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

後上 正樹[1,a)]　松田 裕貴[1,2]　荒川 豊[3]　安本 慶一[1]

概要：アンケートにおいて，なるべく楽に早くタスクを完了しようとする「Satisficing（努力の最小限化）」という態度により，結果の信頼性が低下する問題がある．より正確な結果を得るためには，Satisficing を検出して分析対象から除外するなどの前対処が必要となる．これまでに，回答時間に基づく検出手法や，指示違反や矛盾を問う質問群を追加する手法が考案されてきた．しかし，前者では回答時間を故意に水増しした回答を適切に除外することができない．また，後者は回答者を疑ってスクリーニングするようなものであり，回答者のモチベーションを損ねて Satisficing を助長してしまう原因となる．そこで我々は，回答時間だけでなく回答中の画面操作を利用することで，より高精度に Satisficing が検出可能になるのではないかという仮説を立てた．しかしながら，(1) 画面操作が記録可能なアンケートシステムが存在せず，(2) また，記録できたとしてどのような特徴が Satisficing と関連しているのか不明であった．そこで，我々は世界中で利用されているオンラインアンケートシステム LimeSurvey 用の画面操作記録プラグインを開発し，多人数（5692 人）の様々な画面操作ログを収集して，機械学習による Satisficing 検出を行なった．Leave One Out 交差検証による評価の結果，検出率は 85.9%を達成し，同様のタスクに取り組んだ先行研究の検出率 55.6%を大幅に上回るものとなった．また，本稿で新たに提案した特徴量の中では，スクロールに関連する特徴の寄与率が高い結果となった．

Detection Method of Inappropriate Responses in Online Surveys Using Smartphone Screen Operation Log

1. はじめに

近年，オンラインアンケート調査が社会科学関連の研究にも用いられている．また，2015 年からは総務省が実施している国勢調査にも用いられ始めた [1]．オンラインアンケートは一度に大量のデータを取得できる等の利点がある．しかし，アンケート調査では回答者が必ずしも正確に回答するとは限らない．この問題に関して，Simon ら [2] は，人間は何らかの目的を達成するための認知的努力を最低限に抑えるという認知心理学の概念を考案した．Krosnick ら [3] は，これをアンケート調査の領域に適用し，回答要求に対して努力を最小化しようとする傾向を “Satisficing” と定義した．以降，Satisficing がアンケート結果から導かれる知見の妥当性を損うことが指摘されてきた [4], [5]．そこで，より真実に近い知見の獲得のために，Satisficing を検出する手法が考案されてきた．

例えば，Oppenheimer ら [6] は，教示文の中に「何も回答せず次のページに移動してください」などの指示文を挿入する IMC（Instructional Manipulation Check）を提案し，これを用いたアンケート実験を行なった．この指示に違反した場合は Satisficing であると見なされる．2 回の実験の結果，Satisficing の割合は 46%，35%であったと報告されている．これらの値を見ると，Satisficing によってアンケート結果から得られる知見が歪む可能性は容易に想像できる．

IMC の他にも，DQS（Directed Question Scale）や ARS（Attentive Responding Scale）という，指示違反や矛盾を問う質問を追加する手法がある．しかし，これらの方法では回答者を疑うような質問を用いるため，回答者に心理

[1] 奈良先端科学技術大学院大学
Nara Institute of Science and Technology
[2] 国立研究開発法人科学技術振興機構 さきがけ
JST PRESTO
[3] 九州大学
Kyushu University
[a)] gogami.masaki.gg8@is.naist.jp

的負担を与えてしまう．これにより，適切に回答してアンケートに協力している回答者の内発的動機が損なわれてしまい，その質問自体が Satisficing を引き起こしてしまう可能性が考えられる．さらに，Pei ら [7] は，IMC と DQS の質問に自動で回答するディープラーニングモデルを構築した．これは，従来の Satisficing 検出手法の信頼性を損なうことを意味すると言える．

尾崎ら [8] は，このような検出用の質問を追加する必要がない，機械学習を用いた検出を試みた．様々な機械学習アルゴリズムを試した結果，不適切回答の検出率が最も高いモデルで 55.6%であったと報告されているが，実用の観点から十分な精度とは言えない．いくつものアルゴリズムを試した結果としておおよそ 40%後半 ～ 50%前半の検出率となっている点から，検出率向上に対するボトルネックは，アルゴリズムの種類および性能ではなく特徴量の質である可能性が考えられる．

そこで我々は，近年オンラインアンケートの回答に利用されることが増加しつつあるスマートフォン [9] の画面操作に着目し，不適切回答を高精度に検出することを目指す．後上ら [10] が開発した回答時の画面操作が記録できるアンケートシステムを用いてクラウドソーシング上でアンケートを実施し，画面操作データを収集した．このデータから特徴量を生成し，教師あり学習による不適切回答検出モデルを構築した．正解ラベルは，後述する Satisficing 指標である DQS と ARS の両方が Satisficing であるサンプルを正例，それ以外のサンプルを負例とする二値とした．Leave One Out 交差検証によるモデルの評価を行なったところ，不適切回答の検出率 85.9%という結果が得られた．この結果から，尾崎ら [8] と大差のない不適切回答の定義上で，本手法が検出率を大幅に向上（55.6%→85.9%）可能であることを確認した．

さらに，質問数と検出率の関係を検証するために，特徴量生成に用いる画面操作データの対象ページ数を 3，9 ページと制限した結果，検出率はそれぞれ 79.7%，80.9%であった．全 17 ページを対象とした場合の 85.9%という結果も踏まえると，検証した範囲内においては，質問数が多いほどより高精度に検出できる可能性が示唆された．一方で，17 問（3 ページ）程度の質問数でも約 80%の検出率を維持することが確認された．

本稿の構成は次の通りである．2 章で関連研究について紹介する．3 章で Satisficing と関連すると考えられる画面操作について述べる．4 章では，検討した画面操作を記録するアンケートシステムについて述べる．5 章では，そのシステムを用いたアンケート実験と，Satisficing 指標および不適切回答の定義について説明する．6 章では，実験で収集したデータを用いて構築した不適切回答検出モデルについて説明し，検出結果とそれに対する考察について述べる．最後に，7 章で本稿のまとめと今後の展望を述べる．

