Vol.2023-MBL-108 No.18
Vol.2023-UBI-79 No.18
Vol.2023-CDS-38 No.18
Vol.2023-ASD-27 No.18
2023/9/25

クラウドソーシングを用いたアノテーションにおける不良回答の検出手法

福光 嘉伸[1,2,a)]　松田 裕貴[1,2,b)]　諏訪 博彦[1]　安本 慶一[1]

概要：アノテーション作業をクラウドソーシングで行うことにより，低コストで機械学習のための学習データを収集できる．しかし，対価として報酬を付与すると可能な限り速く回答を行おうとする行動が起こり，不良回答が発生する問題がある．そこで，本研究ではアノテーションタスクを対象として不良回答をリアルタイムで検出することを目指し，作業中の画面操作から得られるカーソル移動量や操作時間などの特徴量を用いた検出手法を提案する．本稿では，手法実現のための基礎的な分析として，提案する特徴量と不良回答との関係性を調査した結果，および構築した適切回答・不良回答の分類モデルの評価結果について報告する．

キーワード：アノテーション，不良回答検出，機械学習

1. はじめに

マイクロタスク型クラウドソーシングとは，インターネット上で不特定多数の群衆に短時間で遂行可能な業務を水平分散的に委託することで，低コストで大規模のデータを取得および解析することを可能とする方法であり，様々な用途での活用が進んでいる．特に機械学習の分野においては，モデル構築・精度向上のために大量の学習データが必要となることから，データのアノテーション作業をクラウドソーシングで行うことで低コストで学習データを収集することが重要となる．しかし，コストと引き換えに得られるデータの品質に大きなばらつきがあり，品質管理が課題となっている [1]．クラウドソーシングでは，アノテーションを行うユーザが必ずしも正確に回答するとは限らず，回答の対価として報酬を付与する場合に可能な限り速く回答を行おうとするように行動すること（努力の最小限化）が考えられる．人間は思考に時間を使うことで様々な誤りを減らすため [2]，努力の最小限化によって思考時間が減少することにより誤ったラベリングを行い不良回答が発生するという問題がある．誤ったラベルが多量に含まれる学習データを用いた場合，機械学習モデルの精度が低下する恐れがあるため，そうしたノイズとなりうるデータの発生を検知・防止することが求められている．

社会心理学といった質問紙調査（アンケート調査）を多く取り扱ってきた分野では，より正確な回答結果を得るために努力の最小限化の傾向を検出する手法が考案されている．三浦ら [3] は The Attentive Responding Scale（ARS）という矛盾を問う評価尺度を質問紙に取り入れることで，努力の最小限化の傾向を示す個人を検出する方法を提案している．しかし，回答者を疑うような質問を回答者自身が認識可能な形で提示する方法は，回答者に心理的負担を与え，回答者の内発的動機が損なわれることで，その質問自体が努力の最小限化の傾向を引き起こす可能性がある．そこで，後上ら [4], [5] らは，評価尺度を取り入れずに努力の最小限化の傾向を検出することを目的として，スマートフォンの画面操作記録を特徴量とする機械学習による検出手法を提案している．しかし，この手法ではアンケートの回答が全て終わってからしか検出を行うことができないことが問題となりうる．同一人物が一度しか実施することができないタスク（調査内容を知る前と知った後で回答が変化する性質があるもの）など，データの母集団が限られている場合，これまでの方法では不良回答の影響によって最終的に得られるデータが不足してしまう恐れがある．この観点はアノテーションについても同様に問題となりうる．本人しかアノテーションができないようなタスクの場合，対象者が不良回答を行った場合にはそのデータを利用できなくなったり，最終的に構築するモデルの精度に悪影響を及ぼしたりすることが懸念される．このことから，アノ

[1] 奈良先端科学技術大学院大学
Nara Institute of Science and Technology
[2] 共同第一著者，Co-first author
[a)] fukumitsu.yoshinobu.ft5@is.naist.jp
[b)] yukimat@is.naist.jp

