修士論文

模擬ワーカとクラスタリングを利用した
クラウドソーシングのデータ品質向上と計算時間削減

太田 奈那

2025 年 1 月 29 日

岐阜大学大学院 自然科学技術研究科 知能理工学専攻 知能情報学領域
鈴木研究室

本論文は岐阜大学大学院自然科学技術研究科に
修士（工学）授与の要件として提出した修士論文である.

太田 奈那

指導教員：
鈴木 優 准教授

模擬ワーカとクラスタリングを利用した クラウドソーシングのデータ品質向上と計算時間削減*

太田 奈那

内容梗概

本研究では，クラウドソーシングにおけるデータ品質の向上と，モデル学習時の計算時間削減を目的とする．クラウドソーシングを利用するワーカには一定数スパムワーカが存在するため，作業依頼者の求める回答を得ることが難しい．そこで，ワーカの作業を模倣する機械学習モデルを利用し，擬似データを作成する手法を提案する．一つのタスクに対してより多くのラベルが付与されたデータが集まれば，作業依頼者の求める回答を得られやすくなる．しかし，各ワーカに一つの機械学習モデルを構築するには相当な計算時間が必要である．そこで，ワーカの作業特性に基づくクラスタリングをし，クラスタ毎に機械学習モデルを構築する手法を提案する．これにより学習時の計算時間を削減することができる．実験の結果，単純多数決と比べ手法を適用させたときのほうがデータの品質が高く，クラスタリングしない場合よりクラスタリングした場合のほうが計算時間を削減できることが確認できた．

キーワード

クラウドソーシング，BERT，クラスタリング，結果集約

*岐阜大学大学院 自然科学技術研究科 知能理工学専攻 知能情報学領域 修士論文, 学籍番号: 1234525023, 2025 年 1 月 29 日.

目次

図目次

表目次

第 1 章　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ワーカに作業を委託するシステムである．クラウドソーシングを利用することによって，複雑な作業や膨大な量の作業を複数の人に作業を分担することができるため，一人当たりの作業負担を減らすことができる．また，ワーカは好きな時間に好きな量の作業を行うことができるという利点がある．そのため，教師あり学習における教師データ作成のために使用される．

一方，クラウドソーシングの欠点として，専門家によるデータ収集や作業に比べてデータの品質が低くなることがあげられる．データ品質とは，作業依頼者による正解との一致率のことである．クラウドソーシングにおいて，ワーカの性格や知識量などを知らなければ，作業精度の低いワーカと高いワーカの区別ができない．ワーカの作業精度とは，作業依頼者による正解の結果とワーカの作業結果を比較した時の一致率のことである．クラウドソーシングで収集するデータは，一般的に一つのオブジェクト（対象）に対して複数のワーカがラベルを付与し，それぞれのラベルが投票として扱われる．これらの投票結果を集約し，多数決を用いて最も多く選ばれたラベルを最終的なラベルとして定める．そのため，作業精度の低いワーカの影響を受けて間違ったラベルが付与されたとき，データ全体として品質が低くなる可能性がある．

そこで，ワーカの作業精度に基づいてワーカの投票結果に重みを付与することによって，各ワーカの結果に対する影響力を変化させる手法を考える．作業精度が低いワーカの投票結果の重みを小さくし，作業精度が高いワーカの投票結果の重みを大きくすることによって，作業精度が高いワーカの作業結果が最終的な結果に大きな影響を与え，より正確な結果を得ることができる．

しかし，ワーカごとに作業数や作業したタスクが異なるため，ワーカが実際に作業したデータのみを利用してワーカの作業精度を公平に求めることは難しい．そこで，模擬ワーカを利用し公平な作業精度を求める方法 [2] を提案する．ワーカが実際に作業したデータを学習データとして使用し，各ワーカの作業を模倣する機械学習モデル (模擬ワーカ) を構築する．この模擬ワーカに，ワーカが作業していないオブジェクトのラベルを付与させる．模擬ワーカを利用した模擬ラベルの付与によ

り，全ワーカが同じ作業をしたという環境を擬似的に作り上げることができる．そのため，作業数やタスクの難易度を考慮することなく公平にワーカの作業精度を算出することが可能であると考える．

各ワーカにつき一つの模擬ワーカを構築することは，ワーカ 600 人のときおよそ 280 時間が必要となる．そこで，ワーカの作業結果から類似の作業特徴をもつワーカをグループ化し，各グループに一つの機械学習モデルを構築する手法を提案する．この手法により，構築する模擬ワーカの数が減るため，模擬ワーカ構築のための計算時間を削減することができる．また，ワーカ個人の作業データのみを使用した場合では，作業データ数が少なすぎるワーカの模擬ワーカは精度が実用的とならないため，一定の作業数を超えないワーカの作業データを破棄することとなる．そのため，各グループに一つの模擬ワーカを構築することによって，グループに属するワーカ全ての作業データを学習データとして使用できるため，機械学習モデルの精度が高くなり破棄するデータが減るため作業データを有効活用できる．

ワーカをグループ化して模擬ワーカを構築する手法の中で，ワーカの作業特徴を抽出する手法を二つ提案する．一つ目は，ワーカの作業データと単純多数決で集計したデータによる混同行列の要素を作業特徴とする手法である．混同行列は，ワーカの正答率やどのラベルを付与する傾向があるのかを表しており，ワーカの得意な分類パターンや誤分類の傾向を分析することができるため，ワーカの作業特徴として利用できると考える．二つ目は，模擬ワーカのモデルに含まれるパラメータを作業特徴とする手法である．ワーカ個人の作業データを使用して構築した機械学習モデルは，そのワーカに特化した機械学習モデルであるため，モデルはワーカの作業データから学習した傾向やパターンを反映している．これにより，ワーカの正答率や誤分類の傾向，回答の偏りなどがモデルのパラメータに組み込まれているため，機械学習モデルのパラメータはワーカの作業特徴を表していると考える．

本手法の有効性を検証するために実験を行った．ワーカの作業精度やデータ品質を算出する方法としてそれぞれ二つの方法がある．単純多数決で集計したデータとワーカの作業データを比較して算出したワーカの作業精度を単純多数決基準作業精度，作業依頼者がラベル付したデータとワーカの作業データを比較して算出したワーカの作業精度を依頼者基準作業精度と呼ぶ．また，単純多数決で集計したデータと手法を適用して収集したデータを比較して算出したデータ品質を単純多数決基

準データ品質，作業依頼者がラベル付けしたデータと手法を適用して収集したデータを比較して算出したデータ品質を依頼者基準データ品質と呼ぶ．本実験ではワーカの作業精度として単純多数決基準作業精度，データの品質として依頼者基準データ品質を使用した．提案手法を適用することによってデータ品質が向上することを確認するため，ワーカ個人のデータでそれぞれ模擬ワーカを構築する手法を適用した結果と，ワーカを作業特徴をもとにグループ化してグループごとにそれぞれ機械学習モデルを構築する手法を適用した結果と，単純多数決で集計した結果をそれぞれ比較する．

また，グループ化をする手法では，グループの数をいくつか変化させて実験を行った．その結果，グループ数が 4 のときに最もデータ品質が高くなることが確認できた．また，ワーカの作業特徴をもとにグループ化する手法を用いたとき，グループ化しない場合の手法と比較して模擬ワーカの構築時間を削減することができた．この結果から，類似の作業特徴をもつワーカが存在することがわかった．また，ワーカの作業精度に基づいて投票結果に重みを与えることによって，作業精度が高いワーカの作業結果が集計結果に大きな影響を与えることができ，データの品質が向上した．

本論文における貢献は以下のとおりである．

- 模擬ワーカをニューラルネットワークにより構築して精度を確かめた．
- 模擬ワーカ構築手法の適用によりクラウドソーシングにおけるデータ品質が向上した．
- 作業特徴が類似するワーカが存在することを確認した．

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，環境，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　機械学習

機械学習とは，大量のデータからパターンや規則性を自動的に見つけ出す技術である．未知のデータで過去に見つけたパターンや規則性と同じような課題に直面した際に，見つけ出したパターンや規則性を用いることによって予測や意思決定の精度を上げることができる．

機械学習には主に三つの学習方法がある．一つ目は，教師ラベルが付与されたデータを用いて学習を行う教師あり学習である．二つ目は，教師ラベルを用いずに学習を行う教師なし学習である．三つ目は，コンピュータ自身が試行錯誤を繰り返しながら，報酬が最大となるような行動を学ぶ強化学習である．

