クラウドソーシングにおける対話型インタフェースを用いたアノテーションタスク遂行時の回答品質低下の検知手法

平良 繁幸[1]　松田 裕貴[2,3,a)]　福光 嘉伸[1]　諏訪 博彦[1,3]　安本 慶一[1,3]

概要：クラウドソーシングベースのアノテーションタスクは，低コストかつ広範囲で実施可能であることから，機械学習における学習データの収集手法として広く活用されている．一方で，クラウドソーシングによって得られる回答の品質にはばらつきがあり，品質管理の難しさが課題となっている．そこで本研究では，回答者の回答品質の低下傾向をリアルタイムに検出することで，品質低下を防止する手法の実現を目指している．本稿では，アノテーションタスクの実施中に取得される端末の姿勢角や画面操作といった端末操作情報から特徴量を抽出し，教師あり機械学習により回答品質の低下を推定する二値分類モデルを構築する手法を提案する．本手法の有効性を検証するために，アノテーションタスクの実施と端末操作情報の取得機能を備えたアプリケーションを開発し，被験者実験を行った．実験では，画像に対するキャプションの正誤評価をアノテーションタスクとして依頼し，被験者の作業中における端末操作情報をバックグラウンドで定常的に収集した．収集した端末操作情報を用いて回答品質を推定する二値分類モデルを構築・評価した結果，個人内データを対象とした時系列バリデーションによるデータセット分割を用いたモデルにおいて，Precision が 0.722，Recall が 0.741，F1-score が 0.731 という結果が得られた．これにより，アノテーションタスクにおける回答者の品質低下の推定において，本手法の有効性が示唆された．

キーワード：クラウドソーシング，アノテーション，Satisficing，不良回答検出，機械学習

1. はじめに

クラウドソーシングは，インターネットを介して不特定多数の群衆に業務を委託するビジネス形態であり，低コストで大規模なデータを取得できるという利点から，さまざまな用途で活用が進んでいる．特に機械学習の分野では，モデル構築および精度向上のために大量の学習データが必要とされるため，アノテーション作業をクラウドソーシングによって外部委託することで，低コストかつ効率的に大規模なデータを収集する手法として用いられている．しかしながら，クラウドソーシングによって得られるデータには品質のばらつきが大きく，品質管理が困難であることが重要な課題として挙げられる [1]．これは，クラウドソーシングに参加するアノテータが常に正確に回答するとは限らないためである．特に，タスクへの報酬が対価として提示されている場合，アノテータは回答の正確性よりも回答時間の短縮を優先し，できるだけ短時間で報酬を得ようとする傾向が見られる．また，タスクの設問内容が十分に理解されていない場合や，作業中の疲労などによりアノテータの集中力が低下している場合には，アノテータに悪意がなくとも不正確な回答（不良回答）が生じる可能性がある．このような傾向について，Simon らは人間の認知資源の限界に着目し，回答者が与えられた要求に対して支払う努力を最小化しようとする傾向を *Satisficing*（**努力の最小化**）と定義している [2]．このような不良回答が多く含まれるデータを機械学習モデルの学習に用いた場合，モデルの精度が著しく低下するおそれがある．そのため，不良回答の発生をリアルタイムで検出・防止することは，モデル品質の確保および実用的な応用において非常に重要である．

社会心理学のように紙によるアンケート調査を多く取り扱ってきた分野では，より正確な回答を得ることを目的として，不良回答の検出手法がさまざまに考案されてきた．代表的な手法としては，IMC（Instructional Manipulation Check）[3]，ARS（Attentive Responding Scale）や DQS（Directed Question Scale）[4] などが挙げられる．これらは，被験者の矛盾や不注意を可視化する質問を設問紙に組み込むことで，Satisficing を検出するものである．しかしながら，このように被験者の注意を試す目的の設問を明示

[1] 奈良先端科学技術大学院大学，
Nara Institute of Science and Technology
[2] 岡山大学，Okayama University
[3] 理化学研究所，RIKEN AIP
[a)] yukimat@okayama-u.ac.jp

的に提示する方式では，被験者に心理的な負担を与える可能性がある．また，回答者の内発的動機を損なうことによって，かえって不良回答を誘発するリスクも指摘されている．加えて，検出用の質問を追加することで全体の設問数が増加し，調査の負担が増大する点も課題である．このような背景を受けて，近年では質問項目を追加することなく不良回答を検出する手法の研究が進められている．後上ら [5,6] は，スマートフォンを用いたアンケート調査において，アンケート完了後に取得した画面操作ログから特徴量を抽出し，機械学習により Satisficing の有無を分類する手法を提案している．このアプローチでは，アンケート完了後に不良回答者を特定し，その回答をデータセットから除外することで全体のデータ品質を向上させている．しかし，この手法ではデータ収集の完了後に不良回答を検出するため，回答のやり直しなどのリアルタイムな対処が困難である．また，収集対象に制限がある場面では，不良回答を除外することにより最終的に得られるデータ量が不足する可能性がある．さらに，不良回答者の削除によって，得られたデータセットが母集団全体の性質を適切に反映しなくなる懸念もある．福光らは，クラウドソーシングを活用したアノテーションタスクにおいて，作業中のクリックやマウスカーソルの移動など，PC 端末上の画面操作ログから特徴量を抽出し，機械学習によって不良回答をリアルタイムで検出する手法を提案している [7]．しかし，この手法もまた，各タスクの回答後に品質判定を行う方式であり，回答品質の事前推定や早期介入には対応していない点が課題として残されている．