2. 関連研究

三浦ら [11] は，Oppenheimer ら [6] が考案した IMC を用いたアンケート実験を日本で実施した．2 つの調査会社で同一のアンケートを行なった結果，Satisficing の割合は 51.2%，83.8%であったと報告されており，日本においても Satisficing が問題であることは明らかである．

IMC の他にも，いくつかの Satisficing 検出手法が考案されている．リッカート形式の質問の中に検出用の質問を追加して用いられる手法として，DQS（Directed Question Scale），ARS（Attentive Responding Scale）がある [12]．DQS は，IMC と同じく，回答指示に対する違反を基に Satisficing を検出する．IMC では教示文の中に回答指示を記載する一方，DQS では質問項目の中に回答指示を記載する．ARS は，回答の矛盾・非常識度を基に Satisficing を検出する．さらに，SC（Seriousness Check）という “真面目に回答したかどうか” を 2 択で直接的に尋ねるという手法も存在する [13]．また，三浦ら [14] は，Satisficing を効率よく，かつ正確に検出するために，Lasso によって ARS や DQS で用いられる質問項目から最低限必要な項目を絞り込むことを試みた．しかしながら，当該実験結果においては確定的な絞り込みは実現できなかったと述べられている．

これらの検出手法では，回答者が適切に回答しているかどうかを疑うような質問を用いるため，回答者がアンケートに協力する内発的動機が損なわれてしまい，検出手法自体が Satisficing の原因となる可能性が考えられる．さらに，Pei ら [7] は，IMC と DQS の質問に自動で回答するディープラーニングモデルを構築し，約 78.5%の精度で正解したと報告した．この結果は，上述の検出手法の信頼性を脅かす結果であると言える．

アンケート調査会社 [15] では，回答時間が極端に短い回答，規則性のある回答，矛盾する回答などを無効回答とする例がある．また，深井ら [16] は，ページ数，質問数，文字数，スクロール速度，読速度などから，不適切回答であると決定づける回答時間の閾値を算出し，閾値以下の所要時間の回答を除去する手法を提案した．ただし，スクロール速度や読速度など個人に依存する要素は回答者によらず定数としており，事前実験によって計測した被験者群の平均値が用いられた．この手法によって閾値以下の回答時間のサンプルを除外した結果，連続同一回答の含有率が 0.66 倍に減少したと報告されている．これらの検出手法で用いられるデータであれば，余分な質問を追加せずとも取得できる．しかし，故意に回答時間を水増しした回答や，無作為な回答を検出することができない問題がある．

このように，様々な Satisficing を検出するための手法が検討されているが，検出手法自身が回答者のモチベーションを低下させる原因となったり，悪意のある回答者に対する脆弱性があったりするため，ストレスフリーかつより堅

牢な Satisficing 検出手法が求められている．

尾崎ら [8] は，機械学習を用いることで，検出用の質問を挿入することなく検出する手法を試みた．対象とした回答用端末は PC であり，特徴量には性別，年齢，回答時間，連続同一回答数，ハマラノビス距離およびその p 値が用いられた．様々な機械学習アルゴリズムを試した結果，不適切回答の検出率が最も高かったブースティングアルゴリズムの検出率が 55.6%であったと報告されている．いくつものアルゴリズムを試した結果としておおよそ 40%後半 ∼50%前半の検出率となっている点から，精度向上を阻むボトルネックはアルゴリズムではなくデータの質である可能性が考えられる．また，性別や年齢は実際のアンケートでは対象者の性質として限定される可能性があるため，特徴量として用いるのは望ましくない．

また，近年オンラインアンケートの回答に用いられる端末として，スマートフォンが PC に取って代わってきている [9]．これを受けて，Roger ら [17] は，PC およびタブレットとスマートフォンでアンケート結果の質にどのような変化があるのかを調査した．この際，評価基準としたのは回答時間や未回答率および連続同一回答数である．結果として，スマートフォンは PC およびタブレットに比べて回答時間が長い傾向が観察された．しかし，結果の信頼性についてはどの端末についても特に差はないと結論づけている．

PC と同等の信頼性があり，携帯性が高いスマートフォンは今後オンラインアンケートに回答する端末としての使用が増加することが見込まれる．そこで本稿では，スマートフォンの画面操作データを用いた機械学習による不適切回答検出手法を提案し，検出率の向上を目指す．

3. Satisficing に関連する画面操作の検討

Satisficing に関連するデータとして考えられる項目を表 1 に示す．表中の「質問形式」欄では，各画面操作が本稿で扱うリッカート形式もしくは自由記述形式のどちらに関するものなのかを表している．なお，項目「全体」は質問形式に関係なくアンケート全体に関する特徴量であることを表す．また，「独自追加」欄では，– 印の今までに考案されていたデータに加えて，独自に追加したデータを ◯ 印で示す．

回答時間は，これまでの研究でも不適切回答検出に用いられてきた．アンケート調査会社等でも，アンケート全体の回答時間が短すぎるサンプルを調査結果から除外する例がある [15]．このようなフィルタリングに引っかからない不適切回答者や，アンケートのある部分のみ不適切な回答をする回答者なども存在し得る．しかしながら，依然として回答時間は Satisficing に強く関係する特徴量であると考える．本稿で扱う回答時間は，「リッカートの回答時間」と「自由記述の回答時間」に分け，それぞれの形式の質問に対する平均回答時間とした．

表 1 Satisficing に関連すると考えられるデータ

特徴量	単位	質問形式	独自追加
リッカートの回答時間	s	リッカート	–
自由記述の回答時間	s	自由記述	–
選択肢の変更回数	回	リッカート	◯
テキストの削除回数	回	自由記述	◯
スクロール長	px	全体	◯
スクロール時間	s	全体	◯
スクロール速度	px/s	全体	◯
逆スクロール回数	回	全体	◯
非操作時間が長すぎる回数	回	全体	◯
最大連続同一回答数	問	リッカート	–
中間回答数	問	リッカート	–
文字数	文字	自由記述	–

