DEIM2022 E34-5

クラウドソーシングのマルチクラス分類タスクにおける票の割れ方に着目したスパマー検知手法の提案

渡邊　綾仁†　　田島　敬史††

† 京都大学大学院情報学研究科　〒 606–8501 京都府京都市左京区吉田本町
†† 京都大学大学院情報学研究科　〒 606–8501 京都府京都市左京区吉田本町
E-mail: † watanabe.a@dl.soc.i.kyoto-u.ac.jp, †† tajima@i.kyoto-u.ac.jp

あらまし　クラウドソーシングでのマルチクラス分類タスクにおける従来のスパマー検知手法では，ワーカーの回答のランダム度合いを混同行列から定量化する方法が利用されてきた．この方法では，クラス数が増加すると，ワーカーへのアイテムの割り当て数や真のラベルの偏り方によっては混同行列を推定するためのデータがスパースになってしまうため，推定のためにはアイテム数を多く割り当てる必要がある．本研究では，混同行列を利用せず，同じアイテムに回答したワーカの投票割合のデータを用いて，票の割れ方に依存してペナルティの重みが変わるスコアリング手法によって，ワーカーのスパマー度合いを定量的に評価する手法を提案する．実験において，割り当てたアイテム数が少ないときに従来のスパマー検知手法よりも高い精度でスパマー検知を行うことが確かめられた．

キーワード　クラウドソーシング，スパマー検知，マルチクラス分類，項目反応理論

1　はじめに

近年，Amazon Mechanical Turk (MTurk)[1] などのクラウドソーシングサービスの台頭により，不特定多数のワーカにオンラインで作業を依頼することができるようになっている．例えば，画像に写っている被写体のラベリング，被写体の領域抽出，テキストデータや音声データへのラベルの付与などを依頼することができる．クラウドソーシングサービスを利用することで，計算機を用いて自動生成することが難しいデータや，大規模なデータセットを構築するのが困難なときに，安価で効率よくデータを収集することが可能となった．

その便利さの一方で，クラウドソーシングならではの問題点もいくつか存在する．クラウドソーシングサービスに登録しているワーカは，専門家と比較してワーカごとの能力や意欲の分散が大きく，極端な場合，意図的に手抜きをしたり，bot などを利用し機械的に作業を行うことによって作業時間を短縮し，効率よく報酬だけを獲得しようとするワーカも存在する．このような悪質なワーカを特定して排除することは，クラウドソーシングによって得られる成果物の品質管理の観点から重要である．

ワーカの能力を推定する既存手法としては，混同行列を用いた手法が存在する．この手法では，推定された混同行列から，ワーカの回答のランダム度合いを定量的に評価する．しかし，ワーカに割り当てられたアイテムの真のクラスの分布の偏りが大きい場合やワーカへの割り当てアイテム数が十分でない場合，混同行列推定のためのデータがスパースになってしまうクラスが存在し，推定精度が落ちるという問題点がある．また，混同行列を用いた手法では，各アイテムごとに他のワーカがどのようなラベルを付与したかの情報や，各アイテムの分類の難易度の情報は利用していない．

そこで，本研究では，混同行列は利用せずに，多数決における票の割れ方に着目した手法を提案する．マルチクラス分類タスクでは，真面目なワーカが回答しても誤分類しやすいアイテムや，真面目にタスクを遂行していれば誤分類しづらいアイテムが存在する．これらの傾向を利用し，ワーカへのスコアリングを行うことで，各ワーカへのアイテムの割り当て数が少ない場合や，割り当てられたアイテムの難易度の差がある場合でも，それらの影響を小さくし，ワーカの能力をより正しく推定できるようにする．

本研究では，MTurk を利用して，画像分類タスクを依頼し依頼したタスクの回答データを用いて，ワーカをランキング付けする実験を行い，提案手法の有用性を評価した．

1：https://www.mturk.com/

2　関連研究

2.1　真のラベル推定とワーカの能力の推定

クラウドソーシングにおいて真のラベル推定を行う際は多数決を用いる方法が主流である．Snow ら [1] は，5 種類のタスクにおいて，専門家が作成した既存の Gold Standard ラベルと，MTurk によって収集された非専門家によるラベリングの比較を行い，5 種類のタスク全てにおいて，少人数の非専門家によって付与されたラベルと，専門家が作成した Gold Standard ラベルとの間で高い一致度を示したと報告している．

多数決の問題点としては，真面目にタスクを遂行するワーカも，手抜きをする悪質なワーカも，同じ一票の重みを持ってしまうことが挙げられる．この問題点の対処法としては，ワーカの能力に応じて票の重みを変動させることが考えられる．Dawid ら [2] は，EM アルゴリズムを利用し，ワーカの混同行列と，アイテムの真のラベルを同時に推定する手法を提案した．この手

法は，推定された混同行列に基づいて多数派の回答を選ぶことが多いワーカの票の重みを大きくした重み付き多数決を行っているものとみなすことができる．

クラウドソーシングサービスには登録して間もなく，能力値が不明なワーカも多数存在する．能力値が不明なワーカについては，ある程度タスクを割り当てることによって，能力値を推測していく必要がある．Joglekar ら [3], [4] は，真面目なワーカにタスクを依頼した際，ワーカの回答のデータから，ワーカの回答の誤り率の信頼区間を推定するアルゴリズムを提案した．Donmez ら [5] は，最適なワーカを選択し，能動学習を行うための手法を提案している．能力の高いワーカを積極的に活用しながら，能力が未知のワーカを早期に登用し，信頼度が低いワーカをできるだけ早い段階で除去することで，ラベリングの精度を向上させている．

タスクの性質に着目するというアプローチによってラベルの推定精度を向上させるという研究もある．Whitehill ら [6] は，項目反応理論を用いて，2 クラス分類タスクにおいて，タスクの難易度を考慮してワーカの能力とタスクの正解を同時に求めることができる GLAD 法という手法を提案している．Bachrach ら [7] は，ワーカの回答データ，タスクの難易度，ワーカの能力，真のラベルの値をグラフィカルモデルを利用することで，一方のデータから他のデータを推定することが可能なモデルを提案している．

2.2 スパマー検知に関する研究

ワーカの提出物やその行動特性からワーカをいくつかのタイプに分類し，悪意のあるワーカに共通する傾向を把握することは，スパマー検知の手法を考案する上で重要である．Gadiraju ら [8] は，成果物の品質が悪いワーカや，悪意のあるワーカについて，その典型的な行動例や特性に応じて，5 種類のワーカに分類している．そして，各ワーカーの分類ごとに回答の内容，タスクを完了するのにかかった時間，悪意のある行動が一番最初に確認される設問はどこであるか，といった指標で評価を行っている．