本研究では，スマートフォン上で実施されるアノテーションタスク，特に画像に対するキャプションの正誤評価を対象として，回答品質の低下をリアルタイムに推定する手法を提案する．提案手法では，アノテーションタスク実施時におけるアノテータのタップやスクロールなどの画面操作情報に加え，端末の姿勢角や加速度といったセンサ情報を含む端末操作情報をリアルタイムで取得し，それらを特徴量として用いることで，機械学習による回答品質低下の推定を行う．本手法の実現には，スマートフォン上でアノテーションタスクを提示する機能に加え，タスク実施中に発生するアノテータの端末操作情報をリアルタイムで取得するシステムの構築が必要となる．本研究では，これらの要件を満たすスマートフォンアプリケーションを開発し，学内学生を対象に学習データの収集実験を実施した．実験により得られた端末操作情報から特徴量を抽出し，回答品質の低下傾向を推定するための二値分類モデルを構築することで，回答品質の推定性能を評価した．評価においては，個人内での推定精度と標本全体に対する汎化性能の 2 つの観点から，Precision，Recall，F1-score を指標としてモデル性能を検証した．個人内の推定では，被験者が回答した任意のタスクをテストデータとし，それ以前に収集されたすべての時系列データを訓練データとすることで，時系列順に訓練・評価を繰り返すエクスパディング型時系列バリデーションによりモデルを構築し，各被験者ごとに評価を行った．汎化性能の評価では，ある被験者のデータをテストデータとして除外し，その他すべての被験者のデータから構築された分類モデルを用いてアンサンブル学習を行い，その予測結果に基づいて評価を行った．学内学生を対象とした学習データの取得実験では，被験者 30 名から端末操作情報および回答内容を含むデータが収集された．個人内での推定精度評価においては，各被験者のデータに対して時系列バリデーションを適用し，回答品質低下の推定モデルを構築した．その結果，Precision は 0.722，Recall は 0.741，F1-score の平均は 0.731 という性能が得られた．これにより，スマートフォン端末操作情報を用いた回答品質の低下推定に関する本手法が，実運用において有効に機能する可能性が示唆された．

2. 関連研究

2.1 不良回答の検出に関する研究

不良回答の検出を目的とした先行研究としては，Oppenheimer らによる IMC（Instructional Manipulation Check）[3]，および Maniaci らによる ARS（Attentive Responding Scale）や DQS（Directed Question Scale）[4] が広く知られている．これらの手法は，回答者の不注意や矛盾を可視化するための設問を質問紙内に挿入することで，Satisficing の発生を検出することを目的としている．一方で，後上ら [5,6] は，検出用の質問を用いずに不良回答の検出を実現する手法を提案している．彼らは，タスク完了後に取得したスマートフォン上のタップやスクロールといった画面操作ログをもとに特徴量を抽出し，それらを用いた機械学習モデルによって Satisficing の有無を分類するアプローチを採用している．さらに，福光ら [7] は，機械学習を用いてリアルタイムに不良回答を検出する手法を提案している．具体的には，固有表現アノテーションタスクを対象に，ラベルの付与状況に加えてクリックやマウスカーソルの移動などの PC 端末における画面操作情報をバックグラウンドで取得し，それを特徴量として用いることで，タスク実施中に不良回答の兆候を検出する枠組みを構築している．

IMC や ARS に代表される検出用の設問を用いる手法では，回答者の注意力や誠実性を確認することが可能である一方で，回答者を疑うような意図を含む設問を明示的に提示することから，回答者に心理的な負担を与える可能性がある．また，これらの設問を追加することにより，質問数が増加し，回答全体にかかる負担が増す点も問題となる．後上らの手法では，アノテーションタスクの完了後に取得された画面操作情報に基づいて不良回答の検出を行っており，検出された不良回答のデータを除外するといった事後的な対応に限定される．そのため，タスク実施中や実施前

にリアルタイムで介入することは困難である．福光らの手法は，リアルタイムで不良回答の検出を行う点において先進的であるが，不良回答が実際に発生したタイミングでの検出を前提としている．そのため，回答品質が低下する兆候を事前に捉えて介入するような，予測的な制御には対応しづらいと考えられる．たとえば，回答者に対してタスク実施前に休憩を促すなどの措置を講じることは難しい．これらの課題を踏まえると，回答品質の低下傾向をあらかじめ推定し，品質の低下が顕在化する前にその兆候を検出する手法には大きな意義があると考えられる．

2.2 回答品質の改善に関する研究

回答品質の改善に関する研究としては，ユーザのエンゲージメントを向上させることや，行動変容を促すことを目的とした介入手法が，HCI（Human-Computer Interaction）の分野を中心に多数提案されている．

