NAIST-IS-MT1651099

修士論文

クラウドワーカの品質推定のための タスク介入による振る舞い拡張

松田　義貴

2018 年 3 月 12 日

奈良先端科学技術大学院大学
情報科学研究科 情報科学専攻

本論文は奈良先端科学技術大学院大学情報科学研究科に
修士（工学）授与の要件として提出した修士論文である．

松田　義貴

審査委員：

中村 哲 教授	（主指導教員）
安本 慶一 教授	（副指導教員）
鈴木 優 特任准教授	（副指導教員）
田中 宏季 助教	（副指導教員）

クラウドワーカの品質推定のための タスク介入による振る舞い拡張*

松田　義貴

内容梗概

マイクロタスク型のクラウドソーシングは，安価に大量のタスクを発注できるというメリットがある．一方で，能力が無いワーカや意図的に不適切な作業を行うワーカが存在するため，依頼者が期待するような結果を得ることができるとは限らない．このような低品質なワーカを排除もしくは指導することができれば，作業結果の品質向上が期待できる．そこで，重要な課題となるのが自動的なワーカの品質推定である．本研究では，ワーカの振る舞いを用いた品質推定手法に着目する．振る舞いを用いることで，特に，意図的に不適切な作業を行うワーカを検知することが可能であると考えられる．既存の研究では，振る舞いの分析ばかりに焦点が当てられ，振る舞いの取得に関しては議論されていない．しかし，単純なタスクでは取得できる振る舞いが限られ，取得できる振る舞いが少ないとワーカの振る舞いに差が見られないことが予想される．また，取得するワーカの振る舞いがより詳細であることが望ましいと考えられる．そこで本研究では，タスクに対してワーカの振る舞いを取得する仕組みを追加し，多くの振る舞いを取得することによって，ワーカの品質をより高精度に推定する手法を提案する．提案手法では低品質なワーカの推定精度が向上した．また，タスク介入によって取得可能となったコンテンツの閲覧時間や回数などの振る舞いがワーカの品質推定に重要な役割を果たしていることが分かった．一方で，処理時間とワーカの離脱率が増加するというトレードオフの関係があった．

キーワード

*奈良先端科学技術大学院大学 情報科学研究科 情報科学専攻 修士論文, NAIST-IS-MT1651099, 2018 年 3 月 12 日.

クラウドソーシング，ワーカの品質推定，振る舞い，タスク設計，トレードオフ

Behavior expansion by task interference for the quality estimation of the crowd worker*

Yoshitaka Matsuda

Abstract

The micro-task type crowdsourcing has the advantage of ordering a lot of tasks inexpensively. On the other hand, it is not always possible to get results that the requester expects because there are workers who cannot provide correct answers and workers who intentionally perform poor works. The elimination and instruction of such low-quality workers will improve the work qualities. Thus, a significant task is the estimating the worker quality automatically. In this research, we focus on estimation method by workers' behavior. By using workers' behavior, we can discover workers who intentionally perform poor work. Existing studies focus on only how to analyze the behavior and don't discuss how to get the behavior. However, we suppose that there will be no difference in worker behavior if a type of behavior is few. Furthermore, it is desirable to get more detailed behaviors. In this research, we propose a method to estimate worker quality by adding a mechanism to get many types of workers' behavior to tasks. This method improves the estimation accuracy of low-quality workers. In addition, we find that the behaviors such as browsing time and count that can be gotten by task interference play an important role in estimating worker quality. However, there is a trade-off between accuracy improvement and influence on workers such as processing time and motivation.

Keywords:

crowdsourcing, worker quality estimation, behavior, task design, trade-off

*Master's Thesis, Department of Information Science, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-MT1651099, March 12, 2018.

目次

図目次

表目次

第 1 章　はじめに

1.1　研究の背景

計算機で解くには難しい問題を人間の処理能力を利用して解決するヒューマンコンピュテーションが注目を浴び，成果を挙げている [25]．近年では，このヒューマンコンピュテーションにおける人間の労働力を確保する場として，クラウドソーシングがよく利用されている [31]．クラウドソーシングとはインターネットを通して不特定多数の人々にタスクを依頼する仕組みである [19]. 本研究では, 誰でも簡単に作業を行うことができ，一つの作業単位が数秒から数分程度で完了するようなマイクロタスク型と呼ばれるクラウドソーシングを対象とする．例えば，データ収集やデータ入力のようなタスクが挙げられる．収集したデータは直接利用される場合もあれば，機械学習における教師データとして利用される場合もある [26][49].

クラウドソーシングが出現する以前，データ入力作業は組織内の従業員など十分に訓練された人により行われていた．そのため，金銭的，時間的なコストの制約があり，短期的に大量のデータを処理することが困難であった．しかし，マイクロタスク型のクラウドソーシングには，インターネットを通して多数の作業者を募ることで，安価に大量のタスクを発注できるという利点があり，従来の課題を解決できる．

一方で，通常の労働環境とは異なり，タスク依頼者が監視できない環境下で不特定多数のワーカが作業を行う．このような労働環境によって不適切なワーカがタスクに従事することがある．例えば，タスクの難易度によっては正しい答えを導くことができないワーカがタスクに従事すると，正しい作業結果を得ることができない．他にも，金銭目的など悪意を持って不適切な作業を行うワーカが存在することが知られている [24]．このような能力が無いワーカや意図的に不適切な作業を行うワーカの存在によって，作業結果の品質が低下するという問題がある．

1.2　研究の目的

品質が高いワーカを確保することができれば，作業結果の品質を向上させることができる．ところが，品質が高いワーカを確保するためには，ワーカの品質を算出しなければならない．依頼者が作業結果を一つひとつ確認し，ワーカの品質を算出することはクラウドソーシングを行うメリットがなくなるため，自動的にワーカの品質を算出する仕組みが必要となる．

本研究では，マイクロタスク型のクラウドソーシングにおいて，ワーカの振る舞いからワーカの品質を自動的に推定する手法を提案する．振る舞いを用いることで，特に，意図的に不適切な作業を行うワーカを検知することが可能であると考えられる．既存の研究では，振る舞いの分析ばかりに焦点が当てられ，振る舞いの取得に関しては議論されていない．しかし，振る舞いの取得方法を工夫することによってワーカの品質推定精度が向上することが期待できる．本研究では，多くの振る舞いを取得するためにタスクに介入すること, すなわち，振る舞いを取得するための仕組みをタスクに導入することによって，ワーカの品質推定精度の向上を目指す．また，既存の研究では扱われなかったが，タスク介入によって取得可能となった振る舞いがワーカの品質推定に重要な役割を果たすことを検証することで，ワーカの品質推定のためのタスク設計の重要性を示す．さらに，タスク介入によるワーカへの影響のトレードオフを調査することを本研究の目的とする．

1.3　本論文の構成

本論文の構成について述べる．第 2 章では，本論文に関する基本的な事項をまとめる．まず，クラウドソーシングについて説明し，提案手法で用いる機械学習やその評価尺度について説明する．第 3 章では，作業結果の品質管理手法について紹介する．特に，ワーカの品質を推定するアプローチについて詳しく述べる．第 4 章では，我々が行ったタスクについて説明する．さらに，既存の研究を用いたワーカの品質推定手法について説明する．第 5 章では，ワーカの品質を推定する提案手法について説明する．まず，提案手法のアイデアについて述べ，具体的な手法について説明する．第 6 章では，提案手法の有効性を検証するための評価実験について説明し，結果を述べる．最後に第 7 章では，本論文をまとめ，今後の課題を示す．

第 2 章　基本的事項

本章では，本論文に関して基本的な事項についてまとめる．はじめに，クラウドソーシングに関して説明する．次に，提案手法で用いる機械学習の概念やランダムフォレストと呼ばれる手法とオーバーサンプリング手法の 1 つである SMOTE アルゴリズムについて説明する．本研究では，ランダムフォレストによってワーカの品質を推定する．また，不均衡データを学習する対策として SMOTE アルゴリズムを用いている．最後に，提案手法の性能評価に用いる評価尺度に関して説明する．

2.1　クラウドソーシング

クラウドソーシングとは，インターネットを通して不特定多数の人に仕事を依頼すること，またはその仕組み（図 2.1）のことである．クラウドソーシングという名前は群衆（crowd）と業務委託（outsourcing）を組み合わせた言葉であり，2005 年に Howe ら [19] によって命名された．特定の人や企業などに作業を委託するアウトソーシングに対して，不特定多数の人々に作業を委託することがクラウドソーシングの特徴である．この作業を委託された不特定多数の人々はワーカと呼ばれ，作業はタスクと呼ばれる．

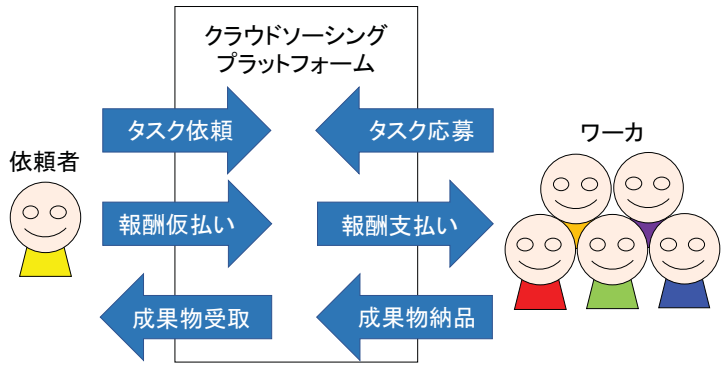


図 2.1: クラウドソーシングの仕組み

タスクの内容はホームページの作成やロゴ作成，記事執筆，翻訳，データ収集，データ入力など多岐にわたる．これらのタスクは大きく三つに分けることができ

る．この三つのタスク形式について表 2.1 にまとめる．一つ目はプロジェクト型である．プロジェクト型では依頼者が条件に合うワーカと契約を結び，ホームページ作成や記事執筆など比較的大きなタスクが扱われる．二つ目はコンペティション型である．コンペティション型の特徴は，最も良い成果物を提出したワーカのみに作業報酬が支払われることである．タスクはロゴのデザインやチラシ作成など比較的難易度が高く，専門的な知識やスキルを持つワーカが従事することが多い．三つ目はマイクロタスク型である．マイクロタスク型のクラウドソーシングは誰でも簡単に作業でき，一つの作業単位が数秒から数分程度で完了する．例えば，企業や大学が大量のデータを収集したい時などに利用され，作業報酬も低く設定される．

