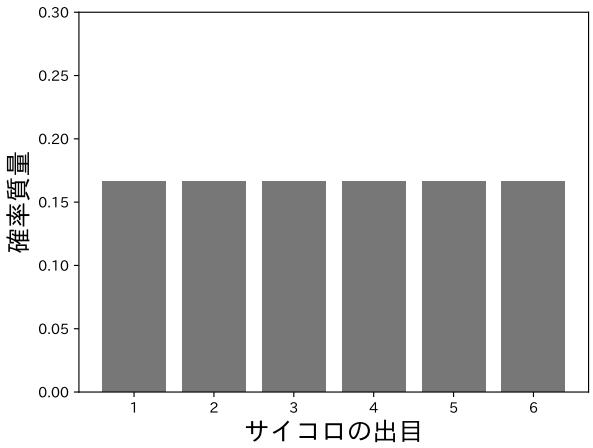

図 2.1　離散型確率分布の例

2.2.2　連続型確率分布

連続型確率分布とは，確率変数が連続値である確率分布である．連続型確率分布において，確率変数 X がある値 x を取るときの確率は 0 である．これは確率変数 X が連続値であるため，発生し得る事象が無限に存在するからである．連続型確率分布における確率変数 X がある値 x を取るときの確率密度 $f(x)$ は確率密度関数と呼ばれる．確率密度関数は，出力がある事象の発生確率を表しているわけではないことに注意が必要である．連続確率分布は，確率変数 X がある範囲内にどれくらいの確率で存在し得るかを求める際に用いられる．確率密度関数 $f(x)$ において，$a \leq X \leq b$ となる確率 P は以下の計算によって求められる．

$$P(a \leq X \leq b) = \int_a^b f(x)dx \tag{2.2.1}$$

連続型確率分布には様々な種類が存在するが，その中の一例としてベータ分布について説明する．

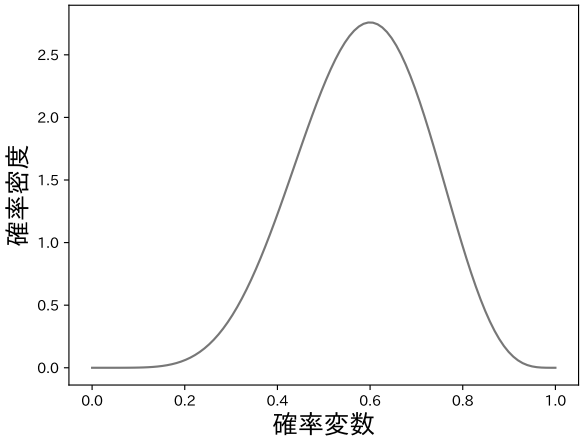

図 2.2　ベータ分布の例

2.2.3　ベータ分布

ベータ分布とは，連続型確率分布の一つである．ベータ分布は，確率変数の区間が 0 から 1 の間で定義されるため，確率や比率などを確率変数とする場合に用いられることがある．ベータ分布が用いられる例として，成功数 a 回と失敗数 b 回が分かっている試行に関して、成功率 p の分布を表す。ベータ分布の従う確率密度関数はベータ関数を用いることで求められる．確率変数 p，成功回数 a, 失敗回数 b のベータ関数 $\beta(a,b)$ を式 (2.2.2) に，ベータ分布の従う確率密度関数 $f(p)$ を式 (2.2.3) に示す．

$$\beta(a,b) = \int_0^1 p^{a-1}(1-p)^{b-1}dp \tag{2.2.2}$$

$$f(p) = \frac{p^{a-1}(1-p)^{b-1}}{\beta(a,b)} \tag{2.2.3}$$

a=6,b=4 の時のベータ分布を図 2.2 に示す．

2.3 バンディットアルゴリズム

経験を蓄積する「探索」と経験を活用して行動する「実践」の組み合わせによって，限られた行動の中で報酬を最大化するアルゴリズムである．複数の事前情報のない確率機（アーム）がある場合に期待値の高いアームを見つけ出し，報酬を最大化するために考えられた方法である．バンディットアルゴリズムは，web の広告システムなどで用いられるアルゴリズムである．バンディットアルゴリズムには，様々な種類の手法が存在するが，本稿では ε -greedy とトンプソンサンプリングについて説明する．

2.3.1 ε -greedy

ε -greedy とは，ある一定の確率でランダムに探索を行い，その経験を元に実践を繰り返す手法である．平常時は過去の経験から最適だと考えられる行動をとり続けるが，ある一定の確率で過去の経験に依存しない完全にランダムな行動をとり，経験を蓄積する．これによって，ある偏った経験に基づく実践を，期待値の高い方向へ修正しながら行動を選択することができる．ランダムな探索を行う確率の最適な値は定まっていないが，様々な知見からおよそ 10 %の確率で探索を行うと良い傾向にあると考えられている．ε -greedy のデメリットは，探索の際に過去の経験から高確率で期待値の低い行動だと分かっているとしても，ランダムな探索によって選ばれる可能性がある点である．これは，探索時の報酬を減らすことに繋がるのに加え，探索の効率が悪いため，実践の報酬の最大化が遅くなってしまう．

2.3.2 トンプソンサンプリング

トンプソンサンプリングとは，各アームにおいて，経験から得られる情報を元に確率分布を生成し，その確率分布に従うサンプリング値が高いアームを選択する方法である．サンプリングとは，母集団や確率分布から標本を抽出する操作のことをいう．サンプリングを行うことによって，分布に従う乱数を生成させることができる．確率分布に従うサンプリングを行う場合，抽出される値は確率変数 X の取り得るある値 x である．トンプソンサンプリングの特徴は，全ての行動が探索であ

り，実践であるということである．総試行回数の多いアームはサンプリング値が安定しているが，総試行回数が少なく，正確な推定が行えていないアームはサンプリング値が不安定である．そのため，基本的には過去の経験から期待値の高いアームを選択しようとするが，たまに期待値から外れたサンプリング値を出力するアームを選択する．これによって総試行回数が少なく，正確な推定が行えていないアームに対して探索を行うことができ，各アームの期待値を高精度で比較することができる． ε-greedy のデメリットであった，探索の際に過去の経験から高確率で期待値の低い行動だと分かっているとしても，ランダムな探索によって選ばれる可能性がある点を克服している点がトンプソンサンプリングの優れている点である．また，探索を行う最適な確率が不明な ε-greedy に比べて，トンプソンサンプリングはそのような変数がないため，様々なシチュエーションに応用しやすい点でも優れているといえる．

2.4 評価指標

本論文では，手法の評価のために Accuracy と F−score の二つの評価指標を用いる．F−score を算出するのに必要となる recall,precision に関しても説明する．予測ラベルと正解ラベルの混同行列を表 2.1 に示す．表中の TP は True Positive の略で，陽性ラベルであるものを正しく陽性であると予測したものである．FN は False Negative の略で，陽性ラベルであるものを誤って陰性であると予測したものである．FP は False Positive の略で，陰性ラベルであるものを誤って陽性であると予測したものである．表中の TN は True Negative の略で，陰性ラベルであるものを正しく陰性であると予測したものである．

表 2.1 正解ラベルと予測ラベルのの混同行列

		予測ラベル	
		陽性	陰性
正解ラベル	陽性	TP	FN
	陰性	FP	TN

2.4.1 Accuracy

Accuracy とは, 正解のラベルを付与できた割合を示す評価指標である．Accuracy は，正解のラベルを付与した回数を総作業数で割ることによって算出できる．表 2.1 を用いて算出の式を示す.

