2M5-OS-07b-2

クラウドソーシングにおけるワーカーの確信度を用いた高精度なラベル統合

Accurate Integration of Crowdsourced Labels Using Workers' Self-reported Confidence Scores

小山 聡[*1] Satoshi Oyama　馬場 雪乃[*2] Yukino Baba　櫻井 祐子[*3] Yuko Sakurai　鹿島 久嗣[*2] Hisashi Kashima

[*1]北海道大学 Hokkaido University　[*2]東京大学 The University of Tokyo　[*3]九州大学 Kyushu University

We have developed a method for using confidence scores to integrate labels provided by crowdsourcing workers. We extended the Dawid-Skene model and created two probabilistic models in which the values of unobserved true labels are inferred from the observed provided labels and reported confidence scores by using the expectation-maximization algorithm. Results of experiments using actual crowdsourced data for image labeling and binary question answering tasks showed that incorporating workers' confidence scores can improve the accuracy of integrated crowdsourced labels.

1. はじめに

Amazon Mechanical Turk に代表されるクラウドソーシングサービスを用いて，インターネット上でワーカーと呼ばれる不特定多数の人に作業を依頼することが盛んに行われるようになってきている．特に，機械学習や情報検索，データベースなどの計算機科学の諸分野において，機械単独では解くことが困難な問題を人間の能力と組み合わせることで解決するヒューマンコンピュテーション [Law 11] の実現手段として，クラウドソーシングの有効性が認識されつつある．例えば，これまで教師付き学習においてはラベル付きの訓練データを準備することがボトルネックになっていたが，クラウドソーシングを用いることで，比較的安価に大量のラベル付きデータを得ることが可能になり，適用範囲の拡大や精度の向上に寄与している．一方，クラウドソーシングによって不特定多数のワーカーから得られたラベルは，従来の専門家らから得られたラベルと異なり，ワーカーの能力不足や意図的な手抜きなどにより，誤りが多く含まれる可能性がある．品質管理の問題は，クラウドソーシングにおける重要な研究課題となっている．

ワーカーが付けた誤りの可能性のあるラベルから真のラベルを推定する最も単純な方法は，同じデータに対して複数のワーカーにラベル付けを依頼することで冗長性を持たせ，その結果の多数決を取ることでラベルの統合を行うことである．この方法は，全てのワーカーのラベルは，同じ誤りの確率を持つと仮定することに対応している．しかし，クラウドソーシング環境のように，ワーカーの能力や誠実さに大きな差がある場合には，ワーカーごとに誤りの確率は異なると考えられるため，必ずしも効率的な方法ではない．

統計学の分野で提案された Dawid と Skene[Dawid 79] の方法は，ワーカーごとに異なる誤り率を想定することができ，クラウドソーシングにおいて真のラベルを推定する場合に，多数決よりも優れた方法であると考えられている．具体的には，真のラベルが与えられたときの，ワーカーが付けるラベルの条件付き確率を規定するパラメータが，ワーカーごとに異なっていると仮定する．このパラメータはワーカー能力（ラベル付けの正確さ）を表現しており，EM 法を用いてこのパラメータと真のラベルを推定値を交互に更新することで，真のラベルの推定を行うことができる．その他にも，ワーカーの専門性や問題の難しさを考慮するなど，様々な方法が提案されている [Whitehill 09, Welinder 10a].

これらの方法は，複数のワーカーによるラベルを比較することで，ワーカーやラベルの信頼度を自動的に推定する方法と考えることができる．一方，本研究では，「機械にとっては困難だが人間にとっては比較的容易な作業は人間に依頼する」というヒューマンコンピュテーションの考えをより推し進めたアプローチを取る．すなわち，ワーカー自身に，自分が付けたラベルが正しいかどうかの自信（確信度）を申告させる．ワーカーに確信度を申告させることは，ワーカー自身に自分が付けたラベルの信頼性を評価させていることに相当し，ワーカーが付けたラベルの信頼性は，その作業を行ったワーカー自信が判断できるという仮説に従っている．これまでに Ipeirotis はワーカーにタスクの難しさを申告させる実験を行っており [Ipeirotis 09]，自己申告させたタスクの難しさとラベルの品質に相関があることを示し，アルゴリズムで推定するのではなく直接ワーカーに問題の難しさを聞くことの有効性を示唆している．ワーカーが感じたタスクの難しさは，基本的にはワーカーの自信と逆の関係にあると考えられるので，自己申告した確信度も正解率と相関があると予想される．また，Kazai [Kazai 11] も文書の適合性を判定する問題において，ワーカーの確信度とラベルの精度の関係について調査している．