悪質なワーカを推定するスパマー検知手法については，様々なアプローチから手法が提案されているが，多数決における票の割れ方に着目した本研究とはアプローチが異なる．

Ipeirotis ら [9] は，マルチクラス分類タスクにおいて，回答に特有のバイアスが存在するワーカの回答について，そのバイアスを考慮したスパマー検知手法を提案した．EM アルゴリズムによって，ワーカの回答データから，各アイテムの真のラベルと，各ワーカの混同行列の推定を行い，混同行列の各列に対してハードなラベルからソフトラベルへの変換を施す．それらのソフトラベルからそのユーザの誤答により生じるコストを計算し，閾値を超えたワーカをスパマーとみなして排除を行う．

Raykar ら [10] は，ランダムに回答をするワーカや，全てのアイテムに対して全く同じラベルを回答するワーカは，真のラベルが何であるかに関わらず回答を決定している点に着目した．推定された混同行列からワーカのスパマー度合いを「スパマースコア」と呼ばれる指標を用いて定量化し，一定品質以下の回答を行ったワーカを排除する過程を，収束するまで交互に繰り返すアルゴリズムを提案している．

Ipeirotis ら [9] の手法と，Raykar ら [10] の手法の問題点を挙げる．ワーカにアイテムを割り当てる際に，真のラベルに偏りがある場合や，特定のクラスにおいてアイテムの割り当て数が少ない場合，混同行列を用いてワーカの能力を推定すると推定の精度が落ちると考えられる．例えば，A，B，C のいずれかに分類する 3 クラス分類タスクについて，あるワーカに割り当てられたアイテムの真のラベルの内訳が，A が 9 個，B が 1 個，C は 0 個だった場合を考える．このとき，クラス B に対応する部分の推定は 1 個のアイテムのみで行われるため，0 または 1 の値となる．また，クラス C の正解率は未定義となるため，混同行列において対応する行の対角成分を 1 としたり，クラス C に回答したワーカの混同行列の平均値を利用するなどの補完が必要になる．

混同行列を用いた手法では，各ワーカが真のラベルに対してどのような回答をしたかの傾向をつかむことはできるが，多数決に参加したほかのワーカがどのような回答をしているかの票の情報や，個々のアイテムにおける分類の難しさといったアイテム特有の情報は利用していない．アイテム数の割り当てが少ないときは，全ての回答データを集約した混同行列よりも，個々のアイテムの情報を利用したほうがより高い精度で推定ができるものと考えられる．

3 提 案 手 法

3.1 ワーカへ付与するスコアの計算

はじめに，本研究で取り扱う問題を定式化する．N 個のアイテムが用意されており，各アイテムはそれぞれ C (C は 2 以上の整数) クラスの中から最も適合するクラスを一つ選ぶマルチクラス分類タスクであるとする．N 個のアイテムには 1 から N までの整数が通し番号として一意に割り当てられているものとする．ワーカの人数を M と表記し，M 人のワーカにそれぞれ 1 から M までの整数が通し番号として一意に割り当てられているものとする．番号 i $(i = 1, 2, ..., N)$ のアイテムに対して，番号 j $(j = 1, 2, ...M)$ のワーカが付けたラベルを $L_i^{(j)}$ $(1 \leqq L_i^{(j)} \leqq C, L_i^{(j)} \in \mathbb{N})$，番号 i のアイテムに回答したワーカの番号の集合を W_i と表記する．

番号 i のアイテムに対して，多数決の票を集約する方法を述べる．番号 i のアイテムに対して，ラベル k を付けたワーカが何人いるかを $\eta_{i(k)}$ と表すことにすると，$\eta_{i(k)}$ は，

$$\eta_{i(k)} = \sum_{w \in W_i} \delta(L_i^{(w)}, k),\ \delta(L_i^{(j)}, k) = \begin{cases} 1 & (L_i^{(j)} = k) \\ 0 & (L_i^{(j)} \neq k) \end{cases}$$

と書くことができる．番号 i のアイテムの投票結果をベクトル $\vec{\eta_i} \in \mathbb{N}_0^C$ （$\mathbb{N}_0$ は 0 以上の整数の集合）と書き，そのベクトルの第 k 成分が，番号 i のアイテムにラベル k を付けた人数とすると，$\vec{\eta_i} = (\eta_{i(1)}, \eta_{i(2)}, ..., \eta_{i(C)})$ と表すことができる．このベクトル $\vec{\eta_i}$ を成分の総和が 1 になるように L1 ノルムで各成分を割ったベクトルを，

$$\vec{v_i} = \frac{\vec{\eta_i}}{||\vec{\eta_i}||_1} = (v_{i(1)}, v_{i(2)}, ..., v_{i(C)}) \tag{1}$$

と書くと，$\vec{v_i}$ の第 k 成分 $v_{i(k)}$ は，番号 i のアイテムに対して m 人のワーカのうちラベル k を付けた人の割合を表す．

ワーカの回答に基づいて，ワーカをスコアリングする方法を述べる．基本的な方針としては，多数派の回答から逸脱するほどペナルティが大きくなるようにスコアをつけることを考える．(1) 式のように，ベクトルでワーカーの投票結果を集約したとき，真面目なワーカーでもクラス分類が難しいアイテムでは，ベクトルの成分がいくつかの選択肢に分散する場合や，あるいは，真のラベルは A であるが，多くのワーカーは B とラベリングし，それ以外の選択肢に間違える人はまずいない，というような場合が考えられる．一方，分類が容易なアイテムにおいては，真のラベルに対応するベクトルの成分が大きくそれ以外の成分は小さくなると考えられる．

アイテムの真のラベルが未知であると仮定する．このとき，番号 i のアイテムでの番号 j のワーカのペナルティ $p_i^{(j)}$ を，

$$p_i^{(j)} = \sum_{k=1}^{C} f(i,j,k)$$

と定義する．ただし，

$$f(i,j,x) = \begin{cases} v_{i(x)} & \left(v_{i(x)} > v_{i(L_i^{(j)})}\right) \\ 0 & \left(v_{i(x)} \leqq v_{i(L_i^{(j)})}\right) \end{cases} \tag{2}$$

である．$p_i^{(j)}$ は直観的には，番号 i のアイテムにおいて，番号 j のワーカがつけたラベル $L_i^{(j)}$ よりも多数派のラベルについて，投票割合の値の和をとっている．ただし，正解が既知の場合は，正解を最多数派の選択肢とみなす．番号 j のワーカが回答したアイテムの番号の集合を $I^{(j)}$ すると，番号 j のワーカのペナルティスコア $P^{(j)}$ は，

$$P^{(j)} = \frac{1}{|I^{(j)}|} \sum_{i \in I^{(j)}} p_i^{(j)}$$

と定義される．この値は，0 以上 1 以下の値をとり，番号 j のワーカが，多数派のクラスに投票をしているほど 0 に近づき，多数派から外れた回答をしているほど 1 に近づく．

