機械学習による不適切なクラウドソーシングタスクの検出

馬場　雪乃† 　鹿島　久嗣† 　木下　慶†† 　山口　豪志†† 　秋好　陽介††

† 東京大学大学院情報理工学系研究科　〒 113–0033 東京都文京区本郷 7–3–1
†† ランサーズ株式会社　〒 248–0006 神奈川県鎌倉市小町 2-7-32
E-mail: †{yukino_baba,kashima}@mist.i.u–tokyo.ac.jp,
††{kinoshita.kei,yamaguchi.goushi,akiyoshi.yosuke}@lancers.co.jp

あらまし　クラウドソーシングサービスでは，利用規約に違反する不適切なタスクが投稿されることがある．サービス運営会社は不適切なタスクを見つけ次第，当該タスクの依頼を停止している．しかし，投稿されるタスクが大量になるにつれ，運営会社による全てのタスクの常時監視は人的・時間的コストの点で困難になると予想される．本稿では，運営会社によるタスク監視を支援することを目的とし，機械学習手法による不適切タスク検出の実験結果を報告する．タスクや依頼者の情報を用いて構築した分類器が高い検出精度を示すことを，実際のクラウドソーシング運営会社のデータを用いた実験により明らかにした．また，クラウドソーシングワーカーにも監視作業を依頼し，運営会社による監視とワーカーによる監視を組み合わせることで検出精度が向上することを示した．
キーワード　クラウドソーシング, ヒューマンコンピュテーション, スパム検出, 機械学習

Yukino BABA†, Hisashi KASHIMA†, Kei KINOSHITA††, Goushi YAMAGUCHI††, and Yosuke AKIYOSHI††

† Graduate School of Information Science and Technology, The University of Tokyo
7–3–1 Hongo, Bunkyo-ku, Tokyo, 113–0033 Japan
†† Lancers Inc.
2–7–32 Komachi, Kamakura-shi, Kanagawa, 248–0006 Japan
E-mail: †{yukino_baba,kashima}@mist.i.u–tokyo.ac.jp,
††{kinoshita.kei,yamaguchi.goushi,akiyoshi.yosuke}@lancers.co.jp

1. はじめに

Amazon Mechanical Turk(注1)に代表されるクラウドソーシングサービスがビジネスや研究で広く用いられるようになってきた．クラウドソーシングサービスは，インターネットを通じて不特定多数の人々に仕事を依頼する仕組みを提供する．この仕組みを利用すると，大量の人々に対する作業発注が容易に実現できる．クラウドソーシングで依頼される仕事の種類は，画像や文章へのタグづけや，文書作成，翻訳，グラフィックデザインなど多岐に渡る．

クラウドソーシングでは，作業者（ワーカー）が作成した成果物の品質管理が重要な課題のひとつである．素性がわからない相手に仕事を依頼するクラウドソーシングでは，事前にワーカーの能力や信頼性を知ることが難しい．ワーカーによって能力にばらつきがあり，さらには短時間で報酬を獲得するためにいい加減に作業を行うワーカーすら存在するため，品質の高い成果物を得るための工夫が必要となる．たとえばAmazon Mechanical Turk では，ワーカーに事前テストを受けさせて成績によってフィルタリングする機構を導入している．また，CrowdFlower(注2)は，正解があらかじめわかっているGold standard data と呼ばれる問題を紛れ込ませ作業させながらワーカーの能力を測る仕組みを提供している．同じ作業を複数のワーカーに依頼し，機械学習手法を利用して作業結果からワーカーの能力を推定する手法も提案されている [1], [10], [11].

一方，クラウドソーシングにおいては依頼される作業（タスク）の品質管理も重要な課題である．Amazon Mechanical Turk や CrowdFlower などのクラウドソーシングサービスでは作業依頼自体もインターネットを通じて行うことができ，依頼

(注1)：https://www.mturk.com/

(注2)：http://crowdflower.com

者は多くの場合匿名である．クラウドソーシングサービス運営会社は，サービスを健全な環境に保つために反社会的・非倫理的なタスクの依頼を防ごうとしている．たとえば，10 万人以上のワーカーを抱える国内最大規模のクラウドソーシングサービスであるランサーズ(注3)では，利用規約の中で「依頼内容において，提案時にユーザ自身の詳細な個人情報の記載を要求する行為」「成果報酬を得ることを目的とする依頼（アフィリエイト，メルマガ登録等）を行う行為」などを禁止している(注4)．サービス運営側は利用規約に違反する「不適切なタスク」を発見次第，ワーカーが作業しないように当該タスクの依頼を停止している．しかし将来，投稿されるタスクが大量になると(注5)，運営側が全てのタスクを常時監視することは人的・時間的コストの面で困難になると予想される．クラウドソーシング運営会社にとってタスクの品質管理は重要な課題であるが既存研究では対象とされてこなかった．

本稿では，ランサーズに実際に投稿されたタスクのデータを利用した，機械学習手法による不適切タスク検出の実験結果について報告する．我々は，機械学習手法の利用により運営側の不適切タスク監視作業を支援することを目的としている．運営側が確認しなければならないタスクの数を削減するために，以下のような機械学習の利用手順を提案する．(1) あらかじめ，いくつかのタスクについてのみ運営側で不適切か否かの判定をおこない，それを訓練データとしてタスク分類器を構築する．(2) 新しいタスクが投稿されると分類器によって不適切か否かの推定を行い，不適切とされたタスクのみを運営側に提示する．(3) 運営側は，提示されたタスクについてのみ確認作業を行い，不適切である場合にはそのタスクの依頼を停止する．

さらに我々は，分類器の精度を向上するために，クラウドソーシング上のワーカーにタスクの監視作業を行わせ，分類器の訓練時にワーカーによる監視結果を組み込む手法を提案する．決められた利用規約と照らしあわせて，あるタスクが不適切かどうかを正しく判定できる運営側（エキスパート）と異なり，ワーカーは能力にばらつきがあり常に正しく判定できるとは限らない．我々は，ワーカーの品質管理手法を導入することでこの問題を解決した．