しかしながら，これらの研究ではワーカーの申告した確信度を真のラベルの推定に用いる方法は提案されていない．ワーカーが申告した確信度は真のラベルを推定する際に有益な情報であると考えられるが，ラベルと同様に確信度にもワーカーによってその品質に差があることが予想される．例えば，あるワーカーは自信過剰で実際には間違っている場合も高い確信度を申告し，逆に別のワーカーは控えめで，実際には正しいラベルを与えている場合でも低い確信度を申告するかもしれない．また，良く考えずに適当に確信度を申告するワーカーが存在する可能性もある．

図 1 に画像分類の実験（詳細は後で述べる）における，ワーカーの確信度の平均と正解率の関係を示す．各点がワーカーに対応し，横軸は 10 個の画像分類タスクでの確信度の平均，縦軸は実際の正解率である．確信度の平均と正解率の相関係数は

連絡先: 〒 060-0814 札幌市北区北 14 条西 9 丁目　北海道大学大学院情報科学研究科　小山聡　oyama@ist.hokudai.ac.jp

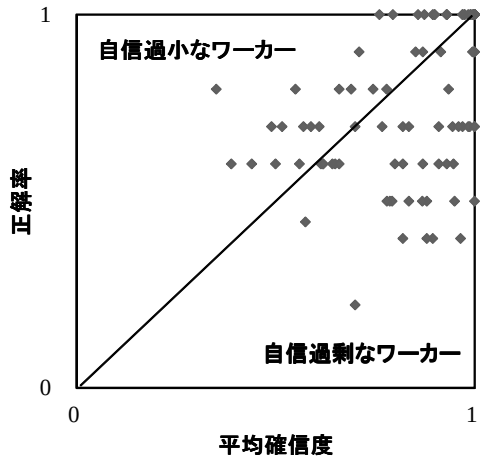

図 1: 確信度と正解率の相関

0.455 であり，たしかに正の相関が存在し，確信度が真のラベルを推定する際に有用な情報を含んでいることを示唆している．しかし，正解率に比べて確信度が高い自信過剰（overconfident）なワーカーや，正解率に比べて確信度が低い自信過小（underconfident）なワーカーも多く存在していることが分かる．人間が自分の正解率を見積もる能力は認知心理学でメタ認知 [Dunlosky 09] の一種とされ，人によってその正確さが異なることが知られている．この実験結果も，各ワーカーの付けた確信度を一様に扱うのではなく，ワーカーの自己評価の正確さの違いを考慮することの必要性を示している．

そこで本研究では，不正確だが有用な情報を含む確信度を用いて，ラベル統合の精度を向上させるための確率的手法を提案する．詳細は Oyama *et al.* [Oyama 13] を参照されたい．

2. 提案手法

2.1 問題設定

本研究では与えられた N 個のアイテムに対して J 人のワーカーがラベル付けを行う問題を考える．ただし，一人のワーカーが全てのアイテムにラベル付けを行う必要はない．$\mathcal{J}_i \subseteq \{1,\ldots,J\}$ をアイテム i にラベルを付けたワーカーの集合とする．$t_i \in \{0,1\}(i \in \{1,\ldots,N\})$ は i 番目のアイテムの真のラベル，$y_{ij} \in \{0,1\}$ $(j \in \mathcal{J}_i)$ はワーカー j がアイテム i に付けたラベルである．Dawid と Skene [Dawid 79] と我々の問題設定との違いは，ワーカーに自分の付けたラベルに対する確信度も入力してもらうことである．$c_{ij} \in \{0,1\}$ $(j \in \mathcal{J}_i)$ はワーカー j がアイテム i に付けたラベルに対する確信度であり，自信がある場合が 1，ない場合が 0 に対応する．ここでは単純化のために確信度を二値で扱っているが，多段階評価や実数値で表現される場合への拡張も容易である．

我々の目標は，ワーカーの付けたラベルの集合 $\{y_{ij}\}$ および確信度の集合 $\{c_{ij}\}$ $(i \in 1,\ldots,N, j \in \mathcal{J}_i)$ が与えられた時に，真のラベルの集合 $\{t_i\}(i \in \{1,\ldots,N\})$ を推定することである．