(2) 式の $f(i,j,x)$ の定義では，投票割合が僅差の場合において，真面目なワーカ間で大きくスコアに差がついてしまう可能性がある．例えば，4 クラス分類において，投票割合が $(0.5, 0.4, 0.1, 0.0)$ のときと $(0.5, 0.2, 0.15, 0.15)$ のときについて，(2) 式を用いてペナルティを計算すると，一つ目の例は $0, 0.5, 0.9, 1.0$，二つ目の例は $0, 0.5, 0.7, 0.7$ となる．このとき，一つ目の例で 2 番目の選択肢を選んだワーカと，二つ目の例で 2 番目の選択肢を選んだワーカに与えられるペナルティは同じ 0.5 であるが，1 番目の選択肢と 2 番目の選択肢の投票割合に着目すると，一つ目の例の方が，二つ目の例よりも投票割合が僅差であるため，より 1 番目と 2 番目の選択肢のどちらを選ぶかで迷いやすかったことが想定される．

そこで，自分がつけたラベルの投票割合と，自分より多数派の回答の選択肢の投票割合との差をとる方法も考えられる．具体的には，$f(i,j,x)$ を，

$$f(i,j,x) = \begin{cases} v_{i(x)} - v_{i(L_i^{(j)})} & \left(v_{i(x)} > v_{i(L_i^{(j)})}\right) \\ 0 & \left(v_{i(x)} \leqq v_{i(L_i^{(j)})}\right) \end{cases} \tag{3}$$

と定義する．投票割合が $(0.5, 0.4, 0.1, 0.0)$ の例で (3) 式を用いてペナルティを計算すると $0, 0.1, 0.8, 1.0$ となり，2 番目の選択肢を選んだワーカのペナルティが (2) 式より抑えられる．

3.2 段階反応モデルを用いた推定

3.1 節のペナルティスコアは，多数派の回答から逸脱するほど大きなペナルティを付与する，というアイデアを直観的に表現し，ワーカーの能力を推定するモデルである。本節では，より数理的なモデルを用いて票の割れ方からワーカの能力を推定する手法を述べる．

段階反応モデルは，Samejima ら [11] が考案したモデルである．段階反応モデルを用いることで，リッカート尺度のような順序尺度を選択肢とする項目の回答データから，各項目の項目識別力や難易度，回答者の能力を推定することができる．

参考文献 [12] に即して段階反応モデルを説明する．r_j は項目 j における反応を表す離散変数で，G を 2 以上の整数として，$r_j = 0, 1, 2, ..., G-1$ の値のいずれかをとるものとする．このとき，能力が θ の被験者が $r_j = k$ と反応する確率 $p_{j,k}(\theta)$ を，

$$p_{j,k}(\theta) = p^*_{j,k}(\theta) - p^*_{j,k+1}(\theta)$$

$$p^*_{j,k}(\theta) = \begin{cases} 1 & (k=0) \\ \dfrac{1}{1+\exp\left(-Da_j(\theta - b^*_{j,k})\right)} & (1 \leqq k \leqq G-1) \\ 0 & (k=G) \end{cases}$$

と表す．ここで，$p^*_{j,k}(\theta)$ は，能力が θ の被験者が $r_j \geqq k$ と反応する確率である．D は定数で $D = 1.0$ または $D = 1.7$ が用いられる．a_j は，項目 j の識別力を表す．値が大きいほど $p^*_{j,k}(\theta)$ $(1 \leqq k \leqq G-1)$ のロジスティック曲線の変化が大きくなり，被験者の能力 θ が少し変わるだけで反応確率が大きく変わることを表す．$b^*_{j,k}$ は困難度と呼ばれ，項目 j に対して $r_j \geqq k$ と反応する確率が 0.5 となる被験者の能力の値を表す．

はじめに，周辺最尤推定法を用いて，各項目の識別力，各項目の困難度を推定を行う方法を述べる．被験者の反応データを記録した行列を R とする．行列 R の第 (i,j) 成分 $r_{i,j}$ は，被験者 i の項目 j における反応を表し，0 以上 $G-1$ 以下のいずれかの整数値をとるものとする．例えば，ある一人の被験者がそれぞれ 0 から 3 の 4 段階のいずれかの値をとる 4 個の項目に全て回答したとすると，$R = (0 \quad 3 \quad 2 \quad 1)$ というような形式のデータが得られる．このデータを，各被験者 i ごとに，$r_{i,j} = k$ ならば $u_{j,k+1} = 1$，$r_{i,j} \neq k$ ならば $u_{j,k+1} = 0$ という変換を施し，第 $(j, k+1)$ 成分に要素 $u_{j,k+1}$ を持つ反応パターン行列 U_i を作る．このとき，尺度値が θ_i である被験者 i が，反応パターン行列 U_i を得る確率は，後述する局所独立を仮定すると，

$$p(U_i|\theta_i, \mathbf{a}, \mathbf{B}^*) = \prod_{j=1}^{n} \prod_{k=0}^{G-1} p_{j,k}(\theta_i)^{u_{j,k}} \tag{4}$$

と表せる．ここで，$\mathbf{a}$，$\mathbf{B}^*$ は，

$$\mathbf{a} = (a_1, a_2, ..., a_n)^\mathsf{T},\ \mathbf{B}^* = \begin{pmatrix} b^*_{1,1} & ... & b^*_{1,G-1} \\ ... & ... & ... \\ b^*_{n,1} & ... & b^*_{n,G-1} \end{pmatrix}$$

である．

局所独立とは，能力 θ_i の値が与えられている場合，各項目への反応は互いに独立であるという仮定である．受験者集団全体では，通常の場合，項目間の反応や成績には相関があり，互いに独立ではない．そのため，複数の項目への反応の同時確率は，各項目に対する反応の確率の総積とはならない．局所独立の仮定は，まったく同じ能力を持った被験者ばかりを集めてきたら，項目間の反応に相関がなくなり，互いに独立であるということを仮定している．

被験者の能力 θ が未知であるので，θ の分布に標準正規分布

$$g(\theta) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\theta^2\right)$$

を仮定して，各項目の識別力と，各項目の困難度を推定する．θ が標準正規分布 $g(\theta)$ に従っているとすると，周辺分布は，

$$h(U_i|\mathbf{a}, \mathbf{B}^*) = \int_{-\infty}^{\infty} g(\theta) p(U_i|\theta, \mathbf{a}, \mathbf{B}^*) d\theta$$

と表現できる．m 人の被験者について，各被験者の回答は独立であるから，m 人分の同時分布は，

$$L(\mathbf{a}, \mathbf{B}^*) = \prod_{i=1}^{m} h(U_i|\mathbf{a}, \mathbf{B}^*)$$