本研究では，ランサーズに実際に投稿されたタスクに対する運営側の不適切タスク登録結果を用いて，エキスパートによる監視結果（エキスパートラベル）のデータを構築した．エキスパートラベルとタスク及びその依頼者の情報を訓練データとして利用し，平均 Area Under Curve (AUC) 値 0.950 を達成する精度の高い分類器を構築できることを確認した．さらに，実際にクラウドソーシングワーカーにタスクの監視作業を依頼しワーカーによる監視結果（ワーカーラベル）を獲得した．ワーカーラベルをエキスパートラベルと統合して訓練時の正解ラベルにすることで，分類器の精度が向上（平均 AUC 0.962）することを明らかにした．また，エキスパートラベルとワーカーラベルを統合することでエキスパートが監視するタスクの数を 25% 程度削減しても，エキスパートが全て監視した場合と同等の分類器精度を達成できることが確認できた．

本研究の貢献は以下となる．

- 本研究は，実際のクラウドソーシングサービス運営会社が持つ内部データを活用し，また，クラウドソーシングにおけるタスクの品質管理問題に取り組んだ初めての研究である．
- タスクの品質管理問題に機械学習手法を適用し，不適切タスク検出に機械学習が有効であることを，実際のクラウドソーシングサービスのデータを用いて示した（3 章）．
- ワーカーによる監視結果をエキスパートによる監視結果と組み合わせることで，機械学習による不適切タスク検出の精度が向上することを示した（4，5 章）．

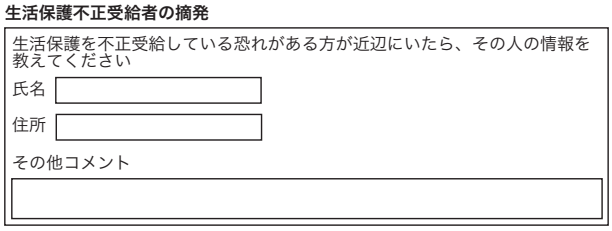

図 1　サービス運営会社によって不適切と判定されるタスクの例（他者の個人情報の入力依頼 ）

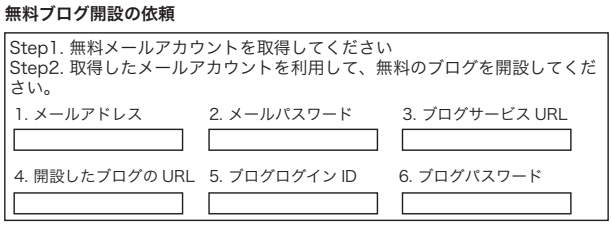

図 2　サービス運営会社によって不適切と判定されるタスクの例（ブログ開設依頼）

2. 不適切なクラウドソーシングタスクの検出

2.1 問題設定

我々の目的は，エキスパートラベルとワーカーラベルを用いて，タスクが不適切か適切か推定する分類器を作ることである．この問題を以下のように定式化する．

N 個の訓練用タスクがあり，それぞれ D 次元の特徴ベクトル $\boldsymbol{x}_i \in \mathbb{R}^D$ で表されている．訓練用タスク集合を $\mathcal{X} = \{\boldsymbol{x}_i\}_{i\in\{1,2,\cdots,N\}}$ とする．各タスク i について，クラウドソーシング運営会社は定められた基準に従って不適切なタスクか否かを判定する．エキスパートが与えるラベルを $y_{i,0} \in \{0,1\}$ とする．ここで，1 が不適切タスク，0 が適切タスクを示すものとする．エキスパートラベルの集合を $\mathcal{Y}_0 = \{y_{i,0}\}_{i\in\{1,2,\cdots,N\}}$ とする．ワーカーの数は J 人で，タスク i の判定を行ったワーカー集合を $\mathcal{J}_i \subseteqq \{1,2,\cdots,J\}$ で表す．ここで，各ワーカーが全てのタスクの判定をしているとは限らない．ワーカーは，各タスクについて K 個の設問（今回の例では，「アフィエイトの

(注3)：http://www.lancers.jp
(注4)：http://www.lancers.jp/help/terms（2012 年 12 月 20 日時点）
(注5)：たとえばランサーズでは，2012 年 11 月時点でも月間約 6,000 件の依頼が投稿されている

恐れがある」「個人情報の記載を要求している」などの 4 つの設問）に二値で回答している．ワーカー j がタスク i に与えたラベルを $\boldsymbol{y}_{i,j} \in \{0,1\}^K$ とする．ワーカーラベルの集合を $\mathcal{Y} = \{\boldsymbol{y}_{i,j}\}_{i\in\{1,2,\cdots,N\},j\in\mathcal{J}_i}$ とする．我々の目的は，訓練データ $(\mathcal{X}, \mathcal{Y}, \mathcal{Y}_0)$ を用いて，二値分類器 $f : \mathbb{R}^D \to \{0,1\}$ を構築することである．

2.2 データセット

タスクの情報と，エキスパートラベル及びワーカーラベルを利用した不適切タスク分類器を構築するために以下のデータセットを用意した．

2.2.1 タスクデータ及びエキスパートラベルデータ

2012 年 6 月から 11 月の間にランサーズに投稿されたタスク方式の依頼の中から，2012 年 12 月時点において (1) 依頼が削除されておらず，(2) 非公開に設定(注6)されていない，(3) 作業プレビュー(注7)が公開されているタスクをまず選んだ．うち，96 件がエキスパートによって不適切と判定されている．96 件の不適切タスクとランダムに選んだ 2,904 件の適切タスクから成る計 3,000 件のデータセットを構築した．データセットには，各タスクの情報（タイトル，概要文，作業画面の HTML，依頼者，単価など）と依頼者の情報（年齢，性別，職業，これまでの実績など）が含まれている．