Vol.2023-MBL-108 No.18
Vol.2023-UBI-79 No.18
Vol.2023-CDS-38 No.18
Vol.2023-ASD-27 No.18
2023/9/25

テーション作業中にリアルタイムに不良回答の検知を行う重要性が高いと考える．

本研究では，データのアノテーションタスクを対象として不良回答をリアルタイムに検知する手法を提案することを目指す．提案手法では，アノテーション作業中のクリックやマウスカーソルの移動等の画面操作をリアルタイムに記録し，得られる特徴量を用いた不良回答検出を行う．これに対し本稿では，不良回答の分類モデルを作成するために，固有表現ラベリングのタスクを対象として学習データの取得実験を実施する．実験により得られた画面操作記録から特徴量を抽出し，正常回答群と不良回答群 (アノテーション結果が基準に満たない) それぞれで各特徴量について母平均の差の検定を行う．その後，分類モデルを作成しモデルの評価を行う．さらに，得られた結果についての分析結果について報告する．

2. 関連研究

不良回答を検出する既存研究としては，リッカート式やテキストベースのアンケートを対象としたものが一般的である．後上ら [4], [5] は，一回のスクロール操作による画面移動量やスクロール速度などの画面操作を記録するシステムを作成し，画面操作を特徴量として用いることで，85.9%の検出率を達成している．加えて，中川ら [6] は，アンケート回答中の迷いが反映されたタッチ操作ログを取得するべく，スライドバーや拡大鏡を活用した 2 種類の回答 UI を提案している．

アノテーションに関して客観的な品質を測る指標についても研究されている．Otani ら [7] は，物体検出精度の品質を測る指標として，一般的な mAP（Mean Average Precision）に加え，誤差を含む検出結果を正解に修正するためのコスト OC-cost（Optimal Correction Cost）を提案している．

また，アノテーションやマイクロタスクに対するユーザのモチベーションや作業及びデータの質を向上させるための研究は多数行われている．Sihang ら [8] は，クラウドソーシングでマイクロタスクを作業として与える際に，従来の Web インターフェースの代わりに会話型インターフェースを用いることで，ユーザのやる気を向上させる手法を提案している．Jeffrey ら [9] は，長時間の作業中に適度な休憩を与えることで，ユーザの定着率を大幅に改善し，作業に対する関与を高めることを明らかにした．また，香川ら [10] は，膀胱鏡画像が異常か正常か判断するタスクにおいて作業前に 1 秒待ち時間を設定することでアノテーション品質を向上できることを明らかにした．

このように，不良回答を検出する既存研究 [4], [5] では，リアルタイム性がなく，回答が全て終わってからしか不良回答の検出を行うことができないという課題を残している．また，不良なクラウドワーカーへの対応策を講じておらず，取得したデータの一部を不良データとして削除しなければならない可能性がある．アノテーションやマイクロタスクに対する既存研究 [8], [9] では，やる気や定着率を向上させる点に留まっており，出力結果（データセット）の質を向上させることは難しい．待ち時間を作業前に設ける研究 [10] では，数%の改善にとどまっている，かつ長い思考時間が必要となるタスクではその効果は小さいと考えられる．これらのことから，リアルタイムに不良回答の検出を行い，不良回答を改善することのできる方法を検討する重要性は高いと考えられる．

表 1 抽出する特徴量

特徴量	単位
回答時間	s
非操作時間	s
別ウインドウ移動回数	回
カーソル移動量 (x 軸, y 軸, 合計)	px
クリック回数	回
カーソル n 秒以上停止回数 {n=1,3,5,10,30}	回
クリック間隔 平均時間	s
クリック間隔 分散	s^2
最長/最短クリック間隔	s
文字選択回数	回
n 秒毎のカーソル移動速度の最大/最小値 {n=0.5,1,3,5,10}	px/s
n 秒毎のカーソル移動速度の分散 {n=0.5,1,3,5,10}	$(px/s)^2$