表 2.1: タスクの分類*

形式	特徴	報酬	タスクの例
プロジェクト型	依頼者とワーカは契約を結ぶ．数日から数ヶ月で完了．	数千円から数百万円	ホームページ作成・記事執筆
コンペティション型	最も良い成果物を提出したワーカのみ報酬を獲得．数分から数日で完了．	数千円から数十万円	ロゴデザイン・チラシ作成
マイクロタスク型	簡単な作業を大勢のワーカに依頼．数秒から数分で完了．	数円から数百円	データ収集・データ入力

*「2014 年版中小企業白書」を参考に作成

代表的なクラウドソーシングのプラットフォームとして Amazon Mechanical

Turk*がある．他にも CrowdFlower†があり，これら二つのサービスは世界展開されている．また，日本で有名なプラットフォームにはクラウドワークス‡やランサーズ§などがある．

近年では，インターネットの普及や機械学習の台頭によって，データ収集を目的としたクラウドソーシングがよく利用されている [50]．その理由の一つとして安価で大量のデータが手に入るという点が挙げられる．しかしその反面，専門家でないワーカや悪意のあるワーカが従事することで，提出された作業結果の品質が担保されないという問題がある．

2.2 機械学習

本節ではまず，機械学習の中でも教師あり学習について説明する．さらに，提案手法で用いたランダムフォレストと SMOTE アルゴリズムについて簡単に説明する．ランダムフォレストは，分類を行う教師あり学習アルゴリズムの一つである．SMOTE アルゴリズムは，不均衡データによる学習の問題をオーバーサンプリングによって解決する手法の一つである．

2.2.1 教師あり学習

機械学習の中でも教師あり学習について説明する．教師あり学習では，入力と出力からなるデータからモデルを学習し，未知のデータに対してモデルを当てはめ，出力を予測する．入力は説明変数や特徴量と呼ばれ，出力は目的変数や従属変数と呼ばれる．教師あり学習は回帰と分類の大きく二つに分けることができる．回帰は連続する値を予測するのに対して，分類はクラスに分ける，すなわちラベリングを行う．

*https://www.mturk.com/mturk/
†https://www.crowdflower.com/
‡https://www.crowdworks.jp
§https://www.lancers.jp/

2.2.2 ランダムフォレスト

ランダムフォレスト [4] は教師あり学習アルゴリズムの一つである．ランダムフォレストは回帰にも用いられるが，本節では分類問題に焦点を当てて説明を行う．ランダムフォレストの説明の前に，決定木 [36] について説明する．決定木はツリー構造のグラフを作成し，分類を行う教師あり学習アルゴリズムの一つである．決定木のアルゴリズムの概念図について図 2.2 に示す．決定木はトップダウンで再帰的にデータを分岐させ，最終的なクラスを決定する手法である．例えばワーカの品質を予測する場合，図 2.2 のように，作業数が 100 以上かどうか，平均処理時間が 10 秒以上かどうかなどの条件でワーカを分岐し，最終的に高品質か低品質かを決定する．

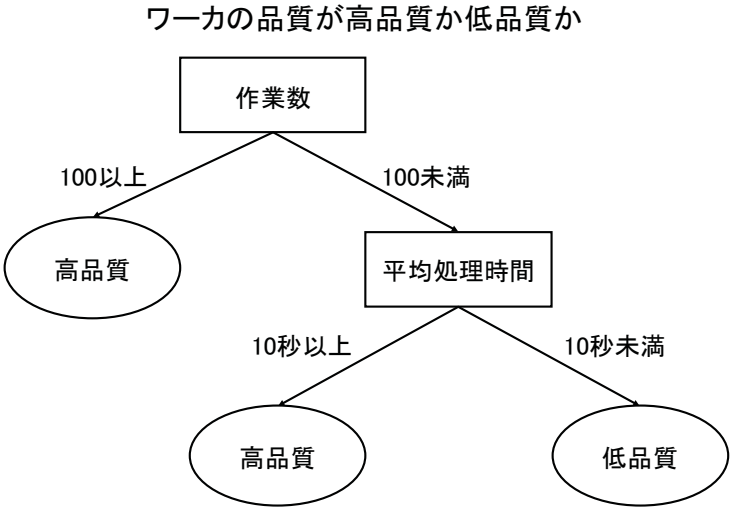


図 2.2: 決定木の概念図

ランダムフォレストはこの決定木を弱学習器としたアンサンブル学習の一つである．弱学習器とは理想的な真の学習器と若干の相関がある学習器であり，一つの弱学習器の精度はあまり良くない．アンサンブル学習とは複数の弱学習器が予測した複数の結果を統合して最終的な予測結果を決める手法である [11]．ランダムフォレストの概念図を図 2.3 に示す．まず，教師データからランダムに選択された教師データの一部で決定木を作成する．この際，決定木を作成するために用いる特徴量もランダムに選択する．複数の決定木が出した予測結果の多数派のクラスをランダ

ムフォレストの予測結果とする．アルゴリズムが単純であるのに関わらず，性能が良いことが知られており [14]，特にコンピュータビジョンの分野へ応用されている [39].

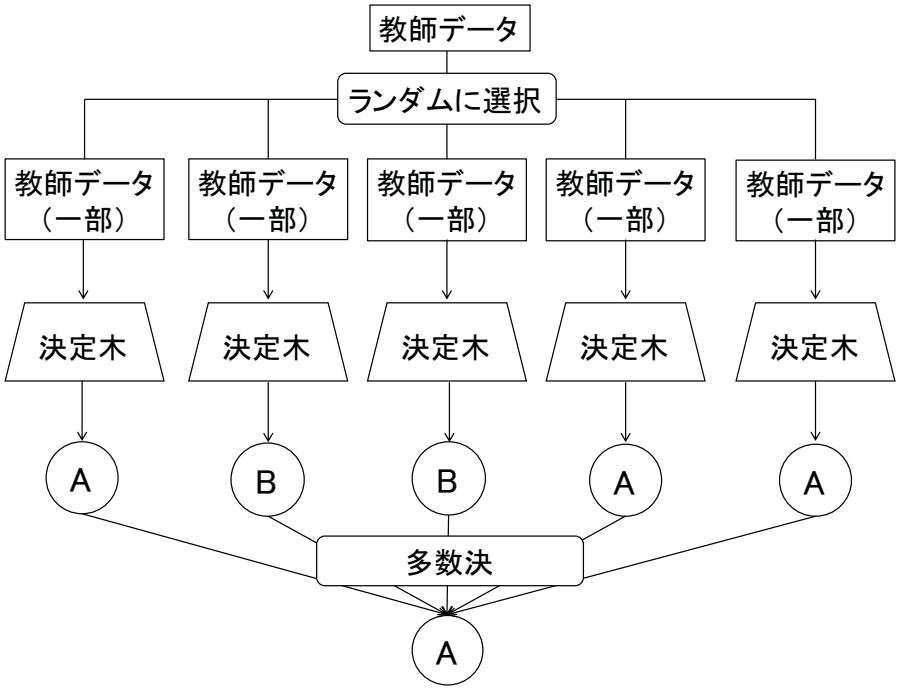


図 2.3: ランダムフォレストの概念図

2.2.3 SMOTE アルゴリズム

オーバーサンプリング手法の一種である SMOTE アルゴリズム（Synthetic Minority Over-sampling Technique）[8] について説明する．分類問題を考えた時，各クラスに属するサンプル数に偏りがあるデータを不均衡データと呼ぶ．不均衡データに対して何も工夫せず，分類器を作成すると，多数派に分類しやすくなり少数派のクラスの分類精度が低くなることが知られている [20].

そこで，少数派のデータを人工的に増加させるオーバーサンプリングという手法が用いられる．最も単純なオーバーサンプリング手法として，ランダムオーバーサンプリングがある [3]．少数派のデータをランダムに選択し，複製することでデー

タ数を増やす手法である．しかし，ランダムオーバーサンプリングは既存のデータを複製することになり過学習 [44] を引き起こしやすい．

この問題を解決した手法が SMOTE アルゴリズムである．SMOTE アルゴリズムでは既存のデータの複製ではなく，既存のデータにノイズを加えてデータを増やす手法である．図 2.4 に $k = 2$ の時のアルゴリズムの概要を示す．以下のステップで少数派のデータのオーバーサンプリングを行う．

1. 少数派のデータをランダムに一つ選択する．
2. そのデータの k-最近傍にあるデータを探索する．
3. ステップ 1 で選んだデータとステップ 2 で探索したデータの間でランダムにデータを生成する．

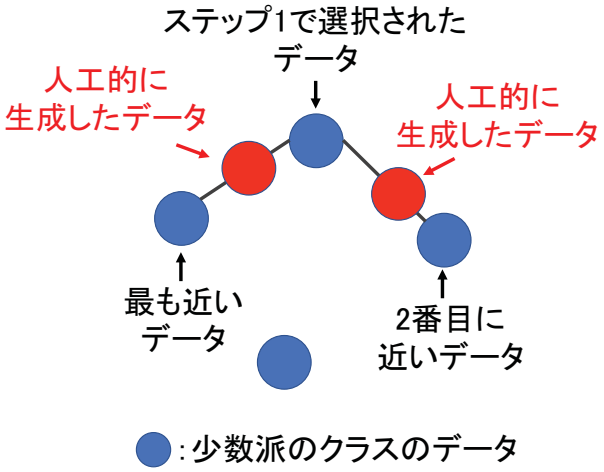


図 **2.4:** SMOTE アルゴリズムの概念図

2.3 評価尺度

本節では，提案手法の評価に用いた混同行列と適合率や再現率，F 値などの各指標について説明する．

2.3.1 混同行列

混同行列（Confusion Matrix）とは，クラス分類の結果をまとめた表のことである．n クラス分類の場合，$n \times n$ 行列で表される．2 クラスの場合の混同行列を表 2.2 に示す．表 2.2 中の各項目は以下のような名前が付いている．

表 2.2: 混同行列

		予測	
		陽性（Positive）	陰性（Negative）
正解	陽性（Positive）	真陽性（True Positive）	偽陰性（False Negative）
	陰性（Negative）	偽陽性（False Positive）	真陰性（TrueNegative）

- 真陽性（True Positive）：正解が陽性であり，分類結果も陽性の場合．
- 偽陰性（False Negative）：正解が陽性であり，分類結果は陰性の場合．
- 偽陽性（False Positive）：正解が陰性であり，分類結果が陽性の場合．
- 真陰性（True Negative）：正解が陰性であり，分類結果が陰性の場合．

このように，陽性のサンプルのうち正しく陽性と分類された数と誤って陰性と分類された数を分かりやすく見ることができる．同様に，陰性のサンプルのうち正しく陰性と分類された数と誤って陽性と分類された数が一目で分かる．