本研究において，ワーカの作業を模倣する機械学習モデルを構築する際は教師あり学習を用いる．また，ワーカの作業特性をもとにクラスタリングする際には教師なし学習を用いる．

2.1.1　ニューラルネットワーク

ニューラルネットワークとは，人間の脳内にあるニューロンという神経細胞と神経回路網の仕組みを数学的にモデル化した機械学習技術である．この技術は大量のデータから複雑なパターンや規則性を学習し，未知のデータに学習したパターンを用いることによって，予測や分類などのタスクを実行することが可能である．

ニューラルネットワークは主に，入力層，隠れ層および出力層の三つの層で構成されている．入力層は分析の基となるデータを受ける役割があり，隠れ層は入力データの処理や分析を行う役割を持つ．さらに，出力層では分析結果に基づく最終的な判断が出力される．

層と層の間にはニューロン同士のつながりの強さを表す重みがあり，この重みは学習過程において最適化される．具体的には，順伝播により得られた出力と正解データとの誤差が計算され，誤差逆伝播法と最適化アルゴリズムを通じて重みが調整される．このプロセスを繰り返すことで，ニューラルネットワークはデータから

パターンや規則性を効率的に学習することができる．

2.1.2 BERT

BERT[3] とは，自然言語処理のための Transformer[4] モデルをベースとしたニューラルネットワークモデルのことである．事前学習として MLM(Masked Language Model) と NSP(Next Sentence Prediction) を行っている．MLM は入力の一部の単語を隠して元の単語を予測するタスクである．NSP は長い文章の中から選んだ 2 文が連続しているかどうかを予測するタスクである．事前学習した後のモデルをファインチューニングすることにより，文章分類や感情分析，チャットボットなど様々な自然言語処理タスクに応用することができる．

2.1.3 ファインチューニング

ファインチューニングとは，訓練済みモデルをベースに出力層などを変更したモデルを構築し，利用者自身で用意したデータを使用してモデル全体のパラメータを再学習させる手法である．利用者自身が用意したデータが少量の場合でも精度の高いモデルを構築することができる．また，初期状態から学習するよりも短時間でモデルを構築することができる．

2.1.4 k-means 法

k-means 法は，データを事前に指定した k 個のクラスタに分割する教師なし学習の手法である．このアルゴリズムでは，まずクラスタ数 k を設定し，データ空間内にランダムに k 個の初期セントロイドを配置し，それぞれのセントロイドをクラスタとして設定する．セントロイドとは，各クラスタの中心を表す点で，クラスタ内のデータの位置を代表する役割を持つ．次に，各データ点を最も近いセントロイドと同じクラスタに割り当てることで，クラスタを形成する．その後，各クラスタ内のデータの平均を計算し，平均値を新しいセントロイドの位置として更新する．

このプロセスは，セントロイドの割り当てが変化しなくなるか，セントロイドの

更新が十分に小さくなるまで繰り返される．最終的に，各データ点が特定のクラスタに属する状態が確定し，各クラスタはそのセントロイドを中心としたグループとして定義される．

k-means 法は，クラスタ内のデータ点とセントロイド間の距離の二乗和を最小化する目的関数に基づいて動作する．この目的関数は，アルゴリズム全体を通じて単調に減少するため，アルゴリズムの収束が保証されている．

2.2 結果集約手法

本研究において結果集約手法とはアンケートなどの回答を集計するための方法を指す．集計方法には，単純な加算で計算できるため全体像の把握に適している「単純集計」や，単純集計よりも詳細な分析ができるため属性ごとの傾向を知るのに向いている「クロス集計」が存在する．クロス集計は，詳細なアンケート分析によって細かい傾向を見つけ出すことができるため，アンケートの回答を集計する主な方法として使用される．また，回答が自由記述で行われたデータを集計する「自由記述集計」もあり，類似単語などを見つけてカテゴライズしたり関連性を見つけて集計したりする方法である．さらに，多数決をとることによって一つのタスクに対してただ一つの評価を与えるといった集約方法もある．本研究で使用するデータセットを作成する際は，多数決をとる方法を用いた．

2.3 評価指標

評価指標とは，モデルやシステムの結果や性能を数値的に評価するための基準を指す．機械学習モデルにおいては，モデルが出力した予測値と正解値を比較して算出される指標がよく利用される．本研究で用いている分類タスクにおける代表的な評価指標として，Accuracy や F1 スコアなどが挙げられる．これらの指標は，タスクの種類や目的に応じて適切に選択されるべきである．例えば，クラスの不均衡が大きい場合には Accuracy よりも適合率（Precision）や再現率（Recall）を重視することがある．さらに，手法を適用して集約されるデータの品質に関しても，集約データと正解データを比較して算出される指標を評価指標として利用している．

表 2.1　評価値導出の混同行列

		予測値	
		正例ラベル	負例ラベル
真の値	正例ラベル	TP	FN
	負例ラベル	FP	TN

TP　正例と予測されたデータのうち，実際のラベルも正例であるデータの数
FP　正例と予測されたデータのうち，実際のラベルは負例であるデータの数
FN　負例と予測されたデータのうち，実際のラベルは正例であるデータの数
TN　負例と予測されたデータのうち，実際のラベルも負例であるデータの数

Accuracy や Precision，Recall，F1 スコアは表 2.1 の混同行列をもとに (2.3.1)〜(2.3.4) 式を用いて導出する．

また，本研究では，モデルの性能評価に加えて，計算効率の評価にも着目している．計算効率の指標として，モデル構築に要する時間を採用しており，これによりアルゴリズムや手法の計算コストを定量的に評価している．適切な評価指標を選ぶことは，モデルの性能や効率を正確に評価し，タスクの有用性を反映させる上で不可欠である．

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{2.3.1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2.3.2}$$

$$Recall = \frac{TP}{TP + FN} \tag{2.3.3}$$

$$F1score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{2.3.4}$$

2.3.1　機械学習モデルの精度に関する評価指標

本研究で構築した機械学習モデルの精度を評価するため，Accuracy を評価指標として採用する．ワーカの作業を模倣する機械学習モデルを構築することが目的であるため，機械学習モデルの予測結果がワーカが実際に作業した結果とどれほど一致しているかが機械学習モデルの精度になると考える．ワーカが実際に付与したラベルが真の値，機械学習モデルが出力したラベルを予測値として，2.3 節に示した混同行列に当てはめて *Accuracy* を算出し，機械学習モデルの精度とする．

2.3.2　データの品質に関する評価指標

様々な手法を用いて集約したデータの品質を評価するため，Accuracy を評価指標として採用する．本研究の目的は，作業依頼者が求める結果を得ることである．そのため，手法を適用して得られた結果が作業依頼者が求める結果とどれほど一致しているかがデータの品質になると考える．作業依頼者が付与したラベルを真の値，手法を用いてデータを集約して付与されたラベルを予測値として，2.3 節に示した混同行列に当てはめ，*Accuracy* を算出し，データの品質とする．

2.3.3　計算効率に関する評価指標

ワーカの作業を模倣する機械学習モデルの計算効率を評価するため，モデル構築に要する時間を評価指標として採用する．この際，全ワーカのモデル構築にかかる計算時間の延べ時間を利用する．延べ時間とは，同時並行で行われた作業も含めて，各作業にかかった時間を単純加算して算出できる時間である．例えば，1 つの機械学習モデル構築の計算時間が 1 時間として，50 個の機械学習モデルを構築したときの全体の構築にかかる計算時間は 50 時間となる．この計測方法を用いて，各手法での機械学習モデルの構築にかかる計算時間を計測し，計算時間を比較することによって手法の評価を行う．

第 3 章　関連研究

クラウドソーシングで収集するデータの品質向上を目的とする研究は，いくつか行われている．西らの研究 [5] では，ワーカ相互の関係を用いたデータの品質向上を目指している．作業を行うワーカは別のワーカ一人に作業を委託することができ，報酬は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．

芦川らの研究 [6][7] 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行うことによって，クラウドソーシングの質の向上を目指している．

Halpin らの研究 [8] では，ワーカの作業数や平均作業時間などの作業特徴を用いたスパムワーカの検出手法を提案している．そこで，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作業時間などの作業特徴を用い，機械学習モデルを構築する．そして，ここで構築した機械学習モデルを利用してスパムワーカを除外し，高品質なデータを収集している．