3. 実験

3.1 画面操作記録システム

画面操作が記録可能なシステムについて述べる．オープンソースソフトウェアのアノテーションツールである LabelStudio [11] に対して，ユーザのブラウザ上の操作を記録する機能を実装することで，アノテーション中の画面操作データを随時データベースに格納する．

次に，画面操作データから抽出する特徴量について述べる．本研究では，アノテーションタスクに対応でき，リアルタイムに抽出することができる特徴量を追加する．さらに，各タスクの実施をトリガーとし特徴量を抽出する．

表 1 に用いる特徴量とその単位を示す．特徴量は，対象とするアノテーションタスクの操作内容に合わせて考案されている．今回対象とする固有表現ラベリングタスクは，文章中から固有名詞（人名や書籍名など）や日時表現など固有表現を抽出し，ラベルを付与するタスクである．タスクには，マウスカーソルを移動させることによる文字選択（該当単語をドラッグ操作で囲う）や，選択した領域へのラベル付与（指定されたリストから対応するラベルを選択する）といった画面操作が含まれる．タスクの実施時には，文章を見てどの単語が固有表現であり，どのカテゴリに属するかを判断する認知処理が必要とされる．不良回答では，速く回答を行うために十分な数の単語にラベルを付与していないことや，文章中の固有表現に適切なラベルが

Vol.2023-MBL-108 No.18
Vol.2023-UBI-79 No.18
Vol.2023-CDS-38 No.18
Vol.2023-ASD-27 No.18
2023/9/25

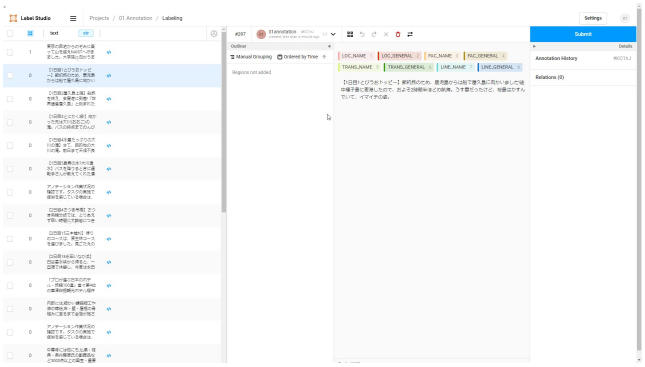


図 1　ラベル付与を行う画面例.

図 2　ラベル付与の様子.

付与されていないこと，回答に迷ってマウスカーソルの速度にばらつきがあることが考えられる．そのため，適切/不良回答間で回答時間，文字選択回数やカーソル移動量・速度等に違いが出ると考えられる．

3.2　学習データ取得実験

3.1 節で述べた画面操作記録システムを用いて実施した学習データ取得実験および取得データセットについて説明する．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）．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募集を行い，被験者数は 61 人であった．被験者に，実際にアノテーション作業を実施してもらいその際の画面操作を記録することで学習データを取得する．被験者には，報酬として 1000 円相当のギフトカードを付与した．実験で用いたアノテーション作業の設計と取得データについて以下で述べる．

3.2.1　アノテーション作業の設計

本実験におけるアノテーション作業としては，地球の歩き方旅行記データセット[*1] [12], [13] を用いた固有表現ラベリングを対象とした．アノテーション作業は**図 1** に示すように LabelStudio 上で行い，被験者は 80〜120 字の日本語文章がシステムから順に与えられる．この文章中の固有表現の内，地名・施設名・乗り物名・路線名や道・橋等の経路に対して固有名詞/一般名詞を区別してラベルを付与する．この 4 種類に該当しない表現や単語はラベル付与の対象外であり，各タスクには 3〜9 個の付与対象の固有表現が含まれる．LabelStudio を用いたラベル付与は，**図 2** に示すように，ラベルをクリックで選択して，任意の文字を範囲選択することで行うことができる．対象となる文章，およびラベル付の例を以下に示す．