「選択肢の変更回数」「テキストの削除回数」は，少なくとも雑な回答をしようとした場合には発生しないと推測される．したがって，適切に回答するために検討している状態を表現する画面操作であると捉え，これらも Satisficing に関連すると考えた．

「スクロール長」は，一回のスクロール操作による画面移動量と定義した．「スクロール時間」も同様に一回のスクロール操作にかかった時間とし，「スクロール速度」は「スクロール長」を「スクロール時間」で除算した値とした．これらは質問間の移動という行動を表す一つのパラメータとして捉えることができる．Satisficing 状態では，早く終わらせたいという思いから質問間の移動操作が粗くかつ速くなると考えられる．

「逆スクロール回数」は，100px 以上の逆向きのスクロールを 1 回と定義した．100px とした理由は，LimeSurvey のアンケートを一般的なスマートフォンで回答する際に前の質問に戻るために最低限必要な移動量であるためである．アンケートの回答中に前の質問の回答を変更する場合や，ページ冒頭の質問文を読み直す際の行動を表していると考えられる．このような行動は丁寧に回答しようとしている状態を裏付けるものであるため，逆スクロール回数がほとんどないような場合は Satisficing 状態にある可能性が考えられる．

「非操作時間が長すぎる回数」は，画面に触れていない時間が基準値以上の回数である．この基準値はリッカート形式では 10 秒，自由記述形式では 40 秒と定義した．この基準値を上回る非操作時間は，回答にかかるであろう想定時間の範囲を超えているため，何か他の作業をしながら回答している状態であると見なす．Gould ら [18] はアンケートページからの一時離脱を捕捉したが，本画面操作ではスマートフォン上で発生しないアナログななから操作にも対応できる．ながら操作は質問への注意を逸らす要因であるため，この回数が多い場合は Satisficing 状態にある可能性が高いと考えた．

「最大連続同一回答数」は，リッカート形式の質問において同じ選択肢を連続で回答する回数の最大値である．Satisficing でない場合でも同一回答になることはあるが，Satisficing 状態ではその最大値が大きくなる可能性が考えられる [19].

「中間回答数」は，リッカート形式の質問において中間の「どちらでもない」のような選択肢を選択する回数である．これに関しても，Satisficing でない状態でも中間回答を選択する場合はある．一方で，Satisficing が発現しており，自分の意見を確認して表明するという認知的コストを払わず実質的に回答を放棄するような場合に，中間の選択肢を選択する傾向がある [20]．そのため，中間回答数は Satisficing に関連すると考えられる．

「文字数」は，自由記述形式の質問 1 問あたりの文字数とする．質問文で文字数や内容の具体度の指定がない場合，回答者は一文で回答する場合と，数文に渡って具体的に回答する場合がある．このような差が Satisficing に関連しており，文字数少ない方が Satisficing 傾向が強いのではないかと推測される．

4. 画面操作記録アンケートシステム

我々は，3 章で述べた画面操作を記録するシステムを開発した [10]．4.1 節では，システム構成の検討の流れを簡単に説明し，開発したシステムの構成について述べる．4.2 節では，本システムが画面操作を記録する方法について述べる．

4.1 システム構成

検討段階では，Web ページの改善のためにページ上でのユーザの行動をヒートマップ等で可視化する ClickTale [21] などのサービスについても検討した．しかしながら，汎用性が高い反面，アンケートに特化した画面操作データの記録はできない点が挙げられた．システムの普及性を考慮し，新たにアンケートシステムを構築せず，既存のシステムを拡張する方法を採用した．そこで，オープンソースの Web アンケートシステムである LimeSurvey に着目した．LimeSurvey は，Google Form [22] や Survey Monkey [23] 等とは異なり，JavaScript を用いて独自のプラグインを作成することが可能である．我々はこの仕組みを用いて，表 1 に示す画面操作データを記録するプラグイン（以降，Operation Logger と呼ぶ）を独自に開発し，画面操作が記録可能なアンケートシステムを構築した．本システムの概観を図 1 に示す．Operation Logger の導入方法は，LimeSurvey をホスティングするアプリケーションサーバ上に JavaScript と PHP のファイルを 3 つ配置し，サービス内の質問設定画面で JavaScript ファイルの読み込み設定をするのみである．これにより，標準機能で記録される回答結果データと共に画面操作データがデータベースに格納

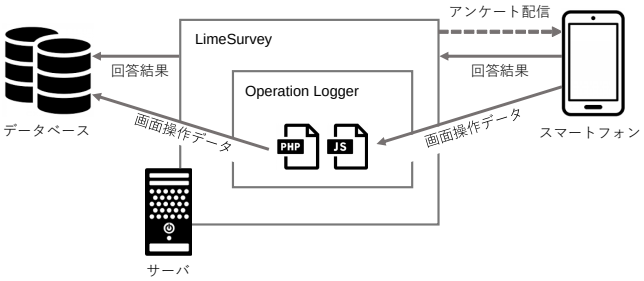

図 1 画面操作記録アンケートシステムの概観

される．なお，回答者側はソフトウェアの追加や設定の変更が一切必要ない．

4.2 画面操作の記録方法

Operation Logger においてデータを取得するトリガーとなるイベントを大別すると，タッチイベント，選択肢のタップ，テキストの入力の 3 種類である．

タッチイベントの種類は，touchstart，touchmove，touchend の 3 種類であり，それぞれスクリーンが指を検知した瞬間，指がスクリーン上で動いている間，指がスクリーンを離れた瞬間のタイミングで発火する．これらのタイミングにおける時刻，画面内の座標，ページ上端からの移動量，タッチイベントの種類を取得し，データベースに格納する．これにより，「スクロール長」，「スクロール時間」，「スクロール速度」，「逆スクロール回数」，「非操作時間が長すぎる回数」を検出することができる．

選択肢をタップしたタイミングでは，回答時刻，質問の ID，選択肢の ID を取得する．これにより，リッカート形式の「質問単位の回答時間」と「選択肢の変更」を検出することができる．

