DBSJ Japanese Journal, Vol.13-J, No.2
February 2015

クラウドソーシングの回答品質向上のための既知ラベル数決定手法

An Approach to Deciding the Appropriate Number of Ground Truths to be Inserted into Labelling Tasks in Crowdsourcing for Accuracy

久保田 琢也♡ 眞鍋 雄貴♢ 有次 正義♠

Takuya KUBOTA Yuki MANABE
Masayoshi ARITSUGI

クラウドソーシングによるラベリングタスクにおいて，ラベリングの精度は重要な問題の 1 つである．これまで，各回答者の正答率に基づいて，回答結果から正しいラベルを推定する研究がなされてきた．しかし，回答者集団の多数がスパマーで占められている状況では正確にラベル推定を行うことが困難である．ラベルの推定精度の向上のために回答データ中にラベルの真値が既知のデータを加える手法がある．だが，用いる既知ラベル数が少なくては推定精度を向上させることができない．一方で多くの既知ラベルを用いることはコストの増大につながる．本論文では，用いるべき既知ラベルの適切な数を推定する手法を提案する．

Labelling accuracy is an important problem in labelling tasks using crowdsourcing. There have been many studies of estimating true labels based on quality of labellers. However, it is still difficult to estimate consensus labels correctly when most of labellers are spammers. There are methods of improving labelling accuracy using ground truths. Note that if the number of ground truths to be inserted into labelling tasks is small, it would not be possible to improve the accuracy. In this paper, we propose an approach to deciding the appropriate number of ground truths to be inserted into labelling tasks.

1. はじめに

近年，インターネットを通じて不特定多数のクラウド (群集) にタスクを依頼して解決してもらうクラウドソーシングが世界規模で使われている [1]．Amazon Mechanical Turk をはじめとするクラウドソーシングサービスを使って解決するタスクの 1 つとして，画像の分類分けなどをするラベリングタスクがあげられる．ラベリングを専門家に依頼する手法は時間とコストがかかることが課題であるが，クラウドソーシングすることで安価で，時間をかけずにラベリングできるようになる [2]．データにラベリングすることにより，例えば分類器の教師データを作成することができる [3]．

クラウドは熟達者，初心者，スパマーなどを含む様々な人々の集合として成り立っているため，回答が必ず正しいという保証はない．正しい回答を得るために，複数の回答者から回答を得てラベルの真値を推定する研究が行われてきた [4][5][6]．これらの研究では，各回答者の回答の傾向を探った上でラベルの真値を推定している．複数の回答から正答を単純な多数決によって決める手法では，回答者の半数以上が正しい回答を行うという前提がなければ真値を推定することはできないためである．

高精度にラベリングする手法の 1 つとして，EM アルゴリズムを用いる Dawid と Skene が提案した手法がある [7]．これは，各回答者の正答率の推定と，正答率の高い回答者の回答を重要視したラベルの真値の推定とを EM アルゴリズムを用いて交互に行う手法である．その研究は，元来医療分野に用いられるように提案されたが，以下に述べる研究等においてクラウドソーシングの分野に応用されている．Yan らは，データ毎によって各回答者の正答率の値を変えてラベルの真値を推定する手法を提案した [4]．また，Ipeirotis らは，各回答者の正答率をもとに，各回答者の回答の有用性を点数化する手法を提案した [5]．有用性を点数化することにより，ランダムに回答するスパマーはほとんど誤答する回答者よりも有用性が低いことを示した．Raykar と Yu は，回答者がスパマーである可能性を点数化することにより，回答者の集団からランダムにラベリングを行うスパマーを検出して，スパマーを回答者集団から除外することでラベルの真値をスパマーのいない環境で推定する手法を提案した [6]．以上の既存研究により高い精度でスパマーを除外しラベルの真値を推定できるようになった．しかし，回答者集団に占めるスパマーの割合が高くなるほど，スパマーを検出できる精度が下がり，ラベルの真値をうまく推定できなくなるという問題がある (3 章を参照)．

また，EM アルゴリズムを用いたラベリングの精度を向上することを目的に，ラベルの真値が既知のデータ (以下，既知ラベルと呼称する) を用いる研究がある．既知ラベルを用いることにより，各回答者の正答率をより正確に推定できるようになるため，ラベルの真値が未知のデータ (以下，未知ラベルと呼称する) のラベルの真値をより正確に推定できるようになるためである．Kajino らは回答者集団の中に完璧な回答をするエキスパートを導入することによってラベルの推定精度を向上する手法を提案した [3]．また，Tang と Lease は，未知ラベルと既知ラベルの両方を用いて正答率を推定することにより，ラベルの推定精度を向上させる手法を提案した [8]．Tang と Lease の手法を用いることでスパマーを検出できる精度が上がり，ラベルの真値の推定精度を向上できる．しかし，既知ラベルを多くタスクに加えることはコストの増大につながる．この問題を解決するためには，既知ラベル挿入数を適切に設定する必要がある．

本論文では，これらの問題を解決するために，適切な既知ラベル挿入数を決定する手法を提案する．併せて，Condorcet の陪審定理 [9] を拡張して，ラベル推定精度の上限を計算する手法も提案する．また，提案手法を用いて，ラベルの推定精度がよくなるか評価する．本論文の貢献は，以下の 3 つである．