2.2.2 ワーカーラベルデータ

クラウドソーシングワーカーによるタスク監視作業を実際にランサーズ上で依頼した．3,000 件のタスクそれぞれについて，不適切タスクの判定を 2 人または 3 人のワーカーに依頼した．ワーカーには，一回の作業で 15 件のタスクをまとめて判定させた．表 1 に収集したワーカーラベルデータの詳細を示す．

図 3 にワーカーの監視作業画面を示す．各ワーカーには，タスクが「アフィリエイトの恐れがある」「個人情報の記載を要求している」「直接の連絡手段が掲載されている」「ステルスマーケティングの恐れがある」のそれぞれについて，該当するか否かを回答させた．これらの設問は，ランサーズ上で提供されている違反タスク申告フォームと同じものである．結果，各タスクに対して「アフィリエイト」「個人情報」「直接取引」「ステルスマーケティング」の 4 個の二値ラベルが監視ワーカーの数（この場合は約 3 人）の分だけ付与される．

3. エキスパートラベル利用による分類器の構築

まずは機械学習による不適切タスク検出が有効であることを確認するために，エキスパートラベルを用いて分類器を作成した．本節ではまず，分類器の構築に利用したタスクと依頼者の特徴量を紹介する．次に，実際のクラウドソーシングデータを用いた分類器の評価実験の結果を示し，分類器の精度と有効な特徴量について述べる．

(注6)：非公開設定の場合，ランサーズにユーザ登録していない人は依頼を閲覧できない

(注7)：ランサーズでは，実際にワーカーが作業を行う際に見る画面（作業の説明文や入力フォームなどが含まれる）のプレビューを公開するかどうか依頼者側が選択できる．プレビューが非公開の場合，依頼を受けたワーカーだけが作業画面を閲覧できる

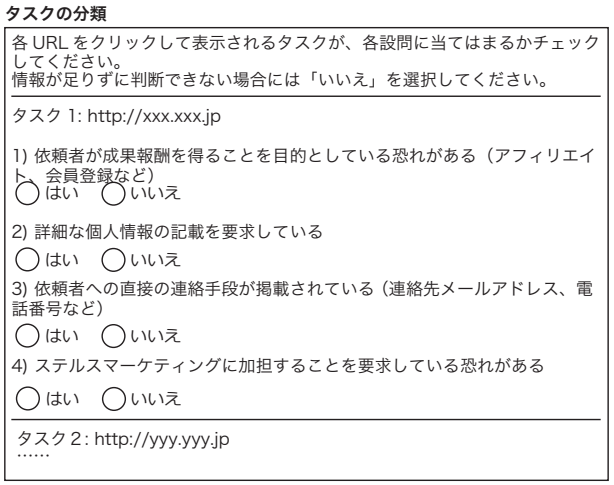

図 3 ワーカーに依頼した不適切タスク監視の作業画面

3.1 特徴量設計

ランサーズにおける実際のタスクと依頼者の情報から，以下 4 種類の特徴量を構築した．

- タスクのテキスト特徴量

各タスクが持つテキスト情報（タイトル，概要文，作業画面中の文章）を bag-of-words で表現した．対象とする単語は，(1) 二つ以上のタスクで使われている，(2) 記号・数字以外の単語とした．各単語特徴の値は単語が出現しているか否か二値で表現した．単語分割及び品詞推定には MeCab(注8)を使用した．結果，6,975 次元の二値ベクトルの特徴量となった．

- タスクの非テキスト特徴量

文章以外のタスクの情報（単価や作業可能ワーカーの条件など）からタスクの非テキスト特徴量を作成した．表 2 に詳細を示す．

- 依頼者 ID 特徴量

「誰がタスクを依頼したのか？」という情報も不適切タスクを検出する上で重要だと考えられるため依頼者 ID 特徴量を作成した．結果，417 次元の二値ベクトルの特徴量となった．

- 依頼者の非テキスト特徴量

「どのような人がタスクを依頼したのか？」という情報を捉えるため依頼者の非テキスト特徴量を作成した．この特徴量は，依頼者の属性（性別，生まれ年など），信頼性（本人確認の状況など），これまでの実績等から成る．詳細を表 3 に示す．

表 2 タスクの非テキスト情報特徴量の詳細

特徴量	タイプ	次元数
タスク内の作業数	整数	1
同じ作業を依頼する人数	整数	1
ワーカーあたりの最大許可作業数	整数	1
作業単価	整数	1
作業可能なワーカーの条件	二値	4
タスク公開オプション	二値	3
タスクの状態	二値	5
作業結果ダウンロード状況	整数	2

(注8)：http://mecab.sourceforge.net/

表 1　ワーカーラベルデータの詳細

対象タスク数	うち，不適切タスク数	のべ判定数	平均判定数／タスク	総作業ワーカー数	平均判定タスク数／ワーカー	1 回の作業での判定タスク数	総報酬額（円）
3000	96	8990	2.997	97	92.68	15	8598

表 3　依頼者の非テキスト特徴量の詳細

特徴量	タイプ	次元数
生まれ年	整数	1
性別	二値	1
居住国		71
居住都道府県	二値	48
法人個人の区分	二値	1
本人確認書類提出済みか	二値	1
機密保持確認済みか	二値	1
ランサーズチェック済みか	二値	1
メールアドレス確認状況	二値	2
利用用途（ワーカーか依頼者か）	二値	1
状態	二値	3
招待設定	二値	2
最低招待金額	二値	2
職業	二値	19
得意なカテゴリ	二値	6
作業承認率	整数	1
平均評価値	実数	1
合計評価値	整数	1
ランク	整数	1
当選回数	整数	1
メール受信設定	二値	7