となる．これが能力 θ に関して，識別力，困難度の周辺尤度関数となる．この対数周辺尤度関数を各パラメータに関して最大化することによって識別力，困難度の最尤推定値を得る．

次に，得られた識別力，困難度のパラメータから，被験者の能力値を推定する．能力が θ_i である被験者が，反応パターン行列 U_i を得る確率は (4) 式で表される．今，項目の識別力，困難度，および反応パターン行列がデータとして得られているため，$p(U_i|\theta_i, \mathbf{a}, \mathbf{B}^*)$ は，θ_i を変数とした尤度関数とみなすことができる．両辺の対数をとり，θ_i に関して偏微分をすると，$q^*_{j,k}(\theta_i) = 1 - p^*_{j,k}(\theta_i)$ として，

$$\frac{\partial \log p(U_i|\theta_i, \mathbf{a}, \mathbf{B}^*)}{\partial \theta_i} = D \sum_{j=1}^{n} a_j \sum_{k=0}^{G-1} u_{j,k} \frac{p^*_{j,k}(\theta_i) q^*_{j,k}(\theta_i) - p^*_{j,k+1}(\theta_i) q^*_{j,k+1}(\theta_i)}{p_{j,k}(\theta_i)}$$

が導かれる．この値が 0 となる θ_i が被験者 i の能力の最尤推定値である．

クラウドソーシングにおけるマルチクラス分類タスクに段階反応モデルを適用する方法を述べる．マルチクラス分類タスクでは，通常，ワーカが各アイテムに付与したラベルがデータとして得られるが，ラベルは名義尺度である．そこで，(1) 式を用いて，各ワーカの回答を，各アイテムごとに C 段階の離散変数 $r_j = 0, 1, 2, ..., C-1$ に変換することを考える．

m 人のワーカが全員共通で回答しているアイテムの番号の集合を Item と表記し，m 人のワーカを 1 から m までの整数で一意に識別するものとする．Item の要素数を n とし，元のアイテムの番号とは別に新たに 1 から n までの整数が通し番号として与えられているものとする．番号 i $(i = 1, ..., n)$ のアイテムに対して，番号 j のワーカが付けたラベルである $L^{(j)}_i$ と (1) 式を用いて，ラベル $L^{(j)}_i$ を以下のように変換する．

$$\lambda^{(j)}_i = \sum_{k=1}^{C} \phi(L^{(j)}_i, k),\ \ \phi(L^{(j)}_i, k) = \begin{cases} 1 & (v_{i(L^{(j)}_i)} \geqq v_{i(k)}) \\ 0 & (v_{i(L^{(j)}_i)} < v_{i(k)}) \end{cases}$$

$\lambda^{(j)}_i$ は 1 以上 C 以下の値をとる離散変数であり，直観的には，番号 i のアイテムに対して，番号 j のワーカは何番目に少数派のラベルをつけたかの値になる．$\lambda^{(j)}_i$ が 1 に近いほど少数派の回答をしており，C に近いほど多数派の回答に投票していることを表す．この $\lambda^{(j)}_i$ の値を順番に n 個並べて，j 番目のワーカのデータを，$\Lambda^{(j)} = \left(\lambda^{(j)}_1, \lambda^{(j)}_2, ..., \lambda^{(j)}_n\right)$ と $1 \times n$ の行列の形で表すことにする．$\mathbf{1}$ を全ての要素が 1 である $1 \times n$ 行列とすると，$\Lambda^{(j)} - \mathbf{1} = \left(\lambda^{(j)}_1 - 1, \lambda^{(j)}_2 - 1, ..., \lambda^{(j)}_n - 1\right)$ は，各成分が 0 以上 $C-1$ 以下の値となる．これを m 人分並べた $m \times n$ 行列である，

$$\begin{pmatrix} \Lambda^{(1)} - \mathbf{1} \\ ... \\ \Lambda^{(m)} - \mathbf{1} \end{pmatrix} = \begin{pmatrix} \lambda^{(1)}_1 - 1 & \lambda^{(1)}_2 - 1 & ... & \lambda^{(1)}_n - 1 \\ ... & ... & ... & ... \\ \lambda^{(m)}_1 - 1 & \lambda^{(m)}_2 - 1 & ... & \lambda^{(m)}_n - 1 \end{pmatrix}$$

を，段階反応モデルで説明したワーカの反応データ R として用いることによって，段階反応モデルでの推定が可能となる．

クラス数 C が多いと，パラメータ量が多くなり，計算量が増える．また，クラス数 C が多いが，真面目に回答をするワーカが多い場合は，特定のクラスに票が偏ってしまうため，C 段階より少ない C' 段階で表したほうが良い場合もある．このような場合は，次に示す C 次元ベクトル $\vec{S}$ を任意に一つ決めて，C 段階の反応から C' 段階への反応への変換を行う．

$$\Lambda'^{(j)} = \left(s_{\lambda^{(j)}_1}, s_{\lambda^{(j)}_2}, ..., s_{\lambda^{(j)}_n}\right)$$

$$\vec{S} = (s_1, s_2, ..., s_C)$$

$$s_k \in \mathbb{Z},\ 0 \leqq s_k \leqq C' - 1\ (k = 1, ..., C,\ C' < C)$$

$$s_1 = 0,\ s_C = C' - 1,\ 0 \leqq s_{k+1} - s_k \leqq 1\ (k = 1, ..., C-1)$$

である．この変換によって生成された $\Lambda'^{(j)}$ を，$\Lambda^{(j)} - \mathbf{1}$ の代わりに用いる．例えば 7 クラス問題において $\vec{S} = (0, 0, 0, 0, 1, 2, 3)$ とすると，各アイテムについて最も多数派の回答を選んだワーカは 3，二番目に多数派の回答を選んだワーカは 2，三番目に多数派の回答を選んだワーカは 1，それ以外の回答をしたワーカは 0 というように 7 段階の反応を 4 段階の反応に圧縮して変換することができる．

4 実　　験

4.1 データ収集

MTurk にマルチクラス分類タスクを 3 種類投稿し，実験で

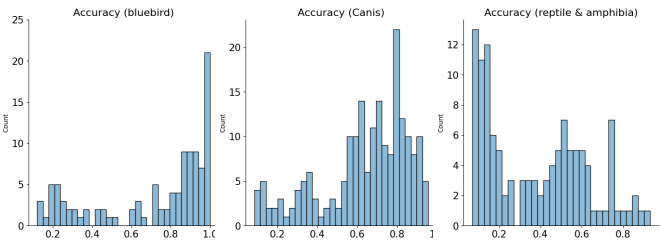
利用するためのデータ収集を行った．

1 種類目のデータセットは，「鳥類データセット」である．Welinder ら [13] が作成した Caltech-UCSD Birds 200 を利用し，画像アイテムセットを構築した．画像の真のラベルは既知であり，ラベルの番号と写っている鳥の種類の対応は，1: Black Throated Blue Warbler，2: Blue Grosbeak，3: Blue Headed Vireo，4: Blue Jay，5: Blue Winged Warbler である．各動物について 15 枚ずつ画像が用意されている．これまで一つ以上の HITs が承認されており，かつ Lifetime Approval Rate が 90% 以上のワーカーを集めた．タスクのデザインは，75 枚の写真を一つのページに並べており，すべての写真にラベル付けが完了したら提出できるようになっている．鳥類データセットでは，108 人のワーカーから回答を集めることができた．