松原らの研究 [9] では，ワーカの特性に適した作業の割り当てを行っている．複数の異なる種類のタスクをワーカに割り当てる場合において，まず各ワーカが希望するタスクの優先順位をつける．こ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って精度の高い結果を得ている．これらの研究では作業精度の高いワーカを採用したり，ワーカが得意としている分野の作業を依頼したりすることによって，データ品質の向上を目指している．

本研究では，ワーカの作業を模倣する機械学習モデルである模擬ワーカを構築し，この模擬ワーカを用いて仮想的にラベル付与を行うことによって，クラウドソーシングのデータ品質の向上を目指す．一つの対象に対する作業データ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結果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

のデータ品質の向上を目指す点では同じであるが，データ品質を向上させるためのアプローチとして，機械学習モデルを使用してデータを増やす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ある．Dawid らの研究 [10] 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．EM アルゴリズムとは，確率モデルの潜在変数とパラメータの最尤推定を行う手法で，E ステップで潜在変数の期待値を計算し，M ステップでパラメータを更新する操作を繰り返す．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．小山らの研究 [11] では，高精度なラベル統合方法について提案している．行った作業の処理結果をどの程度確信しているのかをワーカに申告してもらい，その確信度をもとにワーカの作業結果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ワーカの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．Venanzi らの研究 [12] では，ベイズ推定を利用して効率的にラベルを集約する手法について提案している．ベイズ推定を利用して混同行列からワーカの信頼性や偏りを学習し，ラベルを推定する．このとき，コミュニティに分割して類似の作業特性を持つワーカを集約し，各コミュニティの平均値を用いる．これよって，少量のデータでも高い精度でラベルを予測することができる．

本研究では，作業精度に基づき票の重みを設定して集計を行っている．作業精度の高いワーカの作業結果が最終的な結果により影響を与えるという点で同じである．しかし，ワーカの作業精度を求める方法として，機械学習モデルを利用してデータを増やすことによって全ワーカが同じ作業を行った状況を擬似的に作るという点で異なっている．

第 4 章　提案手法

本研究は，ワーカの作業を模倣する機械学習モデル（模擬ワーカ）を用いたクラウドソーシングのデータ品質向上と，クラスタリングを利用した機械学習モデル構築時の計算時間の削減の二つを目的とする．

クラウドソーシングやアンケートなどでデータを収集する際，単純多数決を用いられることが多い．単純多数決のメリットは全ての人の意見が平等に反映される点にある．ところが，ワーカの質によらず平等に反映されるため，質の低いワーカの作業により作業依頼者の求める回答にならない場合がある．本研究は作業依頼者の求める回答を得ることが目的である．単純多数決によって集計したデータと作業データの一致率が高いワーカは作業精度が良いと考え，ワーカの作業精度を用いて票の重みを付与する．ワーカの作業精度を考慮したデータ集計によって作業精度の高いワーカの結果が反映されやすくなり，作業依頼者の求める回答を得ることができると考える．また，ワーカの作業精度を算出するにあたって，ワーカごとに作業タスクが異なるため，平等に作業精度を算出できない問題がある．例えば，難易度の高い作業ばかりが偶然割り当てられたワーカが，作業精度を不当に低く算出される．そこで，各ワーカの作業データを用いて作業を模倣する機械学習モデルを構築し，機械学習モデルに未作業オブジェクトのラベル付けを行う．機械学習モデルを利用することにより全てのワーカが平等に全ての作業を行うため，ワーカの作業精度を平等に求めることが可能であると考える．

しかし，全ワーカの作業を模倣する機械学習モデルを構築するためには計算時間がかかる．これを解決するために，ワーカグループごとに模擬ワーカを構築する．まず，ワーカの作業特性に基づくクラスタリングを行い，作業を模倣する機械学習モデルを各クラスタに一つずつ構築する．その後，各ワーカの作業データでファインチューニングすることによって，全ワーカの機械学習モデルを構築する．クラスタリングにより，各クラスタに含まれる全ワーカの作業を学習に用いることができるため，学習に使用するデータ数を増やすことができ，モデルの性能精度向上が期待できる．また，クラスタの作業特性をもつ機械学習モデルをワーカ個人の作業データでファインチューニングをすることにより，ファインチューニングしない場合よりパラメータの更新幅が小さくなり，モデルの学習に必要な反復回数が減少す

るため，短時間でワーカの作業を模倣する機械学習モデルを構築することが可能であると考える．

以降，本稿では次のように用語を定義する．

- **模擬ワーカ**：ワーカの作業を模倣する機械学習モデル
- **オブジェクト**：ラベルを付与する対象物 (例：数字の画像・ニュース記事)
  - **作業オブジェクト**：ワーカが実際に作業を行ったオブジェクト
  - **未作業オブジェクト**：ワーカが作業を行っていないオブジェクト
- **ラベル**：オブジェクトが属するカテゴリ（例：0〜9 の数字・スポーツやトピックニュースなど）
- **データ**：オブジェクトとラベルが一対一でセットになっているものの集まり
  - **実作業データ**：ワーカ本人がオブジェクトにラベルを付与したデータ
  - **擬似作業データ**：模擬ワーカがオブジェクトにラベルを付与したデータ
  - **作業データ**：ワーカの実作業データと擬似作業データを結合したデータ
  - **仮正解データ**：ワーカの作業精度を算出する際に使用する，手法を用いて集計したデータ

本手法の流れを以下に示す．

Step1 単純多数決データ D^{MV} を作成する．

Step2 各ワーカの作業特性から特徴量をそれぞれ抽出する．

Step3 各ワーカの特徴量にもとづき，類似した特徴量を持つワーカごとにクラスタリングする．

Step4 クラスタごとにクラスタの特徴量を模倣する機械学習モデルをそれぞれ構築したあと，クラスタに属するワーカ個人の作業データを用いて各ワーカの模擬ワーカをそれぞれ構築する