2.3.2 各種指標

混同行列に基づいて，真陽性，偽陰性，偽陽性，真陰性の数から分類の性能を評価する指標が存在し，適合率（Precision），再現率（Recall），F 値（F-measure）の三つの指標について説明する．説明のために，真陽性の数を TP，偽陰性の数を FN，偽陽性真の数を FP，真陰性の数を TN として表す．

適合率とは，陽性と分類されたデータのうち、真に陽性のデータの割合である．適合率は以下の式で定義される．

$$Precision = \frac{TP}{TP + FP} \tag{2.3.1}$$

再現率とは，陽性のデータのうち、陽性と分類できたデータの割合である．再現率は以下の式で定義される．

$$Recall = \frac{TP}{TP + FN} \tag{2.3.2}$$

F 値とは，適合率と再現率を総合的に判断するための指標であり，調和平均によって表される．F 値は以下の式で定義される．

$$F\text{-}measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{2.3.3}$$

第 3 章　関連研究

本章では，マイクロタスク型のクラウドソーシングでは作業結果の品質が保証されないという課題に対する研究について紹介する．最初に，品質管理全体に関する様々なアプローチを紹介する．次に，本研究で注目したワーカの振る舞いを用いたワーカの品質推定に関する研究を紹介する．最後に，既存の研究に対する本研究の位置付けについて述べる．

3.1　品質管理に関する研究

1.2 節で述べた，マイクロタスク型のクラウドソーシングでは作業結果の品質が保証されないという課題に対して以下の三つのアプローチが提案されている．これら三つのアプローチは併用することが可能であるが，コストとのトレードオフを考慮して作業結果の品質を向上させることが望ましい．

1. タスク設計
2. 作業結果のフィードバックや統合
3. 高品質なワーカの確保

一つ目のタスク設計によるアプローチでは，ワーカの能力を最大限引き出すようなタスクを設計することで，作業結果の品質を向上させることを目指している．翻訳や音声の書き起こしタスクなど一つひとつの作業が大きいタスクに対して作業をできるだけ小さい単位に分割する研究 [27][13][1][24]，ワーカのタスクへの割り当てに関する研究 [2][12][47][46]，タスクのインタフェースに関する研究 [32]，ワーカ同士にコミュニケーションを取らせる研究 [29] などがある．これらの研究ではワーカの能力を最大限引き出し，一つひとつの作業結果の品質を向上させている．タスクを処理する潜在的な能力があり，最大限の努力をするワーカには有効なアプローチであると考えられる．しかし，理想的なタスクが設計できたとしても，意図的に不適切な作業を行うワーカは，タスクの設計によらず意図的に不適切な作業を行うことが予想できる．

二つ目の作業結果のフィードバックや統合によって作業結果の品質を向上させるアプローチでは，間違った作業結果が生成されたとしても，全体として正しい作業結果を得ることを目指した研究である．作業結果のフィードバックとは，クラウドソーシングで得た結果を再度クラウドソーシングを行い確認する手法である [27][15]．また，よく利用されるのが作業結果を統合する冗長化と呼ばれる手法である [41]．冗長化では，複数のワーカに対して同じ作業を割り当てる．例えばデータへのラベル付け作業において，一つのデータに対して複数人のワーカにラベル付けを依頼する．この作業によって得られた複数のラベルを多数決 [38][42] や EM アルゴリズム（Expectation–Maximization Algorithm）[9][30]，機械学習 [7][43][21][40] によって統合し，最終的なラベルを導く．冗長化は，ワーカの大多数が正確な作業を行うという仮定のもと成立している．そのため，不適切なワーカが多く存在する場合には機能しない．また，依頼者にとっては，冗長化させるワーカ数に比例して，金銭的や時間的なコストが増大してしまう．また，能力がないワーカや意図的に不適切な作業を行うワーカのような正しい作業結果を納品できないワーカの存在を黙認することになり，今後のクラウドソーシング全体としては，作業結果の品質が保証されないという課題の解決にはならない．

三つ目の高品質なワーカの確保によって作業結果の品質を向上させるアプローチでは，低品質なワーカを特定しタスクに従事することを禁止したり，タスク内容を正しく理解していないワーカに対して指導を行なったりすることで，高品質なワーカだけを雇用する．ここで言うワーカの品質とは，正しい作業結果を納品する確率のことを指す．低品質なワーカの存在という作業結果の品質が低下する原因を直接的に解決する唯一のアプローチである．低品質なワーカを特定するためには，ワーカの品質を算出しなければならない．依頼者が作業結果を一つひとつ確認し，ワーカの品質を算出することはクラウドソーシングを行うメリットがなくなるため，自動的にワーカの品質を算出する仕組みが必要となる．ワーカの品質を算出するアプローチとして以下の三つが提案されている．

1. ゴールドタスクによる測定
2. 作業結果の分析による推定
3. 振る舞いの分析による推定

一つ目のゴールドタスクによる測定では，あらかじめ答えの分かっているゴールドタスクをワーカに課し，ワーカの品質を測定する．実際のタスクに混入して非明示的にワーカの品質を測定する場合や，明示的に事前にゴールドタスクを課し，作業前にワーカをフィルタリングする場合がある [22]．実際のタスクにゴールドタスクを非明示的に混入してワーカの品質を測定する場合，答えの分かっているゴールドタスクへの回答に対しても作業報酬をワーカに支払う必要がある．事前にゴールドタスクを課してワーカの品質を測定する場合，この明示的な試験には真面目に取り組むが，実際のタスクでは不真面目に作業を行うワーカもいるかもしれない．また，個々のタスクに対して，ワーカの品質を正しく見極めるゴールドタスクを作成する必要があり，タスク依頼者の手間とコストが発生する．

二つ目の作業結果の分析による推定では，冗長化によって複数の作業結果を獲得し，その作業結果からワーカの品質を推定する．EM アルゴリズムによって作業結果の統合と同時にワーカの品質を推定する手法 [9][34][30]，作業結果を入力とした機械学習による手法として，不適切なワーカを排除する研究 [16][45] やワーカのスコアリングやランキングを行う研究 [37][33][5] がある．これらの手法は冗長化が前提となっており，ワーカに対する金銭的なコストや依頼者の時間的なコストの削減が見込めない．

正しい作業結果を納品しない低品質なワーカの中には，作業遂行能力がないワーカ，タスクの内容を正しく理解していないワーカ，注意力が散漫なワーカのような非意図的に低品質となってしまっているワーカもいれば，金銭目的などの悪意を持って不適切な作業を意図的に行うワーカもいると考えられる．しかし，一つ目と二つ目のアプローチでは，意図的な低品質ワーカかどうかは考慮することができない．例えば翻訳タスクや記事執筆タスクでは，翻訳文や執筆された記事から意図的に不適切な作業を行っているかどうか判断できるかもしれない．しかし，作業結果が数値となるようなタスクやデータ分類のような作業結果が極めて単純なタスクでは，作業結果から意図的かどうかを判断することは困難である．さらに既存の研究では，低品質なワーカを排除することを目的としている．今後ますますクラウドソーシングが発展し，一つの職業として確立されることを考えたとき，非意図的な低品質ワーカを排除することは望ましくない．非意図的な低品質ワーカは適切な指導やサポートによって能力を向上させる可能性があり，貴重な労働力を失うことに

なるからである．

そこで，三つ目のアプローチとして，ワーカの振る舞いに着目した研究が近年行われている [35][17][23][28]．金銭目的なワーカは，できるだけ素早くタスクを処理することを考えるだろう．例えばツイート分類タスクでは，ツイートを読まずに分類を行うかもしれない．一方で，能力は低いが真面目に作業を行うワーカがツイートを読まないということは考えられない．このようなワーカの振る舞いの違いに着目することで，特に，意図的に不適切な作業を行うワーカを検知することが可能である．また，ゴールドタスクを作成する必要がなく，冗長化を前提としないため，コストを削減することも可能である．そこで本研究では，このアプローチに注目し，次の 3.2 節では，ワーカの振る舞いを用いたワーカの品質推定に関する研究を紹介する．

3.2　振る舞いを用いたワーカの品質推定に関する研究

本節では，ワーカの振る舞いからワーカの品質推定に関する研究について紹介する．Zhu ら [48] は，クラウドソーシングにおける作業中の振る舞いはワーカによって異なることを示した．ワーカの振る舞いをワーカの品質推定に初めて用いたのは Rzeszotarski ら [35] である．Rzeszotarski らは，マウスやキーボード操作のタイミングや回数，処理時間などをワーカの振る舞いとして用いた．他にも，Hirth ら [17] は，ページのスクロールやラジオボタンのクリック間隔からワーカがどの問題を解いてるかを推定し，問題ごとの処理時間をワーカの振る舞いとして用いた．そして彼らは，ワーカの振る舞いを入力とした機械学習アルゴリズムによってワーカの品質を推定した．Cao ら [6] は，振る舞いからワーカをリアルタイムにスコアリングする手法を提案した．Kazai ら [23] は，充分に訓練されたワーカと通常のワーカの振る舞いに差があることを示した．さらに，ワーカの品質推定のための機械学習モデルの学習時に，訓練者の振る舞いを用いることによってワーカの品質推定精度を向上させた．

これらの既存の研究では，ワーカの振る舞いの扱い方や分析方法ばかりに焦点が当てられてきた．しかし，振る舞いの取得に関する議論はされていない．その理由としては，豊富な種類の振る舞いが存在するタスクが研究対象となっているためだ

と考えられる．しかし，マウスカーソルの動きやページのスクロールがほとんど発生しない単純なタスクでは，取得される振る舞いの種類が限定され，既存の研究をそのまま適用することは困難である．取得できる振る舞いの種類が少ないと，ワーカの品質がその取得した振る舞いと関係があるとは限らないためである．さらに，少数の振る舞いだけでは優秀なワーカの品質を正確に判断できない．例えば，処理時間だけでワーカの品質を判断しようとすると，早く正確に処理できるワーカの品質を低く見積もってしまう可能性がある．そこで本研究では，既存の研究では議論されてこなかった振る舞いの取得に着目し，ワーカの品質推定に与える影響を検証する．

第 4 章　ベースライン

本章では，本研究で行ったクラウドソーシングのタスクについて説明する．さらに，既存の研究に倣ったワーカの品質を推定する手法を説明し，その手法を本研究でのベースラインとする．

4.1　タスク概要

本節では我々が行なったクラウドソーシングのタスクについて説明する．図 4.1 に作業画面を示す．タスク内容は，与えられたツイートが以下の 2 点を満たすかどうかの分類である．