- ラベル推定精度の上限を計算する手法を提案する．Condorcet の陪審定理を EM アルゴリズムに適用できるよう拡張することにより，EM アルゴリズムによるラベルの真値推定精度の上限を求めることができる．
- 適切な既知ラベル挿入数を決定する手法を提案する．提案手法では，はじめに各回答者の正答率からラベル推定精度の上限を計算し，許容精度を決める．次に，各回答者の正答率をもとに，既知ラベル挿入数を 0 個として作成した回答候補からラベルの真値を推定した場合の精度を許容精度と比較し，許容精度に達していない場合は，許容精度に達するまで既知ラベル挿入数を増やしてラベルの真値を推定する．
- これらの提案手法を人工データを用いて評価する．提案手法が，ラベル推定精度を許容精度まで改善するために必要な既知ラベル挿入数を求めることができることを示す．

本論文は以下のように構成される．次章で，準備として本研究で用いた既存研究を紹介する．はじめに Raykar と Yu の既存研究を，回答者の正答率をモデル化する手法，ラベルの真値の推定手法，スパマーの検出手法を紹介する．その後，Condorcet の陪審定理を紹介する．3 章で，ラベル推定精度の上限を求める手法

♡ 学生会員 熊本大学大学院自然科学研究科
kubota@dbms.cs.kumamoto-u.ac.jp
♢ 正会員 熊本大学大学院自然科学研究科
y-manabe@cs.kumamoto-u.ac.jp
♠ 正会員 熊本大学大学院自然科学研究科
aritsugi@cs.kumamoto-u.ac.jp

と，適切な既知ラベル挿入数を求める手法を提案する．4 章で，提案した手法の評価を行う．5 章で，まとめと今後の課題を述べる．

2. 準備

各回答者の正答率とラベルの真値とを推定する手法 [7] を，スパマーを検出し除外することができるように拡張した Raykar と Yu の手法 [6] を本研究で用いる．それに加え，Condorcet の陪審定理 [9] を拡張して用いる．この章では，2.1 節で回答者の正答率をモデル化する手法を紹介し，2.2 節で正答率の尤度関数と事前確率分布を定義する．さらに，2.3 節で EM アルゴリズムを用いたラベルの真値と正答率の推定手法を紹介し，2.4 節でスパマー除外手法を紹介する．これらは，Raykar と Yu が用いた手法である．加えて，2.5 節で Condorcet の陪審定理を紹介する．

2.1 回答者の正答率のモデル化

本論文では，各データに与えたいラベルは二択である状況を想定し，ラベルの値を 0 または 1 と定義する．ラベルの真値が未知の N 個のデータが与えられ，データ $i \in \{1,...,N\}$ に対して未知の真のラベル $y_i \in \{0,1\}$ が存在する．回答者は M 人与えられ，データ i に対して回答者 $j \in \{1,...,M\}$ が $y_i^j \in \{0,1\}$ という形式で回答ラベルを与える．回答者 j が持つパラメータは，下記の式でモデル化される．

$$\alpha^j = Pr(y_i^j = 1 \mid y_i = 1), \beta^j = Pr(y_i^j = 0 \mid y_i = 0) \tag{1}$$

すなわち，α^j が真のラベルの値が 1 である場合の正答率であり，β^j が真のラベルが 0 である場合の正答率である．

スパマーはランダムな回答をすると考えられるため [6]，スパマーは回答ラベル y_i^j を真のラベル y_i に関係なく与える．すなわち，下記の式が成り立つ．

$$Pr(y_i^j = 1 \mid y_i = 1) = Pr(y_i^j = 1 \mid y_i = 0) \tag{2}$$

式 (2) を式 (1) を用いて式変形すると，$\alpha^j + \beta^j - 1 = 0$ となる．したがって，

$$\alpha^j + \beta^j - 1 \tag{3}$$

の値が 0 に近いほどスパマーである可能性が高いことがいえる．

2.2 尤度関数と事前確率分布

各回答者のパラメータに関する尤度関数と事前確率分布を定義し，ベイズの定理を応用した最大事後確率推定法を用いてパラメータを推定する．計算を簡単にするため，尤度関数の対数をとった対数尤度関数を尤度関数として用いる．

M 人の全回答者が N 個の全データに回答した回答データを $D = \{y_i^1,...,y_i^M\}_{i=1}^N$，ラベルの真値の集合を $\boldsymbol{y} = \{y_1,...,y_N\}$，全データに占めるラベルの真値が 1 であるデータの割合を p，各回答者のパラメータの集合と前述の p を $\theta = \{\alpha^1,\beta^1,...,\alpha^M,\beta^M,p\}$，とすると，対数尤度関数は，以下の式で与えられる．

$$\log Pr(D, \boldsymbol{y} \mid \theta) = \sum_{i=1}^{N} \{y_i \log pa_i + (1 - y_i) \log(1 - p)b_i\} \tag{4}$$