自宅を出てのぞみに乗り山を越え新大阪駅に到着しました． 新大阪駅からJR 京都線とバスを乗り継ぎ，奈良駅に着きました．猿沢池までの道は歩きました．

凡例
地名（固有名詞）・地名（一般名詞）
施設名（固有名詞）・施設名（一般名詞）
乗り物名（固有名詞）・乗り物名（一般名詞）
路線名（固有名詞）・路線名（一般名詞）

本実験はデータセットのうち日本語文章の 23 文を選択した．全ての文へのアノテーションの終了，もしくは作業実施から 30 分を経過した時点でのタスクへのアノテーションの終了をもって，本実験の作業終了とした．

3.2.2　データセット

旅行記データセットのグラウンドトゥルースと，実際に被験者がラベリングを行った各タスクを比較し，適切回答であるか不良回答であるかを判別する．被験者 61 人のうち，指示通りに作業を実施した 56 人のデータを分析対象とする．また，前処理としてタスクの回答時間が 10 秒未満，5 分以上のデータに関しては異常値として取り除く．

ラベル範囲・ラベル種類の両方がグランドトゥルースと一致したもののみ正解とし，そのときの F1 score（F1 値）を評価値とする．以下の式 1，2，3 より F1 値を算出する．不良回答とみなす F1 値の閾値について，付与対象が最も少ない 3 つの場合に，1 つのラベル付与間違いまでは許容するように設定する．そのため，F1 値が 0.7 以上のラベル済みデータを適切回答，0.7 未満を不良回答として扱いデータセット $D_{\pm 0}$ を作成する．

また，自然言語領域においてラベル範囲に関しては，文字数の ±N% までは正解とみなすとして部分一致を許容する場合もある．明確に許容範囲の値が決められているわけではなく，対象とするアノテーション作業にあわせて設定する必要がある．本アノテーション作業でラベル付与の範囲不足が発生する原因に，複合名詞に対して片方の名詞にのみラベルを付与する場合が挙げられる．そのため，文字数の ±50% までの部分一致を正解とみなして，F1 値を算出し，データセット $D_{\pm 50}$ を作成する．

作成した各データセットに含まれる適切回答・不良回答のデータ数を**表 2** にそれぞれに示す．

表 2　各データセットに含まれるデータ数

	データセット $D_{\pm 0}$	データセット $D_{\pm 50}$
適切回答	472	572
不良回答	616	516

[*1] https://www.nii.ac.jp/dsc/idr/arukikata/

Vol.2023-MBL-108 No.18
Vol.2023-UBI-79 No.18
Vol.2023-CDS-38 No.18
Vol.2023-ASD-27 No.18
2023/9/25

$$Presition = \frac{\text{正解ラベル数}}{\text{付与ラベル数}} \quad (1)$$

$$Recall = \frac{\text{正解ラベル数}}{\text{グランドトゥルースのラベル数}} \quad (2)$$

$$F1score = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

3.3 分析

実験により取得した画面操作データから 3.1 節で説明した特徴量を抽出したデータセット $D_{\pm0}$，$D_{\pm50}$ それぞれに対して本節で示す各分析を行う．適切回答群と不良回答群の各特長量を，母平均の差を有意水準 5%とするウェルチの t 検定によって検定する．

その後，特徴量を用いて機械学習による分類を行う．機械学習のアルゴリズムとして，LightGBM（LGBM）[14] を用いる．LightGBM は一般的に少ないサンプルデータでも高い精度が出やすい，かつ決定木ベースのアルゴリズムであるため不要な特徴量が含まれていても精度が低下しにくいという特徴がある．汎化性能を考慮した上で分類精度を算出するために，すべての学習データ 1 件ずつに対して，その被験者のものでないデータを集め，検証用データとして分類させる．分類の評価指標として Precision，Recall，Accuracy を用い，モデルの評価を行う．