テキストを入力したタイミングでは，回答時刻，質問の ID，入力内容，レコード生成トリガの種類を取得し，データベースに格納する．これにより，自由記述形式の「自由記述の回答時間」，「テキストの削除回数」が検出できる．テキスト入力中の記録単位を検討した結果，1 レコードを生成するタイミングは「入力のない時間が 1 秒経過」，「デリートから入力への切り替わり」，「入力からデリートへの切り替わり」，「フォーカスアウト」とした．

5. クラウドソーシングを用いたアンケート実験

4 章で述べた Operation Logger を用いて，機械学習による不適切回答検出モデルを構築するためのデータ収集を目的としてアンケート実験を実施した．5.1 節では，実験の手続きについて述べる．5.2 節では，Satisficing 指標および，それに基づく不適切回答の定義について述べる．5.3 節では，Satisficing 指標の質問を含むアンケート内容について述べる．

5.1 実験手続き

本実験は，オンラインアンケートが実際に実施されているクラウドソーシングの環境として，Yahoo!クラウドソーシングを用いた．回答者の募集に際して，ワーカーをプラットフォーム上の指標（ブラックリストなど）でフィルタリングせず，全てのワーカーが 1 回のみ回答できるものとした．回答者には報酬として 5 円相当のポイントを付与した．

本稿で提案する手法はタッチスクリーンが搭載されていれば，スマートフォンに限らずタブレットやノート PC にも対応可能であると考えられる．ただし，画面サイズが一定以上異なると，質問文や選択肢の画面上の配置が大幅に変わり，画面操作の各特徴量の分布に大きな影響を及ぼす．本実験ではそういった回答環境による特徴量の分布のばらつきを抑制するために，回答に用いる端末をスマートフォンに限定することとした．なお，スマートフォンには多様な画面サイズの機種が存在するが，実用性の観点から特に機種の制限は設けなかった．スマートフォンに限定するために，クラウドソーシングサイト上のタスク説明欄とアンケートのスタート画面でスマートフォンで回答するように指示する文章を記載した．また，Operation Logger 側で JavaScript によって回答に用いた端末タイプを記録し，スマートフォン以外の回答は分析対象から除外した．アンケート回答後のページでは，画面操作データの使用に対する同意を問う質問を設け，同意しなかった場合，データは使用せずに破棄する旨と，その場合でも報酬は支払われる旨を記述した．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）．

5.2 不適切回答

本稿では，「不適切回答」を DQS（Directed Question Scale）と ARS（Attentive Responding Scale）の 2 つの指標に基づいて定義した．指標を 2 つ用いたのは，これらが異なる視点から Satisficing を検出するためである．

DQS [12] は，回答の指示文を質問と同列に設置し，その指示に従わなかった場合 Satisficing であるとする指標である．本稿では，3 問の DQS 質問のうち 1 問以上指示に反した回答をした場合に「Satisficing」とした．ARS [12] には，Inconsistency と Infrequency という 2 種類がある．Inconsystency は，内容が同じで文章を微妙に変更した質問対に対する回答の差分に注目するものである．11 の質問対に対する差分の合計が 11 以上であれば Satisficing であるとされる．Infrequency は，常識的に誰もが選択すると想定される選択肢が存在する質問を設け，その想定選択肢と実際に選択された選択肢の差分に注目するものである．11 問の差分の合計が 12 以上であれば Satisficing であるとされる．本稿では，Inconsistency と Infrequency のどちらか 1 つでも Satisficing であるサンプルを ARS 全体の「Satisficing」とした．

表 2　Satisficing 指標および正解ラベルの各クラスの該当者数と割合

Satisficing 指標	クラス	該当者数 [人]	割合 [%]
DQS	not Satisficing	4,520	91
	Satisficing	420	9
ARS	not Satisficing	4,066	82
	Satisficing	874	18
正解ラベル	適切	4693	95
	不適切	247	5

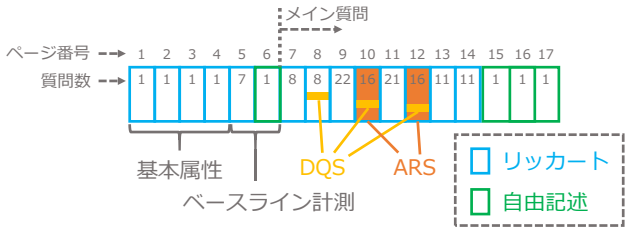

図 2　質問項目の概略

これら 2 つの指標を基に，本稿では正解ラベルである「不適切回答」を「DQS，ARS 両方の指標で Satisficing である回答者」と定義した．5692 人の回答者のうち，画面操作データの利用に同意した 4940 人について，各指標および正解ラベルの各クラスの該当者数を表 2 に示す．最終的に正解ラベルは，「不適切回答（正例）」が 247 件，「適切回答（負例）」が 4693 件であった．これを次章の機械学習による二値分類の正解ラベルとして用いる．

5.3 アンケート内容

アンケート実験で用いた質問票の概略を図 2 に示す．この質問票は三浦ら [14] が公開している質問票 [24] のうち，Big5 尺度，自尊感情尺度，認知欲求，アンケートへのモチベーションおよび DQS，ARS から成るリッカート形式部分をベースとし，次の 3 点を変更した．1 点目は，後述する 1~6 および 15~17 ページの追加である．2 点目は，ARS の質問対を回答者に悟られにくくするためにダミー質問を 11 問追加した点である．3 点目は，DQS の質問が 5 ページ連続してページ末尾に配置されていたため，DQS の質問箇所を悟られないために 3 問に減らし，ページ末尾や冒頭を避けて配置した点である．最終的に，全 17 ページ，128 問（5 段階リッカート形式 124 問，自由記述形式 4 問）で，回答目安時間が約 15 分の質問票とした．

次に，各ページの質問について説明する．1~3 ページは回答者の基本属性を尋ねる質問である．4 ページ目は回答者 ID をプラグインのシステムと共有するためのものである．5 ページ目は回答者ごとのスクロール操作のベースラインを計測するための質問である．例として，ベースライン計測ページとメイン質問ページの実際のアンケート