ただし，$a_i = \prod_{j=1}^{M} Pr(y_i^j \mid y_i = 1, \alpha^j) = \prod_{j=1}^{M} (\alpha^j)^{y_i^j}(1 - \alpha^j)^{1-y_i^j}$，$b_i = \prod_{j=1}^{M} Pr(y_i^j \mid y_i = 0, \beta^j) = \prod_{j=1}^{M} (\beta^j)^{1-y_i^j}(1 - \beta^j)^{y_i^j}$ である．

次に，各回答者のパラメータの分布を与える事前確率分布を定義する．回答者集団からスパマーを検出したいため，回答者がスパマーである可能性を示す式 (3) に関する正規分布として確率分布を与える．回答者集団のほとんどがスパマーである環境を想定して平均を 0，分散の逆数を表す精度パラメータを λ^j として，以下の式で定義する．

$$Pr(\alpha^j, \beta^j \mid \lambda^j) = \frac{1}{N(\lambda^j)} exp(-\frac{\lambda^j(\alpha^j + \beta^j - 1)^2}{2}) \tag{5}$$

ただし，$N(\lambda^j)$ は正規化関数であり，$N(\lambda^j) = \int_0^1 \int_0^1 exp(-\frac{\lambda^j(\alpha^j+\beta^j-1)^2}{2}) d\alpha^j d\beta^j$ で与えられる．全データ中のラベルの真値が 1 であるデータの出現率を考慮したベータ関数を $Beta(p \mid p_1, p_2)$，各回答者のパラメータ θ を制御するパラメータを $\boldsymbol{\lambda} = \{\lambda^1,...,\lambda^M, p_i, p_2\}$ と置くと，各回答者のパラメータの事前確率分布は以下の式で与えられる．

$$Pr(\theta \mid \boldsymbol{\lambda}) = Beta(p \mid p_1, p_2) \prod_{j=1}^{M} Pr(\alpha^j, \beta^j \mid \lambda^j) \tag{6}$$

2.3 EM アルゴリズム

最大事後確率推定を，対数尤度関数の式 (4) と事前確率分布の式 (6) の対数を取った式を用いた以下の式の値を最大にする θ を推定することで行う．

$$\log Pr(D \mid \theta) + \log Pr(\theta \mid \boldsymbol{\lambda}) \tag{7}$$

欠損データがある際に最尤法を用いて解とパラメータを反復して推定する EM アルゴリズムを，最大事後確率推定法を用いて反復推定するように応用することで各回答者のパラメータ推定とラベルの真値の推定を行う．この計算は尤度関数の期待値を求める E ステップと尤度関数の期待値を最大化するパラメータの値を求める M ステップを，値の変化が収束するまで反復実行して行う．

- E ステップ
  あるデータ i の真値が 1 である確率を $\mu_i = Pr(y_i = 1 \mid y_i^1,...,y_i^M, \theta)$ とすると，対数尤度関数の式 (4) の期待値は

  $$\sum_{i=1}^{N} \{\mu_i \log pa_i + (1 - \mu_i) \log(1 - p)b_i\} \tag{8}$$

  で表され，μ_i はベイズの定理を用いて以下の式で更新する．

  $$\mu_i = \frac{a_i p}{a_i p + b_i(1 - p)} \tag{9}$$

  E ステップで求めた μ_i を用いて，ラベルの真値は，μ_i が 0.5 未満であれば 0，0.5 以上であれば 1 が推定されたと考えることができる．既知ラベルの場合，式 (9) の推定を行わずに $\mu_i = y_i$ に固定する．これによって，M ステップにて回答者の正答率を誤って推定することを防ぐ．
- M ステップ
  式 (7) の値が最大になるようなパラメータを推定する．対数関数であるため，式 (7) を各パラメータに関して微分して 0 になるときのパラメータを求めればよい．p は式 (10) によって求めることができる．α^j と β^j は式 (11) と式 (12) を 2 元 3 次方程式として解くことで求められる．

  $$p = \frac{p_1 - 1 + \sum_{i=1}^{N} \mu_i}{p_1 + p_2 - 2 + N} \tag{10}$$

  $$\lambda^j(\alpha^j)^3 + (\beta^j - 2)\lambda^j(\alpha^j)^2 + (\lambda^j - \beta^j\lambda^j - \sum_{i=1}^{N} \mu_i)\alpha^j + \sum_{i=1}^{N} \mu_i y_i^j = 0 \tag{11}$$

  $$\lambda^j(\beta^j)^3 + (\alpha^j - 2)\lambda^j(\beta^j)^2 + \{\lambda^j - \alpha^j\lambda^j - \sum_{i=1}^{N}(1 - \mu_i)\}\beta^j + \sum_{i=1}^{N}(1 - \mu_i)(1 - y_i^j) = 0 \tag{12}$$

2.4 スパマー除外手法

式 (5) の事前確率分布の精度パラメータ λ^j を制御することにより，式 (3) が 0 に近い値を取る確率を制御することができる．λ^j の値が小さければ式 (3) が 0 に近い値を取る確率が低くなり，スパマーと判断される可能性が低くなる．一方，λ^j の値が大きければ式 (3) が 0 に近い値をとる確率が高くなり，スパマーと判断さ