Step5 模擬ワーカを使用してワーカの未作業オブジェクトにラベルを付与し，オブジェクトに擬似ラベルをそれぞれ付与する．

Step6 仮正解データとワーカの作業データを比較して，ワーカの作業精度をそれぞれ算出する．

Step7 ワーカの作業精度に基づきワーカごとに票の重みを与え，再集計データ D^{re} を作成する．Step2 に戻る．

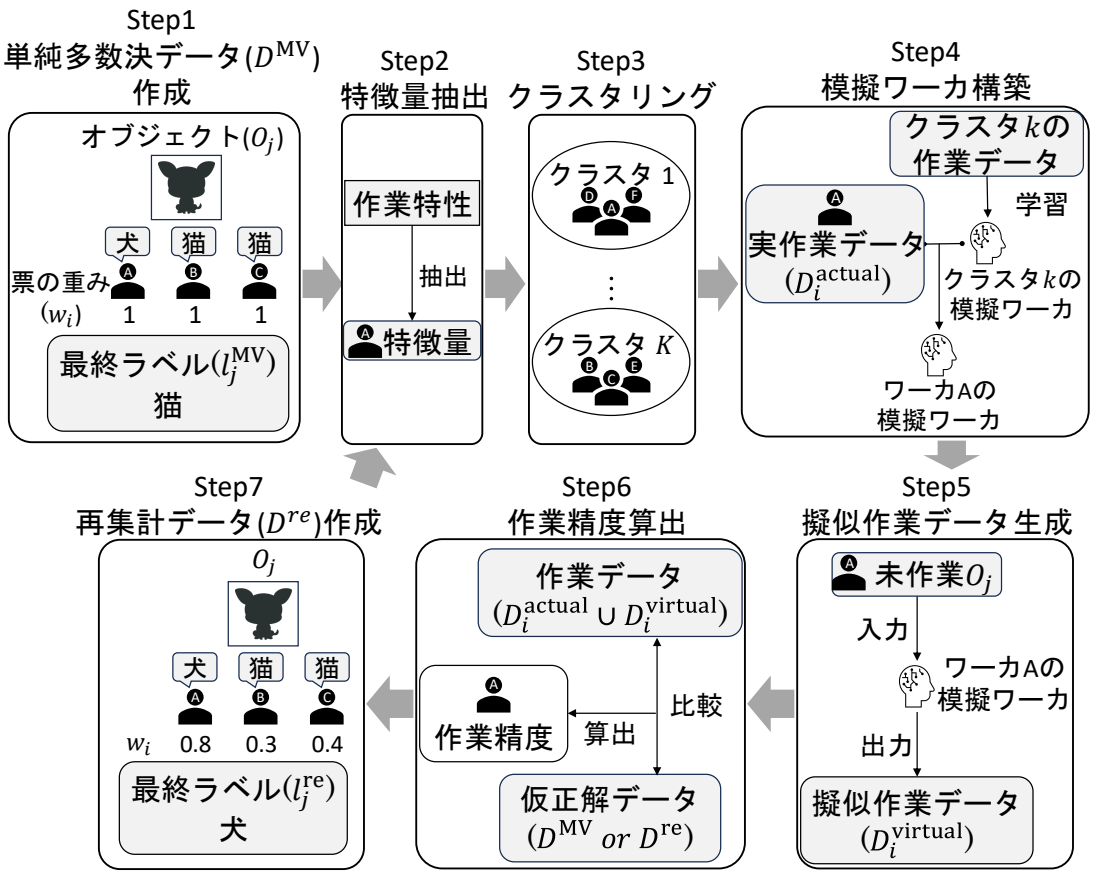

図 4.1　提案手法の概要

図 4.1 は提案手法の概要図である．本手法はワーカの作業特性から抽出した特徴量をもとにワーカをクラスタリングし，クラスタごとに機械学習モデルを構築する手法である．Step2 にてワーカの特徴量を抽出する際，Step7 にて作成された再集計データをもとに作業特性を生成することが可能である．そのため，本手法では Step2 から Step7 を繰り返し行うことによってデータの品質を向上させることを試みる．以降本稿では，Step2 から Step7 を 1 プロセスとし，繰り返し行った回数をプロセス回数と呼ぶ．しかし，Step2 で特徴量を抽出する際に Step7 の結果が必要であることから，事前準備として，Step2 を行わず Step3 にてクラスタ数をワーカの人数として以降 Step4～7 を行う．この事前に行うプロセスをプロセス回数 0 とする．プロセス回数が 0 のときの結果を得て，Step2～Step7 を繰り返していき，プ

ロセス回数を重ねてデータ品質の向上を図る．

Step2 に関しては 4.1 節で，Step3 に関しては 4.2 節で説明する．Step4 に関しては 4.3 節で，Step5 に関しては 4.4 節，Step6，7 に関しては 4.5 節で説明する．

4.1 ワーカの作業特性

本節では，ワーカの作業特性を表す特徴量を抽出する方法について説明する．前節で説明したように，プロセス回数が 0 のときはワーカの作業特性から特徴量を抽出する操作は行わず，プロセス回数が 1 以上のときにこの操作を行う．クラウドワーカの中には似たような作業をするワーカが一定数存在すると仮定する．作業が類似しているワーカをグループ化し，各グループの作業を模倣する機械学習モデルを構築することを考える．ワーカをグループ化する際はワーカの作業特性を知る必要がある．

我々は以下の二種類の特徴量がワーカの作業特性を表していると考える．

特徴量 1 ワーカの作業データと仮正解データを比較して生成した混同行列の要素
特徴量 2 ワーカの実作業データを用いて構築した機械学習モデルのパラメータ

ワーカの作業特性とは，ワーカの作業の正確性やどのようなデータにどのようなラベルを付与したかなどの作業傾向である．作業データと仮正解データをもとに生成した混同行列は，ワーカの作業傾向を表しているため，混同行列の各要素がワーカの作業特性として利用できると考える．また，機械学習モデルの最終層から全結合層に繋がるパラメータはその学習データの特徴を表していると考える．そのため，個人の作業データを使用して構築した機械学習モデルのパラメータはワーカの作業特性を表していると考えることができる．以降，ワーカの作業特性を表す配列をワーカの特徴量とし，ワーカを $i(i = 1, 2, \cdots, n)$，ワーカの特徴量配列を $\mathbf{F}_i = [f_1, f_2, \cdots, f_M]$ と表す．このとき n はワーカの総数を表しており，M は特徴量配列の要素数を表す．ワーカの作業特性を抽出する二種類の方法について次節以降で詳しく説明する．

4.1.1 混同行列を用いた特徴量抽出

ワーカの作業データと仮正解データを比較して生成した混同行列からワーカの特徴量を抽出する方法について説明する．まず，ワーカの作業データと仮正解データを比較して混同行列を生成する．ここで，混同行列の各要素の値がワーカの作業特性を表していると考える．しかし，ワーカによって回答に偏りが存在し，各要素間の値に大きな差が生じる場合がある．このような場合，クラスタリングを行う際に，特徴量間のスケールの違いがクラスタリング結果に影響を及ぼす可能性がある．そこで，各要素の値を全要素の値の合計で割ることによって算出された値をワーカを表現する特徴量として用いる．このように各要素の割合を算出することで，ワーカの特徴量の値は最小値が 0，最大値が 1 となり，各要素の値の偏りを抑えた特徴量を得ることができる．

表 4.1〜4.3 にワーカの混同行列の例を示す．表 4.1〜4.3 における総データ数は 1,000 件であることから，ワーカ A の特徴量は $\mathbf{F}_A = [0.478, 0.316, 0.123, 0.74]$ であり，ワーカ B の特徴量は $\mathbf{F}_B = [0.530, 0.349, 0.110, 0.011]$ である．ワーカ A とワーカ B は，それぞれの特徴量の値がおおよそ同じ範囲 (0.1〜0.5 程度) に収まっている．そのため，距離計算時に特定の要素が強く影響を及ぼすことがなく，正規化を行わなくてもクラスタリングの結果に大きな影響を与えにくい．一方，ワーカ C の特徴量は $\mathbf{F}_C = [0.970, 0.005, 0.01, 0.015]$ であり，左上の特徴量の値が 0.970 と他の特徴量の値よりも極端に大きくなっている．この場合，距離計算において左上の要素の値が大きな重みを持つため，他の要素が十分に考慮されない可能性が高い．そこで，正規化を行い全ての要素の値を $[0, 1]$ に揃えることによって，全ての要素が均等に距離計算に寄与することが可能となる．また，利用するクラスタリング手法によっては，初期クラスタ中心に結果が強く依存する場合がある．特定の要素の値が他の要素の値より極端に大きい場合，初期値がその要素に大きく偏り，最終的な結果が正確に反映されない．そのため，正規化をすることによって初期値の影響が均等化し，特定の特徴量に影響されることを防ぎつつ，安定したクラスタリングを実現できる．

また，ワーカの作業データと単純多数決データを比較して生成した混同行列はワーカの正答率や誤答率だけでなく，誤答の仕方や各ラベルにおける正答数を表している．そのため，混同行列の各要素の値を総作業数で割った値はワーカの

作業特性として使用することができる．例に用いた混同行列の場合，特徴量配列 $\mathbf{F}_i = [f_1, f_2, \cdots, f_M]$ で用いられる要素数は $M = 4$ である．

ラベルの種類が a 種類であったとき，混同行列の要素数は $a \times a$ となる．そのため，特徴量 $\mathbf{F}_i = [f_1, f_2, \cdots, f_M]$ において $M = a^2$ となり，混同行列の要素を左上から右下へ向かう Z 型の順に $f_1, f_2, \cdots, f_{a^2}$ となる．

4.1.2　機械学習モデルのパラメータを用いた特徴量抽出

次にワーカの実作業データを用いて構築した機械学習モデルから，ワーカの特徴量を抽出する方法について説明する．プロセス回数が p（p は自然数）のときはプロセス回数が $p-1$ の際に構築した機械学習モデルのパラメータを特徴量として利用する．

機械学習モデルを構築する際，ワーカごとにパラメータの初期値を変えると，最終パラメータが学習データだけでなく初期値に依存してしまう．そのため，ワーカの作業特徴を表すパラメータ同士の比較が適切に評価することができなくなる．そこで，全ワーカの機械学習モデルの初期パラメータに同一の値を設定する．これにより，学習結果が初期値に影響されなくなり，パラメータの比較を正当化することができる．