- 京都の様子が分かる．
- その様子は京都観光において有益である．

ワーカは，はい，いいえ，ツイートが表示されないのいずれかを選択する．我々の構築したクラウドソーシングプラットフォームでは，ワーカの作業時に twitter からツイート内容を WebAPI により取得するため，ツイート投稿者や twitter 社がツイートを削除した場合，当該ツイートを作業画面に表示することができない．そのため，ツイートが表示されないという選択肢を用意した．ワーカはツイートを読み，三つの選択肢から答えを選択する．送信ボタンを押すと，次のツイートが表示され，この作業を繰り返す．

例えば，「人身事故の影響で京都駅が混雑しています．」というツイートが与えられた場合，ワーカは，はいを選択しなければならない．なぜなら，ツイートから「京都駅が混雑している」という京都の状況が分かり，「京都駅」は京都観光者が利用する可能性があり，有益であると考えられるためである．

4.2　品質推定手法

本節では，4.1 節で説明したタスクにおいてワーカの振る舞いからワーカの品質を推定する手法について説明する．手法の概念図を図 4.2 に示す．この手法は

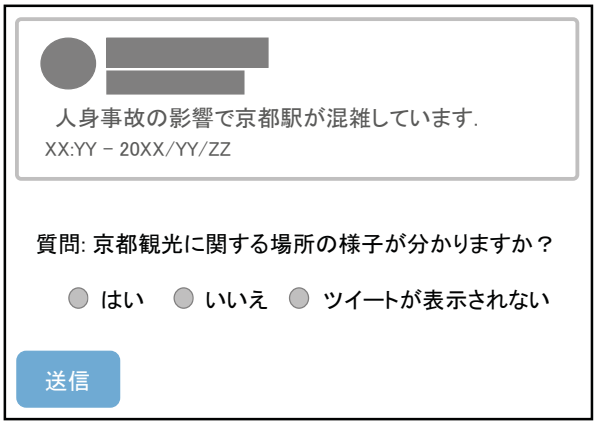


図 4.1: ベースラインの作業画面

Kazai ら [23] の手法を可能な範囲で模倣している．ワーカの品質推定には機械学習を用いる．つまり，複数人のワーカの振る舞いを取得し，そのワーカの品質を求め，振る舞いと品質の関係を学習し分類器を作成する．そして，新たなワーカが作業を行った時に作成した分類器を用いて，そのワーカの振る舞いからそのワーカの品質を予測する．

振る舞いや分類器の作成について述べる．まず，我々はワーカのタスク処理をプログラムにより自動的に監視し，表 4.1 に示すワーカの振る舞いを取得する．これらの振る舞いはワーカが一つのツイートを分類するごとに取得する．つまり，100 回ツイートを分類したワーカにはそれぞれの振る舞いが 100 個ずつ存在する．

次に，このワーカの振る舞いを用いて分類器を構築し，ワーカの品質が高品質か低品質かの分類を行う．ここで，分類の対象となるワーカは 100 回以上のタスク処理をしたワーカに限定する．まず，分類器への入力となる特徴量の構成について説明する．ワーカごとに取得したそれぞれの振る舞いの平均値，中央値，最大値，最小値，標準偏差，エントロピーをそれぞれ計算し，ワーカの特徴量とする．さらに，ワーカのタスク総処理回数と総処理時間も特徴量として用いる．

ここで，ワーカ w の品質 Q_w を式 (4.2.1) で定義する．N_w はワーカ w の総タスク処理回数，T_w はワーカ w の総タスク処理回数のうち正しい答えを選んだ回数である．つまり，ワーカの品質をそのワーカの正答率で定義する．我々は各ツイートに対して 10 人のワーカを割り当て，多数決によって各ツイートの正解ラベルを付

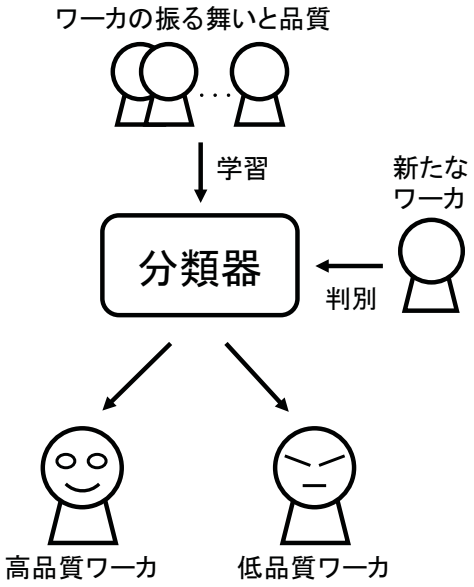


図 **4.2:** ワーカの質推定手法の概念図

表 **4.1:** 取得するワーカの振る舞い

ID	振る舞い	説明	サンプル
B1	処理時間	作業ページがロードされてから送信ボタンを押すまでの時間（秒）	5, 10, 60
B2	1 文字あたりの処理時間	処理時間をツイートの文字数で割った値（秒）	0.05, 0.1, 1
B3	回答時間	作業ページがロードされてから回答を選択するまでの時間（秒）	5, 10, 60
B4	1 文字あたりの回答時間	回答時間をツイートの文字数で割った値（秒）	0.05, 0.1, 1
B5	回答回数	回答を変更した回数（回）	1, 2, 5

与している．

$$Q_w = \frac{T_w}{N_w} \times 100 \tag{4.2.1}$$

ワーカの品質が下位 $\alpha\%$ のワーカを低品質ワーカ，それ以外のワーカを高品質ワーカと定義し，分類器の出力を低品質ワーカか高品質ワーカかのどちらかとする．すなわち，下から $\alpha\%$ に位置するワーカの正答率を β とすると，$Q_w \leq \beta$ ならばワーカ w は低品質ワーカとなり，$Q_w > \beta$ ならばワーカ w は高品質ワーカとなる．

ワーカの品質を分類する分類器としてランダムフォレスト [4] を用いる．本研究において特徴量となるワーカの振る舞いには外れ値だと考えられる値が存在する．例えば，ワーカが作業画面を開いたまま作業を放置した場合，処理時間が 1 時間を超えることがある．ランダムフォレストはこのような外れ値に対して影響を受けにくいという性質がある [14]．さらに，ワーカの品質に関係がない振る舞いが特徴量に含まれていても影響を受けにくい [14]．ランダムフォレストのハイパーパラメータはグリッドサーチによって決定した．

サンプル数が不均衡なデータをそのまま学習に用いた場合，全てのワーカが多数派のクラスに分類されてしまう可能性がある．そこで，SMOTE アルゴリズム [8] を用いて比率が 1:1 になるように少数クラスのサンプルを仮想的に増やすことにより，学習サンプル数の不均衡を解消する．

第 5 章　提案手法

本章では，ワーカの品質を推定する提案手法について説明する．まず，提案手法となるタスク介入アイデアについて述べる．次に，タスク介入の具体的な手法について説明する．最後に，ワーカの品質を推定する手法について説明する．

5.1　アイデア

3.2 節で述べたように既存の研究では，ワーカの振る舞いの扱い方や分析方法ばかりに焦点が当てられ，振る舞いの取得に関する議論がされていない．本研究ではその振る舞いの取得に注目する．

まず，4.1 節で説明したツイート分類タスクを例に作業の流れを考える．最初に，ワーカはツイートを読む．次に，そのツイートが京都観光に有益かどうかを考え，最後に回答を選択するという行動に移る．つまり，作業段階は 1) 読む，2) 考える，3) 答えるの 3 段階に分けられる．ベースラインの作業画面（図 4.1）のようなタスクを設計した場合，ワーカの振る舞いとして取得できるのは 3) 答えるの段階だけである．タスク内容は異なるが，既存の研究では，実際にワーカが行動に移す 3) 答えるの段階の振る舞いしか取得していない．そこで，1) 読むの段階の振る舞いを取得することができれば，より高精度にワーカの品質を推定することが可能であると考えた．本研究のタスクにおける 1) 読むの段階の振る舞いを取得することは，ツイート閲覧時間と閲覧回数を取得することと同値である．

さらに，クラウドソーシングでは，誰にも監視されていない状況でインターネット上で作業ができるという特徴から，他のことを行いながら作業が行われている場合があることが予想される．例えば，パソコンで他の作業を行ったりテレビを見ながら作業を行ったりしていることが考えられる．そのため，作業ページが開かれてから作業結果を提出するまでの時間は必ずしも実際に作業を行った時間と一致しない．したがって，より正確な作業時間を獲得することが必要であり，閲覧時間と閲覧回数を取得することがその一つとなり得る．

また，注意深いワーカは回答を選択した後，ツイートをもう一度読み直すかもしれない．しかし，読み直さないワーカもいることが予測される．閲覧時間と閲覧回

数を取得することができれば，このような 2 人の振る舞いの差も考慮することができる．

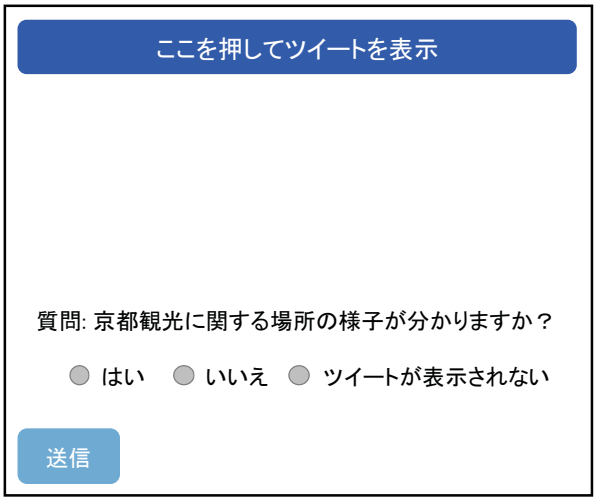
単純なタスクでは，3) 答えるの段階の振る舞いに差が出にくく，1) 読むの段階の振る舞いを取得することが特に必要であると考えたため，簡単なツイート分類タスクを例に実験を行った．しかし，作業段階を 3 段階に分けることができるタスクには適用可能である．また，既存の研究と併用することも可能である．