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{2.4.1}$$

正解ラベル間に偏りがある場合，Accuracy を用いた評価では，正当な評価が行うことができない点に注意が必要である.

2.4.2 precision

precision とは，適合率のことをいい，例えば，予測ラベルが陽性であったもののうち，実際に正解ラベルが陽性であったものの割合を示す．表 2.1 を用いて算出の式を示す.

$$Precision = \frac{TP}{TP + FP} \tag{2.4.2}$$

precision は予測がどれくらい間違えていたかを評価する際に用いられる.

2.4.3 recall

recall とは，再現率のことをいい，例えば，正解ラベルが陽性であったもののうち，実際に予測ラベルが陽性であったものの割合を示す．表 2.1 を用いて算出の式を示す.

$$Recall = \frac{TP}{TP + FN} \tag{2.4.3}$$

recall は予測した際に，ある正解ラベルがどれくらい取りこぼされるかを評価する際に用いられる.

2.4.4 F-score

F-score とは，precision と recall の調和平均である．precision と recall を用いて算出の式を示す．

$$F = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{2.4.4}$$

precision や recall のみで判断すると，予測に偏りがある場合でも値が高くなる場合があり，正しい判断ができない．F-score を用いることによって，precision や recall のバランスを評価し，予測の偏りに依存しない分類精度の評価を行うことができる．

第 3 章　関連研究

タスク型クラウドソーシングの作業結果の質の向上には，様々なアプローチが存在する．例えば，Zhang ら [2] や Ho[3] らの研究のように, ワーカにタスクを割り当てる際に工夫を施す方法や，Chandler[4] らや Gillier[5] らのようにインストラクションを改善することによってワーカの作業精度を向上させる方法などがある．様々なアプローチのうち，我々は集まった複数の作業情報の集約の際に，作業精度の高いワーカの作業情報を元に結果集約することによって，高品質な作業結果の出力を目指す．

複数の作業情報の集約において，ワーカの作業精度を考慮した結果集約手法の研究はいくつか行われている．Dawid と Skene ら [6] はワーカの作業精度にばらつきがあることと作業情報の信頼度を考慮した集約手法を紹介している．彼らの手法は，EM アルゴリズムを用いることによって，ワーカの作業精度と作業情報の信頼度の両方を推定する．この方法で推定した各ワーカの作業精度を元に作業結果を決定することで，高品質な作業結果が得られることを示した．Whitehill ら [7] は，タスクの難易度を考慮した集約手法を紹介している．タスクの難易度とワーカの作業精度の関係性に着目し，作業情報を決定することによって，最終的な集約結果の品質が向上することを示した．Welinder ら [8] は，ワーカとタスクの相性を考慮した手法を紹介している．ワーカの作業の判断基準の傾向とタスクの潜在特徴との関係性を捉えることが集約結果の品質向上につながることを示した．本手法は，作業精度の高いワーカを選出し，その作業情報を用いた多数決によって作業結果を決定する．作業精度の高いワーカを選出するうえで，ワーカの作業精度の推定誤差を考慮したワーカの選出を行うという点で上記の研究とは異なる．

ワーカの作業精度を推定する上で，総作業数が少ないワーカの正確な作業精度を推定できない問題を考慮した集約に関する研究も存在する．Venanzi ら [9] は，各ワーカはいくつかのカテゴリーに分類できると仮定し，ワーカをカテゴリー分類した結果を用いる集約手法について紹介している．ワーカの作業精度を考慮した集約は有効である傾向が強いが，総作業数が少ないワーカの推定される作業精度は信頼度の高いものであるとは言えない，そこで，各ワーカはいくつかのカテゴリーに分類できると仮定し，混同行列によって各ワーカをそのカテゴリーごとに分類した．

総作業数が少ないワーカであっても他のワーカと同等レベルで作業精度を推測でき，集約結果の品質向上に有効であることを示した．小山ら [10] らは，確信度を用いた集約手法を紹介している．ワーカが報告した各作業に対する確信度をパラメータとして用いることで，ワーカの作業精度の推定精度が向上することを示した．本手法は，各ワーカに対して，作業精度を確率変数とする確率分布を生成し，その確率分布に従うサンプリング値を用いた集約を行う．確率分布を用いることで，作業精度の推定誤差を考慮したワーカの選出が可能になると考える．本手法は，作業精度の高いワーカの選出に，確率分布に従うサンプリング値を用いるという点で上記の研究とは異なる．本研究は，複数の作業情報を用いた多数決結果と，ワーカの作業情報から確率分布を生成し，その確率分布に従うサンプリング値を用いた集約の有効性を検証する．

第 4 章　提案手法

我々は，確率分布から生成する乱数を用いた結果集約手法を提案する．個人の作業情報と作業依頼者による正解の作業結果との一致率が高いワーカは作業精度の高いワーカである．本研究の目的は，作業精度の高いワーカの作業情報を集約することによって，高品質な作業結果を出力することである．しかし，作業精度の高いワーカを選出する上で，総作業数の少ないワーカの作業精度の推定が不正確である問題がある．そこで我々は，確率分布から生成する乱数を用いたワーカの選出に着目した．本手法の手順を以下に示す．

Step1　各ワーカの正解作業数（正解の作業結果と同じ作業情報である作業数）と不正解作業数（正解の作業結果と異なる作業情報である作業数）を求める．

Step2　Step1 で求めた結果を元に，各ワーカの確率分布を生成する．

Step3　あるタスク n を作業した複数のワーカに対して，各ワーカの確率分布から乱数を生成をする．その乱数の値が高い数人の作業情報を用いた多数決によってタスク n の作業結果を決定する．この操作を各タスクに対して行い，全てのタスクの作業結果を決定する．