Sihang ら [8] は，クラウドソーシングにおけるマイクロタスクに対して，従来の Web インターフェースではなく会話型インターフェースを用いることで，ユーザのエンゲージメントを向上させる手法を提案している．また，Zhang ら [9] は，歩数の増加を促す行動変容を対象とし，情報提示の対話スタイル（情報の粒度や婉曲表現の程度）が，行動変容の効果に影響を与えることを明らかにしている．これらの研究は，ユーザとの自然なインタラクションが動機づけや集中力維持に寄与することを示唆しており，本研究においても介入方法の柔軟性を考慮し，会話型インターフェースを採用している．

ユーザの作業継続性や集中力の維持に着目した介入としては，Jeffrey ら [10] が，長時間作業中に適切な休憩を挿入することで，ユーザのエンゲージメントを大幅に改善できることを示している．さらに，Peng ら [11] は，クラウドソーシングにおけるマイクロタスクの合間に娯楽要素を取り入れることで，回答品質を維持しつつエンゲージメントを向上させる効果を報告している．これらの手法は，疲労や注意力低下への対処に有効である．

一方で，作業開始前や作業中の態度形成に働きかける介入も提案されている．大山ら [12] は，参加型センシングを対象とし，タスク開始時にボタンタップや端末を振る動作によって貢献の意思表示を行わせることで，不誠実な回答（回答速度の優先）を抑制する手法を提案している．また，中川ら [13] は，回答中の迷いを可視化するため，スライドバーや拡大鏡を活用した UI を用いて，タッチ操作ログからユーザの反応を詳細に取得する手法を提案している．さらに，Mara ら [1] は，職場環境におけるストレスの検出を目的として，マウス動作，キーボード操作，心拍変動などから特徴量を抽出することで，ストレス状態を識別する手法を提示している．

これらの研究は，アノテーションタスク実施プラットフォームの設計段階で回答品質の改善を図るものであり，回答品質の低下に対して応答的な介入（リアクティブな対応）を行うものではない．したがって，リアルタイムで回答品質の低下傾向を検出することにより，適切なタイミングで応答的かつ個別化された介入を行う仕組みの構築は，行動変容を通じた品質改善の実現において重要といえる．

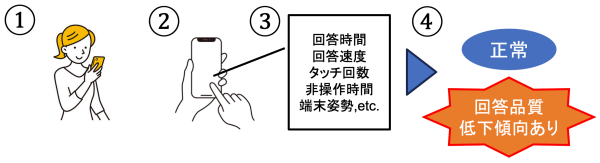


図 1: 提案手法の概要

3. スマートフォンの端末操作情報を用いた回答品質低下推定手法

本研究では，アノテーションタスクとして画像に対するキャプションの正誤評価を取り上げ，タスク中における回答品質の低下傾向をリアルタイムに推定することを目的としている．本章では，タップやスクロールなどの画面操作に加え，端末の姿勢角や加速度といったセンサ情報を含むスマートフォン端末操作情報の取得方法，およびそれらに基づく回答品質低下推定モデルの構築手法を提案する．

3.1 想定シナリオと提案手法の概要

本研究におけるアノテーションタスク実施の想定シナリオと，提案手法の全体的な流れを**図 1** に示し，以下に各ステップの詳細について説明する．

(1) アノテーションタスクの依頼者は，クラウドソーシングサービス等を通じてアノテータを募集する．本研究では，アノテータは専門家ではなく，アノテーション対象言語を母語とする一般市民を想定している．
(2) アノテータは，依頼されたアノテーションタスクを遂行する．各タスクは，アノテータが保有するスマートフォンにインストールされた専用アプリケーション上で実施されることを想定する．
(3) タスクの実施中，アプリケーションに搭載された本研究による開発システムが，アノテータのタップ・スクロール・端末の姿勢角・加速度などを含む端末操作情報をバックグラウンドで定常的に収集する．
(4) 収集されたログデータから特徴量を抽出し，これをもとに構築した機械学習モデルによって回答品質の低下傾向をリアルタイムで推定する．

3.2 想定するアノテーションタスクの概要

本研究では，アノテーションタスクとして「画像に対するキャプションの正誤評価」を対象としている．アノテーション作業は，本研究において開発したスマートフォンア

図 2: アノテーション実施画面例

プリケーション上で実施されることを想定している．

タスクでは，「画像」と「その画像に対するキャプション」のセットが提示され，アノテータはキャプションの内容が画像の内容を正しく記述していると判断した場合は「Yes」，誤っていると判断した場合は「No」と回答することが要求される．具体的なアノテーション作業の手順はいかに示すとおりである．

(1) ユーザは，自身の ID およびパスワードを用いてアプリケーションにログインする．
(2) タスク選択画面から対象タスクを選択すると，アノテーション実施画面に遷移する．
(3) アノテーション実施画面では，会話型インタフェースを介してエージェントからタスクが提示される．
(4) アノテータは，提示された画像とその下部に表示される説明文（キャプション）を確認し，内容が画像を正しく説明していると判断した場合は「Yes」，誤っていると判断した場合は「No」を選択する．
(5) すべてのタスクを完了すると，エージェントから完了メッセージが送信され，アノテーションタスクは終了する．