2.2 モデル

ここで，クラウドソーシングにおいてワーカーがラベルと確信度を与える際の確率的生成モデルを導入する．このモデルを用いることで，ワーカーのラベルと確信度から，真のラベルを推定することが可能となる．そこでは，例えばあるラベルに対するワーカーの確信度が高ければ，そのラベルが真のラベルと一致する可能性も大きくなる．

我々のモデルは，同時分布の以下のような分解で与えられる．

$$p(\{t_i\},\{y_{ij}\},\{c_{ij}\}) = \prod_{i\in\{1,\ldots,N\}} \prod_{j\in\mathcal{J}_i} p(c_{ij}|y_{ij},t_i)p(y_{ij}|t_i)p(t_i)$$

まず，アイテム i の真のラベルは p_i をパラメータとする以下のようなベルヌーイ分布により生成され，確率 p_i で値 1 を，確率 $1-p_i$ で値 0 を取るとする．

$$p(t_i) = p_i^{t_i}(1-p_i)^{(1-t_i)}$$

元の Dawid-Skene モデルでは，真のラベルの事前分布は全てのデータで共通であるとしている．これはある患者が特定の疾患を持つか否かを判定する医療診断のように，質問が均質な場合には適切であるが，異なった種類の質問を含むようなタスクにおいては適切でない．そこで，問題ごとに異なったパラメータ p_i を導入している．p_i はアイテム i にラベル 1 を与えたワーカーの割合で推定可能である．

アイテム i へのワーカーのラベル $\{y_{ij}|j \in \mathcal{J}_i\}$ は真のラベル t_i が与えられたときに条件付き独立であるとする．図 2 の $\boldsymbol{\alpha}^{(j)} = \{\alpha_0^{(j)}, \alpha_1^{(j)}\}$ はワーカー j のパラメータの組を表しており，$\alpha_0^{(j)}$ は真のラベルが 1 のときにワーカーがラベル 1 を付ける確率，$\alpha_1^{(j)}$ は真のラベルが 1 のときにワーカーがラベル 1 を付ける確率をそれぞれ表している．すなわち，$t_i = 1$ のとき，ワーカー j がアイテム i に付けるラベル y_{ij} は $\alpha_1^{(j)}$ をパラメータとする以下のベルヌーイ分布に従う．

$$p(y_{ij}|t_i = 1) = (\alpha_1^{(j)})^{y_{ij}}(1-\alpha_1^{(j)})^{(1-y_{ij})}$$

同様に，$t_i = 0$ のとき，ワーカー j がアイテム i に付けるラベル y_{ij} は $\alpha_0^{(j)}$ をパラメータとする以下のベルヌーイ分布に従う．

$$p(y_{ij}|t_i = 0) = (\alpha_0^{(j)})^{y_{ij}}(1-\alpha_0^{(j)})^{(1-y_{ij})},$$

ワーカー j のアイテム i のラベルに対する確信度 c_{ij} は真のラベル t_i とワーカーのラベル y_{ij} に依存する．以下では，確信度を生成する二つのモデルを提案する．

ワーカーに依存する確信度モデルでは，$\boldsymbol{\beta}^{(j)} = \{\beta_{00}^{(j)}, \beta_{01}^{(j)}, \beta_{10}^{(j)}, \beta_{11}^{(j)}\}$ はワーカー j に固有のパラメータの組である．たとえば，$\beta_{00}^{(j)}$ は真のラベルが $t_i = 0$ かつワーカー j のラベルが $y_{ij} = 0$ のときに，確信度 $c_{ij} = 1$ を付ける確率である．すなわち，確信度は

$$p(c_{ij}|t_i = 0, y_{ij} = 0) = (\beta_{00}^{(j)})^{c_{ij}}(1-\beta_{00}^{(j)})^{(1-c_{ij})}$$

により生成される．$t_i = 0$ かつ $y_{ij} = 1$ の場合には，確信度は以下の分布により生成される．

$$p(c_{ij}|t_i = 0, y_{ij} = 1) = (\beta_{01}^{(j)})^{c_{ij}}(1-\beta_{01}^{(j)})^{(1-c_{ij})}$$

残りの二つの場合の条件付き分布，$p(c_{ij}|t_i = 1, y_{ij} = 0)$ および $p(c_{ij}|t_i = 1, y_{ij} = 1)$ も同様に定義される．