4. 実験結果・考察

各データセットにおける特徴量の検定結果を**表 3**，**表 4** に示す．「母平均の大小」欄では，適切回答群と不良回答群の母平均を比較したとき，不良回答群の方が大きい（High）か小さい（Low）かを示す．「有意差」欄では，統計的な有意差があるか否かを示している．「−」印は有意差がないこと，「*」印は有意差があることを表している．表 3，表 4 からどちらのデータセットにおいても，クリック回数・文字選択回数は有意差があり，不良回答群は適切回答群に比べてこれらの特徴量の平均が小さいことが分かる．また，データセット $D_{\pm0}$ に比べてデータセット $D_{\pm50}$ は，より多くの特徴量で有意差がある．このことより，ラベル付与範囲の部分一致を許容することで，適切回答群と不良回答群の一部特徴量の違いが大きくなるといえる．

どちらのデータセットでも有意差が認められたクリック回数と文字選択回数について，適切回答群と不良回答群との比較を箱ひげ図で**図 3**，**図 4** に示す．図より，データセット間で大きな傾向の違いはないといえる．また，適切回答群に比べて不良回答群は，クリック回数の中央値が小さい値であり，文字選択回数の第二四分位数から最大値にかけての値のばらつきが小さいことが分かる．これらの特徴は機械学習での不良回答で有用な特徴であると考えられる．

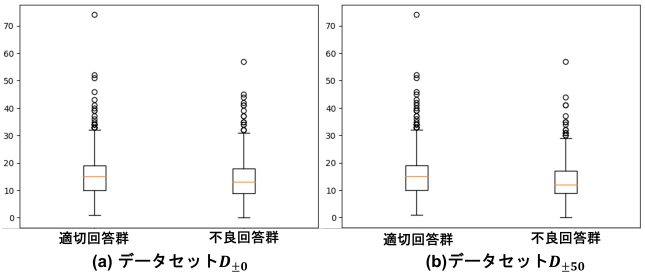
データセット $D_{\pm0}$ で有意差が見られない，かつデータセット $D_{\pm50}$ で有意差がある特徴量のうち，適切回答と不良回答間で違いが出ると仮定していた回答時間とカーソル移動速度の分散の箱ひげ図をそれぞれ**図 5**，**図 6** に示す．

表 3 適切回答群に対する不良回答群の母平均の大小と検定結果（データセット $D_{\pm0}$）

特徴量	母平均の大小	有意差
回答時間	Low	-
非操作時間	Low	-
別ウインドウ移動回数	High	-
カーソル移動量 (x 軸, y 軸, 合計)	Low	-
クリック回数	Low	*
カーソル n 秒以上停止回数 {n=1, 5}	High	-
カーソル n 秒以上停止回数 {n=3, 10, 30}	Low	-
クリック間隔 平均時間	High	-
クリック間隔 分散	High	-
最長クリック間隔	Low	-
最短クリック間隔	High	-
文字選択回数	Low	*
n 秒毎のカーソル移動速度の最大値 {n=0.5, 1, 3, 5, 10}	Low	-
n 秒毎のカーソル移動速度の最小値 {n=0.5, 1, 3, 5, 10}	High	-
n 秒毎のカーソル移動速度の分散 {n=0.5, 1, 3, 5, 10}	Low	-

表 4 適切回答群に対する不良回答群の母平均の大小と検定結果（データセット $D_{\pm50}$）

特徴量	母平均の大小	有意差
回答時間	Low	*
非操作時間	Low	*
別ウインドウ移動回数	Low	-
カーソル移動量 (x 軸, y 軸, 合計)	Low	*
クリック回数	Low	*
カーソル n 秒以上停止回数 {n=1}	High	-
カーソル n 秒以上停止回数 {n=3, 5, 10, 30}	Low	-
クリック間隔 平均時間	High	-
クリック間隔 分散	Low	-
最長クリック間隔	Low	-
最短クリック間隔	High	-
文字選択回数	Low	*
n 秒毎のカーソル移動速度の最大値 {n=0.5, 1, 3, 5, 10}	Low	*
n 秒毎のカーソル移動速度の最小値 {n=0.5, 1, 3, 5, 10}	High	-
n 秒毎のカーソル移動速度の分散 {n=0.5, 1, 3, 5, 10}	Low	*



(a) データセット$D_{\pm0}$　(b)データセット$D_{\pm50}$

図 3 クリック回数.