3.2　実験結果

前節で紹介した特徴量を用いて分類器を構築した．二値の特徴量は $\{0,1\}$ で表し，整数あるいは実数の特徴量は $[0,1]$ で正規化した．構築の際にはデータセット中のランダムに選んだ 60%（1,800 件）のタスクを訓練データとして利用し，残りをテストデータとした．分類器として，線形カーネル SVM の実装 liblinear(注9) を利用した．分類器の精度評価指標として 100 回の試行での AUC の平均値と標準偏差を用いた．AUC は，ランダムに選んだ正例と負例について分類器が，負例よりも正例を「正例らしい」と推定する確率を表している．今回の実験では，不適切なタスクを正例としている．

表 4 に各特徴量を組み合わせて構築した分類器の，AUC の平均値と標準偏差を示す．単独で用いた場合に最も AUC の平均値が高くなるのはタスクのテキスト特徴量であり (0.902)，次いで依頼者 ID 特徴量 (0.848)，依頼者の非テキスト特徴量 (0.771)，タスクの非テキスト特徴量 (0.734) となった．このことからまず，タスクが持つテキスト情報が不適切タスク検出に最も有効であることがわかる．タスクのテキスト特徴量とそれ以外の特徴量をそれぞれ組み合わせた場合，AUC の平均値は高い方からタスクの非テキスト特徴量 (0.946)，依頼者の非テキスト特徴量 (0.910)，依頼者 ID (0.904) であった．タスクの非テキスト特徴量は単独で用いた場合には他の特徴量よりも低い AUC 値であったが，タスクのテキスト特徴量で捉えきれなかった傾向を補完していると言える．最も高い AUC 値を示したのは，全ての特徴量を用いた場合，あるいは依頼者 ID 特徴量以外を組み合わせた場合 (0.950) であった．依頼者 ID 特徴量は，単独で用いた場合にはタスクのテキスト特徴量に次ぐ平均 AUC 値だったが，他の特徴量と組み合わせた場合には精度向上に寄与しなかった．

表 5，6 に不適切なタスクあるいは適切なタスクと判断するのに有用な単語の代表例を示す．「不適切タスクらしい」単語の中には，「アカウント」「パスワード」のような外部サイトへの登録依頼に含まれることが多い単語や「メールアドレス」のような個人情報の入力を示唆する単語，図 2 のようなブログ開設依頼を表す単語が見られた．適切なタスクと判断するのに有用な単語には，たとえば「文字」「以上」「記事」といった記事執筆タスクに含まれる単語があった．また，タスクの非テキスト特徴量からは，相場から逸脱した高額単価のタスクが不適切タスクと推定される傾向や，「本人確認済み」「作業承認率 95% 以上」などの作業者制限を設けて信頼できるワーカーにのみ作業を依頼するタスクが適切タスクと推定される傾向が確認された．また，依頼者の非テキスト特徴量から，ランサーズにおいて他のワーカーからの評価が高い依頼者が投稿したタスクは適切タスクと推定される傾向が見られた．

図 4，5，6，7 に，正しく推定されたタスクの例及び誤って推定されたタスクの例を示す．図 5 のような他サイトのアカウントを要求するタスクは正しく不適切と推定されていた．図 6 のタスクは，外部の飲食店口コミサービスへの投稿を依頼している不適切なタスクである．しかし，「記事」「オススメ」「文字」「以上」といった単語に分類器は負の重みを与えているため，適切タスクと誤推定してしまった．逆に図 7 は，ブログ開設時の挨拶文執筆を依頼する適切タスクであるが，図 2 のような「ブログ開設依頼」タスクに含まれることが多い「ブログ」「開設」といった単語があるため，不適切タスクだと誤推定されてしまった．

以上，分類器が誤推定する例はいくつかあるものの，3.1 節で示した全ての特徴量を用いることで平均 AUC 0.950 という高い分類精度を達成し，機械学習が不適切タスク検出に有効であることが確認できた．

4.　ワーカーラベル利用による分類器の構築

本節では，エキスパートラベルの代わりにワーカーラベルを訓練時の正解として分類器を構築し，その分類精度をエキスパートラベルを用いた場合と比較する．2.1 章で示した通り，

(注9)：http://www.csie.ntu.edu.tw/~cjlin/liblinear/

表 4　エキスパートラベルを用いて構築した分類器の，訓練に用いる特徴量を変化させたときの AUC の平均値と標準偏差の比較

使用した特徴量				AUC
タスク特徴量		依頼者特徴量		
テキスト	非テキスト	ID	非テキスト	
			✓	0.771 (±0.040)
		✓		0.848 (±0.032)
		✓	✓	0.840 (±0.036)
	✓			0.734 (±0.042)
	✓		✓	0.835 (±0.036)
	✓	✓		0.890 (±0.030)
	✓	✓	✓	0.854 (±0.040)
✓				0.902 (±0.032)
✓			✓	0.910 (±0.030)
✓		✓		0.904 (±0.031)
✓		✓	✓	0.911 (±0.030)
✓	✓			0.946 (±0.016)
✓	✓		✓	**0.950** (±0.015)
✓	✓	✓		0.947 (±0.015)
✓	✓	✓	✓	**0.950** (±0.015)

表 5　不適切タスクだと推定するのに有用な単語の例

アカウント	パスワード	メールアドレス	開設	blog

表 6　適切タスクだと推定するのに有用な単語の例

文字	以上	記事	依頼	まとめ

写真の説明文を記述してください

10 個の写真が提示されます。それぞれについて、50 文字以上で日本語の説明文を記述してください。

写真 1：

説明文 1：

写真 2:
……

図 4　正しく「適切」だと推定されたタスクの例

ワーカーラベルはエキスパートラベルと異なり (1) 複数設問への回答から成る K 次元ベクトル（今回の例では $K = 4$）で表現され，(2) 一つのタスクについて複数ワーカーによるラベルが与えられている．そのため，エキスパートラベルと同様に扱うためには工夫が必要となる．特に，エキスパートと異なりワーカーは能力にばらつきがあるため，能力を考慮した取り扱いが有効だと考えられる．本節ではまず複数設問への回答の統合方法について延べ，次に複数ワーカーのラベルの統合方法について述べる．統合したラベルを正解として分類器を構築し，エキスパートラベルを用いて構築した分類器と分類精度を比較する．