一方，ワーカー独立のモデルでは，すべてのワーカーは同一のパラメータ $\boldsymbol{\beta} = \{\beta_{00}, \beta_{01}, \beta_{10}, \beta_{11}\}(= \boldsymbol{\beta}^{(j)})$ を共有すると仮定する．ワーカー依存の分布を導入する方が，自信過剰なワーカーは自分のラベルが正しくない場合でも高い確信度を付

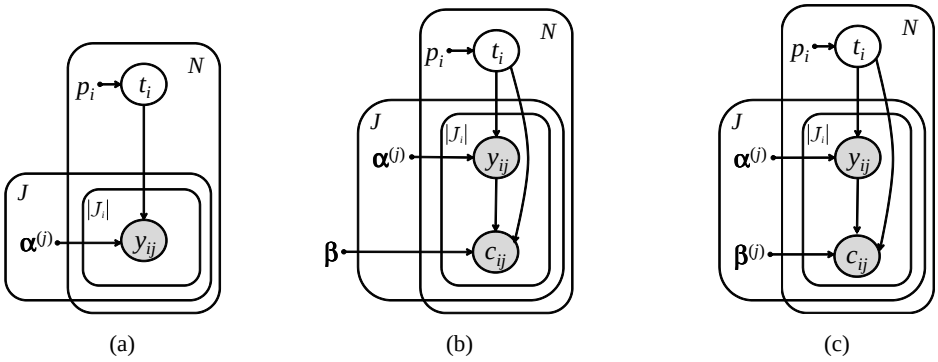

図 2: グラフィカルモデル: (a) Dawid-Skene モデル, (b) ワーカー独立な確信度モデル, (c) ワーカー依存な確信度モデル

け，逆に自信過小なワーカーは自分のラベルが正しい場合でも低い確信度を付ける，といった各ワーカーの傾向を考慮したより詳細なモデル化が可能である．一方で，パラメータの数が多くなるため，データが少ない場合には推定精度が低くなるという問題も生じる．

比較のため，Dawid-Skene モデル，ワーカー独立な確信度モデル，ワーカー依存な確信度モデルを図 2 に示す．提案した確信度モデルは，Dawid-Skene モデルにワーカーの確信度を確率変数として追加することで拡張したモデルになっている．

2.3 推定アルゴリズム

ここで，実際に観測されるワーカーの付けたラベル $\{y_{ij}\}$ および確信度 $\{c_{ij}\}$ から，真のラベル $\{t_i\}$ を推定したいが，モデルのパラメータ $\{\boldsymbol{\alpha}^{(j)}\}$ および $\{\boldsymbol{\beta}^{(j)}\}$ が未知であるので，直接推定することはできない．一方，もし真のラベルが既知であれば，モデルのパラメータを最尤推定によって容易に推定することができる．そこで，EM アルゴリズムによりモデルのパラメータと真のラベルを交互に推定することを行う．ワーカー依存のモデルの場合，以下のステップを繰り返す．

E ステップ パラメータ $\{\boldsymbol{\alpha}^{(j)}\}$ および $\{\boldsymbol{\beta}^{(j)}\}$ の推定値を固定して，隠れ変数 $\{t_i\}$ の期待値を推定する．

M ステップ 隠れ変数 $\{t_i\}$ の期待値を固定して，パラメータ $\{\boldsymbol{\alpha}^{(j)}\}$ および $\{\boldsymbol{\beta}^{(j)}\}$ を推定する．

E ステップにおいて，t_i の期待値は以下で計算される．

$$
\begin{aligned}
E[t_i] =& p(t_i = 1|\{y_{ij}\},\{c_{ij}\}) \\
=& \frac{p_i}{z_i} \prod_{j \in \mathcal{J}_i} \Big\{ (\alpha_1^{(j)})^{y_{ij}} (1-\alpha_1^{(j)})^{(1-y_{ij})} \\
& \times (\beta_{11}^{(j)})^{y_{ij}c_{ij}} (1-\beta_{11}^{(j)})^{y_{ij}(1-c_{ij})} \\
& \times (\beta_{10}^{(j)})^{(1-y_{ij})c_{ij}} (1-\beta_{10}^{(j)})^{(1-y_{ij})(1-c_{ij})} \Big\}
\end{aligned}
$$