3.3 端末操作情報取得システムと抽出する特徴量

本節では，アノテーション作業中に発生する画面操作情報を記録するための開発システムと，そこから抽出される特徴量について述べる．

本研究では，スマートフォンにおけるアノテーションタスク中の端末操作情報を取得するために，iOS のネイティブ開発言語である SwiftUI を用いてアノテーション実施用アプリケーションを開発した．本アプリケーションは，アノテーションタスクの提示機能に加え，ユーザによるタップやスクロールなどの画面操作，さらに端末の姿勢角や加速度といったセンサ情報をバックグラウンドで定常的に取得し，データベースへ記録する機能を有している．

次に，取得された端末操作情報に基づいて抽出される特徴量について述べる．使用する特徴量とその単位は**表 1** に示す．本研究では，後上らの先行研究 [6] で用いられた特徴量に加えて，本研究で対象とする画像キャプションの正誤評価タスクの操作特性に適した特徴量を新たに導入している．特に，端末の姿勢角などのセンサ情報や，リアルタイム処理に適した時系列的な特徴量も考慮し，より柔軟かつ精度の高い推定を可能にすることを目指した．

表 1: 抽出する特徴量

特徴量	単位
データ送信時の時間	s
タスク番号	
タスク識別番号	
回答時間	s
非操作時間	s
画面位置	
タップ回数	回
タップ間隔	s
タップ位置（x 軸，y 軸）	
スクロール回数	回
スクロール長さ	
スクロール時間	s
スクロール速度	
姿勢角（x 軸，y 軸，z 軸）	
ジャイロ加速度（x 軸，y 軸，z 軸）	
加速度（x 軸，y 軸，z 軸）	
重力加速度（x 軸，y 軸，z 軸）	
回答内容（yes，no）	
回答の正誤	
前 10 問での正答率	%

本研究で用いる基本的な特徴量の一覧を表 1 に示す．これらの特徴量は，後述の 4 章で実施する学内学生を対象とした実験においても用いる．本研究で対象とする「画像に対するキャプションの正誤評価」は，画像の視覚的内容を確認し，続けて提示されるキャプションが画像内容と整合しているかを論理的に判断するという認知負荷の高いタスクである．したがって，アノテータには視覚的処理に加え，文理解と照合のための認知処理が求められる．回答品質が低下している状態では，これらの処理を十分に行わないまま，極端に短い時間で回答を完了させたり，判断の正確性が著しく低下したりすることが想定される．このような傾向に基づき，回答品質の低下状態では，平常時と比較して回答時間や正答率に明確な変化が生じる可能性がある．したがって，本研究ではこれらの要素を反映する特徴量の抽出が，品質低下の推定に有効であると考えられる．

3.4 回答品質低下推定モデルの構築

本節では，前述の端末操作情報収集アプリケーションによって取得されたデータから導出された特徴量を入力とし，回答品質の低下傾向を推定するための機械学習モデル

の構築手法について述べる．

本研究では，分類アルゴリズムとして LightGBM（Light Gradient Boosting Machine）[14] を採用する．LightGBM は，後上らの先行研究 [6] において最も高い分類性能を示しており，本研究の目的であるリアルタイム推定においても有効であると判断した．

LightGBM は，勾配ブースティング決定木（GBDT）に基づくアルゴリズムであり，学習および推論の実行時間が短く，他のアルゴリズムと比較して少量のデータでも高い精度が期待できる．また，決定木ベースのアルゴリズムであるため，不要な特徴量や外れ値に対して精度が低下しにくいという性質がある．また，特徴量の中には，画面位置情報やタップ回数など，サンプルごとに値のスケールが大きく異なるものも含まれるため，これらに対しては平均 0・分散 1 での標準化（Z スコア標準化）を適用し，スケールの統一を図る．さらに，ハイパーパラメータの最適化によりモデルの汎化性能向上および過学習防止を行う．本研究では，その最適化プロセスに自動化フレームワークである Optuna を用いる．Optuna は，ベイズ最適化を用いた効率的な探索アルゴリズムを有しており，探索空間が広い場合でも効率よく最適解に収束できる点で優れている．

4. 学内実験

4.1 データ収集実験

3 章で述べた端末操作情報収集アプリケーションを用いて実施した，学習データの収集実験および収集されたデータセットについて説明する．

本実験は，奈良先端科学技術大学院大学の学生を対象として参加者を募集し，最終的に 38 名が参加した．実験では，被験者に対してアノテーションタスクの実施を依頼し，タスク実施中に発生する端末操作情報をバックグラウンドで定常的に収集することで，機械学習モデルの構築に用いる学習用データを取得した．なお，実験への参加が確認された被験者に対しては，謝礼として 1000 円相当のギフトカードを支給した．