れる可能性が高くなる．したがって，良い回答者の精度パラメータ λ^j の値は小さく，スパマーの精度パラメータ λ^j の値は大きく設定したい．そのため，各回答者ごとに適切な λ^j の値を定める必要がある．また，λ^j の値を定めることにより，この値が大きければスパマーであると判断することができる．

λ^j はパラメータ θ を制御するパラメータであるため，経験ベイズ法を用いて下記の周辺尤度を最大にする λ^j を求める．

$$Pr(D \mid \boldsymbol{\lambda}) = \int_{\theta} Pr(D \mid \theta) Pr(\theta \mid \boldsymbol{\lambda}) d\lambda \tag{13}$$

最大事後確率推定によって推定されたパラメータ α^j，β^j，θ をそれぞれ $\hat{\alpha}^j$，$\hat{\beta}^j$，$\hat{\theta}$ と表記すると，λ^j は以下の式で求められる．

$$\lambda^j = \frac{\delta^j}{(\hat{\alpha}^j + \hat{\beta}^j - 1)^2 + \sigma^j} \tag{14}$$

ただし，$\delta^j = 2 - \frac{\sqrt{2\pi\lambda^j} erf(\sqrt{\lambda^j/2})}{\sqrt{2\pi\lambda^j} erf(\sqrt{\lambda^j/2}) + 2exp(-\lambda^j/2) - 2}$，$\sigma^j = Tr(\boldsymbol{H}^{-1}(\hat{\theta}, \boldsymbol{\lambda}) \frac{\partial}{\partial \lambda^j} \boldsymbol{H}(\hat{\theta}, \boldsymbol{\lambda}))$，$erf(x) = 2/\sqrt{\pi} \int_0^x (-t^2) dt$ であり，$\boldsymbol{H}$ はヘッセ行列である．

EM アルゴリズムを用いた最大事後確率推定によってパラメータを推定した後，パラメータをもとに λ^j の値を更新し，λ^j の値が一定以上の回答者をスパマーと判定して回答者集団から除外し，更新後の λ^j の値を用いて EM アルゴリズムによりパラメータを推定する．

2.5 Condorcet の陪審定理

Condorcet の陪審定理とは，多数決で物事を決める際，各投票者が正しい判断をする確率が 50%より高ければ投票者の人数が多いほど多数決により正しい判断が下される確率が高くなる定理である [9]．各投票者の正答率を全員同じ値 p_c，投票者数を n とすると多数決によって正しい判断が下される確率は，n が奇数である場合は

$$\sum_{k=(n+1)/2}^{n} {}_nC_k p_c^k (1 - p_c)^{n-k} \tag{15}$$

で与えられ，n が偶数である場合は

$$\sum_{k=n/2+1}^{n} {}_nC_k p_c^k (1 - p_c)^{n-k} + \frac{1}{2} {}_nC_{(n/2)} p_c^{n/2} (1 - p_c)^{n/2} \tag{16}$$

で与えられる．この式は n 人の投票者による回答として起こりうる 2^n 通りの組合せのうち多数決により正しい判断が下される組合せを選び出し，それらの組合せが起こる確率の和を求めることによって多数決で正しい判断が下される確率を計算している．

3. 提案手法

本論文の提案手法でははじめに，Condorcet の陪審定理を，多数決でなく EM アルゴリズムによってラベルを推定する手法に適用できるように拡張する．これにより，ラベル推定精度の上限を求める．次に，ラベル推定精度が上限に達するまで既知ラベル数を徐々に増やしながら，ラベルの推定を行う．

3.1 ラベル推定精度の上限

Condorcet の陪審定理を EM アルゴリズムによるラベルの推定手法に適用できるよう拡張する．ある回答者 j が正しい判断をする確率を q^j とすると $q^j = \alpha^j p + \beta^j (1 - p)$ である．ここで，あるデータ i に対して正しい回答をすると 1，間違った回答をすると 0 になる変数 r_i^j を導入する．すると，n 人の回答者による回答として起こりうる組合せのうち，EM アルゴリズムによるラベル推定アルゴリズムによって正しい判断が下される組合せは以下の式の値が 0.5 以上になる組合せである．

$$\frac{\prod_{j=1}^{n} (q^j)^{r_i^j} (1 - q^j)^{1 - r_i^j}}{\prod_{j=1}^{n} (q^j)^{r_i^j} (1 - q^j)^{1 - r_i^j} + \prod_{j=1}^{n} (1 - q^j)^{r_i^j} (q^j)^{1 - r_i^j}} \tag{17}$$

また，各回答者の回答がその組合せになる確率は

$$\prod_{j=1}^{n} (q^j)^{r_i^j} (1 - q^j)^{1 - r_i^j} \tag{18}$$

である．正しい判断が下される全ての組合せにおいて，式 (18) の値を求める．それらの和を求めることによって，ラベル推定精度の上限を求めることができる．

上記の手法で求められるラベルの推定精度の上限を用い，Raykar と Yu によって提案された手法を評価した．正答率が 8 割の回答者とスパマーによって構成される回答者集団に 100 個のデータに回答させる実験を，人工データを用いて 100 回実施した．その結果を表 1 に示す．