図 4.2 点線で囲われている，中間層 h^l から出力層 y につながる重みをワーカの作

表 4.1　ワーカ A の混同行列

		作業データ	
		ポジ	ネガ
単純多数決	ポジ	487	316
	ネガ	123	74

表 4.2　ワーカ B の混同行列

		作業データ	
		ポジ	ネガ
単純多数決	ポジ	530	349
	ネガ	110	11

表 4.3　ワーカ C の混同行列

		作業データ	
		ポジ	ネガ
単純多数決	ポジ	970	5
	ネガ	10	15

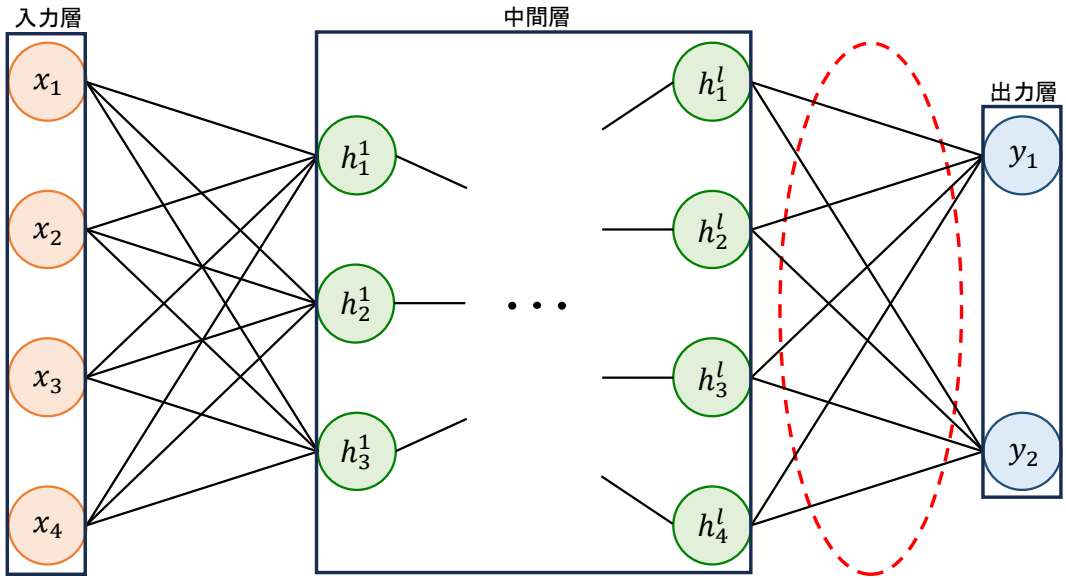

図 4.2　機械学習モデルにおける利用パラメータの概要

業特性を表す特徴量として利用する．ワーカ i の特徴量配列 $\mathbf{F}_i = [f_1, f_2, \cdots, f_M]$ において，図 4.2 の例では $M = 8$ となる．

ラベルの種類が a 種類であったとき，BERT の最終層の出力が 768 次元であることから，特徴量 $\mathbf{F}_i$ において $M = 768 \times a$ となる．このとき，各ノードにつながるパラメータを順に割り当て，m 番目のノードにつながるパラメータは $f_{(m-1)\times 768+1}, f_{(m-1)\times 768+2}, \cdots, f_{m\times 768}$ と定める．

4.2　ワーカのクラスタリング

本節では，作業特性が類似しているワーカをクラスタリングする方法について説明する．ここで，プロセス回数が 0 のとき，クラスタ数をワーカの人数とする．以降の操作はプロセス回数が 1 以上のときに行う．4.1.1 節と 4.1.2 節にて抽出したワーカの特徴量 $\mathbf{F}_i = [f_1, f_2, \cdots, f_M]$ を用い，k-means 法を利用してワーカを K 個のグループにクラスタリングする．以下に k-means 法のアルゴリズムを示す．

Step1 最初のクラスタ中心 (セントロイド) となる $[0, 1]$ の M 次元のランダム配列

をデータ空間内に K 個配置する．

Step2 ワーカの特徴量とセントロイド間の距離を計算し，最も近いセントロイドが属するクラスタラベルを割り当てる．

Step3 各クラスタの重心を，そのクラスタに属するすべての特徴量の平均に更新する．

Step4 セントロイドの位置が変化しなくなるか，変化が微小になるまで，Step2 と 3 を繰り返す．

Step2 ではユークリッド距離を用いてワーカの特徴量とセントロイド間の距離を計算する．クラスタ k のセントロイドを $\mathbf{C}_k = [c_1, c_2, \cdots, c_M]$，$k = 1, 2, \cdots, K$ と表し，2 点 $\mathbf{F}_i = [f_1, f_2, \cdots, f_M]$，$\mathbf{C}_k = [c_1, c_2, \cdots, c_M]$ のユークリッド距離 e は以下の式のように表す．

$$e(\mathbf{F}_i, \mathbf{C}_k) = \sqrt{\sum_{m=1}^{M} (f_m - c_m)^2} \tag{4.2.1}$$

上記のアルゴリズムを実行し，類似の作業特性を持つワーカを K 個のグループにクラスタリングする．

4.3 模擬ワーカの構築

本節では，模擬ワーカの構築方法について説明する．まず 4.2 節にて行ったクラスタリング結果をもとに，各クラスタに一つずつ作業を模倣する機械学習モデルを構築する．そして，ワーカが属するクラスタの機械学習モデルをワーカ個人の実作業データを用いてファインチューニングすることによって，一人のワーカにつき一つの模擬ワーカを構築する．クラスタごとに機械学習モデルを構築することにより，総作業数が少ないワーカでも性能の良い模擬ワーカを構築することができると考える．ただし，プロセス回数が 0 の場合，クラスタ数＝ワーカの人数であるため，ワーカ個人の実作業データだけを用いて模擬ワーカを構築する．クラスタで機械学習モデルを構築する方法とワーカ個人の実作業データでファインチューニングをするる方法については，それぞれ 4.3.1 節と 4.3.2 節にて説明する．

4.3.1　クラスタごとにモデル構築

クラスタの作業特性を模倣する機械学習モデルをクラスタごとにそれぞれ構築する．各クラスタの機械学習モデルは，クラスタに属するワーカがラベルを付与したオブジェクトを入力，ラベルを出力とする．機械学習に用いる学習データには，クラスタに属する全ワーカの実作業データを収集したデータセットを使用する．データセットを作成する際，複数のワーカによってラベルが付与されたオブジェクトが存在する可能性がある．複数のワーカがラベルを付与したオブジェクトに対しては，単純多数決を用いてラベルを一つに決定する．一人のワーカのみがラベルを付与したオブジェクトに関しては，そのワーカが付与したラベルを採用する．

クラスタ k における作業データを訓練データと検証データ，テストデータに分割し，訓練データを用いて学習し，検証データを用いて性能の評価をしてハイパーパラメータを調整する．また，テストデータを用いて，機械学習モデルの性能を評価する．

4.3.2　ファインチューニング

4.3.1 節にて構築したクラスタの作業特性をもつ機械学習モデルを，ワーカ個人の実作業データを用いてファインチューニングをする．模擬ワーカの構築では，入力をワーカが実際に作業したオブジェクトとし，出力をワーカが付与したラベルとして学習する．クラスタの機械学習モデルをワーカ個人の実作業データでファインチューニングする際，クラスタの機械学習モデルの最終層のパラメータのみを更新できるようにする．ファインチューニングを行うことによって，クラスタリングをしない場合と比べて短時間で模擬ワーカを効率的に構築することが可能になると考える．

具体的には，クラスタ k に属するワーカ i の個別モデルを作成する際，クラスタ k の作業特性を反映したモデルをベースとする．その後，このベースモデルをワーカ i の個人作業データでファインチューニングし，ワーカ i の作業特性を反映した個別モデルを構築する．

4.4　模擬ワーカを利用したラベル付け

本節では，模擬ワーカを利用し，ワーカの未作業オブジェクトにラベルを付与する方法について説明する．ワーカ i が実際にラベル付けを行った作業データを実作業データ D_i^{actual} とし，ワーカ i が作業を行っていない未作業オブジェクトの集合を D_i' とする．また，単純多数決データを D^{MV} とし，単純多数決データに含まれるオブジェクトを O_j，$j = 1, 2, \cdots, N$，とすると，$D_i' = \{O_j | O_j \in D^{\mathrm{MV}} \cap O_j \notin D_i^{\mathrm{actual}}\}$ と表せる．このとき，N はオブジェクトの総数を表す．ワーカ i の模擬ワーカに D_i' に含まれるオブジェクト O_j を入力し，ラベルを出力させる．模擬ワーカが出力したラベルをオブジェクトに付与したデータの集合を擬似作業データ D_i^{virtual} とする．ワーカ i の実作業データと擬似作業データを含むデータの集合を作業データと呼び，D_i と表す．