本実験においては，被験者が端末操作情報の取得を意識することによって生じるバイアスを低減するため，実験開始前にはタスクの実施方法のみを説明し，端末操作情報の収集に関する説明は行わなかった．実験終了後に，端末操作情報を収集していた旨を説明し，収集したデータの利用可否について改めて同意を得た．なお，本研究は奈良先端科学技術大学院大学人を対象とする研究に関する倫理審査委員会の承認を受けて実施した（承認番号：2020-I-2）．

以降では，実験で実施したアノテーションタスクおよび得られたデータセットについて述べる．

4.1.1 アノテーションタスクの概要

本実験において採用したアノテーションタスクについて述べる．本実験では，アノテーション作業として GQA（Generative Question Answering）データセット [15] に含まれる画像に対するキャプションの正誤評価を対象とした．

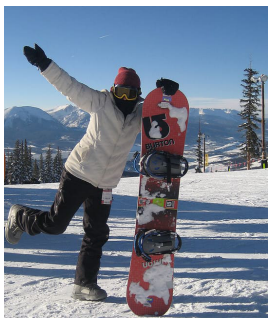


図 3: GQA データセットを用いた画像に対するキャプションの正誤評価の出題例

被験者は，GQA データセット内に含まれる画像とキャプションのペアに対して，キャプションが画像の内容を正しく記述しているかどうかを判断し，正しい場合は「Yes」，誤っている場合は「No」と回答する形式でアノテーションを行った．アノテーション作業の手順は，3.2 節の内容に準拠している．タスクの具体的な出題例を**図 3** に示す．

本実験では，GQA データセットから 150 組の画像とキャプションを選定し，被験者ごとに提示順をランダムにシャッフルして出題した．タスクの終了条件は，すべてのアノテーションタスクを完了した場合，あるいは作業開始から 30 分が経過しその時点で実施中のタスクへのアノテーションを終了した場合，のいずれかを満たした場合とした．

4.1.2 データセット

本節では，実験により得られたデータセットについて述べる．作成したタスクデータセットに対して，画像とキャプションのペアごとに正解ラベルを付与し，被験者が実際に回答した結果と照合することで，各回答の正誤を評価した．本研究で対象とするアノテーションタスクは，2 択（「Yes」または「No」）による選択形式であることから，任意の 10 タスクごとの正答率が 40%未満であった場合を，回答品質が低下している状態と定義した．このように定義することで，一時的な誤答や偶然の正解を排除しつつ，明確な傾向としての品質低下を検出可能にしている．

被験者 38 名のうち，期日までにすべてのタスクを完了し，かつ収集データの研究利用に同意が得られた 30 名分のデータを分析対象とした．作成したデータセットに含まれる適切回答および不良回答の件数の概要を**表 2** に示す．

得られたデータセットから，各個人の端末操作情報から類似した端末操作を行ったユーザー群が確認できるかを DBSCAN クラスタリングによって検証した．**図 4** に eps = 5.0, min_samples = 5 で実行したクラスタリング結

表 2: データセットに含まれるデータ数（学内実験）

	データセット Stu $D_{\pm 0}$
適切回答	2015
不良回答	2159

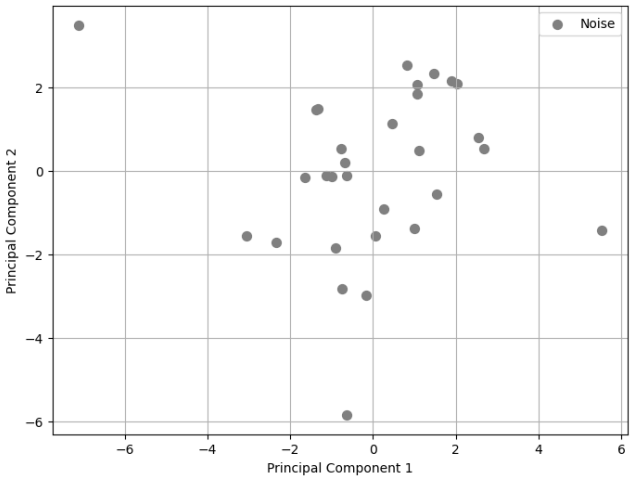


図 4: 学内実験データにおける端末操作情報から抽出した特徴量クラスタ

果を示す．結果より，DBSCAN クラスタリングでは特徴的なクラスタが生成されなかったことがわかる．また，他の被験者とは大きく異なる特徴量を有する被験者のデータが 3 つ含まれていることがわかる．