これらの手順の詳細について各節にて説明する．

4.1　作業情報の集計

本節では，複数の作業情報を用いた多数決結果とワーカの作業情報の比較することによってワーカの作業情報情報を集計する方法について説明する．作業情報の集計で求める出力は，正解作業数（正解の作業結果と同じ作業情報である作業数）と不正解作業数（正解の作業結果と異なる作業情報である作業数）である．作業情報の集計は，ワーカの作業情報と作業依頼者による正解の作業結果と比較する．しかし，十分な精度で作業精度を推定するためには，作業依頼者による正解の作業結果を大量に用意する必要がある．しかし，大量の正解の作業結果を用意することは作業依頼者に大きな負担となる．そこで我々は，作業依頼者の求める作業結果の代わりに，複数の作業情報を用いた単純多数決結果を使用できるのではないかと考え

た．タスク型クラウドソーシングは，同一タスクを複数のワーカに割り当て，それらの作業情報を用いた多数決によって最終的な作業結果を決定する．複数の作業情報を用いた集約は，多くのワーカの意見を反映させた作業結果を出力する．そのため，一人で作業した結果と比べて，複数の作業情報を用いた集約結果のほうが信頼度の高い作業結果となる．同一タスクに割り当てるワーカの数を増やすほど，それらの作業情報を用いた集約結果は作業依頼者の求める作業結果と一致しやすくなると考えられている．以後，複数の作業情報の単純多数決結果を投票結果とする．作業依頼者の求める作業結果の代わりに投票結果を用いることは，作業依頼者の負担を軽減することにつながる．我々は，作業依頼者の求める作業結果の代わりに投票結果を使用することの妥当性を確認するため，5.3.1 節にて紹介するデータセットを用いて，作業依頼者の求める作業結果と投票結果の一致率を求めた．一致率の算出手順を以下に示す．

Step1　各タスクに対して，集まった複数の作業情報を用いた多数決によって投票結果を決定する．

Step2　全てのタスクのうち，無作為に抽出した 500 件のタスクに対して作業依頼者が正解の作業結果を付与する．

Step3　投票結果と Step2 で付与した正解の作業結果を比較する．

Step4　正解の作業結果と投票結果が同じタスクの件数を正解の作業結果を付与した件数である 500 件で割ることにより一致率を求める．

上記の手順により 5.3.1 節のデータセットにおける作業依頼者の求める作業結果と投票結果の一致率（Accuracy）は 0.894 であった．作業精度を推定するうえで，作業依頼者の求める作業結果と投票結果の一致率は十分に高い．結果から我々は，作業依頼者の求める作業結果の代わりに投票結果を使用できると考えた．以上の理由より本手法では，個人の作業情報と投票結果を比較することによって各ワーカの作業情報を集計する．

総ワーカ数を N 人とするとき，ワーカ $i(i = 1, 2, \cdots, N)$ に対して作業情報の集計で求める出力は，正解作業数 a_i（正解の作業結果と同じ作業情報である作業数）と不正解作業数 b_i(正解の作業結果と異なる作業情報である作業数) である．作業情報の集計の概要を図 4.1 に示す．

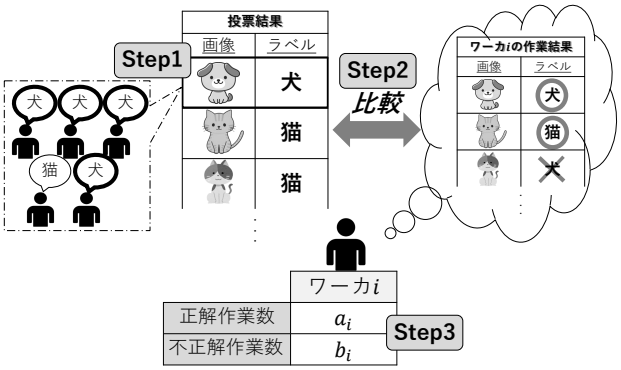

図 4.1　作業情報の集計の概要

Step1　各タスクに対して，集まった複数の作業情報を用いた多数決によって投票結果を決定する．

Step2　ワーカ i が作業した全てののタスクに対して，ワーカ i の作業情報と投票結果を比較する．

Step3　投票結果と同じ作業情報であったワーカ i の作業数を正解作業数 a_i とする．投票結果と異なる作業情報であったワーカ i の作業数を不正解作業数 b_i とする．

Step2 と Step3 の操作をすべのワーカに対して行う．

4.2　結果集約手法

本節では，確率分布から生成する乱数を用いた結果集約手法について説明する．4.2.1 節では，4.1 節にて集計した各ワーカの正解作業数と不正解作業数を元に生成する確率分布について述べる．4.2.2 節では，4.2.1 節にて紹介する確率分布に従うサンプリングとサンプリング値を用いた集約手法について述べる．

4.2.1　確率分布の生成

本手法は作業精度の高いワーカを選出するために，作業精度を確率変数とするベータ分布を用いる．

作業精度の高いワーカを選出には，各ワーカの作業精度を推定する必要がある．ここで我々は，総作業数が少ないワーカの作業精度を推定する場合に注意が必要である．なぜならば，総作業数が少ないワーカの作業情報から推定される作業精度は本来の作業精度とは大きく異なる場合があるからである．本稿では, 推定した作業精度と本来の作業精度の間に生じる誤差を推定誤差と呼ぶ．作業精度の推定誤差が大きい場合，推定される作業精度は信頼度が低いといえる．Accuracy や F 値を用いた作業精度の推定では，ワーカの総作業数を考慮していないため，推定される作業精度の信頼度を評価することができない．そこで我々は，確率分布による作業精度の推定に着目した．

確率分布とは，ある試行で起こり得る全ての事象（確率変数）の発生確率を分布を用いて表したものである．確率分布には，離散型確率分布と連続型確率分布がある．離散型確率分布とは，確率変数が $0, 1, 2$ のような離散値である場合の確率分布である．連続型確率分布とは，確率変数が連続値である場合の確率分布である．連続型確率分布において，確率変数 X がある値 x を取るときの確率は 0 である．これは確率変数 X が連続値であるため，発生し得る事象が無限に存在するからである．連続型確率分布における確率変数 X がある値 x を取るときの確率密度 $f(x)$ は確率密度関数と呼ばれる．確率密度関数は，出力がある事象の発生確率を表しているわけではないことに注意が必要である．連続確率分布は，確率変数 X がある範囲内にどれくらいの確率で存在し得るかを求める際に用いられる．確率密度関数 $f(x)$ において，$a \leq X \leq b$ となる確率 P は以下の計算によって求められる．

$$P(a \leq X \leq b) = \int_a^b f(x)dx \tag{4.2.1}$$

本手法では，ワーカ i の作業精度を確率変数とする連続型確率分布を用いることによって，ワーカ i の作業精度を推定する．ワーカの作業数が多いほど，生成する確率分布は本来の作業精度に収束する．一方，ワーカの作業数が少ないほど，生成する確率分布は分散が大きく，推定される作業精度のばらつきが大きいことを表現できる．上記の理由より我々は，確率分布を用いることによって，ワーカごとの作業精度推定の不確定さを考慮した集約が可能なのではないかと考えた．