4.5　票の重み付けによる集計方法

本節では，ワーカの作業精度をもとに再集計する方法について説明する．

4.5.1　作業精度の算出方法

ワーカ i の実作業データ D_i^{actual} と 4.4 節で作成した擬似作業データ D_i^{virtual} を含む作業データ D_i と仮正解データを比較してワーカ i の作業精度 q_i を算出する．プロセス回数が 0 のときは仮正解データを単純多数決データ D^{MV} とし，プロセス回数が 1 以降の仮正解データは 4.5.2 節にて作成する再集計データ D_{p-1}^{re} を使用する．ワーカ i の作業精度 q_i の算出は以下の手順に従って行う．

Step1 作業データ D_i と仮正解データ (D^{MV} または D_{p-1}^{re}) を比較し，オブジェクト O_j に付与されたラベルが一致するデータ数を数える．

Step2 ラベルが一致したデータの数を総オブジェクト数 N で割り，算出された値をワーカ i の作業精度 q_i とする．

このとき，ワーカ i の作業データ D_i におけるオブジェクト O_j に付与されたラベルを y_{ij} と表す．また，単純多数決データ D^{MV} におけるオブジェクト O_j に付与

されたラベルを y_j^{MV} とし，再集計データ D_{p-1}^{re} におけるオブジェクト O_j に付与されたラベルを $y_{j,p-1}$ とする．

ワーカ i の作業精度 q_i はラベル y_{ij}，y_j^{MV}，$y_{j,p-1}$ を用いて以下の式のように表す．

$$q_i = \frac{\sum_{j=1}^{N} \mathbb{I}(y_{ij} = y_j)}{N}, \quad \text{where } y_j = \begin{cases} y_j^{\mathrm{MV}} & (p = 0) \\ y_{j,p-1} & (p \geq 1) \end{cases} \tag{4.5.1}$$

ここで，$\mathbb{I}(\cdot)$ は指示関数を表し，条件が成立する場合に 1、そうでない場合に 0 を返す．(4.5.1) 式を用いて算出された作業精度 q_i を使用してデータの再集計を行う．

4.5.2 集計方法

4.5.1 節にて算出したワーカの作業精度 q_i をもとに，ワーカごとに票の重みを設定し，データを集計する手法について説明する．この手法では，作業精度の低いワーカが行った作業の結果を集約結果に反映されにくくするために，データを集計する際にワーカごとに票の重みを設定する．

まず，ワーカ i の票の重みを w_i として，最大値 1，最小値 0 となるように作業精度 q_i を正規化する．ワーカ i の票の重み w_i は以下の式を用いて求める．

$$w_i = \frac{q_i - x_{\min}}{x_{\max} - x_{\min}} \tag{4.5.2}$$

$$W = \{x | q_i (i = 1, 2, \cdots, n)\} \tag{4.5.3}$$

(4.5.3) 式は全ワーカの作業精度を含む集合である．．このようにして得たワーカの票の重み w_i を利用してデータの集計を行う．．

次に，オブジェクト O_j に対して実際にラベルを付与した n 人のワーカの作業データに，先ほど求めた票の重みを付与してラベルを決定する．票の重みを用いた集計方法を以下の数式で示す．ラベルの種類は L 種類とし，オブジェクト O_j における各ラベル $l (l = 1, 2, \cdots, L)$ の票数 c_{lj} を以下のように定義する:

$$c_{lj} = \sum_{i=1}^{n} w_i \cdot \mathbb{I}(y_{ij} = l) \quad \text{for } l \in \{0, 1, \ldots, L\} \tag{4.5.4}$$

ここで，$\mathbb{I}(\cdot)$ は指示関数を表し，条件が成立する場合に 1、そうでない場合に 0 を返す．オブジェクト O_j における最終ラベル $y_{j,p}$ は，最も票数が多いラベル l を選ぶことで定義される：

$$y_{j,p} = \arg \max_{l \in \{0,1,\ldots,L\}} c_{lj} \tag{4.5.5}$$

以上の式を用いて全てのオブジェクトにラベルを付与する操作を行う．この操作を行うことによって得られたラベル $y_{j,p}$ がオブジェクト O_j に付与されたデータの集合を再集計データ D_p^{re} とする．

各手法にて集計されたデータの品質は，作業依頼者がラベル付けした正解データと比較することによって計算される．本研究の目的は作業依頼者の求める回答を得ることである．そのため本稿では, 単純多数決データからランダムに抽出した 1,000 件のテキストデータに著者がラベルを付与したデータセットを正解データ D^{correct} とする．正解データ D^{correct} におけるオブジェクト O_j のラベルを y_j^{correct}，各手法のそれぞれの再集計データにおけるオブジェクト O_j のラベルを $y_{j,p}$ とする．

$$Q = \frac{\sum_{j=1}^{N} \mathbb{I}(y_{j,p} = y_j^{\mathrm{correct}})}{N}, \tag{4.5.6}$$

以上の式より，各手法におけるデータの品質 Q を算出する．

第 5 章　評価実験

提案手法の有効性を検証するために本実験を行った．

5.1　データセット

評価実験ではクラウドソーシングで収集したデータセットを使用する．データセットはある製品に関する YouTube のコメントに対してワーカがラベルを付与したものである．このデータセットの作成目的は，製品に関するユーザの感情分析とユーザの使用状況を調査することである．このデータセットを作成するためにクラウドソーシングを利用し，ワーカにいくつかの質問を用意した．本実験では，用意した質問のうち一つの質問に対する回答ラベルを保持するワーカの作業データを使用した．本実験で使用するデータのオブジェクトはテキストデータであり，ラベルは「テキストデータに，ある製品に対する要望や経験・感情が含まれていますか」という質問に対する含まれている，もしくは含まれていないのどちらかの回答である．

本実験で使用したデータセットには 569 人のワーカの作業データが含まれている．作業データにはワーカ ID，テキスト ID，テキスト，ワーカが付与したラベルの四つのカラムがある．データセットに含まれる総作業データは 262,000 件であり，一つのテキストに対し 10 人のワーカがラベルを付与している．この作業データを使用し，単純多数決を一つのテキストにつき一つのラベルを付与する．単純多数決にて同票になった際は含まれているのラベルを付与する．そのため，総テキストデータ数は 26,200 件となる．

5.2　実験環境

本実験では，各手法における機械学習モデル構築の計算時間を測定する．計算時間を公平に計測するため，全ての手法を同一の環境で実行する．具体的には，NVIDIA RTX A4000 を 3 基搭載したサーバーを使用し，全ての手法における機械学習モデル構築を同条件下で実施する．

5.3 節で後述するように，本実験では，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*を使用し，訓練データを用いたファインチューニングによってワーカの作業を模倣する機械学習モデルを構築する．BERT はその構造やハイパーパラメータが統一されているため，手法間で計算時間を公平に比較することが可能である．

5.3 実験手順

本実験では，提案手法の有効性を検証するために四つの手法を用いて実験する．一つ目は，各ワーカの票の重みを 1 として加算する，単純多数決手法である．この手法は一般的にデータ集計に用いられる方法であるため，本実験ではベースラインとする．二つ目は，プロセス回数 $p = 0$ として 4 章の Step2〜Step7 の手法を適用させる，ワーカ別モデル構築手法である．三つ目は，4.1.1 節で示した，プロセス回数 $p \geq 1$ のときに手法を適用させるクラスタ別モデル構築手法のうち，混同行列の要素を作業特徴とする，混同行列特徴抽出法である．四つめは，4.1.2 節で示した，クラスタ別モデル構築法のうち，モデルのパラメータを作業特徴とする，モデル特徴抽出法である．

本実験では，5.1 節にて説明した YouTube のコメントデータセットを利用する．まず，事前準備として，全データからワーカ ID にもとづいてデータを抽出し，各ワーカの実作業データ D_i^{real} を作成する．また，ワーカが作業していないテキストデータを含む未作業データ D_i^{undone} も作成する．これらのデータを使用して Step1 から順に実験を進める．