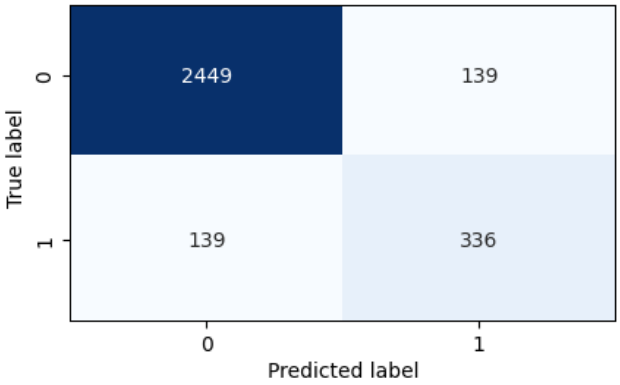
4.2 モデルの評価方法

本研究では，個人内での推定精度と標本内での汎化性能の 2 つの側面から，Precision，Recall，F1-score に基づいて構築モデルの性能を評価する．個人内での推定精度については，被験者が回答した任意のタスクにおけるデータをテストデータとし，それ以前のすべての時系列データを訓練データとした分割を時系列順に繰り返すエクスパディング型時系列バリデーションを適用し，二値分類モデルを構築する．得られたモデルの予測結果に基づいて，推定精度を評価する．標本内での汎化性能については，任意の被験者のデータをテストデータとし，それ以外の被験者のデータを用いて構築したモデルを用いてアンサンブル学習を行い，その予測結果に基づいて評価を行う．これにより，提案手法が他者に対しても有効に機能するかを検証する．

4.3 実験結果・考察

4.3.1 エクスパディング型時系列バリデーションによる個人内モデルの構築

図 5 に，各被験者ごとに端末操作情報から抽出した特徴量を用いてエクスパディング型時系列バリデーションを実施し，構築したモデルの分類結果を示す．表中には，正解データに対して推定された結果をデータ数で示しており，表の下部には分類モデルに対する Precision，Recall，



Precision: 0.722　Recall: 0.741　F1-score: 0.731

図 5: 学内実験データの分類結果

F1-score の評価指標を記載している．提案手法に基づくモデルは，Precision が平均 0.722，Recall が平均 0.741，F1-score が平均 0.731 を示した．この結果から，スマートフォン端末操作情報に基づいてアノテーションタスク中の回答品質低下を推定する本手法は，実運用において有効である可能性が示唆された．一方で，分類に失敗したデータについては，タスクが画像に対するキャプションの正誤を二択で判断する形式であるため，正常に作業を実施していない状態であっても偶然正答する可能性が高いことが影響していると考えられる．このようなデータでは，操作ログと正解ラベルの整合性が低く，学習や推定精度に悪影響を及ぼす可能性がある．

図 6 に，被験者全体において分類精度の向上に寄与した特徴量のうち，重要度が高かった上位 20 個を示す．図の縦軸は各特徴量の名称を，横軸はその重要度を表し，横軸の値が大きいほど分類への寄与が大きいことを示している．図から，表示画面の絶対座標に関する特徴量（`abs(viewPos)`），タスク内容を示す特徴量（`question`），各選択肢の選択回数（`rightCount`，`leftCount`），タスクの進行状況を示す特徴量（`task`），および前問までの正答率（`pre_correct_rate`）が，特に分類精度に寄与していることが分かる．表示位置に関する特徴量やタスクの進行状況が重要であることから，タスクの進行に伴って被験者の疲労が蓄積し，結果として回答品質が低下していく傾向があることが示唆される．また，タスク内容を間接的に表す特徴量が重要であることから，今後はタスクの難易度を定量的に評価・導入することで，分類モデルの精度がさらに向上する可能性がある．加えて，前問までの正答率が高い重要度を示していることから，被験者の回答傾向や信頼度をスコア化し，それをモデルに取り込むことで，品質低下の推定精度が向上すると考えられる．さらに，各選択肢の選択回数が重要であったことは，被験者が回答に迷った際に特定の選択肢を選びやすい傾向が存在すること，あるいは，疲労が蓄積した状態で操作しやすい位置の選択肢（ボタン）を無意識に選択していた可能性があることを示している．

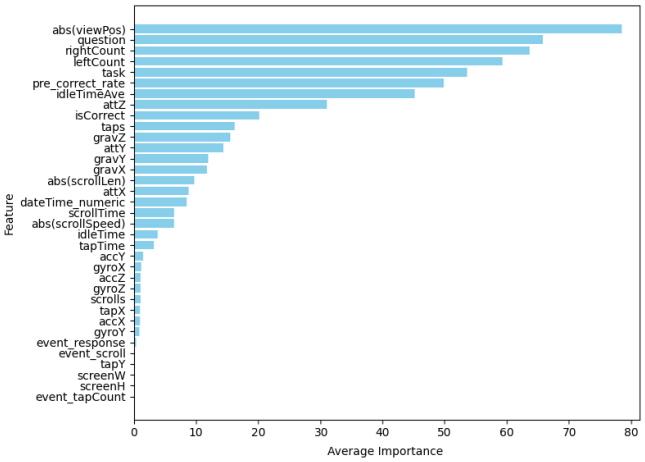
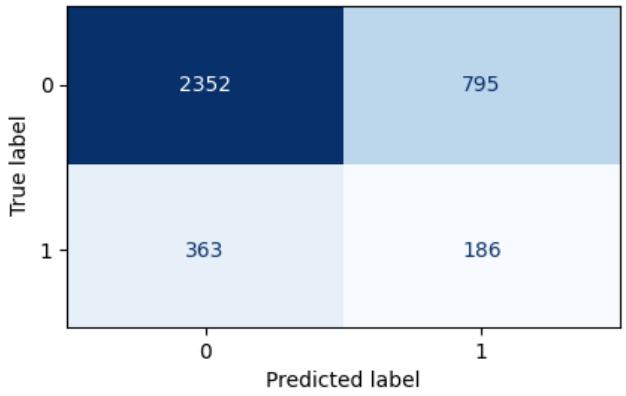