確率分布には様々な種類が存在するが，本手法ではベータ分布を使用する．ベータ分布とは，連続型確率分布の一つであり，確率変数の区間が 0 から 1 の間で定

義される．ベータ分布は，成功回数 a 回と失敗回数 b 回が分かっている試行に関して、成功確率 p の分布を表すことができる。ベータ分布を使用する理由は，4.1 節でワーカの作業情報を正解の作業を行ったか否かの 2 値で分類しており，いかなるタスクであってもワーカの作業情報をベルヌーイ過程で表現できるからである．ベルヌーイ過程とは，2 種類の独立した結果のみを出力する試行のことである．本手法は，ワーカ i の推定される作業精度を確率変数とするベータ分布を用いることによって，ワーカ i の作業精度を推定する．ベータ分布の従う確率密度関数はベータ関数を用いることで求められる．ワーカ i の作業精度を p_i，正解作業数を a_i, 不正解作業数を b_i とするときのワーカ i のベータ関数 $\beta(a_i, b_i)$ を式 (4.2.2) に，ベータ分布の従う確率密度関数 $f(p)$ を式 (4.2.3) 示す．

$$\beta(a_i, b_i) = \int_0^1 p_i^{a_i-1}(1-p_i)^{b_i-1}dp_i \tag{4.2.2}$$

$$f(p_i) = \frac{p_i^{a_i-1}(1-p_i)^{b_i-1}}{\beta(a_i, b_i)} \tag{4.2.3}$$

作業精度 p が 0.6 のワーカに関して，総作業数が 10 回（正解作業数：6 回，不正解作業数：4 回）の場合と，100 回（正解作業数：60 回，不正解作業数：40 回）の場合のベータ分布を図 4.2 に示す．図 4.2 から，総作業数が少ないワーカの推定される作業精度はばらつきが大きく，総作業数が多いワーカの推定される作業精度はばらつきが小さいことが分かる．

4.2.2　サンプリング値を用いた集約

本手法では，4.2.1 節で紹介した確率分布に従うサンプリングを行う．そのサンプリング値をワーカの作業精度と仮定することによって．ワーカ同士の作業精度を比較し，作業精度の高いが考えられるワーカを決定する．サンプリングとは，確率分布から標本を抽出することをいう．つまり，サンプリングを行うことによって確率分布に従う乱数を生成することができる．サンプリングで得られる値は，確率分布に従って抽出される確率変数 x である．4.2.1 節で紹介したベータ分布は，各ワーカの推定される作業精度とその推定誤差を表現することができる．ワーカの作

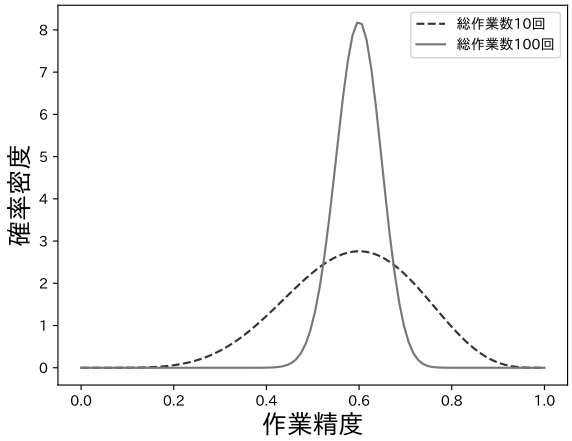

図 4.2　作業精度 $p = 0.6$ のワーカのベータ分布

業数が多いほど，生成する確率分布は本来の作業精度に収束するため，サンプリング値は本来の作業精度付近の値になる．例えば，図 4.2 の実線のような総作業数が多いワーカから生成する確率分布上でサンプリングを行うと，0.60 や 0.65 など本来の作業精度に近い値を出力しやすい．そのため，総作業数が多く，作業精度の高いワーカはサンプリング値が高い値を出力しやすい．一方，ワーカの作業数が少ないほど，生成する確率分布のばらつきが大きくなるため，サンプリング値は本来の作業精度から離れた値になりやすい．図 4.2 の点線のような総作業数が少ないワーカから生成する確率分布上でサンプリングを行うと，0.40 や 0.80 など本来の作業精度から離れた値を出力する確率が高くなる．そのため，総作業数が少なく，推定される作業精度の信頼度の低いワーカのサンプリング値は不安定であり，作業精度の高いワーカとして選出されにくくなる．以上の理由より，確率分布に従うサンプリング値を元にワーカの選出を行うことによって，総作業数が多く，作業精度の高いワーカを優先的に選出でき，高品質な作業結果の出力に繋がると考えた．サンプリング値を用いた集約手法の手順を以下に示す．

Step1　4.2.1 節の式 (4.2.2) から各ワーカの確率分布を得る．

Step2　あるタスク n を作業した複数のワーカに対して，各ワーカの確率分布に従うサンプリングを行う．

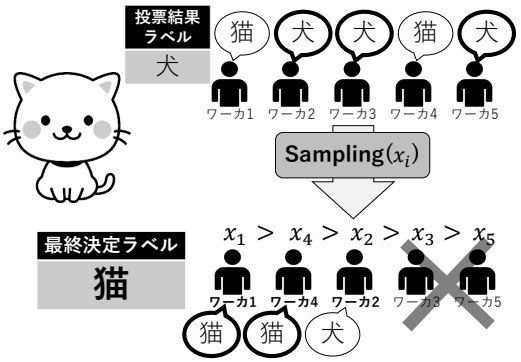

図 4.3　サンプリング値を用いた集約の例

Step3　サンプリング値の高い上位数人のワーカの作業情報のみで多数決を行い，タスク n の最終的な作業結果を決定する．

Step2 と Step3 の操作をすべのタスクに対して行う．Step3 において，多数決を行う際に選ぶ上位人数は，データセットによって適当な人数を用意する必要がある．多数決を行う上位人数を 3 人とした時の集約例を図 4.3 に示す．