5.3.1 単純多数決

一つ目の手法であり，ベースラインとなる単純多数決手法について説明する．この手法は提案手法のうちの Step1 にあたり，ここで作成したデータが単純多数決データとなる．各ワーカの票の重みを 1 として，各ラベルに投票された数をカウントする．投票された数が多かったラベルを，そのテキストデータのラベルとして付与したデータを作成する．各ラベルに投票された数が同票であった場合，「含まれ

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

る」のラベルを付与する．この作成されたデータが単純多数決データ D^{MV} であり，ベースラインとなる．

5.3.2　ワーカ別モデル構築手法

二つ目の手法であるワーカ別モデル構築法について説明する．この手法ではプロセス回数 $p = 0$，クラスタ数 $k = 569$(ワーカの総数) として 4 章の Step2〜Step7 を行う．

クラスタ数はワーカの総数と等しいため，一つのクラスタにつき一人のワーカがそれぞれ属している．そのため，事前準備として作成したワーカの実作業データを用いて，訓練済みモデルをファインチューニングすることによって模擬ワーカを構築する．構築した模擬ワーカに未作業データ D_i^{undone} を入力し，ラベルを出力させる．出力したラベルと入力したテキストデータを対応づけ，擬似作業データ D_i^{virtual} を作成する．ワーカの実作業データと擬似作業データを結合した作業データ $D_i = D_i^{\mathrm{real}} \cup D_i^{\mathrm{virtual}}$ と，単純多数決データ D^{MV} を比較してワーカの作業精度 q_i を算出する．全ワーカの作業精度を，4.5.2 節にて定義した 4.5.2 式を用いて正規化し，正規化後の値をワーカの票の重み w_i とする．各ワーカに票の重み w_i を付与し，各ラベルに投票された票数を数え，最も値が大きいラベルをそのテキストデータのラベルとして付与する．このようにして集計したデータを再集計データ D_1^{re} である．

5.3.3　クラスタ別モデル構築手法

三つ目と四つ目の手法であるクラスタ別モデル構築手法について説明する．この手法ではプロセス回数 $p \geq 1$ のとき，クラスタ数 k を固定して Step2〜Step7 を繰り返し行う．本実験ではクラスタ数を $k = 4$ として行う．

プロセス回数 $p = 0$ のときに作成したワーカの作業データや模擬ワーカなどからワーカの特徴量 $\mathbf{F}_{\mathrm{i}}$ を抽出する．各手法については後述する．抽出した特徴量をもとにワーカを k 個のグループにクラスタリングする．各クラスタに含まれるワーカの実作業データ D_i^{real} を連結し，単純多数決を用いてクラスタ作業データを作成

する．このクラスタ作業データを用いて訓練済み日本語 BERT モデルをファインチューニングすることによって，クラスタの作業特性を模倣する模擬ワーカを構築する．

構築した模擬ワーカに未作業データ D_i^{undone} を入力し，ラベルを出力させる．出力したラベルと入力したテキストデータを対応づけ，擬似作業データ D_i^{virtual} を生成する．ワーカの実作業データと擬似作業データを結合した作業データ D_i と，プロセス回数 $p-1$ で生成した再集計データ D_{p-1}^{re} を比較してワーカの作業精度 q_i を算出する．

全ワーカの作業精度を，4.5.2 節にて定義した 4.5.2 式を用いて正規化し，正規化後の値をワーカの票の重み w_i とする．各ワーカに票の重み w_i を付与し，各ラベルに投票された票数を数え，最も値が大きいラベルをそのテキストデータのラベルとして付与する．このようにして集計したデータを再集計データ D_p^{re} とする．

混同行列特徴抽出法

ワーカの作業データと仮正解データから混同行列を生成しワーカの特徴量を抽出する「混同行列特徴抽出法」について説明する．プロセス回数 p においてワーカの混同行列は，ワーカの作業データ D_i とプロセス回数 $p-1$ で生成した再集計データ D_{p-1}^{re} を比較して生成する．

本実験で使用するデータセットのラベルの種類は 2 種類である．そのため，混同行列の要素数は四つであり，特徴量 $\mathbf{F}_i = [f_1, f_2, \cdots, f_M]$ において $M = 4$ となる．各要素において，要素の値を総データ数の 26,200 で割った値を特徴量とする．このとき，左上，右上，左下，右下の Z 型の順に f_1, f_2, f_3, f_4 とする．

モデル特徴抽出法

ワーカの実作業データを用いて構築した機械学習モデルからワーカの特徴量を抽出する「モデル特徴量抽出法」について説明する．プロセス回数 $p-1$ で構築した模擬ワーカの全結合層につながるパラメータをプロセス回数 p のときのワーカの特徴量とする．

本実験では模擬ワーカの構築に訓練済み日本語 BERT モデルを使用する．そのため，BERT の最終層と全結合層をつなぐパラメータがワーカの特徴量となる．

BERT の最終層の出力は 768 次元であり，最終的なラベルの出力は 2 種類であることから特徴量 $\mathbf{F}_i = (f_1, f_2, \cdots, f_M)$ において $M = 1,536$ となる．このとき，一つ目のノードにつながるパラメータを順に $f_1, f_2, \cdots, f_{768}$ とし，二つ目のノードにつながるパラメータを順に $f_{769}, f_{770}, \cdots, f_{1536}$ とする．

5.3.4　データ品質の算出

上記すべての手法において生成されるデータの品質を算出し，各手法の有効性を検証する．データの品質とは，作業依頼者が付与した正解データとどれくらい一致しているかである．そこで本実験では，単純多数決データからランダムに抽出した 1,000 件のテキストデータに著者がラベルを付与したデータセットを正解データ D^{correct} とする．(4.5.6) 式を用いて，各手法におけるデータの品質 Q を算出し，手法の有効性を比較する．

5.4　評価指標

本実験では，提案手法を評価するために二つの観点で評価を行った．

5.4.1　再集計データの評価

各手法を用いて集計したデータの品質を評価する．本手法の目的は，作業依頼者が求める結果を得ることである．そのため，5.1 節で述べた著者が作成した正解データセットと，各手法を用いて集計した再集計データを比較してデータの品質を評価する．

正解データに含まれるテキスト ID をもつデータを再集計データから抽出して比較する．正解データと再集計データで同じテキストデータに対して付与されるラベルが同じであるデータ数をカウントし，その値を正解データの総数で割った値をデータの品質とする．また，手法の有用性を確かめるために，単純多数決を用いて集計したデータと正解データ間でも比較を行い，データの品質を算出する．

5.4.2　機械学習モデルの計算効率の評価

各手法を用いて機械学習モデルを構築した際の計算効率を機械学習モデルの構築にかかった時間によって評価する．

本実験は，3 基の GPU を搭載したサーバーで実施しており，一つの GPU につき一つの機械学習モデルを構築している．そのため，各機械学習モデル構築にかかった時間を加算して算出する延べ時間を手法の計算効率とする．

5.5　結果・考察

四つの手法にて集計したデータの品質と，単純多数決手法以外の三つの手法にて模擬ワーカ構築にかかった計算時間を表 5.5 に示す．クラスタ別モデル構築手法では，プロセス回数 $p = 2$ まで行った．単純多数決手法と比較すると，提案した三つの手法をそれぞれ用いて集計したデータは全てデータの品質が向上した．さらに，クラスタ別モデル構築手法のうちモデル特徴量抽出法を用いた場合が最もデータの品質が高かった．このことから，ワーカの作業精度を考慮してデータを集計することが，データ品質の向上に有効であることが確認できた．

混同行列特徴量抽出法における $p = 1$ と $p = 2$ のワーカの重みを比較する．$p = 1$ において重みが 0.2 未満のワーカ 46 人のうち，$p = 2$ で重みが小さくなったワーカは 38 人であった．また，$p = 1$ において重みが 0.7 以上のワーカ 441 人のうち，$p = 2$ で重みが大きくなったワーカは 240 人であった．同じように，モデル特徴量抽出法における $p = 1$ と $p = 2$ のワーカの重みを比較する．$p = 1$ において

表 5.1　クラスタ数 $k = 4$ の結果

手法			データ品質	計算時間 (hour)
単純多数決			0.897	-
ワーカ別モデル構築手法			0.922	269
クラスタ別モデル構築手法	混同行列特徴抽出法	p=1	0.919	231
		p=2	0.919	230
	モデル特徴量抽出法	p=1	0.924	253
		p=2	0.920	251