図 6: 学内実験データにおける特徴量の重要度



Precision: 0.190　Recall: 0.339　F1-score: 0.243

図 7: アンサンブル学習による分類結果

4.3.2　アンサンブル学習による汎化性能の評価

4.3.1 項で構築した各被験者ごとのモデルに基づき，アンサンブル学習を実施し，その予測結果から汎化性能の評価を行った．**図 7** に，任意の被験者をテストデータとし，それ以外の被験者のデータを用いて構築されたモデルによるアンサンブル学習を適用した際の予測結果を示す．この評価において，回答品質の低下を推定するモデルは，Precision が平均 0.190，Recall が平均 0.339，F1-score が平均 0.243 を示した．

4.3.1 項で実施したエクスパディング型時系列バリデーションによる個人内モデルと比較すると，これらの評価指標はいずれも大きく低下している．この要因として，個人内データを用いて学習したモデルは，被験者固有の端末操作や行動パターンを強く学習しており，アンサンブル学習では異なる被験者に特有の行動バイアスが混在することで，分類性能が低下した可能性がある．また，後上らの報告 [6] においても指摘されているように，回答品質の推定においては個人間のばらつき（個人間偏差）よりも，個人内における変動（個人内偏差）の把握がより重要であることが示されており，本研究の結果も同様であった．

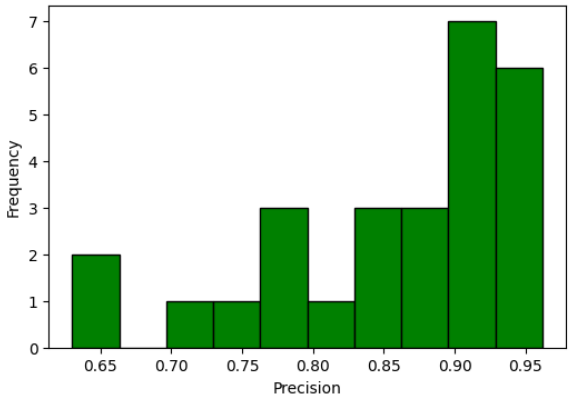


図 8: Precision 分布

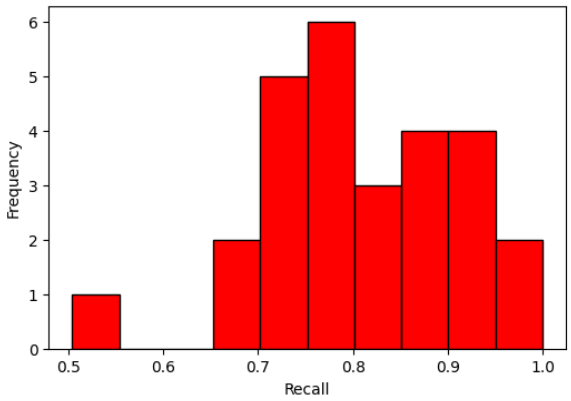


図 9: Recall 分布

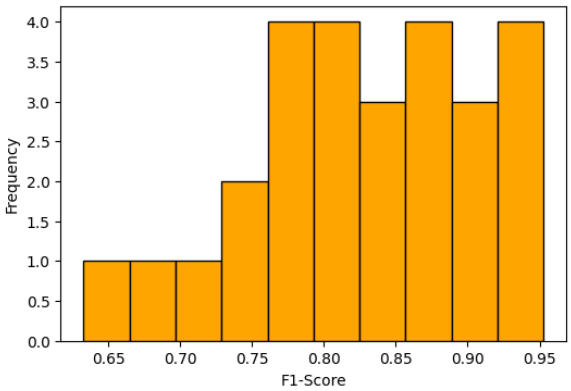


図 10: F1-score 分布

図 8，**図 9**，**図 10** にアンサンブル学習による予測結果から得られた Precision，Recall，F1-score の分布を示す．Precision，Recall の分布において外れ値的に性能が低いサンプルが確認できるが，この要因として，4.1.2 項で示した特徴量のクラスタリング結果において，他の被験者と端末操作の特徴が大きく異なる被験者が存在していたことが考えられる．このような被験者が混在することで，全体のモデル性能にばらつきが生じ，特定の指標において外れ値が現れた可能性がある．したがって，サンプル数を拡張し，有意なクラスタに属する被験者群のみに対して交差検証を実施することで，モデルの精度が改善される可能性がある．