Vol.2023-MBL-108 No.18
Vol.2023-UBI-79 No.18
Vol.2023-CDS-38 No.18
Vol.2023-ASD-27 No.18
2023/9/25

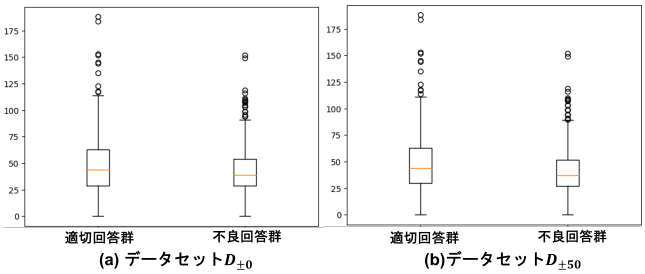


図 4　文字選択回数.

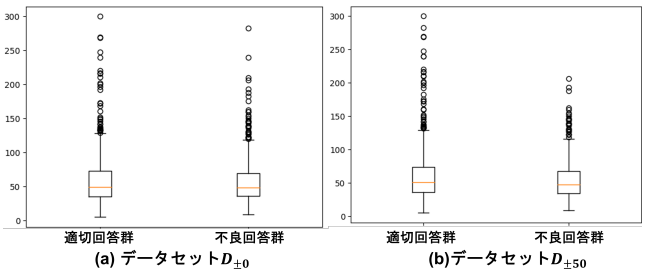


図 5　回答時間.

図より，データセット $D_{\pm0}$ では回答時間とカーソル移動速度の分散共に違いがあまりないといえる．また，どちらの特徴量においても適切回答群，不良回答どちらにも外れ値が多数存在する．データセット $D_{\pm50}$ についても，四分位範囲自体に大きい違いはなく外れ値が多数存在する．回答時間とカーソル移動速度の分散の特徴量で有意差がある要因として，外れ値の影響が挙げられる．図 6 で 2 つのデータセットを比較すると，データセット $D_{\pm0}$ の不良回答群で存在していた外れ値が，データセット $D_{\pm50}$ では一部存在しない点から母平均に違いが出たと考えられる．これらの特徴の適切回答群と不良回答群の母平均の差はデータ全体での特徴ではなく，外れ値の影響が大きいため機械学習での不良回答で有用な特徴であるといえない．

表 5，**表 6** に画面操作記録から抽出した特徴量を学習したモデルの分類結果を示す．表中に正解データの適切回答/不良回答それぞれに対して予測した結果をデータ数で示し，表下部に分類モデルの各評価指標の値を示す．データセット $D_{\pm0}$ は 0.548，データセット $D_{\pm50}$ は 0.528 の Accuracy であった．このことから，今回抽出した特徴量では適切回答/不良回答の間の違いを表すことが難しく分類することができなかったと考えられる．主な原因として，アノテーション作業の実施において作業速度や本人の能力などの個人差や問題毎の難易度が存在し，適切回答/不良回答間で特徴が混在していることが挙げられる．

次に，各タスクで全被験者の F1 値を箱ひげ図にしたものを**図 7** に示す．横軸はタスク番号，縦軸は F1 値を表している．箱ひげ図から各タスクで F1 値の中央値や四分位範囲に違いがあることが分かる．この結果から，アノテーション作業においてタスク間に難易度差が生じていたため，各タスクで F1 値のばらつきに違いが出たと推測できる．