重みが 0.2 未満のワーカ 31 人のうち，$p = 2$ で重みが小さくなったワーカは 15 人であった．また，$p = 1$ において重みが 0.7 以上のワーカ 470 人のうち，$p = 2$ で重みが大きくなったワーカは 312 人であった．この結果から，プロセスを繰り返す中で，作業精度の高いワーカはさらに重みが大きくなり，低いワーカは小さくなっていくことがわかる．しかし，$p = 1$ と $p = 2$ においてデータの品質に差がないことから，$p = 1$ の時点でワーカの作業精度を正しく評価できていると考えられる．

また，提案した三つの手法において，ワーカ個人モデル構築手法よりクラスタ別モデル構築手法の方が模擬ワーカ構築にかかる計算時間が削減した．クラスタリングを活用することで，ゼロから学習を始める必要がある模擬ワーカの数を減らし，ファインチューニング時の学習効率を向上させることができる．そのため，クラスタリングをしないよりする方が，模擬ワーカ構築にかかる時間を削減することが可能であったと考える．

ワーカ別モデル構築手法とクラスタ別モデル構築手法で，それぞれで構築した模擬ワーカの性能精度を比較する．混同行列特徴量抽出法における $p = 1$ で構築した模擬ワーカの性能精度が，ワーカ別モデル構築手法で構築した模擬ワーカの性能精度より高くなったワーカは，569 人中 295 人であった．また，モデル特徴量抽出法における $p = 1$ で構築した模擬ワーカの性能精度が，ワーカ別モデル構築手法で構築した模擬ワーカの性能精度より高くなったワーカは，569 人中 306 人であった．どちらの手法を用いた場合でも，ワーカ別モデル構築手法を用いるより模擬ワーカの性能精度が高くなったことが確認できた．この結果から，クラスタごとに模擬ワーカを構築することによって，学習に使用するデータが増えるため，モデルの性能精度が高まると考えられる．

模擬ワーカの性能精度が向上することによって，模擬ワーカが付与したラベルの信頼性が向上する．そのため，正確なワーカの作業精度を算出することが可能となり，最終的なデータの品質が高くなると考えられる．また，ワーカ別モデル構築手法よりクラスタ別モデル構築手法の方が，模擬ワーカの構築にかかる計算時間を削減することができているため，クラスタごとに模擬ワーカを構築する手法は有効であると考える．

第 6 章　おわりに

本研究では，模擬ワーカを利用してクラウドソーシングにおけるデータ品質の向上と，模擬ワーカ構築にかかる計算時間の削減を目的としている．クラウドソーシングではインターネット上でワーカを募集するため，ワーカの作業精度が不明であり，作業精度が低いワーカの影響でデータの品質が低くなることがある．そこで，作業精度に応じた票の重みをワーカに与えることによって，作業精度の高いワーカの影響が大きくなりデータの品質を向上させることができると考えた．しかし，ワーカによって作業数や作業したタスクが異なるため，公平な作業精度を求めることが難しい．そこで，ワーカの作業データを使用して模擬ワーカを構築し，模擬ワーカにラベル付けを行わせることで，全ワーカが同じ作業を行った擬似環境を作ることができる．そして，作業数やタスク難易度を考慮する必要がなくなり，公平に作業精度を算出することができると考えた．

しかし，全ワーカの模擬ワーカを構築するには，相当な計算時間が必要となる．そこで，作業特徴が類似するワーカが一定数存在すると仮定し，ワーカを作業特徴に基づくクラスタリングを行う．各クラスタに一つの機械学習モデルを構築し，その後ワーカ個人の作業データを用いてファインチューニングし模擬ワーカを構築する．クラスタリングする手法を用いることによって機械学習モデル構築時の学習データ量が多くなり，精度の高い機械学習モデルを構築することが可能であると考えた．また，ワーカ個人の作業データのみを使用して機械学習モデルを構築する場合より，類似の作業特徴をもつクラスタの機械学習モデルをワーカの作業データでファインチューニングを行うことによって，パラメータの更新幅が小さくなると考えられるため，短時間で機械学習モデルを構築することができると考えた．

また，ワーカの作業特徴として混同行列とワーカの機械学習モデルのパラメータを利用した．混同行列はワーカの正答率や回答の傾向を表しており，ワーカの作業特徴として利用可能であると考えた．また，ワーカの機械学習モデルのパラメータはワーカの作業データを使用して学習しており，ワーカ個人に特化した機械学習モデルであるため，パラメータはワーカの作業特徴として利用できると考えた．

本手法の有効性を検証するため，単純多数決手法とワーカ別モデル構築手法，クラスタリング別モデル構築手法を適用して実験を行った．また，クラスタリングを

する手法では，プロセス回数 $p = 2$ まで繰り返して実験を行った．実験の結果，単純多数決手法のデータ品質は 0.897 であり，ワーカ別モデル構築手法のデータ品質は 0.922 であり，クラスタ別モデル構築手法のうち最も高いデータ品質は 0.924 であった．クラスタ別モデル構築手法のうち，モデル特徴量抽出法において $p = 1$ のとき，データの品質が最も高かった．提案した手法全てにおいて，ベースラインの単純多数決手法よりデータの品質が高くなることが確認できた．また，ワーカ別モデル構築手法よりクラスタ別モデル構築手法の方が，短時間で模擬ワーカを構築することができた．クラスタリングをすることで，より効率よく学習することができ，模擬ワーカ構築の時間を削減することができた．

しかし，本稿では一つのデータセットを使用した実験しかしていないため，全てのデータセットにおいて提案手法が有効かどうかは不明である．そのため，他のデータセットを使用して同様の実験を行うことによって，提案手法が有効かどうかを検証する必要がある．また本論文では，模擬ワーカが付与したラベルを，実際にワーカが付与したラベルと同等に扱っている．つまり本研究では，実際にワーカが付与したラベルと模擬ワーカが付与したラベルを同等に扱うことは，模擬ワーカの性能を完全に信頼している．しかし，模擬ワーカの性能には限界があり，誤りが含まれる可能性もあるため，模擬ワーカが付与したラベルをそのまま用いるのはリスクがある．このため，模擬ワーカの性能を考慮した上でデータの集計をすることが必要である．

謝辞

本研究を進めるにあたって，指導教員である鈴木優准教授にはたくさん指導していただきました．深く感謝申し上げます．国内の学会や国際会議に参加する機会をくださり，論文執筆やプレゼンテーション資料の作成など，多くの場面でサポートしていただきました．特に，国際会議では，英語を聞き取ることができず困っていたときや，現地での移動時などに助けていただきました．先生のおかげで無事発表を終え，日本に帰ってくることができました．

事務員の井尾さんには，学会や短期雇用に関する手続きなどの場面で大変お世話になりました．研究の合間にお話しすることもあり，すごく楽しかったです．これからもお仕事頑張ってください．

後輩や同期のみんなとは，研究に関することだけでなく，雑談や飲み会などで楽しい時間を共有することができました．おそらく一番うるさかったのは私だと思いますが，私の話にたくさん付き合ってくれてありがとうございます．後輩たちは，これからも研究を続けていくと思いますが，無理をしすぎない程度に頑張ってください．同期のみんなは，それぞれの新しい場所で頑張っていきましょう．卒業してからも，一緒に飲みに行ったり遊びに行ったりできると嬉しいです．

研究がどうしようもなく嫌になったときやつらいとき，家族には話を聞いてもらったり励ましてもらったり，時にやさしく時にきびしく支えていただきました．皆様のご助力により終わりを迎えられたこと，心より深く感謝申し上げます．

参考文献

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

[2] Nana Ota and Yu Suzuki. A label aggregation method using worker quality in crowdsourcing. In *International Conference on Information Integration and Web Intelligence*, pp. 49–55. Springer, 2023.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, Vol. 30, pp. 5998–6008, 2017.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

発表リスト

[1] 太田奈那，鈴木優『分類時における特徴を用いたインストラクション自動生成手法の検討』東海関西データベースワークショップ，2023.
[2] Nana OTa, Yu Suzuki『A Label Aggregation Method using Worker Quality in Crowdsourcing』The 25th International Conference on Information Integration and Web Intelligence (iiWAS2023)，2023.
[3] Nana OTa, Yu Suzuki『A Method to Improve Crowdsourcing Outcome and to Reduce Calculation Costs Using Machine-Learning』The 13th International Workshop on Advances in Data Engineering and Mobile Computing(DEMoC2024)，2024.