ここで，z_i は正規化定数である．

M ステップにおいて，$\{\boldsymbol{\alpha}^{(j)}\}$ および $\{\boldsymbol{\beta}^{(j)}\}$ の推定値は以下で計算される．

$$
\begin{aligned}
\hat{\alpha}_0^{(j)} &= \frac{\sum_{\{i:j\in\mathcal{J}_i\}} (1-E[t_i]) y_{ij}}{\sum_{\{i:j\in\mathcal{J}_i\}} (1-E[t_i])} \\
\hat{\alpha}_1^{(j)} &= \frac{\sum_{\{i:j\in\mathcal{J}_i\}} E[t_i] y_{ij}}{\sum_{\{i:j\in\mathcal{J}_i\}} E[t_i]}
\end{aligned}
$$

$$
\begin{aligned}
\hat{\beta}_{00}^{(j)} &= \frac{\sum_{\{i:j\in\mathcal{J}_i\}} (1-E[t_i])(1-y_{ij}) c_{ij}}{\sum_{\{i:j\in\mathcal{J}_i\}} (1-E[t_i])(1-y_{ij})} \\
\hat{\beta}_{01}^{(j)} &= \frac{\sum_{\{i:j\in\mathcal{J}_i\}} (1-E[t_i]) y_{ij} c_{ij}}{\sum_{\{i:j\in\mathcal{J}_i\}} (1-E[t_i]) y_{ij}} \\
\hat{\beta}_{10}^{(j)} &= \frac{\sum_{\{i:j\in\mathcal{J}_i\}} E[t_i](1-y_{ij}) c_{ij}}{\sum_{\{i:j\in\mathcal{J}_i\}} E[t_i](1-y_{ij})} \\
\hat{\beta}_{11}^{(j)} &= \frac{\sum_{\{i:j\in\mathcal{J}_i\}} E[t_i] y_{ij} c_{ij}}{\sum_{\{i:j\in\mathcal{J}_i\}} E[t_i] y_{ij}}
\end{aligned}
$$

ワーカー独立な確信度モデルにおいても，同様に EM 法により真のラベルとモデルのパラメータを推定することができる．

3. 実験

ラベル統合における確信度利用の有効性を確認するために，Amazon Mechanical Turk を用いた実験を行った．Caltech-UCSD Birds 200 [Welinder 10b] から選んだ 10 枚の鳥の画像に対し，その画像に写っている鳥の名前を 2 つの選択肢から選ぶ問題を用いた．ワーカーの数と正解率との関係を調べるために，100 人のワーカーに同一の 10 画像に対してラベリングを依頼した．

ワーカーには画像へのラベルと同時に，自分の付けたラベルに関する確信度を 0 から 100 までの数値で入力させた．提案したモデルは確信度が二値であることを仮定しているため，ワーカーごとに確信度の中央値を求め，中央値より大きな値を 1 に，中央値より小さな値を 0 に変換した．中央値と確信度の値が一致した場合は 1 のデータと 0 のデータの数がバランスする方に割り当てた．

アイテムあたりのワーカー数が正解率に与える影響を調べるため，100 人のワーカーを一定数のグループに分割し，各グループでラベル推定を行いその正解率を平均した．5 人，10 人，20 人，50 人ずつのグループにおいて，単純な多数決，Dawid-Skene モデル，ワーカー独立な確信度モデル，ワーカー依存な確信度モデルで統合したラベルの正解率の平均を図 3 に示す．ワーカー数が 50 の場合，高い冗長性により，単純な多数決であっても 3 つの確率モデルと同等の高い正解率が得られる．実際には，予算の制約などから利用できるワーカー数には制約がある．ワーカー数が 5 ないし 10 の場合，確信度を用いた 2 つのモデルは多数決と Dawid-Skene モデルより優れている．特に，ワーカー数が 5 の場合，ワーカー依存な確信度モデルが他の手法を大きく上回っている．