追加募集！【１分作業５０円】確認作業＆いいね！のクリック

【1】Google検索もしくはYahoo検索にて「オリジナルプレゼント　通販」のキーワードで検索をしてください。
【2】サイト名（ショップ名）を確認
【3】ページ下部にあるFacebook「いいね！」をクリックしてください。

下記に【2】でご確認されましたサイト名をご記入いただき、ご自身のFacebookのアカウントURLも合わせてご入力ください。

※本タスクのために新しくFacebookアカウントをつくり、作業されることは禁止とさせていただきます。現行でアカウントをお持ちの方のみ作業をお願い致します。

※こちらは任意で結構ですが、ブックマーク（ヤフーブックマークやはてなブックマーク）へのご登録もしていただけましたら幸いです。

図 5　正しく「不適切」だと推定されたタスクの例

１０文字以上のお店情報のクチコミをお願いします

専門店、雑貨店、飲食店、ホテルなどのお店情報募集
日本国内の企業・店舗に関するクチコミを募集しています。
掲載されているお店はすべて対象となります。

【仕事方法】
＜目的のお店を探す＞
・http://www.xxx.jp/
上記サイトのヘッダ部分にある検索BOXから利用したことのあるお店名＋都道府県などで目的のお店を検索してください。
・お店情報の下段にある「あなたもオススメメッセージを書きませんか？」の欄に直接ご入力下さい。

＜入力内容＞
・口コミを10文字以上で書いてください。（タイトルは5文字程度でも可です）
・記事の内容は店舗の特徴・オススメ・雰囲気など、オススメのポイントを自由に書いていただけます。
　※できるだけ具体的な内容でお願い致します。
「おいしくておすすめです。」だけの口コミなどはカウントされない場合もございますのでご注意ください。
（クレームも不可となっております。おすすめ店舗のみお願いします。）
…

図 6　誤って「適切」だと推定されたタスクの例（他のサービスへの口コミ投稿を依頼しているため，正解は「不適切」）

【簡単】ブログを開設したときの文言を考えてください（70〜100字程度）1030-1

同じ文言を組み替えて投稿しないでください。
ブログ開設時の挨拶文を創作してください！

ブログを新しく開設してはじめての投稿のとき、どんな言葉を書きますか？
70〜100字程度の簡単な文言でかまいません。

どんなブログにも対応できる汎用性のある文言でおねがいします。
…

図 7　誤って「不適切」だと推定されたタスクの例（正解は「適切」）

4.1　複数設問に対する回答統合

今，タスク i に対するワーカー j のラベルは $\boldsymbol{y}_{i,j} \in \{0,1\}^K$ で表されている．本節の目的は，K 個の設問に対する回答から成る K 次元の二値ベクトル $\boldsymbol{y}_{i,j}$ を統合した $y'_{i,j} \in \{0,1\}$ を得ることである．ここで，タスク i に対するワーカー j の各設問「アフィリエイト」「個人情報」「直接取引」「ステルスマーケティング」への回答をそれぞれ $y^{(a)}_{i,j}, y^{(p)}_{i,j}, y^{(d)}_{i,j}, y^{(s)}_{i,j}$ とする．単純には，いずれかの設問に「はい」と答えている場合は，ワーカーが当該タスクは不適切だと判定したものとみなす，つまり $y'_{i,j} = y^{(a)}_{i,j} \vee y^{(p)}_{i,j} \vee y^{(d)}_{i,j} \vee y^{(s)}_{i,j}$ とする方法が考えられる．しかし，設問ごとにタスク分類に寄与する度合いは異なり，設問の中にはその回答が不適切タスク検出に大きく寄与するものと寄与が小さいものがあると考えられる．そこで，どの設問がタスク分類に貢献するのかを調べるために，回答の統合方法を変えた場合の分類器の精度を比較する．

4.2　複数ワーカーのラベル統合

前節のように複数設問に対する回答を統合すると，タスク i に対するワーカー j のラベルが $y'_{i,j}$ で与えられる．次に，エ

キスパートラベルと同様に扱うために，タスク i に対する複数ワーカーのラベル $\{y'_{i,j}\}_{j\in\mathcal{J}_i}$ を統合することを考えたい．単純な統合方法として多数決が考えられるが，多数決では各ワーカーの能力は等しいと考えワーカーごとの重みを考慮しない．しかし実際は，ワーカーの能力にはばらつきが存在する．図 8 にワーカーの判定性能の分布を示す．適合率と再現率の両方において，1.0 に近い能力の高いワーカーもいれば，0.0 に近いワーカーもいるなど，能力にはたしかにばらつきが存在している．ワーカーの能力等を考慮した上で各タスクに対する真のラベル $\{y'_i\}_{i\in\{1,2,\cdots,N\}}$ を統計的に推定する手法がいくつか提案されている [1], [10], [11]．本稿では，多数決及び統計的な統合手法と，統合を行わない手法の三つを比較する．

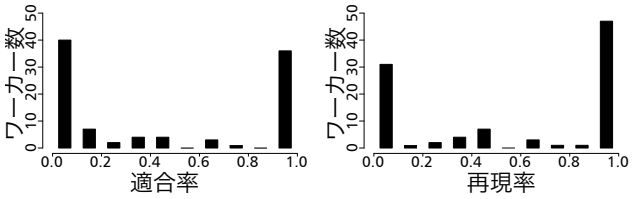

図 8　ワーカーの判定性能の分布（ワーカーが「アフィリエイト」「個人情報」のいずれかに「はい」と回答した場合，ワーカーによる判定は「不適切タスク」だとした）

- 多数決による統合

この手法では，$\{y'_{i,j}\}_{j\in\mathcal{J}_i}$ の多数決を取って統合ラベル y'_i とする．今回は，タスクに対するラベルの数が偶数の場合がある．その場合，0, 1 の数が同数のときにはランダムにいずれかを選択した．得られた統合ラベルを用いて，$\{(x_i, y'_i)\}_{i\in\{1,2,\cdots,N\}}$ を訓練データとする．