5.2 タスク介入手法

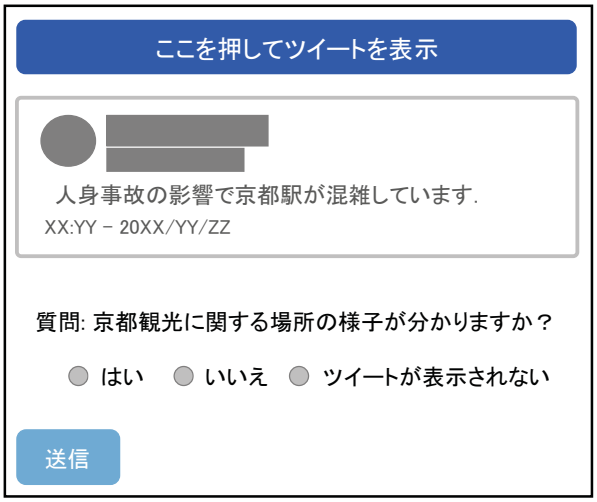
閲覧時間と閲覧回数を取得するための具体的なタスク介入手法について説明する．タスクの内容はツイートの分類であり，ベースラインと変わらない．しかし，提案手法では，初期状態ではツイートを表示しない．その代わりにツイートを表示させるためのボタンを設置する．作業画面を図 5.1 に示す．図 5.1a はツイート表示ボタンを押していない時の作業画面で，図 5.1b はツイート表示ボタンを押している時の作業画面である．ワーカがこのボタンを押下している時だけツイートが表示され，ボタンを離せばツイートは非表示となる．そうすることで，ツイートの閲覧時間や閲覧回数が取得可能となる．しかし，ワーカにとってはタスクの内容と直接関係はない無駄な作業が増えることになる．

5.3 品質推定手法

提案手法では，表 4.1 の振る舞いに加えて表 5.1 に示す振る舞いを取得する．ベースラインと同様に，各振る舞いの統計量をランダムフォレストの特徴量として用いて，ワーカの品質を推定する．つまり，ワーカの特徴量として用いる振る舞いだけがベースラインと異なり，特徴量の設計やワーカの品質推定は同様に行う．



(a) ボタンを押していない時



(b) ボタンを押している時

図 5.1: 提案タスクの作業画面

表 5.1: 追加で取得するワーカの振る舞い

ID	振る舞い	説明	サンプル
B6	閲覧回数	ツイート表示ボタンを押した回数（回）	1, 2, 5
B7	閲覧時間	ツイート表示ボタンを押している間の合計時間（秒）	5, 10, 60
B8	1 文字あたりの閲覧時間	閲覧時間をツイートの文字数で割った値（秒）	0.05, 0.1, 1
B9	1 回あたりの閲覧時間	閲覧時間を閲覧回数で割った値（秒）	5, 10, 60
B10	1 文字 1 回あたりの閲覧時間	1 回あたりの閲覧時間をツイートの文字数で割った値（秒）	0.05, 0.1, 1
B11	閲覧割合	処理時間のうちの閲覧時間の割合	0.1, 0.5, 0.8

第 6 章　評価実験

本章では，提案手法におけるタスク介入によって取得した振る舞いがワーカの品質推定に有益であるか検証するための評価実験について述べる．また，ワーカの品質推定とタスクに介入したことによるワーカへの影響の関係を検証する．

6.1　実験設定

本節では，本研究のために行ったクラウドソーシングの実験手順について説明し，収集したデータの大きさについて述べる．

6.1.1　実験手順

我々は独自のプラットフォームを作成し，ワーカは我々のプラットフォーム上で作業を行う．また，Javascript と jQuery ライブラリを使用してワーカの振る舞いを取得する．ワーカの募集と作業報酬の支払いは既存のクラウドソーシングのプラットフォーム（Crowdworks *）を通して行う．

我々のタスクに応募したワーカは全員雇用する．つまり，我々は雇用するワーカを事前に選別はしない．

ワーカは 100 ツイート分類するごとに 30 円の作業報酬を得ることができる．平均的なワーカは 1 時間に約 450 ツイート分類した．つまり，ワーカは 1 時間で 135 円稼ぐことができる．これは，標準的なクラウドソーシングの賃金（1.38 ドル/時間）とほとんど等しい [18].

ワーカは自身のタイミングで作業の中断や再開，離脱することができる．つまり，作業を行ったツイート数が 100 個以下であったため報酬を得ずに離脱したワーカや，1 万ツイート以上のタスクに従事したワーカも存在する．

*https://www.crowdworks.jp

6.1.2　収集データの概要

提案手法とベースライン手法で我々が雇用したワーカの数やツイートの分類総数などについて表 6.1 にまとめる．ここで，総ワーカ数とは我々のタスクに応募し 1 回でもタスクに従事したワーカの総数のことを指し，対象ワーカ数とは 100 回以上タスクに従事した分類対象となるワーカを指す．対象ワーカはベースラインと提案手法でそれぞれ 250 人を超えている．既存の研究では 100 から 200 ワーカ程度の実験規模であり，本研究の実験規模も十分であると言える．総タスク処理回数とは，全てのワーカでのタスクの処理回数である．

表 6.1: 収集したデータの規模

	総ワーカ数（人）	対象ワーカ数（人）	総タスク処理回数（回）	ツイート数（個）
ベースライン手法	439	250	132,594	–
提案手法	486	255	108,836	–
合計（重複無し）	793	439	241,430	26,713

また，我々はこの全てのタスク処理においてワーカの振る舞いを取得している．ツイート数とはワーカが 1 人でも分類を行ったツイート数である．我々は一つのツイートに対して最大 10 人のワーカを割り当てた．ただし，ツイートが削除され，作業画面に表示できなくなった場合は，ワーカの割り当てを取りやめた．

6.2　ワーカの品質推定

本節では，ベースライン手法と提案手法それぞれのワーカの品質推定結果について説明する．また，提案手法でのツイート表示ボタンによって取得可能となった振る舞いの有効性を検証する．さらに，正しく推定できたワーカとできなかったワーカの性質を分析する．

6.2.1 推定結果

ランダムフォレストを用いて，ワーカの品質が高品質であるか低品質であるかの分類を 5 分割交差検証で評価した．すなわち，本実験で用いるデータを 5 分割し，4 つのデータを学習データとして用い，残りの 1 つのデータで評価することを 5 回繰り返した．

まず，$\alpha = 10$ としてワーカの品質を推定した．すなわち，正答率が下位 10% のワーカを低品質ワーカ，それ以外のワーカを高品質ワーカと定義する．この時，低品質ワーカと高品質ワーカの境界となる正答率 β はベースライン手法で 65.5%，提案手法で 62.7% となった．

ベースライン手法と提案手法それぞれにおいて，5 回分の分類結果をそれぞれまとめた混同行列を表 6.2a，6.2b に示す．また，表 6.2c には，用いる振る舞いをベースライン手法と同じ振る舞いに限定にした場合の提案手法での分類結果も示す．以後，用いる振る舞いをベースライン手法と同じ振る舞いに限定にした場合の提案手法のことを限定的提案手法と呼ぶ．この分類結果を用いてワーカの排除や指導を行うことを考えた時，排除や指導に該当する低品質ワーカをできるだけ見逃さないことが重要である．なぜならば，低品質ワーカが作業結果の品質を低下させる原因となるためである．つまり，低品質ワーカの再現率が重要であると言える．ベースライン手法では低品質ワーカの再現率が 0.72 であった．一方で，提案手法では 0.84 に向上した．また，限定的提案手法では 0.64 であり，提案手法が低品質ワーカをより高精度に検知することができた．また，適合率や F 値を見ても，提案手法が最も良い精度であった．しかし，高品質ワーカを低品質と予測するケースが多く，改善の余地がある．

同様に，低品質ワーカを下位 5% と 15% としてワーカの品質の推定を行った．表 6.3 に β の値をまとめる．また，それぞれの β でそれぞれの手法で分類した時の分類結果を表 6.4 に示す．下位 5% と 15% のワーカを低品質ワーカと定義した時にも，提案手法が再現率だけでなく，適合率や F 値においても最も良い精度であった．これは，α によらず提案手法が効果的であることを示している．しかし，下位 10% のワーカを低品質ワーカと定義した時と同様に，高品質ワーカを低品質と予測するケースが多く存在した．

提案手法によって正しく分類できたワーカと，提案手法でも正しく推定できな

表 6.2: 分類結果の混同行列

(a) ベースライン手法

		予測	
		低品質	高品質
正解	低品質	18	7
	高品質	69	156

(b) 提案手法

		予測	
		低品質	高品質
正解	低品質	21	4
	高品質	72	158

(c) 限定的提案手法

		予測	
		低品質	高品質
正解	低品質	16	9
	高品質	65	165

表 6.3: 低品質ワーカと高品質ワーカの境界

α	手法	境界となる正答率 β
5%	ベースライン	65.5%
	提案手法	62.8%
10%	ベースライン	73.0%
	提案手法	72.6%
15%	ベースライン	80.7%
	提案手法	78.7%

表 6.4: 分類結果

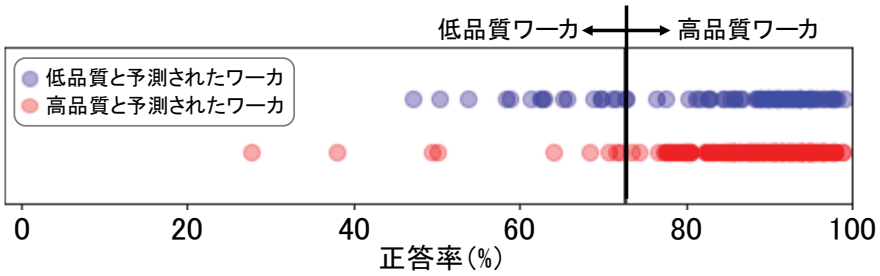
α	手法	適合率	再現率	F 値
5%	ベースライン	0.17	0.55	0.26
	提案手法	0.21	0.75	0.33
	限定的提案手法	0.16	0.58	0.25
10%	ベースライン	0.21	0.72	0.32
	提案手法	0.23	0.84	0.36
	限定的提案手法	0.20	0.64	0.30
15%	ベースライン	0.22	0.62	0.33
	提案手法	0.32	0.79	0.45
	限定的提案手法	0.27	0.71	0.39

かったワーカの特徴を分析する．図 6.1 に $\alpha = 10$ の時の分類結果の詳細について示す．図 6.1a が限定的提案手法での分類結果で，図 6.1b が提案手法の分類結果である．それぞれの点はワーカを表しており，横軸はワーカの正答率である．青で表されたワーカは我々の分類器で低品質と予測されたワーカであり，赤で表されたワーカは高品質と予測されたワーカである．理想的な分類器では，低品質ワーカと高品質ワーカの境界より左側には青い丸が並び，右側には赤い丸が並ぶことになる．また，正答率がより低いワーカほど正しく分類できるほうがより良い分類器であると言える．