2 種類目のデータセットは，「イヌ属データセット」である．このデータセットの画像アイテムセットは，イヌ属の動物 7 種類のうち 1 種類が写った画像 70 枚で構成されている．画像の真のラベルは既知であり，ラベルの番号と写っている動物の種類の対応は，1: Alaskan Malamute，2: Coyote，3: Dhole，4: Gray Wolf，5: German Shepherd，6: Samoyed，7: Siberian Husky となっている．各動物について 10 枚ずつ画像が用意されている．依頼した画像ラベリングタスクでは，1 枚ずつ画像を順番に表示し，写っている動物の種類として最も適合するものを一つ選択する．一度回答を提出すると，遡って前の画像の回答を修正することはできない．各ワーカーは 70 枚すべての画像に対してラベルを付与する．また，1 バッチ 18 人（1 回だけ 19 人）ごとに，タスクで表示する 70 枚の画像の順番を，乱数を用いてプログラムによって順番をランダムに入れ替える処理を行い，各バッチごとに画像の出題順序が異なるタスクを用意した．イヌ属データセットについては，199 人のワーカーから回答を集めることができた．

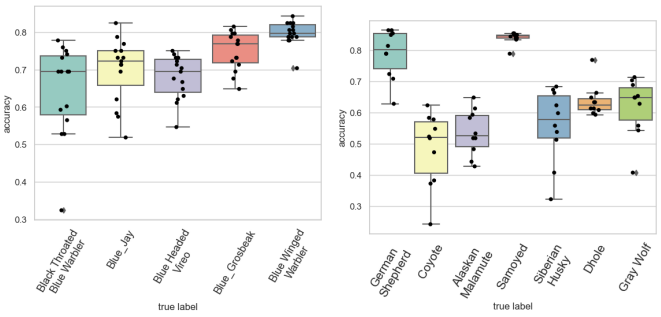
3 種類目のデータセットは，「両生類・爬虫類データセット」である．このデータセットの画像アイテムセットは，両生類・爬虫類の動物 10 種類のうち 1 種類が写った画像 100 枚で構成されている．画像の真のラベルは既知であり，ラベルの番号と写っている動物の種類の対応は，1: Basilisk，2: Chameleon，3: Agama，4: Gecko，5: Iguana，6: Komodo dragon，7: Skink，8: Tuatara，9: Giant salamander，10: Newt となっている．各動物について 10 枚ずつ画像が用意されている．依頼した画像ラベリングタスクでは，1 枚ずつ画像を順番に表示し，写っている動物の種類として最も適合するものを一つ選択する．一度回答を提出すると，遡って前の画像の回答を修正することはできない．また，各ワーカーごとに出題順序を乱数を用いてシャッフルした．両生類・爬虫類データセットでは，118 人のワーカーから回答を集めることができた．

図 1a，図 1b，図 1c は，横軸を画像アイテムセットに含まれる画像のうち何枚の画像に対して真のラベルと同じラベルを付与したかの正解率，縦軸を人数として描画したヒストグラムである．鳥類データセットでは，比較的正答率が高く，半分以上のワーカが 6 割以上正解しており，多くのワーカの正答率が 8 割以上となった．イヌ属データセットでは，正解率が 6 割から

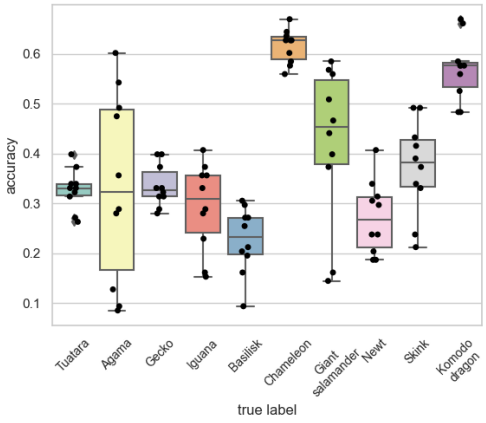


(a) 鳥類　(b) イヌ属　(c) 両生類・爬虫類

図 1　正解率分布



(a) 鳥類　(b) イヌ属



(c) 両生類・爬虫類

図 2　ラベルごとの画像の正解率分布

9 割程度の人が多く，正解率 4 割以下のワーカは全体の 2 割ほどであった．両生類・爬虫類データセットでは，正答率が 2 割程度かそれ以下の人が非常に多いことが読み取れる．

図 2a，図 2b，図 2c は，各画像について正解率を計算し，その画像の真のラベルごとに箱ひげ図を描画したものである．鳥類データセットでは，どのラベルもおおよそ似た傾向を示しているが，5 種類のラベルのうち，Blue Grosbeak と Blue Winged Warbler の正解率が他のラベルより少し高い傾向にある．イヌ属データセットでは，Samoyed，German Shepherd の分類難易度が比較的易しく，それ以外のラベルについては正解率が低くなっていることが読み取れる．Coyote と Alaskan Malamute については，全体的に正解率が他のラベルよりも低い．Coyote，Siberian Husky，Gray Wolf については，正解率が高い画像と低い画像が混在していることが読み取れる．両生類・爬虫類データセットでは，クラスごとに正解率，正解率の分布に大きく差があることが読み取れる．Chameleon と Komodo Dragon が

比較的正答率が高い．Agama，Giant salamander，Skink については，同じクラス内でも画像の正解率に大きな差がある．一方で，Tuatara，Gecko はどの画像も正答率が 3 割から 4 割程度と低い．全体的に正答率も低く，このデータセットではラベリングが全体的に難しかったと考えられる．

4.2 実験 1:割り当てアイテム数を減らした場合

ワーカ全員に同じ画像のラベリングタスクを依頼し，その割り当て枚数の数を変動させると，スパマー検知の精度がどうなるかを実験によって検証した．本実験は「ワーカの分類」「ワーカのスコアリング」「ランキング評価」の 3 段階に分けられる．

「ワーカーの分類」では，ワーカーを良好なワーカーと悪質なワーカーの 2 種類に分類する．各データセットについて，各ワーカーごとに鳥類データセットでは 75 枚すべて，イヌ属データセットでは 70 枚すべて，両生類・爬虫類データセットでは 100 枚すべての画像アイテムへの回答から Rayker ら [10] によるスパマースコアを計算する．その後，4.1 節での分析結果を参考に，鳥類データセットでは下位 24 人，イヌ属データセットでは下位 30 人，両生類・爬虫類データセットでは下位 36 人を悪質なワーカー，それ以外を良好なワーカーに分類する．