- **Dawid と Skene の方法による統合 [1]**

統計的に真のラベルを推定しラベルの統合を行う手法のうち，最も基礎的な Dawid と Skene の方法を採用した [1]．この方法ではワーカーの能力を，「真のラベルが 1 のときに正しく 1 と答える確率」と「真のラベルが 0 のときに正しく 0 と答える確率」の二つのパラメータでモデル化する．EM アルゴリズムを用いてモデルパラメータと真のラベルを推定し，真のラベルを統合ラベル y'_i として，多数決と同様に，$\{(x_i, y'_i)\}_{i\in\{1,2,\cdots,N\}}$ を訓練データとする．

- 非統合

複数ワーカーのラベル統合を行わず，全てのラベルを訓練データとして用いる手法が既存研究で利用されている [7]．この場合，$\{(x_i, y'_{i,j})\}_{i\in\{1,2,\cdots,N\},j\in\mathcal{J}_i}$ が訓練データとなる．

4.3　実 験 結 果

3.2 節と同様の設定で分類器を構築し実験を行った．テストデータの正解ラベルにはエキスパートラベルを用いた．特徴量には，3.1 節で示した全ての特徴量（タスクのテキスト特徴量，タスクの非テキスト特徴量，依頼者 ID 特徴量，依頼者の非テキスト特徴量）を使用した．

各回答統合手法及びワーカーラベル統合手法について分類器を構築し，分類精度を確認した．結果を表 7 に示す．回答統合手法については，いずれのワーカーラベル統合手法においても「アフィリエイト」「個人情報」を組み合わせた場合が最も AUC 平均値が高くなった．各設問を単独で用いた場合を見ても，この二つの設問は他の「直接取引」「ステルスマーケティング」よりも比較的高い AUC 値を示している．ここから「アフィリエイトの恐れがある」「個人情報の記載を要求している」という設問が，特に不適切タスクの検出に有効であると言える．逆に「ステルスマーケティングの恐れがある」という設問は不適切タスクでなくても「はい」と回答されやすい．今回のデータセットでこの設問に「はい」と回答があった例は 8,990 件中 1,030 件であった．他の設問（「アフィリエイト」で 699 件，「個人情報」で 105 件，「直接取引」で 45 件）と比較すると多く，実際はステルスマーケティングに関与しないタスクであっても，ワーカーは誤判定してしまう傾向があるようだ．

ワーカーラベルの統合手法を比較すると，Dawid と Skene の方法を用いて統合したラベルを用いた場合が最も高い AUC 平均値を達成しており (0.817)，次いで多数決で統合した場合，非統合の場合となった．図 8 に示した通り，ワーカーごとの能力に大きくばらつきがあるため，Dawid と Skene の方法が有効に機能したと考えられる．逆に非統合の場合には，能力の低いワーカーのラベルも正解として扱ってしまうために，分類器の精度が低くなったと考えられる．

以上から，回答統合手法では「アフィリエイト」「個人情報」の二つの回答の論理和を取る手法，ワーカーラベル統合手法では Dawid と Skene の方法が有効であることがわかった．また，3.2 節で示した，エキスパートラベルを正解と用いた場合の結果 (AUC 0.950) と比較すると，ワーカーラベルだけではエキスパートラベルに匹敵する精度は得られないことが確認できた．

表 7　ワーカーラベルだけを用いて構築した分類器の，訓練時の正解ラベル作成手法を変化させたときの AUC の平均値と標準偏差の比較

設問の組み合わせ（回答を論理和で統合）				ワーカーラベル統合手法		
アフィリエイト	直接取引	個人情報	ステルスマーケティング	統合 多数決	統合 Dawid-Skene	非統合
			✓	0.616	0.615	0.603
		✓		0.684	0.698	0.748
	✓			0.767	0.763	0.720
✓				0.752	0.812	0.741
		✓	✓	0.647	0.673	0.658
	✓		✓	0.659	0.627	0.626
	✓	✓		0.705	0.717	0.743
	✓	✓	✓	0.634	0.680	0.667
✓			✓	0.737	0.763	0.734
✓		✓		**0.759**	**0.817**	**0.754**
✓		✓	✓	0.744	0.770	0.748
✓	✓			0.752	0.811	0.738
✓	✓		✓	0.740	0.765	0.738
✓	✓	✓		0.758	0.816	0.749
✓	✓	✓	✓	0.745	0.772	0.748

5. エキスパートラベルとワーカーラベルの併用による分類器の構築

本節では，エキスパートラベルとワーカーラベルを組み合わせて分類器を構築する方法について述べ，その分類精度を，エキスパートラベルだけを用いた場合とワーカーラベルだけを用いた場合それぞれと比較する．

5.1 エキスパートラベルとワーカーラベルの併用方法

エキスパートラベルとワーカーラベルを組み合わせる方法を考える．いま，各タスク i に対して，エキスパートラベル $y_{i,0}$ とワーカーラベルを統合する場合には y'_i が，統合しない場合には $y'_{i,j}$ が付与されている．以下簡単のため，エキスパートラベルを e，ワーカーラベルを w（$=y'_i$ あるいは $y'_{i,j}$）とする．エキスパートラベルとワーカーラベルを組み合わせる単純なやり方として，論理積あるいは論理和を用いることが考えられる．この二つの手法は，エキスパートとワーカーの判定が一致しないとき，つまり $e \neq w$ の場合に異なる結果を返す．すなわち，$e \neq w$ の場合，論理積を用いると 0 をラベルとして採用し，論理和を用いると 1 をラベルとして採用する．この二つの方針以外に，ワーカーとエキスパートの判定が異なるような曖昧なサンプルは訓練データに加えないという方針も考えられる．