確率分布に従うサンプリング値の利用は，推定される作業精度の信頼度の高いワーカの選出に有効であると考える．総作業数が多く，作業精度の高いワーカの確率分布に従うサンプリングは，高い値を出力する確率が非常に高い．そのため，Step3 にて選ばれる可能性は高くなる．対して，総作業数が少なく，推定される作業精度の信頼度の低いワーカはサンプリング値にばらつきが大きく，Step3 にて選ばれる可能性が下がる．これらの理由より，サンプリング値を用いた集約は推定される作業精度の信頼度の高いワーカを優先的に選出することができる．

第 5 章　評価実験

本実験の目的は二つある．一つ目は，本手法を用いた集約が高品質な作業結果の出力に有効であるかどうかを検証することである．二つ目は，サンプリング値を用いることによって，作業精度が高いと推定されるワーカのうち，総作業数が多く，作業精度が高いワーカを優先的に選出することができるかどうかを検証することである．本実験では，ベースラインとして単純な多数決による集約手法，F 値により作業情報を評価する集約手法と本手法であるサンプリング値を用いた集約手法の三種類の集約手法を比較する．

5.1　データセット

本実験では，岐阜大学鈴木研究室とソニー株式会社の共同研究にて構築したデータセットを用いる．このデータセットは，クラウドソーシングを用いて構築した．このデータセットを構築した目的は，テキストデータに含まれるある製品に対する要望，体験，感情について分析することである．このテキストデータは，ある製品に関する動画のコメントである．データセットの構築に伴い，一つのテキストデータあたり五つの質問を設定した．質問の項目を以下に示す．

Q1　テキストデータがある製品に対する要望，体験，感情が含まれるか．
　　回答の選択肢は「含む」「含まない」の二種類である．
Q2　テキストデータがある製品の何について書かれているか．
　　回答の選択肢は「本体」「品質・不具合」「アクセサリ・周辺機器」「その他」「該当なし」の五種類である．
Q3　テキストデータを投稿した著者はどのようなある製品に対してどのような感情を持っているか．
　　回答の選択肢は「ポジティブ」「ネガティブ」「ニュートラル」「ポジティブ&ネガティブ」の四種類である．
Q4　テキストデータを投稿した著者はある製品を使用したことがあると考えられるか．

回答の選択肢は「はい」「いいえ・不明」の二種類である．

Q5　テキストデータ内で，ある製品は何の製品と比較されているか．

回答の選択肢は「同社の製品」「他社の製品」「比較なし」の三種類である．

ただし，Q1 で「含まない」と回答した場合，後の四つの質問を行わないこととした．

本実験で用いるデータセットは 606 人のワーカの作業情報をまとめたものである．各作業情報は，「ワーカ ID」「テキストデータ」「質問番号」「ワーカが付与したラベル」の四種類の項目で構成されている．また，データセットの構築において対象としたテキストデータ数は 26,200 件である．データセットの構築にあたり，多くのワーカの意見を反映したラベル決定を行うため，一つのテキストデータに対して 10 人のワーカに作業を依頼した．この 10 人のワーカの作業情報を用いた多数決によって決定するラベルを投票結果とする．

集約精度の評価を行うにあたって，集約結果と作業依頼者による正解の作業結果を比較する．そのため，各テキストデータに対して作業依頼者が正解ラベルを付与したデータセットを作成必要がある．作業依頼者は，投票結果が付与されたテキストデータのうち，無作為に抽出したテキストデータに対して正解ラベルを付与した．無作為に抽出したテキストデータのうち正解ラベルを付与した 500 件のテキストデータを正解データセットとする．全てのテキストデータに対して正解ラベルを用意するべきだが，コストがかかるため本実験では 500 件とした．正解データセットを作成する際，作業依頼者の負担を軽減するため，正解ラベルの決定に迷うテキストデータは取り除いた．本実験では，作業依頼者は著者であるとして，正解データセットを構築した．正解データセットに含まれる 500 件のテキストデータに対して 5.2 節にて紹介する三種類の集約手法を適用し，集約精度を算出する．

5.2　実験設定

本実験では，異なる三種類の手法の集約精度を比較することによって，本手法の集約精度を評価する．手法の集約精度とは，手法を用いて決定したラベルと正解ラベルの一致率である．本実験では，一致率の指標として F 値を用いる．三種類の手法について以下で説明する．

手法 1　各テキストデータに対して，集まった複数の作業情報を用いた多数決によって作業結果を決定する．この時，集まった全作業情報を用いた多数決によって決定した作業結果を投票結果とする．

手法 2　各ワーカに対して，個人の作業情報と投票結果を用いることにより F 値を求め，その値を作業精度とする．各テキストデータに対して，F 値の高い上位数人の作業情報を用いた多数決によって作業結果を決定する．

手法 3　本研究の提案手法であるサンプリング値を用いた集約によって作業結果を決定する．

手法 1 の集約精度を各実験のベースラインとする．ベースラインと提案手法の比較により，提案手法が高品質な作業結果の集約に有効であるかどうか検証する．

手法 2 は，作業精度が高いワーカを選出するうえで，作業精度の推定の不確定さを考慮していない集約方法である．4.2.1 節にて，F 値による作業精度の推定では推定の不確定さを考慮できていない問題があると述べた．対して提案手法は，確率分布に従うサンプリング値の利用により，推定の不確定さを考慮した集約を行うことができる．F 値により作業精度を評価する集約手法と提案手法の比較により，提案手法が推定の不確定さを考慮した集約を行うことができているかどうかを検証する．加えて，推定の不確定さを考慮した集約が高品質な作業結果の出力に有効であるかどうか検証する．手法 2 の集約手順を以下に示す．

Step1　各ワーカに対して，個人の作業情報と投票結果用いることにより F 値を求め，その値を作業精度をする．

Step2　あるテキストデータ n に対して作業したワーカの中で F 値の高い上位数人を選出する．

Step3　選出した数人の作業情報のみで多数決を行い，テキストデータ n の最終的なラベルを決定する．

Step1 から Step3 を全てのタスクに対して行う．

手法の集約精度を求める手順を以下に示す．

Step1　データセットの全てのテキストデータに対し，集約手法を用いて作業結果を決定する．

Step2　Step1 の結果と作業依頼者による正解の作業結果を用いることにより集約精度（F 値）を求める．

上記の手順で，手法 1，手法 2，手法 3 の集約精度を求める．ただし，手法 1 と手法 3 を用いる場合は，上記の手順を 10 回行い，それらの平均を取ることによって集約精度を決定する．手法 2 に関しては，同じ操作をした場合，集約精度に変化がないため，上記のような複数回の操作から平均を求める作業を行わない．