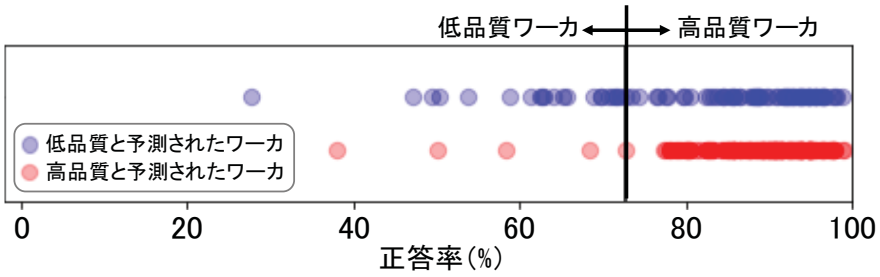
まず，最も正答率が悪いワーカに着目し，このワーカをワーカ A と呼ぶことにする．ワーカ A は限定的提案手法では高品質ワーカと誤って予測され，提案手法では低品質ワーカと正しく分類されていることが図 6.1 から分かる．ワーカ A の処理時間の平均は約 6 秒であり，平均的なワーカに比べ処理時間が短い．しかし，6.2.2 項にも述べたのと同様に，処理時間が短くても高品質なワーカも存在する．そのため，ワーカ A を低品質なワーカだと断定することは難しい．しかし，ワーカ A は 9 割以上のタスクで 1 回もツイートを閲覧していない．さらに，閲覧した場合でも閲覧時間は 2 秒に満たず，ワーカ A は明らかに適切な作業をしているとは言えない．このように，提案手法によって取得した閲覧時間や閲覧回数によって低品質ワーカ

であると判断することが可能になったと考えられる．

次に，2 番目に正答率が悪いワーカに着目する．このワーカをワーカ B と呼ぶことにする．ワーカ A は提案手法でも誤って高品質ワーカと分類してしまった例である．ワーカ B の閲覧回数は平均 1.5 回以上であり，閲覧時間の平均も約 12 秒である．平均的なワーカの閲覧時間の約 6.5 秒であり，約 2 倍ツイートを読んでいることになる．ワーカ B の正答率は低いが，振る舞いが不適切であるとは言い難い．そのため，提案手法でも高品質ワーカと分類してしまったと考えられる．ワーカ B は真面目に作業をしているが，タスクの内容を正確に理解できていない可能性がある．本手法では振る舞いのみでワーカを判断するため，このようなワーカを低品質ワーカと判断することはできなかった．正答率が悪い低品質なワーカを検知するためには，振る舞いだけでなく，作業結果も見る必要があると考えられる．



(a) 限定的提案手法



(b) 提案手法

図 6.1: 分類結果の詳細

6.2.2　重要な振る舞い

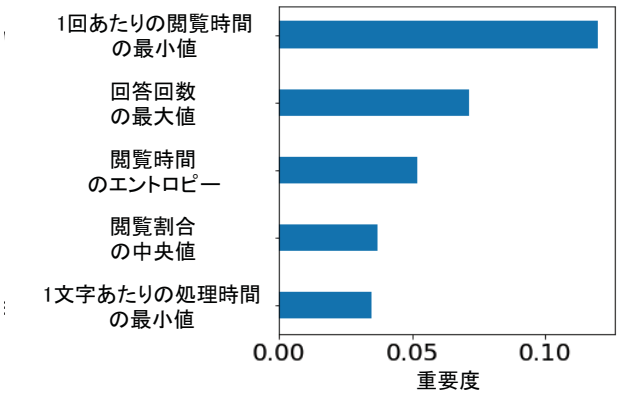
ランダムフォレストでは，特徴量の重要度を計算することができる．本実験では5 分割交差検証を行ったため，5 回の分類それぞれで重要度が計算される．ここでは，5 回の重要度の平均値をその特徴量の重要度として用いる．提案手法における重要な特徴量の上位 5 件を図 6.2 に示す．α の値，すなわち，低品質ワーカと高品質ワーカの境界に関わらず，重要度が高い特徴量の多くはタスクに介入したことによって得ることができた振る舞いの特徴量である．つまり，提案手法でなければ取得できない振る舞いが分類に重要な役割を果たしたと言える．

例えば，1 回あたりの閲覧時間の最小値は 3 種類の分類すべてで重要な振る舞いの Top3 に入った．1 回あたりの閲覧時間の最小値と正答率の関係を図 6.3 に示す．一つひとつの点はワーカを表している．1 回あたりの閲覧時間の最小値が大きいワーカは正答率が高い傾向がある．そのため，1 回あたりの閲覧時間の最小値から高品質なワーカを検出することは容易である．しかし，1 回あたりの閲覧時間の最小値が小さいからと言って正答率が低いとは限らないが，正答率が低いワーカは概ね 1 回あたりの閲覧時間の最小値が小さい．そのため，低品質なワーカを検知することもできるが，高品質なワーカを低品質と誤判別してしまう．これは，表 6.2 の混同行列にも表れている．

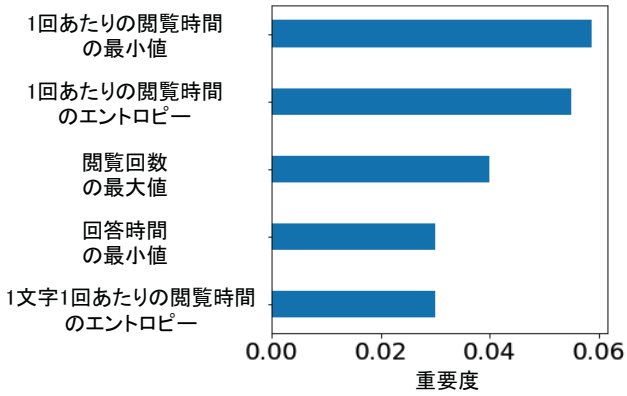
また，閲覧回数も分類に重要な役割を果たしていることが図 6.2 から分かる．高品質なワーカは回答を選択した後にもツイートを閲覧し直すなど，複数回閲覧していると考えられる．

6.3　タスク介入が及ぼすワーカへの影響

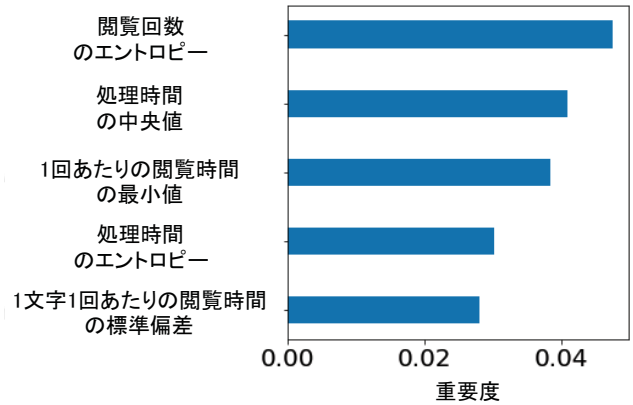
提案手法では，ツイートを読むためにボタンを押し続ける作業をワーカに対して課しているが，この作業自身はタスクの遂行に直接関係無い．これはワーカの品質推定のために行ったことであるが，作業結果の品質を低下させることや，ワーカが集まらないなどの問題があっては意味がない．そこで本節では，提案手法がワーカに与える影響についての分析結果の説明と議論を行う．ワーカへの影響として，タスクの難易度，処理時間とモチベーションの三つの観点から評価する．



(a) $\alpha = 5$



(b) $\alpha = 10$



(c) $\alpha = 15$

図 **6.2:** 重要な特徴量の上位 5 件

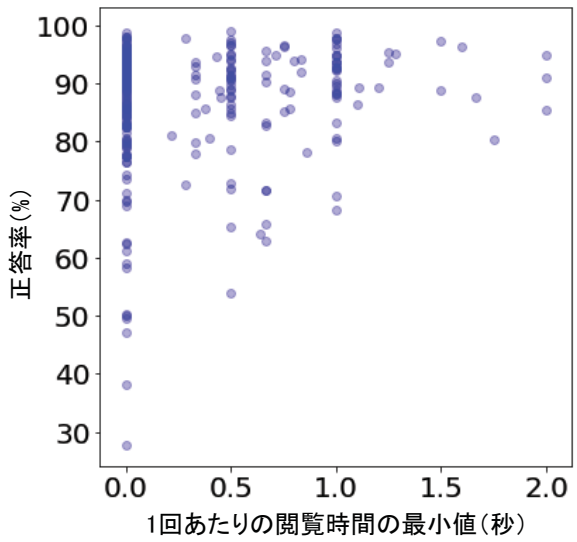


図 **6.3:** 1 回あたりの閲覧時間の最小値と正答率の関係

6.3.1 タスクの難易度

提案手法のタスクとベースライン手法のタスクでの正答率の違いと分布からタスクの難易度を評価する．まず，両方のタスクに従事した対象ワーカ 66 人を対象として，提案手法とベースライン手法でのそれぞれの正答率を図 6.4 に示す．横軸がベースライン手法での正答率，縦軸が提案手法での正答率を表し，一つひとつの点が各ワーカを表している．相関係数は 0.78 となっており，強い正の相関が見られる．

さらに，両方のタスクに従事したワーカの正答率の分布を図 6.5 に示す．横軸が正答率，縦軸がその正答率のワーカ数である．提案手法とベースライン手法でのタスクのそれぞれの正答率をウィルコクソンの符号順位検定 [10] を用いて検定したが有意な差は認められなかった（$p > 0.05$）．つまり，タスクの難易度が異なるとは言えない．

提案手法とベースライン手法での正答率には強い正の相関があり，有意差が認められないことから，タスクの難易度の差は無い．さらに，正答率は，提案手法とベースライン手法のユーザインタフェースによって差がないことから，本実験のタスクにおけるワーカの品質と言うことができる．つまり，ワーカの品質はユーザイ