以上のように，$e \neq w$ の場合，{ 適切ラベル (0) を採用（方針 N），不適切ラベル (1) を採用（方針 P），訓練データに追加しない（方針 S）} という三つの方針が考えられる．さらに，$e \neq w$ となるのは $(e,w)=(0,1)$ と $(e,w)=(1,0)$ の 2 通りがあり，それぞれについて方針を選ぶと $\{\mathsf{N},\mathsf{P},\mathsf{S}\} \times \{\mathsf{N},\mathsf{P},\mathsf{S}\}$ の 9 通りの戦略が考えられる．ただし，$(e,w)=(0,1)$ のときに P，$(1,0)$ のときに N を選ぶ戦略はワーカーラベルだけを用いる方法と同じであり，$(0,1)$ のときに N，$(1,0)$ のときに P を選ぶ戦略はエキスパートラベルだけを用いる方法と同じである．

本節で用いる訓練データ作成手順を示す．サンプル選択の戦略が与えられたとき，各タスク $i \in \{1,2,\cdots,N\}$ について以下の手続きを行う．

（1） エキスパートラベルを $e=y_{i,0}$ とする．ワーカーラベル統合を行う場合，ワーカーラベルを $w=y'_i$ とする．非統合の場合，各 $j \in \mathcal{J}_i$ について $w=y'_{i,j}$ とし以下の処理を行う．

（2） $e=w$ のとき：$(e,\boldsymbol{x}_i)$ を訓練データに加える

（3） $e \neq w$ のとき：

- 方針 N を採用した場合：$(0,\boldsymbol{x}_i)$ を訓練データに加える
- 方針 P を採用した場合：$(1,\boldsymbol{x}_i)$ を訓練データに加える
- 方針 S を採用した場合：当該サンプルを訓練データに追加しない

5.2 実験結果

各サンプル選択戦略それぞれについて分類器を構築し分類精度を比較した．分類器の構築は 3.2 節と同様の設定で行い，特徴量は 4.3 節と同じく，3.1 節で示した全ての特徴量を用いた．回答の統合には「アフィリエイト」「個人情報」に対する回答の論理和を取る手法を用いた．ワーカーラベルの統合方法は，4.2 節で示した 3 種類（多数決による統合，Dawid と Skene の方法による統合，非統合）をそれぞれ用いて比較した．結果を表 8 に示す．

いずれのワーカーラベル統合手法においても方針 (S, P) のときに最も高い平均 AUC 値を示している．方針 (S, P) は，エキスパートが不適切タスクと判定していれば，ワーカーの判定に関わらずにエキスパートラベルを採用して「(エキスパートラベル，タスク)」を訓練データに加える．一方，エキスパートが適切タスクと判定しているにも関わらずワーカーが不適切タスクと判定した場合には，どちらが正しいか判断できないとして訓練データに加えない．この方針にもとづき構築した分類器が高精度となる理由は以下のように考えられる．エキスパートによる判定は適合率が高く，エキスパートがあるタスクを不適切と判定した場合はワーカーの判定に関わらず正しいと見なせる．一方で，再現率についてはワーカーの方が高く，適切という判定に関してはワーカーの意見を考慮することで精度が向上する（事実，図 8 にあるようにワーカーは平均的には適合率よりも再現率が高い）．

最も高い AUC 値を示したのは，Dawid と Skene の方法によって複数のワーカーラベルを統合し，(S,P) 設定を用いた場合である (0.962)．これは，エキスパートラベルだけを用いた場合 (0.950) よりも高い AUC 値となっており，t 検定により統計的優位性 ($p < 0.05$) が確認された．以上から，ワーカーラベルをエキスパートラベルと組み合わせることで，より高精度の分類器が構築できることが確認できた．

さらに表 9 に，訓練時に用いるタスクにエキスパートラベルが付与されている割合を変化させたときの，構築した分類器の平均 AUC 値を示す．エキスパートだけを用いて分類器を構築する場合は，訓練に用いるタスク数を変化させていることになる．結果，エキスパートが監視するタスクが全体の 70%〜100% であるときにはエキスパートラベルとワーカーラベルを組み合わせて構築した分類器が，エキスパートラベルだけを用いるときよりも高い精度となることがわかる．また，エキスパートラベルの割合が 75%〜100% の場合にはエキスパートラベル 100%で用いた分類器の精度 (AUC 0.950) を上回っており，ワーカーラベルを組み合わせることでエキスパートが監視しなければならないタスクの数を 25%程度削減しても，エキスパートが全て監視した場合と同等の分類器精度を達成できることが確認できた．

6. 関連研究

ワーカーの品質管理はクラウドソーシングにおける重要な課題の一つである．ワーカー品質管理手法は，タスク設計によって品質を向上させる手法や [5]，ワーカーをフィルタリングする手法，同じタスクに対する複数の回答結果を統合する手法が提案されている．ワーカーフィルタリング手法は, [4] にまとめられており，たとえばワーカーの居住地やこれまでの作業承認率でフィルタリングする方法，事前テストを用いる手法，Gold standard data と呼ばれる正解がわかっている問題を紛れ込ませワーカーの能力を測る手法などが用いられている．

同じタスクを複数のワーカーに回答させ，多数決や [8]，ワーカーの能力や問題の難易度を考慮して統計的に統合する手

表 8　ワーカーラベルとエキスパートラベルを組み合わせて構築した分類器の，ラベルの組み合わせ戦略を変化させたときの AUC の平均値と標準偏差の比較．戦略はそれぞれ，N「適切 (0) ラベルを採用」，P が「不適切 (1) ラベルを採用」，S が「訓練データに追加しない」を表している．また，9 個の戦略のうち，(N, P) と (P, N) はそれぞれ，エキスパートラベルだけ，ワーカーラベルだけを用いたときと同じであるため，結果から除外した．