(a) ベースライン計測ページ　　(b) メイン質問ページ

図 3　実際のアンケート画面のスクリーンショット例

画面のスクリーンショットをそれぞれ図 3 (a)，(b) に示す．回答者は下方向にスクロールしつつ回答を進めていく形式となっている．ベースライン計測ページの質問内容は，図 3 (a) に示す「あまりあてはまらないを選んでください」といった，ある特定の選択肢を選択するように指示するものである．一般的な質問よりも認知的コストが低く，Satisficing が発現しにくい状態でのスクロール行動を計測する．これは，後述する特徴量であるスクロール速度の回答者内偏差を算出するために計測した．6 ページ目は回答者ごとに，自由記述形式質問におけるテキストの削除回数のベースラインを計測するための質問である．指定した文章を入力する際の削除回数を，各回答者のベースラインとした．これは後述する特徴量であるテキストの削除回数および削除率の回答者内偏差（delete_num_Selfdev, delete_rate_Selfdev）を算出するために計測した．指定する文章は削除が発生するような長さおよび内容にする必要があったため，次に示す内容とした．

「吾輩は猫である。名前はまだ無い。どこで生れたかとんと見当がつかぬ。何でも薄暗いじめじめしたところでニャーニャー泣いていた事だけは記憶している。吾輩はここで初めて人間というものを見た。」

7 ページ目以降がメインの質問であり，7~14 ページは主に先述の三浦らの質問票をベースとして作成した質問で構成されている．そのうち，8, 10, 12 ページ目に DQS の質問を，9, 11 ページ目に ARS の質問を配置した．15~17 ページ目では，簡単な自由記述形式の質問を設けた．なお，メインの内容である 7 ページ目以降は必須回答設定を OFF とした．また，自由記述形式の文字数も一つの特徴量であるため，文字数指定もなしとした．

6. 機械学習による不適切回答検出

5 章で述べたアンケート実験によって収集したデータを用いて，機械学習による不適切回答の検出を行なった．本稿では，「不適切回答の検出」という課題を「適切 or 不適切の二値分類」問題とした．6.1 節では，分類モデルのアルゴリズムおよび検証方法に関して述べる．6.2 節では，検出率向上を目指した特徴量の追加・選択の方法について述べる．6.3 節では，6.2 節の各過程における分類結果について報告する．また，質問数と外れ値に対する分類モデルの頑健性について検証する．

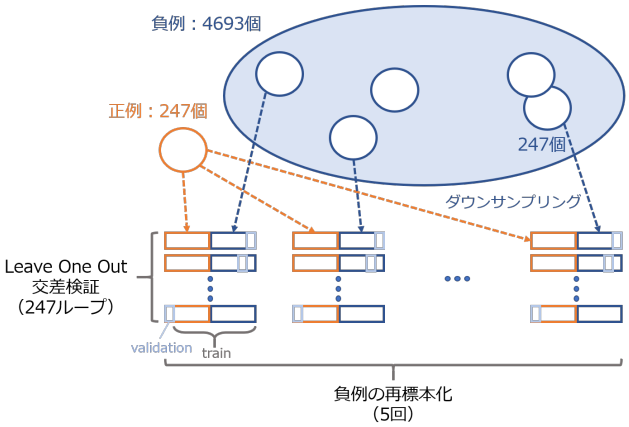

図 4　ランダムダウンサンプリングと Leave One Out 交差検証のデータの分割方法

6.1 機械学習モデル

尾崎ら [8] は数ある機械学習モデルの中で，ブースティングアルゴリズムが最も不適切回答の検出率が高かったと報告している．この知見に倣って本稿でも，決定木をベースとした勾配ブースティングアルゴリズムである Light-GBM [25] を用いた．ハイパーパラメータのチューニングには，ベイズ最適化アルゴリズムを用いた自動最適化ツールである Optuna [26] を用いた．モデルの精度評価には Accuracy，Precision，Recall，F1 Score を用いた．ただし，本稿の目的は不適切回答の検出であることから，その検出率を表す Recall に注目すべきであると考える．

モデルの汎化性能の検証は，対象サンプルを 1 つだけテスト用として交差検証する Leave One Out 交差検証によって行なった．本稿で扱うデータは正例と負例の比率が不均衡であったため，図 4 に示すように負例をランダムにダウンサンプリングし，正例：負例=1：1 として評価した．このとき，汎化性能評価の観点から，ダウンサンプリングは 5 回行ない，精度および特徴量重要度はその平均値とした．

6.2 特徴量の追加・選択

用いた特徴量の一覧を表 3 に示す．また，特徴量の追加・選択を行ない，3 段階でモデルを改良した際の各モデルで用いた特徴量を示す．「大小関係」欄では，各特徴量について，不適切回答群の平均値が適切回答群の平均値よりも大きいか小さいかを表す．まず，オリジナルモデルで