8 割正解する 5 人の回答者がいる環境で推定精度の上限に達するためには，スパマーが 100 人の場合は間違って推定されるラベル数をあと 9.6 個減らさなければならない．これに対し，スパマーが 0 人の場合はあと 1.9 個である．このことから，回答者集団にスパマーが多い場合は推定精度が落ちることがわかる．

また，8 割正解の回答者が 1 人いる環境，3 人いる環境，5 人いる環境を比較すると，8 割正解する回答者の人数が少ない環境では，多い環境と比べて推定精度が下がることがわかる．

これらのいずれの場合も，ラベルの推定精度の上限には達していない．そのため，まだ改善の余地があるといえる．

3.2 既知ラベル挿入数を決める手法

適切な既知ラベル挿入数を決めるために，既知ラベル挿入数を変えながらラベル推定精度を測る．

ここで，既知ラベル挿入数を決める前に，ラベル推定精度の上限に対する許容誤差を予め決めておく．ラベル推定精度が上限に達した際にラベリングを誤ると推定されるデータ数を s 個，許容誤差を t% とすると，ラベリングを誤る個数が $(1 + t/100)s$ 個に達した際に許容する精度に達したとする．許容誤差は，タスクの依頼者がラベルに求める精度によって適切に定めるものとする．

はじめに，各回答者の過去の作業履歴から作業タスク数に占める正答タスク数をラベルの推定値ごとに計算し，正答率 α^j と β^j の値を決定する．過去の作業履歴がない作業者においては，初めはスパマーと仮定して $\alpha^j + \beta^j - 1 = 0$ となるような α^j と β^j を設定する．作業履歴のない回答者をスパマーと仮定することによって，仮に彼らの一部または全員がスパマーでなければラベルの推定値の精度が良い結果が得られるのと同時に，仮に全員が本当にスパマーであるならば全員を排除することが可能になるためである．

次に，各回答者の正答率と，ラベリングしたいデータの数とをもとに，各回答者から期待される回答データの候補を作成する．ラベリングしたいデータの r% は，ラベルの真値が 1 であるものとして，各回答者が持つ固有の正答率 α^j をもとに当該データにラベリングを行う．データの残りの $1 - r$% も同様にラベルの真値が 0 であるものと仮定して各回答者の β^j をもとにラベリングを行う．r の値は，データの種類によって適切に定めるものとする．

それから，この回答候補データをもとにラベルの真値を推定する．推定結果をラベリングしたいデータに仮に置いたラベルの真値と比較し，ラベル推定精度と許容する精度とを比較する [1]．ラベルの推定精度が許容する精度に達していない場合はデータに既知ラベルを 1 つ加えたうえで再び回答候補データを作成する．これをラベルの推定精度が許容する精度に達するまで繰り返す．

ラベル推定精度が許容する精度に達した時点における，データに加えた既知ラベル数が，ラベルの推定精度を改善するために必要な最小の既知ラベル数である．この結果をもとに，実際のクラウドソーシングサービスで各回答者に回答を依頼する際に推定精

[1] ただし，この一連の手法において，各データに仮に定めたラベルの真値は真値推定のアルゴリズムにおいて真値が未知であるものとして計算し，新たに加えた既知ラベルは真値が既知であるものとして計算する．

表 1: 既存研究をもとにした評価結果

8 割正解の回答者数	1	3	5	5
5 割正解のスパマー数	24	75	100	0
正しく推定されたラベル数の平均	50.9	63.6	84.7	92.4
正しく推定されるべきラベル数の上限	80	89.6	94.3	94.3

表 2: 正答率が異なる 5 人の回答者による評価での 5 人の正答率

回答者番号	α^j	β^j
1	0.9	0.9
2	0.8	0.8
3	0.7	0.8
4	0.8	0.7
5	0.7	0.7

度が最適になる既知ラベル数をラベリングタスクの依頼者に提案することができる．

4. 評価

3 章で評価したように，スパマーが多い環境ではスパマーがいない環境と比較して既存の手法ではラベルの推定精度が劣る．スパマーが多く含まれる環境において，提案手法を用いることでラベルの推定精度を上げることができるのか，またスパマーが多い環境と比べて推定精度が高い環境であるスパマーがいない環境において，提案手法によってさらに推定精度を向上させることができるのかを調べるために評価を行った．

スパマーがいない環境での評価では，各回答者の正答率にばらつきがある場合と全く同じ場合で結果に変化があるのかを見るために 2 つの条件を設定した．

本章におけるすべての評価は人工的な環境で実験した．回答者は実在する人物ではなく計算機上で人工的に作成した回答者を用い，回答データは正答率に基づき人工的に作成した．また，ラベリングしたいデータ中のラベルの真値が 1 である割合を 50%，ラベル推定精度の上限に対する許容誤差を 10% と設定して評価した．

4.1 スパマーでない 5 人の回答者のみの環境での評価

まずはじめに，スパマーではない回答者 5 人が，100 個のデータにラベリングをする設定で評価を行った．5 人の正答率を表 2 の通り設定した．この 5 人によるラベルの推定精度の上限は 0.9345 である．100 個のデータにラベリングするとき，平均で 6.55 個のデータに付けられたラベルが誤りである場合に，推定精度が最良であると考えることができる．