「ワーカーのスコアリング」では，画像の割り当て数を減らした状態でワーカーの能力を推定する．K を 5 以上 25 以下の整数として，鳥類データセットでは，108 人のワーカー全員に，75 枚の画像から抽出された同じ K 枚，イヌ属データセットでは，199 人のワーカー全員に，70 枚の画像から抽出された同じ K 枚，両生類・爬虫類データセットでは，118 人のワーカー全員に，ワーカー 118 人の正答率が 25%以上であった 80 枚の画像から抽出された同じ K 枚の画像のラベリング作業を依頼したという想定で，そのラベリング結果から各手法においてワーカーの能力を評価する．評価手法は，以下の六つを用意した．

- Acc: 全員での多数決の結果を正解としたときの正解率.
- Sp: Rayker ら [10] によるスパマースコア.
- Slc: Ipeirotis ら [9] によるソフトラベルコスト.
- Ps: ペナルティスコア（各アイテムのペナルティを (2) 式で定義）.
- PsD: ペナルティスコア（各アイテムのペナルティを (3) 式で定義）.
- GRM: 段階反応モデル.

段階反応モデルでは，3.2 節で説明されている $\vec{S}$ を，鳥類データセットでは $\vec{S} = (0, 0, 0, 1, 2)$，イヌ属データセットでは $\vec{S} = (0, 0, 0, 0, 1, 2, 3)$，両生類・爬虫類データセットでは $\vec{S} = (0, 0, 0, 0, 0, 0, 0, 1, 2, 3)$ とした．また，段階反応モデルについては，Luo ら [14] の論文で紹介されている実装を利用した．

「ランキング評価」では，スコアリングされたワーカをスコアの悪い順にソートし，スコアの悪い順に 1 位から順にワーカが悪質なワーカか良好なワーカかのいずれかを順に確認していく．この際，Average Precision を計算することによってランキングを評価する．スパマースコア，ソフトラベルコストの計算の際に，割り当てられていないラベルが存在した場合，混同行列において対応するラベルの行は対角成分を 1（正答率を 1）として計算する．

以上の 3 段階を，画像の割り当て枚数 K を変化させ，各 K について，割り当てる画像を選択する際の乱数のシード値を 10 回変更し，10 個の Average Precision の平均値 (MAP) で各手法を評価する．

4.3 実験 2:クラウドソーシングの状況に近い設定

よりクラウドソーシングの実際のシチュエーションに近い状況を想定した実験も行った．この実験は,「ワーカの分類」「多数決を行うチームの編成とタスクへの回答」「ワーカのスコアリング」「ランキング評価」の 4 段階に分けられる.「ワーカの分類」と「ランキング評価」は，4.2 節の実験のものと全く同じである．

「多数決を行うチームの編成とタスクへの回答」では，10 人 1 組のチームを，鳥類データセットでは 10 チーム，イヌ属データセットでは 19 チーム，両生類・爬虫類データセットでは 11 チーム作成する．ワーカーの選出と組み合わせはランダムである．各チームにそれぞれ異なる画像を複数枚割り当て，MTurk で得られた回答をもとに画像へのラベル付けを行う．

「ワーカーのスコアリング」は，各チームに対して，鳥類データセットでは 75 枚の画像アイテムセット，イヌ属データセットでは 70 枚の画像アイテムセット，両生類・爬虫類データセットではワーカー 118 人の正答率が 25%以上であった 80 枚の画像アイテムセットの中から一部分を割り当て，その回答内容に応じてスコアリングを行う．各チームごとに，以下の五つの手法でスコアリングを行う．

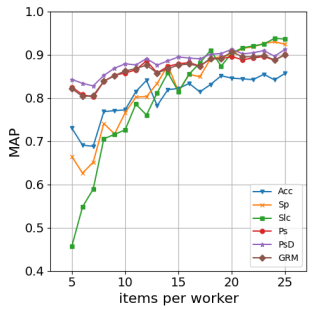
- Acc: チーム内での多数決による解を正解としたときの正解率.
- Sp: Raykar ら [10] によるスパマースコア.
- Slc: Ipeirotis ら [9] らによるソフトラベルのコスト計算.
- Ps: ペナルティスコア（各アイテムのペナルティを (2) 式で定義）.
- PsD: ペナルティスコア（各アイテムのペナルティを (3) 式で定義）.

割り当てられる画像はチームごとに異なる．画像の正解ラベルは未知とし，多数決で最も多い票を獲得したものをその画像のラベルの推定結果とする．最も票を多く獲得したクラスが複数ある場合は，今回は，4.1 節で定義したラベルの番号の若い順に優先度を定めておき推定を行う．スパマースコア，ソフトラベルコストの計算の際に，割り当てられていないラベルが存在した場合，混同行列において対応するラベルの行は対角成分を 1（正答率を 1）として計算する．

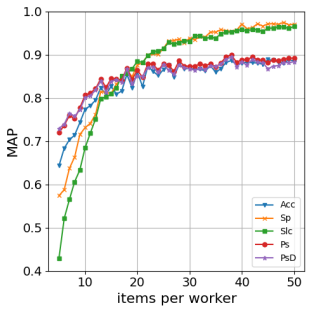
以上の四段階を，乱数のシード値を変更しながら 20 回行い，Average Precision の平均値 (MAP) で各手法を評価する．

4.4 実験結果と考察

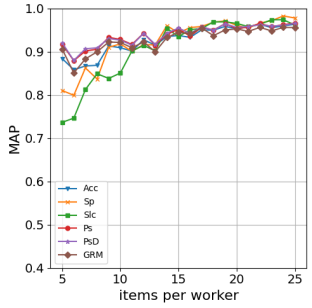
図 3c と表 1 は鳥類データセット，図 3a と表 2 はイヌ属データセット，図 3e と表 3 は両生類・爬虫類データセットでの実験 1 の結果である．また，図 3b と表 5 はイヌ属データセット，図 3d と表 4 は鳥類データセット，図 3f と表 6 は両生類・爬虫