手法を適用をする際，何人のワーカの作業情報を用いると適切なのかを調査する．そのため，多数決に使用するワーカの数を変えて集約を行い，集約精度を比較する．この時，各手法において最も高かった集約精度をその手法の集約精度とする．求めた三種類の手法の集約精度を比較することによって，本手法の有効性を検証する．

5.3　実験 1，「要望，体験，感情が含まれるか」

5.3.1　実験条件

実験 1 では，5.1 節にて用意したデータセットのうち，Q1 の質問に関して作業を行った作業情報のみを抽出し構築したデータセットを用いる．作業によって付与されるラベルは体験・感情・要望を「含む」「含まない」の 2 種類である．上記のデータセットは，一つのテキストデータに対し，10 人のワーカが同じ作業をしている．そこで各テキストデータに付与するラベルを，作業した 10 人のワーカの作業情報を用いた多数決によって決定する．10 人のワーカの作業情報を用いた多数決によって決定したラベルを投票結果とする．あるテキストデータに関して，投票結果が「含む」「含まない」のどちらかに決まる場合は，そのラベルを最終的に付与するラベルして決定する．あるテキストデータに関して，投票結果が「含む」「含まない」のどちらかに決まらない場合は，最終的に付与するラベルを「含む」とした．この操作によって構築されたデータセットのラベル別のテキストデータ総数は，「含む」が 6,829 件，「含まない」が 19,371 件であった．投票結果が付与されたデータセットと各ワーカの作業情報を比較し，4.1 節にて説明した正解作業数と不正解作業数を求める．

このデータセットを用いた実験は，選出するワーカ人数を 1 人，3 人，5 人，7

人，9 人の 5 パターンで行った．これは，2 値分類において作業情報の多数決の結果が一意に決まる人数である．ベースラインである単純多数決は，無作為に抽出した数人のワーカの作業結果を用いた多数決によって作業結果を決定する．F 値によりワーカの作業精度を評価する集約手法では，F 値の高い上位数人の作業情報を用いた多数決によって作業結果を決定する．本手法である，確率分布に従うサンプリング値を用いた集約手法では，サンプリング値の高い上位数人の作業情報を用いた多数決によって作業結果を決定する．

5.3.2　結果・考察

5.2 節にて説明した三種類の手法を用いた集約結果と正解データセットの一致率を F 値で算出し，表 5.3.2 に示す．ベースラインである単純多数決による集約は，選出人数が増えるほど精度が高くなっている．多数決結果が 5:5 に割れた場合，「含む」というラベルを付けたが，実際は正解ラベルが「含まない」であるデータが多かったため，10 人で集約した時に少し精度が下がったと考えられる．以後，ベースラインである単純多数決の集約精度は最も高かった 0.853 を採用し，他の手法の集約精度と比較を行う．

本手法である，サンプリング値を用いた集約手法の集約精度は，選出人数が 5 人の時に最も高く，0.892 であった．ベースラインの 0.853 と比べて本手法の集約精度が高いことから，サンプリング値を用いた集約は高品質な作業結果の出力に有効であることが分かる．これは，サンプリング値が高いワーカが実際に作業精度が高いワーカであり，かつサンプリング値が低いワーカが実際に作業精度が低いワーカであったためだと考えられる．また，本手法を用いた際，選出人数が 9 人の場合の

表 5.1　実験 1 における各手法を用いた際の F 値

	選出人数（人）					
	1	3	5	7	9	10
ベースライン	0.737	0.796	0.827	0.842	**0.853**	0.843
F 値上位	0.839	0.871	**0.874**	0.850	0.855	0.843
サンプリング	0.857	0.890	**0.892**	0.873	0.870	0.843

集約精度と 10 人の場合の集約精度では大きな差がある．これは，サンプリング値が低いワーカが高確率で作業精度の低いワーカであり，本手法を用いることによってそのようなノイズとなる作業情報を取り除くことができるといえる．

F 値により作業情報を評価する集約手法では，選出人数が 5 人の時に最も集約精度が高く，0.874 であった．本手法を用いた場合の作業精度は 0.892 であり，F 値により作業情報を評価する集約手法に比べて優れているといえる．また，本手法と F 値により作業情報を評価する集約手法における作業精度が高いと推測した上位 5 人のワーカに対して作業数の平均を求めた．本手法におけるサンプリング値が高い上位 5 人のワーカの平均作業数は 2,900 件であった．また，F 値により作業情報を評価する集約手法における F 値が高い上位 5 人のワーカの平均作業数は 2,747.7 件であり，本手法を用いた場合のほうが高い値であった．以上の結果から，サンプリング値が高いワーカを選出することによって，総作業数が多く，作業精度が高いワーカを優先的に選出することができたことが分かった．

また，本手法を用いる際，選出する上位人数が 5 人の時に集約精度が最も高いことから，およそ半分の作業情報は多数決の集約においてノイズであることが分かる．このようなノイズとなる作業を行うワーカを事前に取り除くことで，コスト削減とさらなる集約精度向上に繋がると考える．

各手法において選出人数が 1 人の場合の集約精度はあまり高くない．これは，高精度で最も集約精度が高いワーカを選出できたとしても，そのワーカ 1 人の作業結果のみでは，高品質な作業結果の出力は困難であることがいえる．作業精度の高い複数のワーカの作業情報を集約することが高品質な作業結果の出力に有効だと考える．

5.4　実験 2，「どのような感情を持っているか」

5.4.1　実験条件

実験 1 では，5.1 節にて用意したデータセットのうち，Q3 の質問に関して作業を行った作業情報のみを抽出し構築したデータセットを用いる．作業によって付与されるラベルは「ポジティブ」「ネガティブ」「ニュートラル」「ポジティブ&ネガティブ」の四種類である．本実験では，ラベルをポジティブな感情を「含む」「含まな

い」の二種類に変換し，データセットを用いた．「ポジティブ」と「ポジティブ&ネガティブ」はポジティブな感情を含むとし，「ネガティブ」「ニュートラル」はポジティブな感情を含まないとした．上記のデータセットは，一つのテキストデータに対し，10 人のワーカが同じ作業をしている．しかし，Q1 の質問にて「含まない」と回答した場合，Q3 の質問は行われないため，作業情報がない場合も存在する．一つのテキストデータに対し，5 人以上のワーカが Q3 の質問を回答したテキストデータに関して，付与するラベルを，作業した複数の作業情報の多数決によって決定する．複数の作業情報を用いた多数決によって決定したラベルを投票結果とする．あるテキストデータに関して，投票結果がポジティブな感情を「含む」「含まない」のどちらかに決まる場合は，そのラベルを最終的に付与するラベルして決定する．あるテキストデータに関して，投票結果が「含む」「含まない」のどちらかに決まらない場合は，そのテキストデータはラベルなしとした．この操作によって構築されたデータセットのラベル別のテキストデータ総数は，「含む」が 3,096 件，「含まない」が 3,437 件であった．投票結果が付与されたデータセットと各ワーカの作業情報を比較し，4.1 節にて説明した正解作業数と不正解作業数を求める．