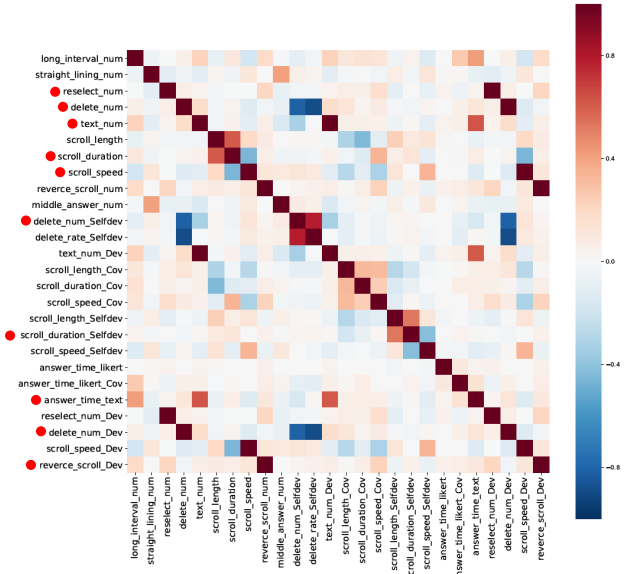

図 5 特徴量同士の相関（図中の丸印は特徴選択モデルで削除対象とした特徴量を表す）

は，表 1 に示した特徴量を基にした絶対的な特徴量のみを用いた．

次に，オリジナルモデルの各種特徴量に対する変動係数，回答者内および回答者間の偏差を特徴量に追加した（このモデルを「相対的特徴量追加モデル」と呼ぶ）．変動係数は，標準偏差を平均値で除した値である．各種画面操作データのばらつきを特徴量とする際に，回答者ごとに異なる平均値で標準化したばらつきとするために標準偏差ではなく変動係数を用いた．回答者内の偏差は，各回答者について，アンケート冒頭に設置したベースライン計測ページで計測したベースラインとアンケート全体の平均値との差である．回答を進めるにつれて不注意な回答が増加するという傾向 [27] を考慮し，アンケート冒頭との差が不適切回答の検出に寄与するのではないかと考えた．回答者間の偏差は，ある回答者の値と回答者全体の平均との差である．これは，回答者全体の平均的な回答行動との違いを表しているため，不適切回答の検出に寄与するのではないかと考えた．

さらに，図 5 に示す特徴量の相関行列のうち，相関係数が 0.5 以上のペアについて，片方を削除する特徴選択を行なった（このモデルを「特徴選択モデル」と呼ぶ）．このとき，図 6 に示す相対的特徴量追加モデルにおける特徴量の寄与率が低い方の特徴量を削除した．該当する特徴量を図 5，6 中に丸印でマークしている．ここで，削除された特徴量は相対的なものよりも絶対的なものが多いことが確認できる．すなわち，本稿で新たに考案した特徴量の中では，回答者内・回答者間偏差などの相対的な特徴量の方が分類に寄与する傾向があると言える．画面操作の種類としては，スクロール長やスクロール速度など，スクロールに関する特徴量の寄与率が高いことが確認された．また，テキストの削除については，削除回数よりも削除率の方が寄

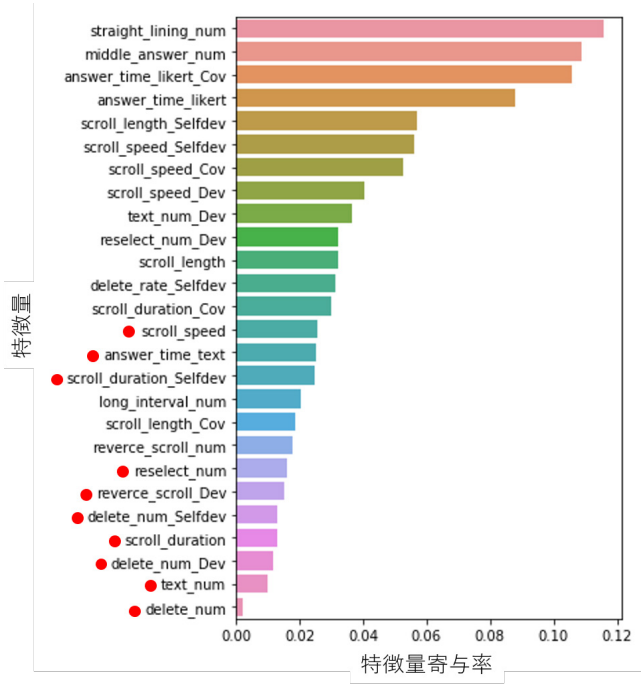

図 6 相対的特徴量追加モデルの特徴量の寄与率（図中の丸印は特徴選択モデルで削除対象とした特徴量を表す）

与率が高い結果となった．これは，削除回数では考慮できない，「文字数に応じて変化する削除の発生確率」を削除率では考慮できているためであると考えられる．

6.3 結果

3 つの不適切回答検出モデルの評価結果を表 4 に示す．特に注目すべき Recall を含め，全ての指標でモデルを改良するにつれて精度が向上していることが確認できる．最終的なモデルの Recall は 85.9%であった．また，本分類問題において Precision が低いということは，適切な回答を不適切であると誤って分類してしまう確率が高いということであるため，コストをかけて収集した適切なデータを無駄にしてしまうことに繋がる．また，偽陽性のサンプルを除くことは調査の内部妥当性を不用意に損ねることになる．これらの観点から，Precision も Recall とともに大切な指標であると言える．Precision はオリジナルモデルから特徴選択モデルにかけて 2.3%向上した．これらの結果から，スマートフォンにおける画面操作データを用いることで，尾崎ら [8] と大差のない不適切回答の定義において，その検出精度を飛躍的に向上（55.6%→85.9%）可能であることを確認した．

さらに，質問数と検出率の関係を検証するために，特徴選択モデルにおいて，特徴量生成の対象ページ数（ベースライン算出に用いたページ数は除く）を 3 ページ（平均 17 問），9 ページ（90 問）に制限した場合，検出率はそれぞれ 79.7%，80.9%であった．先述の全 17 ページ（128 問）を対象とした場合の 85.9%という結果も踏まえると，検証し