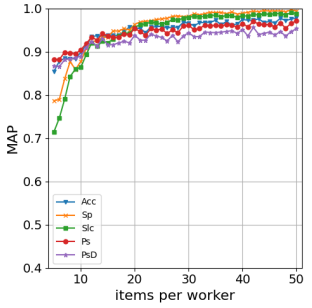
(a) 実験 1: イヌ属



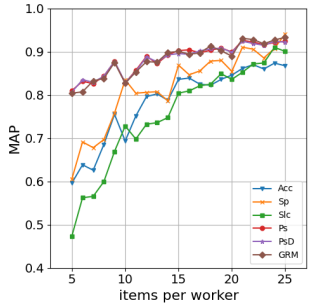
(b) 実験 2: イヌ属



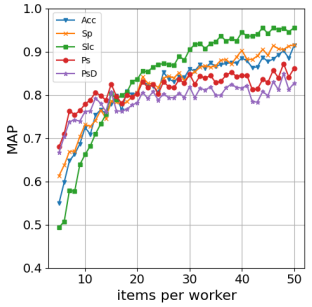
(c) 実験 1: 鳥類



(d) 実験 2: 鳥類



(e) 実験 1: 両生類・爬虫類



(f) 実験 2: 両生類・爬虫類

図 3　画像の割り当て枚数と MAP 値の変化

表 1　実験 1: MAP 値 (鳥類データセット)

K	Sp	Slc	Acc	Ps	PsD	GRM
5	0.8100	0.7365	0.8838	0.9171	**0.9203**	0.9058
10	0.9177	0.8503	0.9088	**0.9280**	0.9263	0.9212
11	0.9003	0.9016	0.9012	**0.9164**	0.9148	0.9087
12	0.9163	0.9143	0.9275	**0.9425**	0.9421	0.9223
13	0.9153	0.9017	0.9146	0.9147	**0.9168**	0.8995
14	**0.9599**	0.9538	0.9347	0.9398	0.9424	0.9324
15	0.9430	0.9343	0.9370	0.9515	**0.9544**	0.9445
16	**0.9554**	0.9506	0.9329	0.9358	0.9417	0.9413
17	**0.9580**	0.9534	0.9503	0.9563	0.9574	0.9544
20	0.9606	**0.9651**	0.9527	0.9530	0.9558	0.9524
21	0.9540	0.9576	0.9570	0.9567	**0.9581**	0.9466
25	**0.9775**	0.9614	0.9632	0.9657	0.9654	0.9556

類データセットにおける実験 2 の結果である．青色の折れ線が Acc の MAP，橙色の折れ線が Sp の MAP，緑色の折れ線が Slc の MAP，赤色の折れ線が Ps の MAP，紫色の折れ線が PsD の MAP，茶色の折れ線が GRM の MAP である．表中の値は，小数第 5 位以下を切り捨て，各画像割り当て枚数について，最も良い性能を示したものの数値を太字で記載している．

どの手法も，1 チームあたりの画像割り当て枚数が増えると，MAP の値が向上していくことが確認できる．これは，ワーカの回答データが増えるにつれて，ワーカに対する情報が増えた

表 2　実験 1: MAP 値 (イヌ属データセット)

K	Sp	Slc	Acc	Ps	PsD	GRM
5	0.6645	0.4571	0.7297	0.8248	**0.8427**	0.8212
10	0.7656	0.7261	0.7722	0.8573	**0.8787**	0.8622
15	0.8147	0.8144	0.8216	0.8787	**0.8944**	0.8758
16	0.8543	0.8548	0.8332	0.8814	**0.8918**	0.8783
17	0.8486	0.8805	0.8138	0.8734	**0.8901**	0.8730
18	0.8901	**0.9095**	0.8307	0.8900	0.9008	0.8911
19	0.8953	0.8733	0.8511	0.8906	**0.9023**	0.8911
20	0.8997	0.9034	0.8455	0.8954	**0.9124**	0.9061
21	0.9142	**0.9154**	0.8438	0.8879	0.9016	0.8948
22	0.9179	**0.9201**	0.8421	0.8921	0.9043	0.8947
23	**0.9258**	0.9240	0.8544	0.8954	0.9090	0.8978
24	0.9295	**0.9376**	0.8415	0.8871	0.8962	0.8875
25	0.9238	**0.9358**	0.8567	0.9002	0.9122	0.8991

表 3　実験 1: MAP 値 (両生類・爬虫類データセット)

K	Sp	Slc	Acc	Ps	PsD	GRM
5	0.6054	0.4735	0.5961	**0.8097**	0.8063	0.8038
6	0.6912	0.5621	0.6382	0.8315	**0.8350**	0.8070
7	0.6779	0.5654	0.6259	0.8257	**0.8290**	0.8322
8	0.6966	0.5994	0.6852	**0.8435**	0.8430	0.8379
9	0.7579	0.6692	0.7548	0.8770	**0.8775**	0.8748
10	**0.8347**	0.7280	0.6933	0.8276	0.8284	0.8265
15	0.8686	0.8040	0.8358	**0.9023**	0.8953	0.9001
20	0.8542	0.8358	0.8455	**0.8992**	0.8985	0.8896
21	0.9102	0.8523	0.8612	0.9259	0.9237	**0.9306**
22	0.9055	0.8712	0.8698	0.9220	0.9187	**0.9278**
23	0.8853	0.8737	0.8598	0.9158	0.9145	**0.9179**
24	0.9073	0.9097	0.8733	0.9210	0.9234	**0.9268**
25	**0.9409**	0.9005	0.8672	0.9239	0.9219	0.9330

表 4　実験 2: MAP 値 (鳥類データセット)

K	Sp	Slc	Acc	Ps	PsD
5	0.7864	0.7143	0.8541	**0.8810**	0.8674
6	0.7893	0.7462	0.8777	**0.8818**	0.8651
7	0.8372	0.7903	0.8846	**0.8984**	0.8812
8	0.8769	0.8425	0.8837	**0.8970**	0.8831
9	0.8585	0.8594	0.8832	**0.8950**	0.8859
10	0.8775	0.8636	0.8987	**0.9040**	0.8892
11	0.9072	0.8932	0.9148	**0.9186**	0.9098
12	0.9212	0.9201	**0.9343**	0.9334	0.9232
13	0.9263	0.9121	**0.9367**	0.9264	0.9126
14	0.9324	0.9239	0.9352	**0.9417**	0.9296
15	0.9337	0.9204	**0.9371**	0.9366	0.9160

表 5　実験 2: MAP 値 (イヌ属データセット)

K	Sp	Slc	Acc	Ps	PsD
5	0.5738	0.4297	0.6430	0.7195	**0.7294**
6	0.5878	0.5212	0.6833	0.7364	**0.7388**
7	0.6371	0.5652	0.7032	0.7590	**0.7646**
8	0.6633	0.6060	0.7138	0.7521	**0.7575**
9	0.7153	0.6336	0.7428	**0.7771**	0.7722
10	0.7310	0.6845	0.7736	**0.8061**	0.8005
11	0.7411	0.7184	0.7820	**0.8116**	0.8045
12	0.7616	0.7511	0.7947	**0.8207**	0.8185
13	0.8143	0.7981	0.8235	**0.8441**	0.8387
14	0.8128	0.8024	0.8075	**0.8238**	0.8145
15	0.8101	0.8091	0.8265	**0.8446**	0.8382