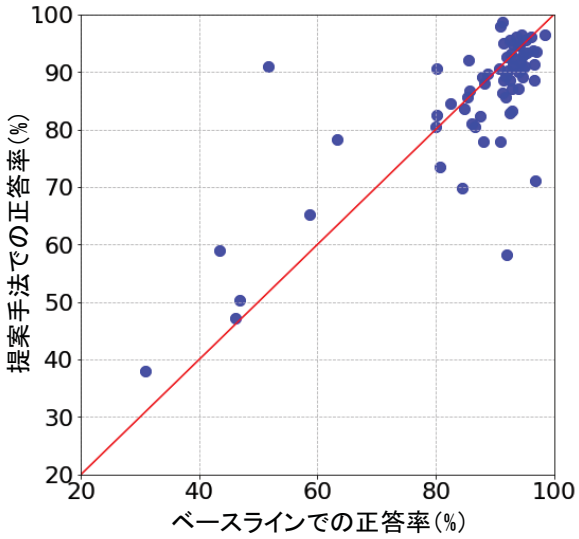


図 **6.4:** 両方のタスクに従事したワーカの正答率

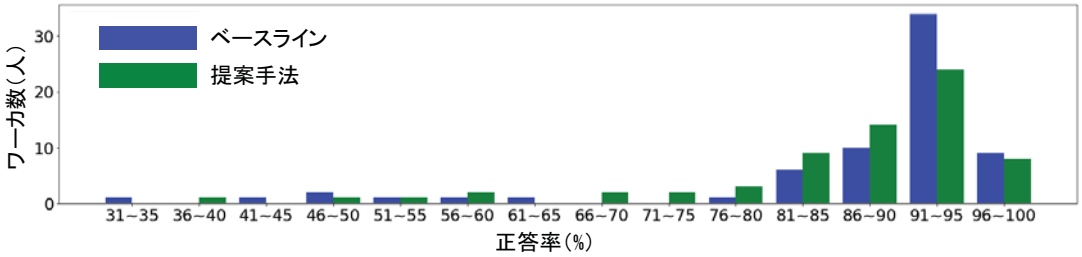


図 **6.5:** 両方のタスクに従事したワーカの正答率の分布

ンタフェースに依存するものではない．

6.3.2 処理時間

提案手法では，ツイートを読むためにはツイート表示ボタンを押し続けなければならないため，ツイートを読む度にマウスカーソルをツイート表示ボタンへ移動させなければならない．その結果，ベースライン手法と比べ多くの処理時間を要することが予想される．

それぞれの手法の処理時間の分布の箱ひげ図を図 6.6 に示す．グラフの縦軸は処理時間であるが，対数目盛となっている．ベースライン手法では処理時間の中央値が 6 秒であった．対して，提案手法では 9 秒となった．我々の予想通り，提案手法では処理時間が 3 秒増加し，ワーカへ時間的な負担を与えることが分かった．さらに，タスク依頼者にとっても依頼する全てのタスクが処理されるまでにより多くの時間が必要となる．

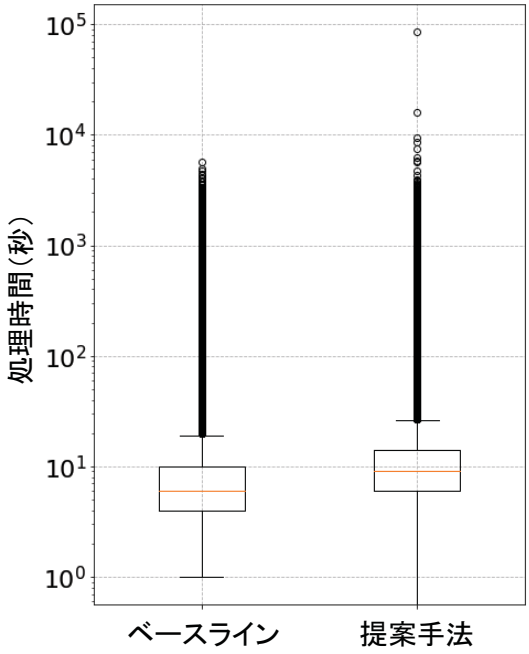


図 6.6: タスク処理時間の分布

また，ベースライン手法と提案手法の双方で，処理時間が 1,000 秒を超えるようなケースも存在する．Twitter の文字数は最大 140 字であり，これほど多くの時間が必要であるとは考えにくい．これは 5 章で述べたように，ワーカは常にタスクに集中している状況ではなく，他のことを行いながら作業を行っていたり，休憩を挟んでタスクに取り組んでいたりすることが表れている．

6.3.3 ワーカのモチベーション

ワーカのモチベーションを測定する指標として，作業をどれだけ続けたかの処理回数と作業単位である 100 ツイートに満たず作業から離脱した割合を用いる．

まず最初に，図 6.7 に対象ワーカのタスク処理回数の分布を示す．提案手法とベースライン手法ではタスクの処理回数の分布には大差がない．

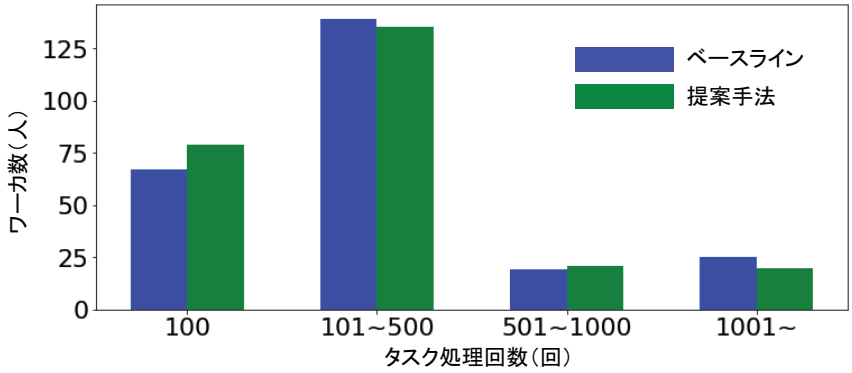


図 6.7: 対象ワーカのタスク処理回数の分布

次に，作業からの離脱について議論する．表 6.1 の総ワーカ数から対象ワーカ数を引いた人数が離脱したワーカの人数となる．つまり，ベースライン手法では 43.1% の 189 ワーカが途中で作業から離脱した．一方で，提案手法では，47.5% の 231 ワーカが途中で作業から離脱した．提案手法では 4.4% 離脱率が増加した．さらに，離脱したタイミングを図 6.8 に示す．横軸がタスク処理回数，縦軸がそのタスク処理回数で離脱したワーカ数である．作業報酬額は同じであったが，提案手法では 20 回以内に離脱するワーカがベースライン手法よりも明らかに多いことが分かる．

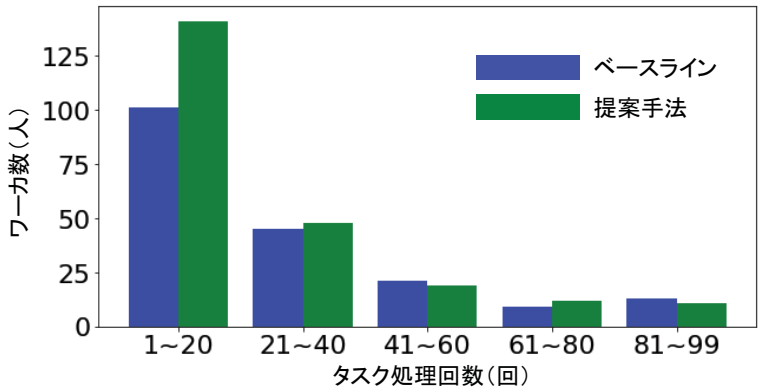


図 **6.8:** ワーカの離脱のタイミング

第 7 章　おわりに

7.1　まとめ

本論文では，マイクロタスク型のクラウドソーシングにおいて，ワーカの品質をワーカ振る舞いから推定する手法を提案した．既存の研究では振る舞いの分析に焦点が当てられてきたが，我々は振る舞いの取得に焦点を当てた．提案手法では閲覧時間や閲覧回数などのより正確な振る舞いを取得するために，コンテンツを表示するためのボタンを設置した．提案手法では，適合率を下げることなく再現率を向上させ，低品質ワーカをより高精度に検出することができた．特に，低品質ワーカの定義を正答率が下位 10% のワーカとした時，再現率は最も高い 0.84 となった．また，1 回あたりの閲覧時間や閲覧回数など，ボタンの設置によって取得可能となった振る舞いがワーカの分類に重要な役割を果たしていることが分かった．処理時間が短いワーカには低品質なワーカも存在すれば高品質なワーカも存在するが，閲覧時間や回数などの処理時間の内訳を見ることで，これらのワーカを正しく分類することが可能となった．しかし，タスク介入によるデメリットとしては，処理時間が増えることや，ワーカの離脱率の増加が挙げられる．

7.2　今後の課題

今後の課題としては，以下のようなことが挙げられる．一つ目の課題は，高品質なワーカを低品質と分類してしまう例が多いことや十分まじめに作業しているが正答率が悪いワーカを高品質と分類してしまうことである．つまり，素早く適切な作業を行えるワーカやタスクの内容を正しく理解していないワーカを誤判別しないように，振る舞いだけでなく作業結果と組み合わせてワーカの品質を判断する工夫が必要である．二つ目の課題は，提案手法の汎用性を検証することである．本論文では，一種類のタスクに対して実験を行った．5.1 節でも述べたように，作業段階を 3 段階に分けることができるタスクには効果があると考えている．また，本研究で行ったタスクは多くのワーカの正答率が 90% 程度あり，容易であった．ワーカの能力によって正答率にばらつきが出るタスクや全体的な正答率が低くなるタスクで

も検証するべきである．三つ目の課題は，他のタスク介入方法を検討することである．提案手法ではコンテンツを表示するボタンを設置したが，より正確な振る舞いが取得でき，よりワーカに負担が少ない方法を検討することが必要であると考えている．

これらの課題を解決すると，低品質ではあるが真面目に作業を行うワーカと意図的に不適切な作業を行うワーカを高精度に分類することが可能となる．この 2 種類のワーカに対する排除や指導方法に関する議論は別途必要である．指導を行う場合，人手で行うのではなく自動的に行うことが望ましい．指導内容やタイミングなどを決定する手法の開発が今後の課題となる．

謝辞

本学情報科学研究科の中村哲教授には，主指導教官として数多くの貴重なご助言をいただいたほか，研究活動を手厚くご支援いただきました．心より感謝を申し上げます．

本学情報科学研究科の安本慶一教授には，副指導教官として様々なご助言をいただきました．心より感謝申し上げます．

本学情報科学研究科の鈴木優特任准教授には，副指導教官として，貴重なご助言をいただいたほか，論文執筆など日頃から数多くのご支援をいただきました．また，学会参加時にも大変お世話になりました．心より感謝申し上げます．

本学情報科学研究科の田中宏季助教には，副指導教官として，貴重なご助言をいただきました．心より感謝申し上げます．

本学情報科学研究科の須藤克仁准教授，Sakriani Sakti 特任准教授，吉野幸一郎助教には，研究における貴重なご助言をいただきました．心より感謝申し上げます．

また，林美保様，松田真奈美様には諸手続きの際に大変お世話になったほか，研究室運営の様々な面でご尽力いただきました．心より感謝を申し上げます．

研究室の学生の皆様には，研究に対するご助言をいただいたほか，公私ともにお世話になりました．心より感謝を申し上げます．

最後に，私をここまで育て，本学に進学させてくれた両親に心より感謝を申し上げます．

参考文献

[1] O. Alonso, "Implementing crowdsourcing-based relevance experimentation: an industrial perspective," *Information retrieval*, vol. 16, no. 2, pp. 101–120, 2013.