表 3　特徴量の説明と各モデルでの使用不使用

特徴量名	説明	単位	大小関係[*1]	オリジナルモデル	相対的特徴量追加モデル	特徴選択モデル
answer_time_likert	リッカートの回答時間の平均	s	小	○	○	○
answer_time_likert_Cov	リッカートの回答時間の変動係数	–	大	○	○	○
answer_time_text	自由記述の回答時間の平均	s	小	○	○	×
reselect_num	選択肢の変更回数の平均	回	小	○	○	×
reselect_num_Dev	reselect_num の回答者間偏差	回	小	×	○	○
delete_num	文字の削除回数の平均	回	小	○	○	×
delete_num_Dev	delete_num の回答者間偏差	回	小	×	○	×
delete_num_Selfdev	delete_num の回答者内偏差	回	大	×	○	×
delete_rate_Selfdev	文字数に対する文字の削除回数の回答者内偏差	回	大	×	○	○
scroll_length	スクロール長の平均	px	大	○	○	○
scroll_length_Cov	スクロール長の変動係数	–	小	○	○	○
scroll_length_Selfdev	scroll_length の回答者内偏差	px	大	×	○	○
scroll_duration	スクロール時間の平均	s	小	○	○	×
scroll_duration_Cov	スクロール時間の変動係数	–	大	○	○	○
scroll_duration_Selfdev	scroll_duration の回答者内偏差	s	小	×	○	×
scroll_speed	スクロール速度の平均	px/s	大	○	○	×
scroll_speed_Cov	スクロール速度の変動係数	–	小	○	○	○
scroll_speed_Dev	scroll_speed の回答者間偏差	px/s	大	×	○	○
scroll_speed_Selfdev	scroll_speed の回答者内偏差	px/s	大	×	○	○
reverce_scroll_num	逆向きスクロール回数の平均	回	大	○	○	○
reverce_scroll_num_Dev	reverce_scroll_num の回答者間偏差	回	大	×	○	×
long_interval_num	非操作時間が長すぎる回数	回	小	○	○	○
straight_lining_num	連続同一回答数の最大値	問	大	○	○	○
middle_answer_num	中間回答数	問	大	○	○	○
text_num	文字数の平均	文字	小	○	○	×
text_num_Dev	text_num の回答者間偏差	文字	小	×	○	○

[*1] 各特徴量の平均値について，不適切回答群が適切回答群よりも大きいか小さいかを表す．

表 4　不適切回答検出モデルの評価結果

評価指標	オリジナルモデル	相対的特徴量追加モデル	特徴選択モデル
Accuracy	0.844	0.850	0.862
Precision	0.841	0.848	0.864
Recall	**0.849**	**0.852**	**0.859**
F1 Score	0.845	0.850	0.862

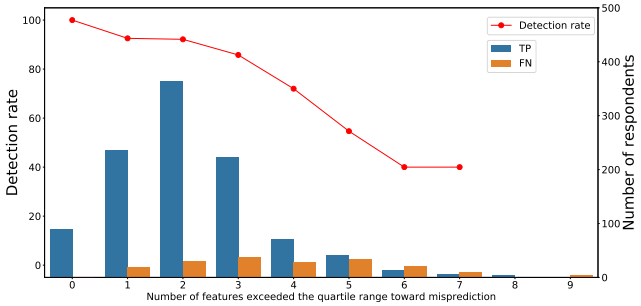

図 7　特徴選択モデルの外れ値の数と不適切回答の検出率

た範囲内においては，質問数が多いほどより高い精度で検出できる可能性が示唆された．一方で，17 問（3 ページ）程度の質問数でも約 80%の検出率を維持することが確認された．

また，本稿で用いた機械学習モデルは決定木をベースとしたモデルであるため，例えば，不適切回答では適切回答よりも少ないはずの自由記述の文字数が多いなど，ミスリーディングな方向の外れ値は検出率を低下させる原因となる．そこで，このような外れ値に対する特徴選択モデルの頑健性を検証した．まず，全データを正例（不適切回答群）と負例（適切回答群）に分け，それぞれの群で各特徴量の平均値と四分位範囲を算出した．その平均値を比較して導かれた適切回答群と不適切回答群の大小関係を表 3 の「大小関係」欄に示す．次に，ミスリーディングな方向，つまり表 3 に示す大小関係に対して逆向きに四分位範囲を超える特徴量（以後，「外れ値」と呼ぶ）の数をサンプルごとに算出し，真陽性と偽陰性のサンプル数との関係を図 7 にプロットした．横軸は外れ値の数を表す．棒グラフは真陽性と偽陰性それぞれのサンプル数，折れ線グラフは不適切回答の検出率を表す．折れ線グラフより，外れ値が多くなるに従って検出率が低下する傾向が確認できる．5 つ以上になると検出率は 50%以下になってしまう一方で，正例の約 80%を占める 3 つ以下の範囲で 86%以上の検出率を維持するモデルとなっており，本稿で用いた特徴量の外れ値に対して一定の頑健性を持つと考える．

7. まとめ

本稿では，回答時の画面操作を観測することで，オンラ

インアンケートの信頼性を毀損する Satisficing に基づく不適切回答の高精度な検出を目指した．この目標に対して，画面操作が記録可能なアンケートシステムが存在しなかったため，既存システムの実装を変更することなく画面操作が記録できる LimeSurvey 用のプラグイン（Operation Logger）を開発した．これを用いてクラウドソーシング上で 5692 人を対象とした大規模なアンケートを実施し，回答結果とともに回答時の画面操作データを収集した．そして，不適切回答検出に寄与すると見込まれる特徴量を検討し，収集した画面操作データから生成して不適切回答を検出する機械学習モデルを構築した．検出率の向上を目指し，回答者内および回答者間の偏差などの相対的な特徴量の追加や，特徴量同士の相関に基づく特徴選択を実施した．結果として，検出率 85.9%という結果が得られ，先行研究の精度を大幅に改善することに成功した．また，検出率は質問数が少ない場合も大幅に低下することはないが，多い方が高くなる可能性が示唆された．さらに，本稿で用いた特徴量の外れ値に対しても一定の頑健性を備えたモデルであることが確認された．新たに提案した特徴量に関しては，スクロールに関する特徴量の寄与率が高い結果となった．また，絶対的な特徴量よりも回答者内・回答者間の偏差などの相対的な特徴量の方が分類に寄与する可能性が高いことが確認された．

今後はさらに実用に向けて，例えば自由記述質問に文字数の制限がある場合や質問文が長い場合など，本稿のアンケートと異なる性質のアンケートに対する検出率を検証し，汎用的なモデルの実現を目指す必要がある．また，本稿で述べた検出手法をもとに，不適切回答を準リアルタイムに検出し，不適切な回答者に注意を促すよう介入するシステムの実現が期待される．そのシステムが実現できれば，介入のない従来システムとの比較による評価を行なう．さらに，アンケートページごとの各特徴量の推移なども特徴量に盛り込むなど，精度向上に寄与する新たな特徴量を模索する．