推定精度が最良であるときの誤推定数は 6.55 個である．予め設定した許容誤差が 10% であるため，誤って推定されるラベルの個数が 7.205 個以下になるまで既知ラベル数を徐々に増やすことで評価を行った．図 1 はその結果を示したグラフである．ただし，許容する推定精度まで改善された後も，既知ラベル数を増やして評価した．ラベルの真値が誤って推定された数をグラフの縦軸，加えた既知ラベル数が横軸である．加えた既知ラベル数それぞれについて 100 回評価した際の平均を示しており，エラーバーは標準偏差である．また，ラベル推定精度が最良であるときの誤推定数を縦軸と垂直な破線で示した．

その結果，既知ラベル挿入数が 23 個のときに許容する推定精度にまで達した．このことから，既知ラベルを適切な数加えることによりラベルの推定精度に改善がみられることがわかる．しかし，既知ラベルを加える数が少ない場合は全く加えない場合と比較して推定精度が悪くなった．ただし，推定精度が悪くなった状況下の誤推定数の標準偏差が大きくなっていることから，100 回の評価における誤推定数にばらつきが多いことがわかる．

加えた既知ラベル数が 23 個のときのラベルの誤推定数の平均が 6.94 個であったのに対して，加えた既知ラベル数が 0 個のときのラベルの誤推定数の平均が 7.61 個であった．その差は 0.67 個であったため，既知ラベルを加えることによって発生するコストを考えたうえで，23 個の既知ラベルを用いる必要があるかをラベリングタスクの依頼者が判断する必要がある．

また，既知ラベル数が 24 個以上 100 個以下の環境での誤推定数の平均は 7.10 個であり，ラベル推定精度が最良であるときの誤推定数との誤差は 8.40%であった．また，誤推定数がラベル推定精度が最良であるときの誤推定数を下回ったケースは既知ラベル数が 100 個の際の誤推定数 6.24 個の時のみであり，その際のラベル推定精度が最良であるときの誤推定数との誤差は 4.73%であった．

次に，全回答者が同じ正答率である条件で評価を行った．スパマーでない 5 人の回答者の正答率は全員同じ値 $\alpha^j = \beta^j = 0.8$ とした．ラベルの推定精度の上限は 0.94208 である．100 個のデータにラベリングする評価を複数回行ったとき，平均して 5.792 個の誤りがある場合，推定精度は最良である．前述と同様に 10%の誤差を許容し，誤推定ラベル数が 6.3712 個以下に減少するまで既知ラベル数を増やして評価を行った．その結果が図 2 である．ただし，許容する推定精度に達した後も既知ラベル数を増やして評価を行った．

すると，既知ラベル挿入数が 2 個の時に許容する推定精度まで達した．各回答者の正答率にばらつきをつけた前回の評価結果と異なり，許容する誤差に達するまでは，ラベルの誤推定数は既知ラベル挿入数が増えるごとに単調減少した．

また，既知ラベル数が 3 個以上 100 個以下の場合において，誤推定数の平均は 6.04 個であり，ラベル推定精度が最良であるときの誤推定数との誤差は 4.28%であった．また，誤推定数がラベル推定精度が最良であるときの誤推定数を下回ったケースは 97 例中 26 例あり，最も誤推定数が少なかったケースは既知ラベル挿入数が 88 個の時の誤推定数 5.35 個であり，ラベル推定精度が最良であるときの誤推定数との誤差は 7.63%である．

この 2 つの結果を比較すると，スパマーがいないという環境では既知ラベル挿入数を増やした際の誤推定数の推移の仕方は回答者集団によって異なることがわかる．したがって，回答者集団ごとにシミュレーションを行い，データに加えるべき既知ラベル数を決める手法をとることが有効である．また，いずれの場合も，許容する精度に達した後も既知ラベル挿入数を増やすとラベル推定精度が上限まで改善できた．

4.2 作業履歴がない回答者による影響

次に，スパマーでない回答者が 1 人だけの環境に，作業履歴がない回答者を追加するとどのような影響が出るのかを調べるために評価を行った．スパマーでない回答者の正答率を $\alpha^j = \beta^j = 0.8$ と設定し，作業履歴がない回答者の人数を変えながら評価した．スパマーと仮定する，作業履歴がない回答者の人数と正答率は表 3 に沿って設定した．表 3 の 2 行目は α^j の値，3 行目は β^j の値であり，これらが 3 行目以降の，3 列目から 11 列目に記されている各回答者の正答率を定義している．2 列目が作業履歴のない回答者の合計人数であり，その同じ行の 3 列目以降に記された人数がその内訳である．