5. 結論

本研究は，クラウドソーシングにおけるマイクロタスク，特に画像に対するキャプションの正誤評価を対象とし，ユーザの回答品質が低下する傾向をリアルタイムに推定する手法を提案し，機械学習による二値分類モデルの構築と評価を行った．構築した機械学習モデルによる分類では，学内実験により得られたデータを用いた個人内の時系列バリデーションにおいて，Precision が平均 0.722，Recall が平均 0.741，F1-score が平均 0.731 となり，提案手法が実運用においてリアルタイムでの回答品質低下の推定に有効である可能性が示唆された．一方，アンサンブル学習を用いた汎化性能の評価では，Precision が平均 0.190，Recall が平均 0.339，F1-score が平均 0.243 にとどまり，今回構築したモデルには汎化性能が十分でないことが確認された．

今後の展望として，端末操作情報から回答者の行動をより詳細に推定し，高レベルの特徴量を設計することで，モデルの分類精度を向上させることが期待される．また，本実験は学内環境において実施されたため，より現実的な使用環境を再現するために，実際のクラウドソーシング環境における実験を行う必要がある．アノテーションタスクの評価指標については，本実験では 10 問ごとの正答率に基づいて回答品質を定義したが，今後は社会心理学で用いられる IMC や ARS に基づいた指標を導入することで，先行研究との比較が可能になると考えられる．また，構築した分類モデルは，端末操作情報を用いた二値分類モデルであったが，本研究で用いたアノテーションアプリケーションは会話型インタフェースによってタスクを提示しており，回答品質の低下が推定された場合に自然言語によるフィードバックを行うエージェントを実装することで，回答品質の低下だけでなくその要因を考慮した多クラス分類への応用が期待される．

謝辞

本研究の一部は，JSPS 科研費（JP24K20763）の助成を受けて行われたものです．

参考文献

[1] Mara Naegelin, Raphael P. Weibel, Jasmine I. Kerr, Victor R. Schinazi, Roberto La Marca, Florian von Wangenheim, Christoph Hoelscher, and Andrea Ferrario. An interpretable machine learning approach to multimodal stress detection in a simulated office environment. *J. of Biomedical Informatics*, Vol. 139, No. C, 2023.

[2] Herbert A Simon. Rational choice and the structure of the environment. *Psychological review*, Vol. 63, No. 2, p. 129, 1956.

[3] Daniel M Oppenheimer, Tom Meyvis, and Nicolas Davidenko. Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of experimental social psychology*, Vol. 45, No. 4, pp. 867–872, 2009.

[4] Michael R Maniaci and Ronald D Rogge. Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality*, Vol. 48, pp. 61–83, 2014.

[5] 後上正樹, 松田裕貴, 荒川豊, 安本慶一. オンラインアンケート回答時のスマートフォン画面操作状況に基づく不適切回答検出. 第 25 回一般社団法人情報処理学会シンポジウム・インタラクション 2021, pp. 11–20, 2021.

[6] Masaki Gogami, Yuki Matsuda, Yutaka Arakawa, and Keiichi Yasumoto. Detection of careless responses in online surveys using answering behavior on smartphone. *IEEE Access*, Vol. 9, pp. 53205–53218, 2021.

[7] Yoshinobu Fukumitsu, Yuki Matsuda, Hirohiko Suwa, and Keiichi Yasumoto. Detecting careless responses in dataset annotation using screen operation logs. In *2024 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom '24)*, pp. 775–780, 2024.

[8] Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. Improving Worker Engagement Through Conversational Microtask Crowdsourcing. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI'20, pp. 1–12, 2020.

[9] Zhihua Zhang, Juliana Miehle, Yuki Matsuda, Manato Fujimoto, Yutaka Arakawa, Keiichi Yasumoto, and Wolfgang Minker. Exploring the Impacts of Elaborateness and Indirectness in a Behavior Change Support System. *IEEE Access*, Vol. 9, pp. 74778–74788, 2021.

[10] Jeffrey M Rzeszotarski, Ed Chi, Praveen Paritosh, and Peng Dai. Inserting micro-breaks into crowdsourcing workflows. In *The First AAAI Conference on Human Computation and Crowdsourcing*, HCOMP'13, pp. 62–63, 2013.

[11] Peng Dai, Jeffrey M Rzeszotarski, Praveen Paritosh, and Ed H Chi. And Now for Something Completely Different: Improving Crowdsourcing Workflows with Micro-Diversions. In *Proceeding of The 18th ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW'15, pp. 628–638, 2015.

[12] Kohei Oyama, Yuki Matsuda, Rio Yoshikawa, Yugo Nakamura, Hirohiko Suwa, and Keiichi Yasumoto. A Method for Expressing Intention for Suppressing Careless Responses in Participatory Sensing. In *18th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*, MobiQuitous'21, pp. 769–782, 2021.

[13] Takaaki Nakagawa, Yutaka Arakawa, and Yugo Nakamura. Augmented Web Survey with enhanced response UI for Touch-based Psychological State Estimation. In *2022 IEEE 4th Global Conference on Life Sciences and Technologies*, LifeTech, pp. 91–95, 2022.

[14] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Advances in Neural Information Processing Systems*, Vol. 30 of *NIPS'17*, 2017.

[15] Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.