謝辞 本研究の一部は，科研費（18H03233）および，JST さきがけ（JPMJPR2039）の助成で行われた．また，大阪大学人間科学研究科の三浦麻子教授が公開している質問票を，本稿のアンケート実験のベース質問票として用いさせて頂いた．さらに，Satisficing 検出質問を含むアンケート設計についてご教示頂いた．ここに同氏に対して謝意を表する．

参考文献

[1] 総務省統計局：国勢調査のあゆみ，`https://www.stat.go.jp/data/kokusei/2015/kouhou/ayumi.html`.
[2] Simon, H. A.: Rational Choice and the Structure of the Environment, *Psychological Review*, Vol. 63, No. 2, pp. 129–138 (1956).
[3] Krosnick, J. A.: Response strategies for coping with the cognitive demands of attitude measures in surveys, *Applied Cognitive Psychology*, Vol. 5, No. 3, pp. 213–236 (1991).
[4] Berinsky, A. J., Margolis, M. F. and Sances, M. W.: Can we turn shirkers into workers?, *Journal of Experimental Social Psychology*, Vol. 66, pp. 20–28 (online), DOI: https://doi.org/10.1016/j.jesp.2015.09.010 (2016). Rigorous and Replicable Methods in Social Psychology.
[5] Steger, D., Schroeders, U. and Gnambs, T.: A Meta-Analysis of Test Scores in Proctored and Unproctored Ability Assessments, *European Journal of Psychological Assessment*, Vol. 36, pp. 174–184 (2020).
[6] Oppenheimer, D. M., Meyvis, T. and Davidenko, N.: Instructional manipulation checks: Detecting satisficing to increase statistical power, *Journal of Experimental Social Psychology*, Vol. 45, No. 4, pp. 867 – 872 (online), DOI: https://doi.org/10.1016/j.jesp.2009.03.009 (2009).
[7] Pei, W., Mayer, A., Tu, K. and Yue, C.: Attention Please: Your Attention Check Questions in Survey Studies Can Be Automatically Answered (2020).
[8] 尾崎幸謙，鈴木貴士：機械学習による不適切回答者の予測，行動計量学， Vol. 46, No. 2, pp. 39–52 (2019).
[9] Lugtig, P. and V., T.: The Use of PCs, Smartphones, and Tablets in a Probability-Based Panel Survey: Effects on Survey Measurement Error, *Social Science Computer Review*, Vol. 34, pp. 78–94 (2016).
[10] 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の回答信頼性検証に向けた回答時画面操作ログ取得システム，情報処理学会研究報告，Vol. 2020-HCI-186, No. 35, pp. 1–7 (2020).
[11] 三浦麻子，小林哲郎：オンライン調査モニタの Satisfice に関する実験的研究，社会心理学研究，Vol. 31, No. 1, pp. 1–12 (2015).
[12] Maniaci, M. R. and Rogge, R. D.: Caring about carelessness: Participant inattention and its effects on research, *Journal of Research in Personality*, Vol. 48, pp. 61–83 (2014).
[13] 増田真也，坂上貴之，森井真広：調査回答の質の向上のための方法の比較，心理学研究，Vol. 90, No. 5, pp. 463–472 (2019).
[14] 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（Satisfice）を検出する技法：大学生サンプルを用いた検討，社会心理学研究，Vol. advpub (2016).
[15] NTT コムオンライン・マーケティング・リサーチ株式会社：回答結果の品質：回答結果の品質向上のための取り組み，`http://research.nttcoms.com/service/qpolicy4.html`.
[16] 深井裕二, 河合洋明：Moodle アンケートに対応した Satisfice 回答の適応的除去システムの開発，工学教育，Vol. 65, No. 3, pp. 60–65 (2017).
[17] Tourangeau, R., Sun, H., Yan, T., Maitland, A., Rivero, G. and Williams, D.: Web Surveys by Smartphones and Tablets: Effects on Data Quality, *Social Science Computer Review*, Vol. 36, No. 5, pp. 542–556 (2018).
[18] Gould, S. J. J., Cox, A. L. and Brumby, D. P.: Diminished Control in Crowdsourcing: An Investigation of Crowdworker Multitasking Behavior, *ACM Transactions on Computer-Human Interaction*, Vol. 23, No. 3 (online), DOI: 10.1145/2928269 (2016).
[19] Kim, Y., Dykema, J., Stevenson, J., Black, P. and Moberg, D. P.: Straightlining: Overview of Measurement, Comparison of Indicators, and Effects in Mail–Web Mixed-Mode Surveys, *Social Science Computer Review*, Vol. 37, No. 2, pp. 214–233 (online), DOI:

10.1177/0894439317752406 (2019).
[20] Baumgartner, H. and Steenkamp, J.-B. E.: Response Styles in Marketing Research: A Cross-National Investigation, *Journal of Marketing Research*, Vol. 38, No. 2, pp. 143–156 (online), DOI: 10.1509/jmkr.38.2.143.18840 (2001).
[21] CONTENTSQUARE: Clicktale, `https://www.ctale.jp/`.
[22] Google, LLC: Google Forms, `https://www.google.com/forms/about/`.
[23] SurveyMonkey: SurveyMonkey, `https://www.surveymonkey.com/`.
[24] 三浦麻子，小林哲郎：Supplemental materials for "Satisficing" studies by Miura, A. and Kobayashi, T., `https://osf.io/6gu3q/`.
[25] 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, *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, Red Hook, NY, USA, Curran Associates Inc., p. 3149–3157 (2017).
[26] Akiba, T., Sano, S., Yanase, T., Ohta, T. and Koyama, M.: Optuna: A Next-Generation Hyperparameter Optimization Framework, *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, New York, NY, USA, Association for Computing Machinery, pp. 2623—-2631 (2019).
[27] Bowling, N. A., Gibson, A. M., Houpt, J. W. and Brower, C. K.: Will the Questions Ever End? Person-Level Increases in Careless Responding During Questionnaire Completion, *Organizational Research Methods*, p. 1094428120947794 (online), DOI: 10.1177/1094428120947794 (2020).