このデータセットを用いた実験は，一つのテキストデータあたり作業したワーカの人数は 5 人から 10 人であることが分かっているため，選出するワーカ人数を 1 人，3 人，5 人ワーカの 3 パターンで行った．これは，2 値分類において作業情報の多数決の結果が一意に決まる人数である．ベースラインである単純多数決は，無作為に抽出した数人のワーカの作業結果を用いた多数決によって作業結果を決定する．F 値によりワーカの作業精度を評価する集約手法では，F 値の高い上位数人の作業情報を用いた多数決によって作業結果を決定する．本手法である，確率分布に従うサンプリング値を用いた集約手法では，サンプリング値の高い上位数人の作業情報を用いた多数決によって作業結果を決定する．

5.4.2　結果・考察

5.2 節にて説明した 3 種類の手法を用いて集約した場合の正解データセットとの一致率を F 値で算出し，表 5.4.2 に示す．ベースラインである単純多数決による集約は，選出人数が増えるほど精度が高くなっている．単純多数決においては，作業

情報が増えるほど，正解データセットとの一致率が高くなることが考えられる．これは，作業情報増えることによって，様々なワーカの意見が反映され，その中に有識者の作業情報が多く含まれようになるからだと考える．以後，ベースラインである単純多数決の集約精度は最も高かった 0.752 を採用し，他の手法の集約精度と比較を行う．

本手法である，サンプリング値を用いた集約手法の集約精度は，選出人数が 3 人の時に最も高く，0.756 であった．ベースラインの 0.752 と比べて本手法の集約精度が高いことから，サンプリング値を用いた集約は高品質な作業結果の出力に有効であることが分かる．これは，サンプリング値が高いワーカが実際に作業精度が高いワーカであり，作業精度の高いワーカの作業情報を用いた多数決によって集約精度が向上したと考えられる．

F 値により作業情報を評価する集約手法では，選出人数が 5 人の時に最も集約精度が高く，0.760 であった．本手法を用いた場合の作業精度は 0.756 であり，F 値により作業情報を評価する集約手法のほうが集約精度が高かった．しかし，F 値により作業情報を評価する集約手法を用いる際，選出人数が 3 人の時の集約精度はベースラインよりも低い 0.751 であった．このことから，F 値により作業情報を評価する集約手法では，F 値の高い順に上位 4 番目と 5 番目に実際に作業精度の良いワーカおり，選出人数が 5 人の時にたまたま精度が良くなったと推測される．対して，本手法は，選出人数が 3 人の時点でベースラインよりも集約精度が高い．このことから実際に作業精度の高いワーカを優先的に選出するという点では，F 値により作業情報を評価する集約手法に比べて優れているといえる．また，本手法と F 値により作業情報を評価する集約手法における作業精度が高いと推測した上位 3 人のワーカに対して作業数の平均を求めた．本手法におけるサンプリング値が高い上位

表 5.2　実験 2 における各手法を用いた際の F 値

	選出人数（人）			
	1	3	5	all
ベースライン	0.712	0.730	0.745	**0.752**
F 値上位	0.755	0.751	**0.760**	0.752
サンプリング	0.749	**0.756**	0.752	0.752

5 人のワーカの平均作業数は 678.36 件であった．また，F 値により作業情報を評価する集約手法における F 値が高い上位 5 人のワーカの平均作業数は 674.61 件であり，本手法を用いた場合のほうが高い値であった．以上の結果から，サンプリング値を用いたワーカの選出は，作業精度が高いと考えられるワーカのうち，総作業数が多く，作業精度推定の信頼度が高いワーカを優先的に選出できると考える．

実験 2 の結果は，本手法を用いた場合の精度向上率があまり良くなかった．また，本手法の集約精度は F 値により作業精度を評価する手法の集約精度より低かった．本手法の集約精度があまり良くなかった理由について二つ考える．一つ目は，作業情報の集約に正解の作業結果との一致率が低い投票結果を用いていることである．正解の作業結果が一致しない投票結果を用いてワーカの作業情報を集計すると，誤った作業精度推定をしてしまう．その結果，実際は作業精度が高いワーカを取りこぼしてしまう可能性が高くなると考える．二つ目は，データセットを構築する際の質問が人によって回答がわかれやすいものであったことである．実験 2 で用いたデータセットを構築する際の質問は，「どのような感情を持っているか」という感覚的な内容であった．そのため，ワーカごとに判断基準が異なり，真面目に作業したワーカだとしても作業依頼者が求める作業結果を出力できない可能性が高くなる．これによって，推定される作業精度と本来の作業精度の誤差が大きくなり，集約精度があまり向上しなかったと考える．

第 6 章　おわりに

本研究は，ワーカの作業精度を考慮した結果集約によって，高品質な作業結果を出力することを目的としている．しかし，ワーカの作業精度を推定するうえで，総作業数が少ないワーカの正確な作業精度を推定できない問題がある．そこで我々は，作業精度が高いワーカの中でも，作業精度が多く，推定される作業精度の信頼度が高いワーカを優先的に選出する手法を提案する．

作業精度を Accuracy や F 値を用いて求める場合，ワーカの作業数を考慮していないため，推定される作業精度の不確かさを評価することができない．そこで我々は，ワーカごとに確率分布を生成し，その確率分布に従うサンプリング値が高い上位数人の作業情報を集約することによって推定される作業精度の不確かさを考慮した集約ができると考えた．実験では，手法の有効性を確認するため，三種類の集約手法を比較した．ベースラインとして，単純な多数決による方法と．ワーカの作業精度を F 値を用いて推定し，推定した作業精度が高いワーカの作業結果を用いた多数決によって作業結果を決定する方法である．それぞれの集約手法を用いて集約した結果と正解の作業結果を比較し，F 値を求めた．

実験の結果，確率分布に従うサンプリング値を用いた結果集約は，高品質な作業結果の出力に有効であることが確認できた．これは，本手法を用いることによって作業精度の高いワーカを選出できており，高品質な作業結果を出力できているからだと考える．また，確率分布に従うサンプリング値を用いたワーカの選出は，総作業数が多く，作業精度の高いワーカを優先的に選出できていることが確認できた．ワーカの作業数が多いほど，推定される作業精度の信頼度は高いため，安定した集約精度向上に繋がっていると考えられる．