[2] V. Ambati, S. Vogel, and J. G. Carbonell, "Towards task recommendation in micro-task markets." *Human computation*, vol. 11, p. 11, 2011.

[3] G. E. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM Sigkdd Explorations Newsletter*, vol. 6, no. 1, pp. 20–29, 2004.

[4] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[5] A. Burnap, Y. Ren, P. Y. Papalambros, R. Gonzalez, and R. Gerth, "A simulation based estimation of crowd ability and its influence on crowdsourced evaluation of design concepts," in *ASME 2013 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*. American Society of Mechanical Engineers, 2013, pp. V03BT03A004–V03BT03A004.

[6] H.-Â. Cao, F. Rauchenstein, T. K. Wijaya, K. Aberer, and N. Nunes, "Leveraging user expertise in collaborative systems for annotating energy datasets," in *Big Data (Big Data), 2016 IEEE International Conference on*. IEEE, 2016, pp. 3087–3096.

[7] B. Carpenter, "A hierarchical bayesian model of crowdsourced relevance coding." in *TREC*, 2011.

[8] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[9] A. P. Dawid and A. M. Skene, "Maximum likelihood estimation of observer error-rates using the em algorithm," *Applied statistics*, pp. 20–28, 1979.

[10] M. H. DeGroot and M. J. Schervish, *Probability and statistics*. Pearson

Education, 2012.

[11] T. G. Dietterich, "An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization," *Machine learning*, vol. 40, no. 2, pp. 139–157, 2000.

[12] P. Donmez and J. G. Carbonell, "Proactive learning: cost-sensitive active learning with multiple imperfect oracles," in *Proceedings of the 17th ACM conference on Information and knowledge management*. ACM, 2008, pp. 619–628.

[13] K. Evanini, D. Higgins, and K. Zechner, "Using amazon mechanical turk for transcription of non-native speech," in *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. Association for Computational Linguistics, 2010, pp. 53–56.

[14] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*. Springer series in statistics New York, 2001, vol. 1.

[15] M. Goto and J. Ogata, "Podcastle: Recent advances of a spoken document retrieval service improved by anonymous user contributions," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[16] H. Halpin and R. Blanco, "Machine-learning for spammer detection in crowd-sourcing," in *Workshop on Human Computation at AAAI, Technical Report WS-12-08*, 2012, pp. 85–86.

[17] M. Hirth, S. Scheuring, T. Hoßfeld, C. Schwartz, and P. Tran-Gia, "Predicting result quality in crowdsourcing using application layer monitoring," in *Communications and Electronics (ICCE), 2014 IEEE Fifth International Conference on*. IEEE, 2014, pp. 510–515.

[18] J. J. Horton and L. B. Chilton, "The labor economics of paid crowdsourcing," in *Proceedings of the 11th ACM conference on Electronic commerce*. ACM, 2010, pp. 209–218.

[19] J. Howe, "The rise of crowdsourcing," *Wired magazine*, vol. 14, no. 6, pp. 1–4, 2006.

[20] N. Japkowicz *et al.*, "Learning from imbalanced data sets: a comparison of various strategies," in *AAAI workshop on learning from imbalanced data sets*, vol. 68. Menlo Park, CA, 2000, pp. 10–15.

[21] E. Kamar, S. Hacker, and E. Horvitz, "Combining human and machine intelligence in large-scale crowdsourcing," in *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*. International Foundation for Autonomous Agents and Multiagent Systems, 2012, pp. 467–474.

[22] G. Kazai, J. Kamps, M. Koolen, and N. Milic-Frayling, "Crowdsourcing for book search evaluation: impact of hit design on comparative system ranking," in *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. ACM, 2011, pp. 205–214.

[23] G. Kazai and I. Zitouni, "Quality management in crowdsourcing using gold judges behavior," in *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*. ACM, 2016, pp. 267–276.

[24] A. Kittur, E. H. Chi, and B. Suh, "Crowdsourcing user studies with mechanical turk," in *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 2008, pp. 453–456.

[25] E. Law and L. v. Ahn, "Human computation," *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 5, no. 3, pp. 1–121, 2011.

[26] W. Ling123, L. Marujo123, C. Dyer, A. Black, and I. Trancoso, "Crowdsourcing high-quality parallel data extraction from twitter," *ACL 2014*, p. 426, 2014.

[27] M. Marge, S. Banerjee, and A. I. Rudnicky, "Using the amazon mechanical turk to transcribe and annotate meeting speech for extractive summarization," in *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. Association for Computational Linguistics, 2010, pp. 99–107.

[28] R. K. Mok, R. K. Chang, and W. Li, "Detecting low-quality workers in QoE

crowdtesting: A worker behavior-based approach," *IEEE Transactions on Multimedia*, vol. 19, no. 3, pp. 530–543, 2017.

[29] R. Munro, "Crowdsourced translation for emergency response in haiti: the global collaboration of local knowledge," in *AMTA Workshop on Collaborative Crowdsourcing for Translation*, 2010, pp. 1–4.

[30] S. Oyama, Y. Baba, Y. Sakurai, and H. Kashima, "Accurate integration of crowdsourced labels using workers' self-reported confidence scores." in *IJCAI*, 2013, pp. 2554–2560.

[31] A. J. Quinn and B. B. Bederson, "Human computation: a survey and taxonomy of a growing field," in *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 2011, pp. 1403–1412.

[32] B. Rahmanian and J. G. Davis, "User interface design for crowdsourcing systems," in *Proceedings of the 2014 International Working Conference on Advanced Visual Interfaces*. ACM, 2014, pp. 405–408.

[33] V. Raykar and P. Agrawal, "Sequential crowdsourced labeling as an epsilon-greedy exploration in a markov decision process," in *Artificial intelligence and statistics*, 2014, pp. 832–840.

[34] V. C. Raykar, S. Yu, L. H. Zhao, A. Jerebko, C. Florin, G. H. Valadez, L. Bogoni, and L. Moy, "Supervised learning from multiple experts: whom to trust when everyone lies a bit," in *Proceedings of the 26th Annual international conference on machine learning*. ACM, 2009, pp. 889–896.

[35] J. M. Rzeszotarski and A. Kittur, "Instrumenting the crowd: using implicit behavioral measures to predict task performance," in *Proceedings of the 24th annual ACM symposium on User interface software and technology*. ACM, 2011, pp. 13–22.

[36] S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," *IEEE transactions on systems, man, and cybernetics*, vol. 21, no. 3, pp. 660–674, 1991.

[37] A. D. Shaw, J. J. Horton, and D. L. Chen, "Designing incentives for inexpert human raters," in *Proceedings of the ACM 2011 conference on Com-*

puter supported cooperative work. ACM, 2011, pp. 275–284.

[38] V. S. Sheng, F. Provost, and P. 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*. ACM, 2008, pp. 614–622.

[39] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore, "Real-time human pose recognition in parts from single depth images," *Communications of the ACM*, vol. 56, no. 1, pp. 116–124, 2013.

[40] E. Simpson and S. Roberts, "Bayesian methods for intelligent task assignment in crowdsourcing systems," in *Decision Making: Uncertainty, Imperfection, Deliberation and Scalability*. Springer, 2015, pp. 1–32.

[41] P. Smyth, U. M. Fayyad, M. C. Burl, P. Perona, and P. Baldi, "Inferring ground truth from subjective labelling of venus images," in *Advances in neural information processing systems*, 1995, pp. 1085–1092.

[42] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng, "Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks," in *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, 2008, pp. 254–263.

[43] W. Tang and M. Lease, "Semi-supervised consensus labeling for crowdsourcing," in *SIGIR 2011 workshop on crowdsourcing for information retrieval (CIR)*, 2011, pp. 1–6.

[44] I. V. Tetko, D. J. Livingstone, and A. I. Luik, "Neural network studies. 1. comparison of overfitting and overtraining," *Journal of chemical information and computer sciences*, vol. 35, no. 5, pp. 826–833, 1995.

[45] P. Wais, S. Lingamneni, D. Cook, J. Fennell, B. Goldenberg, D. Lubarov, D. Marin, and H. Simons, "Towards large-scale processing of simple tasks with mechanical turk," in *Third AAAI Conference on Human Computation and Crowdsourcing*, 2011.

[46] M.-C. Yuen, I. King, and K.-S. Leung, "Taskrec: probabilistic matrix fac-

torization in task recommendation in crowdsourcing systems," in *Neural Information Processing*. Springer, 2012, pp. 516–525.

[47] Y. Zheng, S. Scott, and K. Deng, "Active learning from multiple noisy labelers with varied costs," in *Data Mining (ICDM), 2010 IEEE 10th International Conference on*. IEEE, 2010, pp. 639–648.

[48] D. Zhu and B. Carterette, "An analysis of assessor behavior in crowdsourced preference judgments," in *SIGIR 2010 workshop on crowdsourcing for search evaluation*, 2010, pp. 17–20.

[49] 芥子育雄, 鈴木優, 吉野幸一郎, 大原一人, 向井理朗, and 中村哲, "単語意味ベクトル辞書を用いた twitter からの評判情報抽出," **電子情報通信学会論文誌** *D*, vol. 100, no. 4, pp. 530–543, 2017.

[50] 鹿島久嗣, 梶野洸 *et al.*, "クラウドソーシングと機械学習 (< 特集> 知識の転移)," **人工知能学会誌**, vol. 27, no. 4, pp. 381–388, 2012.

発表リスト

国際会議

[1] Yoshitaka Matsuda, Yu Suzuki, Satoshi Nakamura. A Trade-off between Estimation Accuracy of Worker Quality and Task Complexity. The First IEEE Workshop on Human-Machine Collaboration in BigData (HMData2017). Boston. December 2017.

研究会

[1] 松田 義貴，鈴木 優，中村 哲．クラウドワーカの振る舞いを用いた能力推定とタスク複雑度のトレードオフ．関西データベースワークショップ 2017．兵庫．2017 年 9 月．[学生奨励賞]

[2] 松田 義貴，鈴木 優，中村 哲．タスク介入によるクラウドワーカの品質推定精度の改善．第 10 回データ工学と情報マネジメントに関するフォーラム（DEIM2018）．福井．2018 年 3 月．（発表予定）