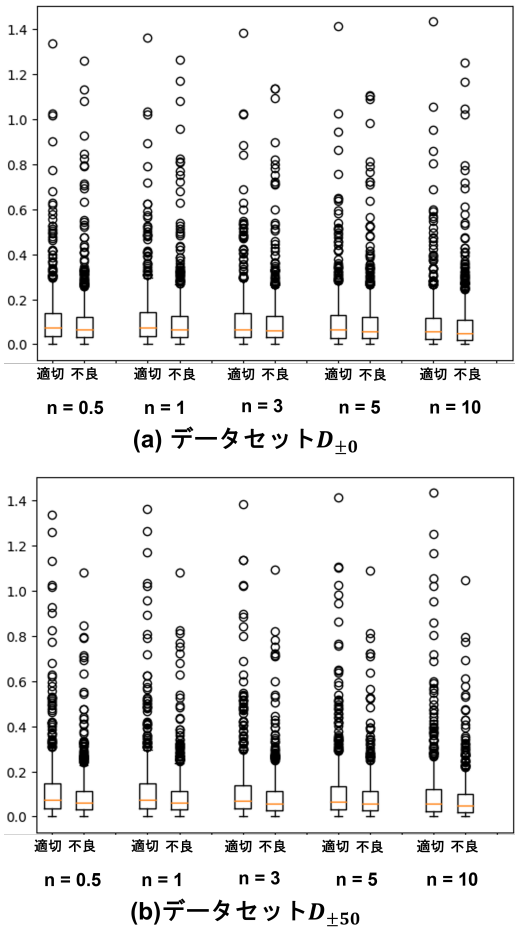


図 6　n 秒毎のカーソル移動速度の分散（n = 0.5, 1, 3, 5, 10）.

表 5　データセット $D_{\pm0}$ の分類結果

		予測	
		適切回答	不良回答
正解	適切回答	191	281
	不良回答	211	405

Presicion: 0.475　Recall: 0.405　Accuracy: 0.548

表 6　データセット $D_{\pm50}$ の分類結果

		予測	
		適切回答	不良回答
正解	適切回答	329	243
	不良回答	270	246

Presicion: 0.549　Recall: 0.575　Accuracy: 0.528

そのため，似た画面操作記録でもタスクによって F1 値が異なってしまい分類が上手く出来なかったと考えられる．

これらの結果から，作業速度や本人の能力に依存せず適切回答/不良回答間の画面操作の違いを表すことのできる特徴量の考案や，適切回答/不良回答の F1 値の基準を一律で設定するのではなくタスク毎に設定をする必要がある．

5. おわりに

本研究はクラウドソーシングにおけるマイクロタスク，特に機械学習のためのアノテーションタスクを対象とし，

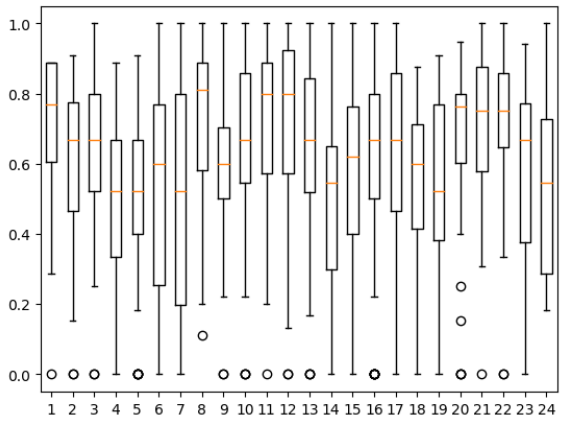


図 7 タスク毎の F1 値の違い.

ユーザが取る不良回答をリアルアイムに検知する手法を提案し，学習データにおける統計量分析を実施した．提案手法はアノテーション作業中の画面操作をリアルタイムに記録し，特徴量を抽出することで作業実施中の不良回答検出を実現する．分析の結果から，固有表現ラベリングのアノテーション作業において，適切回答群に対して不良回答群ではクリック回数と文字選択回数の特徴量で母平均が小さく有意差があることがわかった．機械学習モデルによる分類では，上手く適切回答/不良回答を分類することができず，被験者それぞれの能力や作業速度などの個人差やタスク間の難易度差の影響が原因である可能性が示唆された．今後は，タスク毎に F1 値の基準を設定する方法や，作業速度や本人の能力に依存しない特徴量について模索するとともに，適切回答・不良回答の分類モデルの精度向上を目指す.