エキスパートラベルとワーカーラベルの統合方法		ワーカーラベル統合方法		
(e, w)		統合		非統合
$(0, 1)$	$(1, 0)$	多数決	Dawid-Skene	
N	N	0.786 (±0.087)	0.763 (±0.076)	0.895 (±0.042)
N	S	0.816 (±0.081)	0.791 (±0.070)	0.929 (±0.029)
P	P	0.936 (±0.021)	0.951 (±0.017)	0.891 (±0.034)
P	S	0.790 (±0.047)	0.841 (±0.061)	0.829 (±0.046)
S	N	0.825 (±0.080)	0.816 (±0.075)	0.900 (±0.041)
S	P	**0.959** (±0.013)	**0.962** (±0.013)	**0.950** (±0.016)
S	S	0.877 (±0.033)	0.867 (±0.032)	0.935 (±0.026)

表 9　エキスパートラベルを付与するタスク数を変化させたときの分類器の性能の変化．なお，いずれの場合でも訓練に用いるタスクの数は 1,800 件てであり，この全てにワーカーラベルが付与されている．たとえばエキスパートラベルが付与されているのが 1,260 件のとき，残り 540 件にはワーカーラベルだけが付与されている．ワーカーラベルの統合には Dawid-Skene の方法を用いている．

利用した監視結果	エキスパートラベルが付与されている訓練タスク数	Average AUC
エキスパートラベルとワーカーラベル	1800 (100%)	0.962
	1710 (95%)	0.960
	1620 (90%)	0.958
	1530 (85%)	0.955
	1440 (80%)	0.955
	1350 (75%)	0.951
	1260 (70%)	0.946
エキスパートラベルのみ	1800 (100%)	0.950

法 [1], [10], [11] も提案されている．これらは全てワーカーの品質管理に関する研究であり，実際のクラウドソーシングサービスのデータを用いてタスクの品質管理に取り組んだ研究は，我々の知る限りこれが初めてである．

今回は，複数のワーカーラベルを統合して訓練時の正解ラベルとして用いる手法を使用したが，複数のワーカーラベルから直接分類器を学習する手法もいくつか提案されている [2], [6], [12]．さらに，エキスパートラベルとワーカーラベルが混在する設定でのワーカーラベル統合手法 [9] や分類器学習手法 [3] も提案されている．これらを用いることで，不適切タスク検出の精度向上が期待される．

7. ま と め

本研究では，クラウドソーシング運営会社の不適切タスク監視作業を支援することを目的として，機械学習を用いて不適切タスク検出を行う分類器を構築した．運営会社だけではなくクラウドソーシングワーカーにも監視作業を依頼し，運営会社とワーカーの監視結果を統合して訓練に用いることで，分類器の精度が向上することを示した．また，ワーカーの監視結果を用いることで，エキスパートが監視するタスクの数を 25%程度削減しても精度が保てることを確認した．

今回は，タスクとそれに対する運営側・ワーカーのラベルが与えられた場合にオフラインで分類器を構築する問題を対象とした．しかし，実際に運用する場合にはタスクが逐次的に投稿され，運営側・ワーカーのラベルも次々と追加されていく状況を想定しなければならない．このような設定において有効なオンラインでの分類器学習手法と，運営側のコストを減らすための効率的な利用プロセスを検討することが今後の課題である．

謝　辞

本研究は内閣府最先端研究開発プログラム (FIRST)「超巨大データベース時代に向けた最高速データベースエンジンの開発と当該エンジンを核とする戦略的社会サービスの実証・評価」の助成を受けたものである．

文　献

[1] A. P. Dawid and A. M. Skene. Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statics)*, 28(1), 1979.

[2] H. Kajino, Y. Tsuboi, and H. Kashima. A Convex Formulation for Learning from Crowds. In *Proc. of AAAI*, 2012.

[3] H. Kajino, Y. Tsuboi, I. Sato, and H. Kashima. Learning from Crowds and Experts. In *Proc. of HCOMP*, 2012.

[4] G. Kazai, J. Kamps, M. Koolen, and N. Milic-Frayling. Crowdsourcing for book search evaluation: impact of hit design on comparative system ranking. In *Proc. of SIGIR*, 2011.

[5] A. Kittur, E. Chi, and B. Suh. Crowdsourcing user studies with mechanical turk. In *Proc. of CHI*, 2008.

[6] V. C. Raykar, S. Yu, L. H. Zhao, C. Florin, L. Bogoni, and L. Moy. Learning From Crowds. *Journal of Machine Learning Research*, 11, 2010.

[7] V. S. Sheng, F. Provost, and P. G. Ipeirotis. Get Another Label? Improving Data Quality and Data Mining Using Multiple, Noisy Labelers. In *Proc. of KDD*, 2008.

[8] 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 *Proc. of EMNLP*, 2008.

[9] W. Tang and M. Lease. Semi-Supervised Consensus Labeling for Crowdsourcing. In *ACM SIGIR Workshop on Crowdsourcing for Information Retrieval (CIR)*, 2011.

[10] P. Welinder, S. Branson, S. Belongie, and P. Perona. The Multidimensional Wisdom of Crowds. In *Proc. of NIPS*, 2010.

[11] J. Whitehill, P. Ruvolo, T. Wu, J. Bergsma, and J. Movellan. Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise. In *Proc. of NIPS*, 2009.

[12] Y. Yan, R. Rosales, G. Fung, M. Schmidt, G. Hermosillo, L. Bogoni, L. Moy, J. Dy, and P. Malvern. Modeling Annotator Expertise: Learning When Everybody Knows a Bit of Something. In *Proc. of AISTATS*, 2010.