表 6　実験 2: MAP 値 (両生類・爬虫類)

K	Sp	Slc	Acc	Ps	PsD
5	0.6130	0.4936	0.5495	**0.6796**	0.6666
6	0.6372	0.5072	0.5980	**0.7100**	0.7035
7	0.6690	0.5792	0.6475	**0.7632**	0.7382
8	0.6698	0.5770	0.6630	**0.7540**	0.7420
9	0.7034	0.6389	0.6867	**0.7638**	0.7394
10	0.7315	0.6620	0.7247	**0.7786**	0.7618
11	0.7269	0.6822	0.7088	**0.7882**	0.7626
12	0.7410	0.7100	0.7533	**0.8058**	0.7927
13	0.7647	0.7330	0.7657	**0.7976**	0.7797
14	0.7457	0.7548	0.7566	**0.7878**	0.7615
15	0.7858	0.8031	0.7794	**0.8247**	0.8066

ため，より精度の高いスコアリングができるようになったからであると考えられる．割り当て枚数が増加するにつれて，提案手法よりもスパマースコア，ソフトラベルコストによるランキングのほうが精度が高い傾向がみられる．これは実験の「ワーカの分類」のステップで混同行列を用いた手法であるスパマー

スコアを基準にワーカこれらの手法が有利になるように実験の設定をしているからと考えられる．実験 1，実験 2 の両方において，提案手法の間では大きな性能の差はみられなかったが，既存手法と比較すると，画像の割り当て枚数が少ないときに提案手法のほうが性能が高い傾向がみられる．

アイテムの割り当て数が少ないときになぜ提案手法のほうが精度が高かったかを考察する．多数決による解を正解としたときのワーカの正解率でワーカをランキングする手法では，K 個のアイテムが割り当てられたとき，正解率は $0, 1/K, ..., 1$ の $K+1$ 個の値のいずれかとなり，K が小さいときに粗い評価となる．提案手法では，多数決による票の割れ方を考慮してワーカにペナルティを付与しているため，K の値が小さくても，$K+1$ 段階よりも細かい段階でワーカを評価することができるようになり，より高い精度で推定できるようになったのではないかと推測する．

混同行列を用いた推定手法では，アイテムの分類の難しさを考慮できていないのが問題である．例えば，イヌ属データセットにおいて，分類が容易な German Shepherd と Samoyed を全て Samoyed とラベリングし，それ以外は全て真のラベルと同じラベルを付与した場合と，分類が難しい Alaskan Malamute と Siberian Husky を全て Siberian Husky とラベリングし，それ以外は真のラベルと同じラベルを付与した場合を考える．この二つの場合において混同行列からスパマースコアを計算するとスパマースコアは全く同じ値となる．ソフトラベルコストの計算では，真のラベルの数がどのクラスでも同じときに，全く同じ値となる．提案手法では，真面目にタスクを遂行していても誤分類しやすいアイテムと，真面目にタスクを遂行していれば誤分類しづらいアイテムを区別でき，推定に活用できるため，アイテムの割り当て数が少ない時でもスパマー検知の精度が高くなったと考えられる．

クラウドソーシングにおいてどのように提案手法を応用できるかを考察する．ワーカにラベルを付与してもらいたいアイテムが大量にある場合，一度に全てのアイテムにラベルを付与してもらうのではなく，タスクを適当なサイズに分割し，複数回に分けてタスクを投稿することが適切と考えられる．提案手法は，アイテム数の割り当て数が少ないときに精度が高くなるため，小さいサイズのタスクを複数回投稿する場合と相性が良いと考えられる．

5 おわりに

本研究では，クラウドソーシングにおけるマルチクラス分類タスクにおいて，多数決における票の割れ方に着目したスパマー検知手法を提案した．同じアイテムにラベルを付与したワーカの回答の票の割れ方のデータを利用し，多数派の回答から逸脱するほどペナルティが大きくなる「ペナルティスコア」手法や，段階反応モデルを用いたモデルによってワーカの能力を推定した．これにより，混同行列を用いた手法の問題点であった，アイテムの真のラベルが偏っている場合や，アイテムの分類の難易度による影響を小さくし，アイテムの割り当てが少ない場合にワーカの能力を高い精度で推定することができた．

謝　辞

本研究は JST CREST (JPMJCR16E3) の支援を受けたものである．

文　献

[1] Rion Snow, BrendanO ’connor, Dan Jurafsky, Andrew Y Ng. Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, pp. 254–263, 2008.

[2] Alexander Philip Dawid and Allan M Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, Vol. 28, No. 1, pp. 20–28, 1979.

[3] Manas Joglekar, Hector Garcia-Molina, and Aditya Parameswaran. Evaluating the crowd with confidence. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 686–694, 2013.

[4] Manas Joglekar, Hector Garcia-Molina, and Aditya Parameswaran. Comprehensive and reliable crowd assessment algorithms. In *2015 IEEE 31st International Conference on Data Engineering*, pp. 195–206. IEEE, 2015.

[5] Pinar Donmez, Jaime G Carbonell, and Jeff Schneider. Efficiently learning the accuracy of labeling sources for selective sampling. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 259–268, 2009.

[6] Jacob Whitehill, Ting-fan Wu, Jacob Bergsma, Javier Movellan, and Paul Ruvolo. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. *Advances in neural information processing systems*, Vol. 22, pp. 2035–2043, 2009.

[7] Yoram Bachrach, Thore Graepel, Tom Minka, and John Guiver. How to grade a test without knowing the answers—a bayesian graphical model for adaptive crowdsourcing and aptitude testing. *arXiv preprint arXiv:1206.6386*, 2012.

[8] Ujwal Gadiraju, Ricardo Kawase, Stefan Dietze, and Gianluca Demartini. Understanding malicious behavior in crowdsourcing platforms: The case of online surveys. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pp. 1631–1640, 2015.

[9] Panagiotis G Ipeirotis, Foster Provost, and Jing Wang. Quality management on amazon mechanical turk. In *Proceedings of the ACM SIGKDD workshop on human computation*, pp. 64–67, 2010.

[10] Vikas C Raykar and Shipeng Yu. Eliminating spammers and ranking annotators for crowdsourced labeling tasks. *The Journal of Machine Learning Research*, Vol. 13, No. 1, pp. 491–518, 2012.

[11] Fumiko Samejima. Estimation of latent ability using a response pattern of graded scores. *Psychometrika monograph supplement*, 1969.

[12] 豊田秀樹. 項目反応理論 [入門編][第 2 版]. 朝倉書店, 2012.

[13] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010.

[14] Yong Luo and Hong Jiao. Using the stan program for bayesian item response theory. *Educational and Psychological Measurement*, Vol. 78, No. 3, pp. 384–408, 2018.