謝辞 本研究の一部は，JST さきがけ（JPMJPR2039）の助成を受けたものである.

参考文献

[1] Daniel, F., Kucherbaev, P., Cappiello, C., Benatallah, B. and Allahbakhsh, M.: Quality Control in Crowdsourcing: A Survey of Quality Attributes, Assessment Techniques, and Assurance Actions, *ACM Comput. Surv.*, Vol. 51, No. 1 (online), DOI: 10.1145/3148148 (2018).

[2] Kolvoort, I. R. and van Maanen, L.: Causal reasoning under time pressure: testing theories of systematic non-normative reasoning patterns, *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 43, No. 43 (2021).

[3] 三浦麻子，小林哲郎：オンライン調査における努力の最小限化を検出する技法，社会心理学研究， Vol. 32, No. 2 (2016).

[4] 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).

[5] Gogami, M., Matsuda, Y., Arakawa, Y. and Yasumoto, K.: Detection of Careless Responses in Online Surveys Using Answering Behavior on Smartphone, *IEEE Access*, Vol. 9, pp. 53205–53218 (online), DOI: 10.1109/ACCESS.2021.3069049 (2021).

[6] Nakagawa, T., Arakawa, Y. and Nakamura, Y.: Augmented Web Survey with enhanced response UI for Touch-based Psychological State Estimation, *2022 IEEE 4th Global Conference on Life Sciences and Technologies*, LifeTech, pp. 91–95 (2022).

[7] Otani, M., Togashi, R., Nakashima, Y., Rahtu, E., Heikkilä, J. and Satoh, S.: Optimal Correction Cost for Object Detection Evaluation, *The IEEE/CVF Computer Vision and Pattern Recognition Conference*, CVPR'22 (2022).

[8] Qiu, S., Gadiraju, U. and Bozzon, A.: Improving Worker Engagement Through Conversational Microtask Crowdsourcing, *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI'20, pp. 1–12 (online), DOI: 10.1145/3313831.3376403 (2020).

[9] Rzeszotarski, J. M., Chi, E., Paritosh, P. and Dai, P.: Inserting micro-breaks into crowdsourcing workflows, *The First AAAI Conference on Human Computation and Crowdsourcing*, HCOMP'13, pp. 62–63 (2013).

[10] 香川璃奈，白砂大，池田篤史，讃岐勝，本田秀仁，野里博和：1 秒待つことによるアノテーション品質の向上：作業能力向上と心的負担のトレードオフを考慮した作業環境への介入，第 36 回人工知能学会全国大会 (2022).

[11] HumanSignal, Inc.: Label Studio, `https://github.com/heartexlabs/label-studio` (2019). (Accessed on 2023-09-01).

[12] Ouchi, H., Shindo, H., Wakamiya, S., Matsuda, Y., Inoue, N., Higashiyama, S., Nakamura, S. and Watanabe, T.: Arukikata Travelogue Dataset, *arXiv*, No. 2305.11444, pp. 1–6 (online), DOI: 10.48550/arXiv.2305.11444 (2023).

[13] Higashiyama, S., Ouchi, H., Teranishi, H., Otomo, H., Ide, Y., Yamamoto, A., Shindo, H., Matsuda, Y., Wakamiya, S., Inoue, N., Yamada, I. and Watanabe, T.: Arukikata Travelogue Dataset with Geographic Entity Mention, Coreference, and Link Annotation, *arXiv*, No. 2305.13844, pp. 1–11 (online), DOI: 10.48550/arXiv.2305.13844 (2023).

[14] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. and Liu, T.-Y.: LightGBM: A Highly Efficient Gradient Boosting Decision Tree, *Advances in Neural Information Processing Systems*, NIPS'17, Vol. 30 (2017).