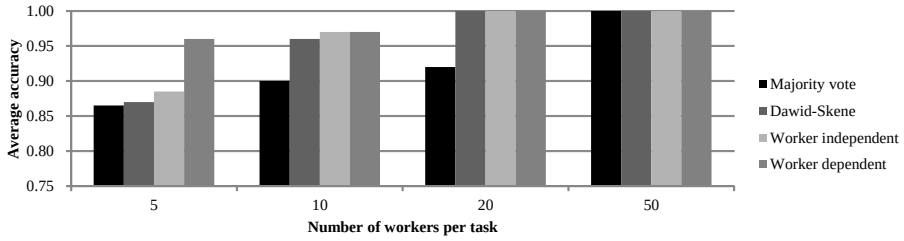

図 3: 画像分類の実験結果

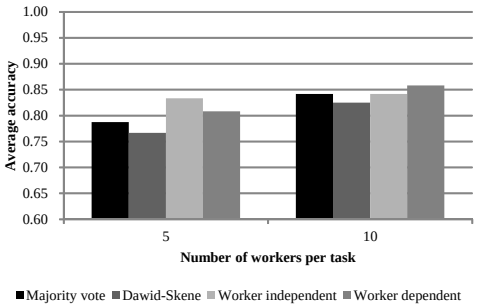

図 4: 一般知識問題の実験結果

さらに，120 個の一般的知識を問う問題に Yes/No で答えさせる実験も行った．ここでは，日本のクラウドソーシングサービスであるランサーズを用いた．各質問は 10 人のワーカーによって回答され，ここでも画像分類の場合と同様な形式でワーカーに確信度を入力させ，二値に変換をした．

この実験では全体として 42 人のワーカーが参加した．図 4 に示すように，全てのワーカーのラベルを用いた場合，単純な多数決でも確信度を用いたモデルと同様の正解率が得られる．これは画像分類の場合と同様に，冗長性が高い場合は多数決で十分な精度が得られることを示している．ワーカー数を半分にし，アイテムあたりのラベル数の平均を 5 とした場合，ワーカー独立な確信度モデルが最も正解率が高く，ワーカー依存の確信度モデルがそれに続く結果となった．ここでワーカー独立なモデルが優れていた理由としては，ワーカーの付けた平均確信度の標準偏差が画像分類の場合は 0.22 であったのに対し，この実験では 0.16 であり，ワーカーごとの確信度のばらつきが小さかったことが考えられる．

4. おわりに

Dawid-Skene モデル，ワーカー独立な確信度モデル，ワーカー依存な確信度モデルは，それぞれこの順番で前のモデルを拡張したものになっており，変数が増えて複雑さも大きくなる．データに応じてモデル選択を行う方法は今後の重要な研究課題である．

項目反応理論においては，被験者に選択肢が正しい確率を答えさせる方式も提案されているが，確信度として 0%や 100%を付けたケースが全体の 6 割を超えたり，自分の確信度を確率として適切に表現するためには事前に被験者の訓練が必要である，といった報告 [Kato 10] がなされている．ワーカーに確信度を入力させる方式には，ユーザインタフェースの検討やワーカーのインセンティブの考慮など，さらなる研究の余地がある．また，回答に要した時間など，問題の難しさに関して自動的に測定可能な情報を用いた方法と比較することも考えられる．

参考文献

[Dawid 79] Dawid, A. P. and Skene, A. M.: Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm, *Journal of the Royal Statistical Society. Series C (Applied Statics)*, Vol. 28, No. 1, pp. 20–28 (1979)

[Dunlosky 09] Dunlosky, J. and Metcalfe, J.: *Metacognition*, SAGE (2009)

[Ipeirotis 09] Ipeirotis, P.: How good are you, Turker? (2009), http://www.behind-the-enemy-lines.com/2009/01/how-good-are-you-turker.html

[Kato 10] Kato, K. and Zhang, Y.: An Item Response Model for Probability Testing, in *International Meeting of the Psychometric Society* (2010)

[Kazai 11] Kazai, G.: In Search of Quality in Crowdsourcing for Search Engine Evaluation, in *ECIR* (2011)

[Law 11] Law, E. and Von Ahn, L.: *Human Compuation*, Morgan & Claypool Publishers (2011)

[Oyama 13] Oyama, S., Baba, Y., Sakurai, Y., and Kashima, H.: Accurate Integration of Crowdsourced Labels Using Workers' Self-reported Confidence Scores, in *IJCAI* (2013)

[Welinder 10a] Welinder, P., Branson, S., Belongie, S., and Perona, P.: The Multidimensional Wisdom of Crowds, in *NIPS 23* (2010)

[Welinder 10b] Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S., and Perona, P.: Caltech-UCSD Birds 200, Technical Report CNS-TR-2010-001, California Institute of Technology (2010)

[Whitehill 09] Whitehill, J., Ruvolo, P., Wu, T., Bergsma, J., and Movellan, J.: Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise, in *NIPS 22* (2009)