スパマーでない 1 人の回答者の正答率から求められるラベルの

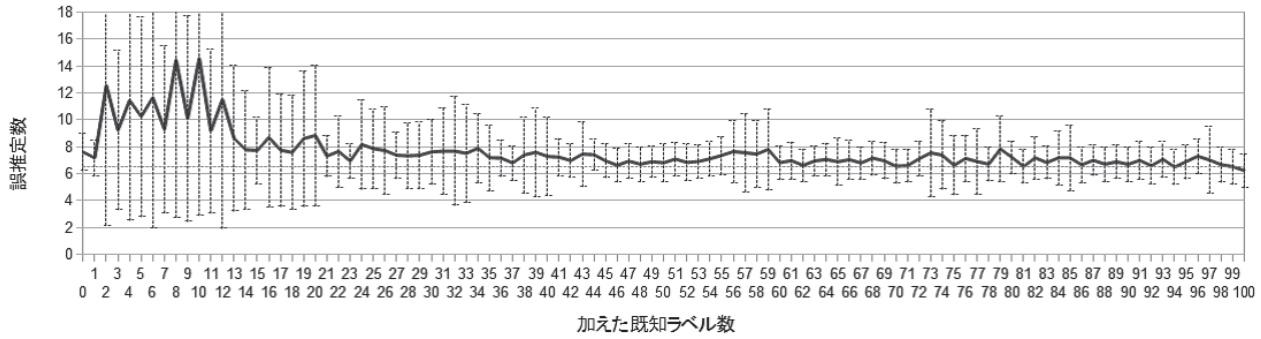

図 1: 正答率が異なる 5 人の回答者による評価でのラベルの誤推定数の推移

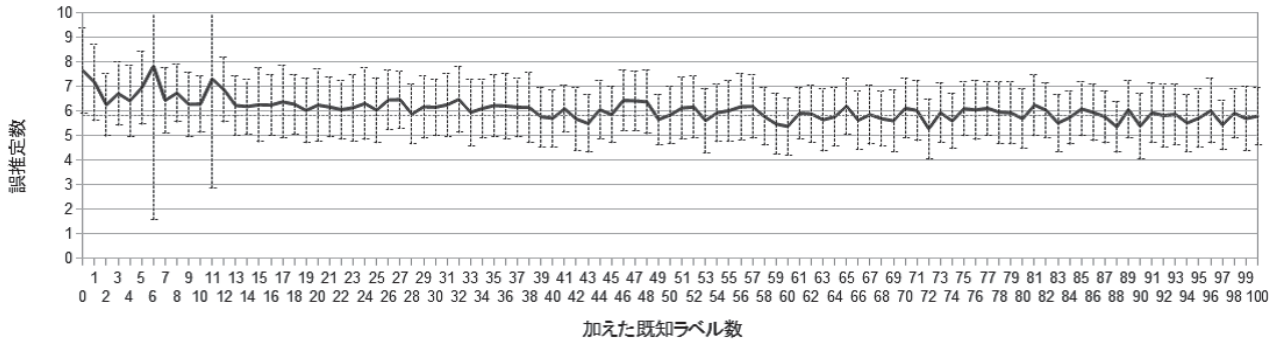

図 2: 正答率が同じ 5 人の回答者による評価でのラベルの誤推定数の推移

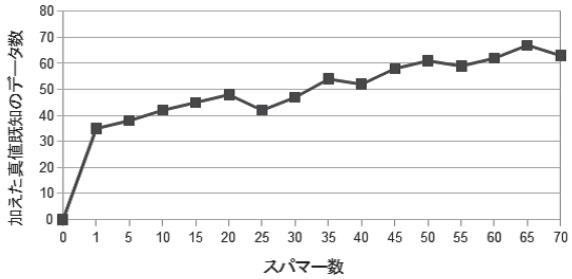

図 3: 作業履歴のない作業者数を変える評価でのラベルの誤推定数の推移

推定精度の上限は 0.8 である．前回の評価と同様に，ラベリングするデータ数は 100 であり，誤推定数が 20 個のときが最良の推定精度であるが，前述の評価と同様に 10%の誤差を容認して誤推定数が 22 個以下になるまで既知ラベル数を増やす評価を行った．作業履歴のない回答者の数を表 3 に沿って変化させながら，それぞれの場合において，データに加えるべき既知ラベルの最適な数を算出した．図 3 はその結果を示したグラフである．縦軸が既知ラベル挿入数であり，横軸が作業履歴のない作業者数である．

結果から，スパマーとおいた人数が増える毎に必要な既知ラベル数は増えていく傾向があることがわかる．ただし，スパマーとおいた人数を 25 人に設定した評価では 20 人の設定と比べて必要な既知ラベル数は減った．同様に，スパマーとおいた人数が 40 人の時，55 人の時，70 人の時はそれぞれスパマーとおいた人数が 5 人少ない時と比較して必要な既知ラベル数が減った．そのため，必ずしもスパマーとおいた人数が増えると必要な既知ラベル数が増えるとは限らない．

5. おわりに

本論文ではラベリングの精度を向上させることを目的とし，ラベリングしたいデータに加える既知ラベルの適切な数を決定する手法を提案した．併せて，その過程で必要となるラベル推定精度の上限を算出する手法を提案した．

データにラベリングをする際に，既知ラベルを適切な数だけ混ぜる手法を用いることでラベルの真値が未知のデータのラベル推定精度を上限まで改善できることが評価によりわかった．