今回行った二つの実験において，本手法を用いた集約がベースラインである単純多数決に比べて，高い集約精度であることを確認した．本手法を用いる際，確率分布に従うサンプリング値が高いワーカの選出人数によらず，集約精度はベースラインと比べて高かった．これは，作業精度の低いワーカのサンプリング値は低いため，選出されにくくなっていることが考えられる．作業数が多く，推定される作業精度の信頼度が高いワーカのうち作業精度の低いワーカは，確率分布に従うサンプリングで低い値が出やすく，選出される確率が低い，そのため，本手法を用いるこ

とによって作業精度の低いワーカの作業情報を高精度で排除することができる．

また，本手法を用いた集約はＦ値による集約と比べて，作業数が多く，作業精度の高いワーカの選出精度が高いことを確認した．作業数が多く，作業精度が高いワーカの確率分布は，本来の作業精度に収束するため，サンプリング値が安定して高い値を出力する．そのため，サンプリング値を元にワーカを選出する際に，選ばれやすくなっているからだと考えられる．本手法を用いることによって推定される作業精度の推定誤差が小さいワーカを優先的に選出でき，推定誤差を考慮してない場合と比べて，安定した集約精度向上に繋がっていると考えられる．

本研究を通じ，確率分布に従うサンプリング値を用いた結果集約手法は，高品質な作業結果の出力に有効であることが分かった．我々は，確率分布に従うサンプリング値を利用したワーカの選出は，ワーカの作業精度を評価する方法の一つになりうると考える．また，本手法を用いる際に，サンプリング値の高いワーカの選出人数が，実験 1 では 5 人，実験 2 では 3 人の時に最も集約精度が高かった．これは，およそ半分の作業情報は多数決によって作業結果を決定する際のノイズになっていることを示している．そこで我々は，このようなノイズなりうる作業情報を事前に生み出さないように工夫することで，さらなる品質改善とコスト削減ができるのではないかと考える．

今後の展望として，データセットを変えたさらなる検証実験と手法の改善を考えている．本実験では，主にテキストデータに対する 2 値分類のデータセットを対象に実験を行った．今後は本手法が，多値分類や画像の分類タスクを目的としたデータセットなどに対しても有効であるかを検証したいと思う．また手法の改善に関しては，主にワーカの作業情報の集計の際にさらなる工夫をしていきたい．作業情報の集計の際，本手法は正解作業数を投票結果と同じ作業情報であった作業数としている．しかし，タスクによっては難しいタスク，簡単なタスクが存在するため，一概に全てのタスクの価値を同じに考えるべきではない．適切な重みを付与した集計を行うことによって，集約精度が向上するのではないかと考える．

謝辞

本研究を進めるにあたって，様々な方に協力していただきました．自分の力だけではここまで来ることはできませんでした．本当にありがとうございました．

鈴木優准教授には，たくさん指導してもらいました．研究を始める際は，私のわがままをたくさん聞いてくださったうえで，まずはバンディットアルゴリズムを勉強してくるのが良いとアドバイスしてくださり，それが今の研究に活かされていることに本当に感謝しています．また，共同研究に学部生として参加させていただき，様々な体験をさせていただきました．この体験を活かし，今後もますます研究に勤しんでいきたいと思います．ゼミや研究会などでは，私の身勝手な行いによってたくさんの迷惑をおかけしましたが，そのたびに時間をかけて指導してくださり謝罪と同時に感謝しております．今後ともよろしくお願いいたします．

この一年間は，ほぼ研究室で活動し，研究室の先輩方には本当に助けられました．論文の添削や発表スライドの作成アドバイスだけでなく，研究室が楽しくなるよう様々なイベントを計画してくださったり，いつも研究室を盛り上げてくださりましたこと本当に感謝しています．来年は，私が今のような居心地の良い研究室を作り上げていこうと思います．また，大学 1 年の頃から仲良くしている同期は，いつも私の悪ふざけに乗ってくれ，時には真剣に研究や院試に向き合うきっかけになってくれたりと本当に感謝しています．これからもお互いに高めあって行きましょう．

最後に，たくさん迷惑をかけたけど，それでも支えてくれた家族に感謝を伝えたいと思います．夜遅く帰ったり，家のことを何もしない身勝手な私でしたが，いつも心配し，経済的にも精神的にも支えてくださった家族には本当に助けられました．

皆様の協力によってここまで来れたこと，心より感謝申し上げます．

参考文献

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

[2] Xiaomei Zhang, Yibo Wu, Lifu Huang, Heng Ji, and Guohong Cao. Expertise-aware truth analysis and task allocation in mobile crowdsourcing. *IEEE Transactions on Mobile Computing*, Vol. 20, No. 3, pp. 1001–1016, 2019.

[3] Chien-Ju Ho and Jennifer Vaughan. Online task assignment in crowdsourcing markets. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 26, pp. 45–51, 2012.

[4] Dana Chandler and Adam Kapelner. Breaking monotony with meaning: Motivation in crowdsourcing markets. *Journal of Economic Behavior & Organization*, Vol. 90, pp. 123–133, 2013.

[5] Thomas Gillier, Cédric Chaffois, Mustapha Belkhouja, Yannig Roth, and Barry L Bayus. The effects of task instructions in crowdsourcing innovative ideas. *Technological Forecasting and Social Change*, Vol. 134, pp. 35–44, 2018.

[6] 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.

[7] Jacob Whitehill, Ting-fan Wu, Jacob Bergsma, Javier Movellan, and Paul Ruvolo. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. *Advances in neural information processing systems*, Vol. 22, , 2009.

[8] Peter Welinder, Steve Branson, Pietro Perona, and Serge Belongie. The multidimensional wisdom of crowds. *Advances in neural information processing systems*, Vol. 23, , 2010.

[9] Matteo Venanzi, John Guiver, Gabriella Kazai, Pushmeet Kohli, and Milad

Shokouhi. Community-based bayesian aggregation models for crowdsourcing. In *Proceedings of the 23rd international conference on World wide web*, pp. 155–164, 2014.

[10] 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.

発表リスト

[1] 城所祥太，鈴木優『クラウドソーシングにおけるラベルの正確性向上のためのラベル付け難易度を考慮したタスク割当手法』，東海関西データベースワークショップ，2023
[2] 城所祥太，鈴木優『クラウドソーシングにおけるラベル付け難易度を考慮したタスク割当手法』，第 176 回データベースシステム・第 149 回情報基礎とアクセス技術合同研究発表会，2023
[3] 城所祥太，鈴木優『クラウドワーカの作業結果と投票結果から生成される確率分布を利用した結果集約手法』，第 16 回データ工学と情報マネジメントに関するフォーラム，2024