ただ，既知ラベルを加える数が適切な数より少ない場合は，既知ラベルを加えない場合よりも精度が劣る場合もあることがわかった．したがって，加えるべき既知ラベルの数を正しく推定することは重要である．そのためには，本論文で示した手法により適切な既知ラベル挿入数を見積もることが有効である．

また，スパマーがいない環境では，スパマーが多い環境と比較して既存研究を用いた場合のラベルの推定精度が良い．そのため提案手法により改善される精度はスパマーが多い環境と比べて低い．そのことから，ラベリング作業の依頼者に評価結果を見せ，既知ラベルを増やすことによるコストと，見込まれる改善精度を比較検討してもらう必要があると考えられる．ゆえに，既知ラベルを加えることによって改善する精度と，既知ラベルを加えることによって生じるコストとを比較して既知ラベルを加える個数を判断する手法を検討することが今後の課題である．

[謝辞]

本研究に関して活発な議論を行い，本研究に有益な情報を頂いた熊本大学大学院自然科学研究科博士前期課程修了生大久保佑紀さんに感謝する．

表 3: 作業履歴のない回答者の正答率

		正解率								
	α^j	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
	β^j	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
合計人数	1	0	0	0	0	1	0	0	0	0
	5	0	0	1	1	1	1	1	0	0
	10	1	1	1	1	2	1	1	1	1
	15	1	1	2	2	3	2	2	1	1
	20	2	2	2	2	4	2	2	2	2
	25	2	3	3	3	3	3	3	3	2
	30	3	3	3	4	4	4	3	3	3
	35	3	4	4	4	5	4	4	4	3
	40	4	4	4	5	6	5	4	4	4
	45	5	5	5	5	5	5	5	5	5
	50	5	5	6	6	6	6	6	5	5
	55	6	6	6	6	7	6	6	6	6
	60	6	6	7	7	8	7	7	6	6
	65	7	7	7	7	9	7	7	7	7
	70	7	8	8	8	8	8	8	8	7

[文献]

[1] Aniket Kittur, Jeffrey V. Nickerson, Michael S. Bernstein, Elizabeth M. Gerber, Aaron Shaw, John Zimmerman, Matthew Lease, and John J. Horton “The Future of Crowd Work”, Proceedings of the 2013 conference on Computer supported cooperative work (CSCW), pp.1301-1318, 2013

[2] Rion Snow, Brendan O’Connor Daniel Jurafsky and Andrew Y. Ng “Cheap and Fast ― But is it good? Evaluating Non-Expert Annotations for Natural Language Tasks”, Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP ’ 08), pp.254-263, 2008

[3] Hiroshi Kajino, Yuta Tsuboi, Issei Sato, and Hisashi Kashima “ Learning from Crowds and Experts”, Proceedings of the 4th Human Computation Workshop, pp.107-113, 2012

[4] Yan Yan, Rómer Rosales, Glenn Fung, Mark Schmidt, Gerardo Hermosillo, Luca Bogoni, Linda Moy, and Jennifer G. Dy “Modeling annotator expertise: Learning when everybody knows a bit of something”, Proceedings of the Thirteenth International Conference on Artifical Intelligence and Statistics (AISTATS 2010), pp.932-939, 2010

[5] Panagoiotis G. Ipeirotis, Foster Provost, and Jing Wang “Quality Management on Amazon Mechanical Turk”, Proceedings of the ACM SIGKDD Workshop on Human Computation (HCOMP’10), pp.64-67, 2010

[6] Vicas C. Raykar and Shipeng Yu “Eliminating Spammers and Ranking Annotators for Crowdsourced Labeling Tasks”, Journal of Machine Learning Research, Vol. 13, pp.491-518, 2012

[7] 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 Statistics), Vol. 28, No.1, pp.20-28, 1979

[8] Wei Tang and Matthew Lease “Semi-Supervised Consensus Labeling for Crowdsourcing”, SIGIR 2011 Workshop on Crowdsourcing for Information Retrieval, pp.36-41, 2011

[9] Nicholas R. Miller “Information, Electorates, and Democracy: Some Extensions and Interpretations of the Condorcet Jury Theorem”, Grofman, Bernard and Guilermo Owen, eds., Information Pooling and Group Decision Making, pp.173-192, 1986

久保田 琢也 Takuya KUBOTA
熊本大学大学院自然科学研究科博士前期課程在学中．2014 熊本大学工学部情報電気電子工学科卒業．クラウドソーシングに興味を持つ．日本データベース学会学生会員．

眞鍋 雄貴 Yuki MANABE
熊本大学大学院自然科学研究科助教．2006 大阪大学基礎工学部情報科学科退学．2011 同大大学院了．博士（情報科学）．同年同大特任助教．2013 より現職．ソフトウェアやソフトウェア開発データの分析等に興味を持つ．

有次 正義 Masayoshi ARITSUGI
熊本大学大学院自然科学研究科教授．1991 九州大学工学部情報工学科卒．1996 同大大学院了．博士 (工学)．同年群馬大学助手．同助教授を経て，2007 より現職．データベースシステム，分散並列データ処理等に興味を持つ．