Crowdsourced Labels from Multiple Contexts

Abstract

Recent probabilistic models can predict ground-truth from crowdsourced labels much more accurately than majority voting, but conventional models group all labels into the same homogeneous context. In contrast, existing crowdsourcing platforms collect labels for multiple aims and datasets, from a shared pool of contributors. This paper presents a method to efficiently extract highly-quality information, from the contributors' annotations, across many heterogeneous contexts, such as qualitatively different domains or different question types. Our model can rapidly and incrementally learn context specific information, such as ground-truth and annotator expertise, as well as information that transfers between contexts, e.g. honesty and exactitude. We describe how to efficiently perform inference on our model, which makes use of Variational-Bayes approximations and we propose a novel approach to approximate problematic hyperparameters. We evaluate our approach, and discuss applications for our model.

1. Introduction

The recent popularity of crowdsourcing, has been accompanied by a growth in methods for inference with the associated data. Particular success has been achieved in labelling tasks, where for task-instances (*objects*) the unknown *ground-truth* is inferred using *labels* from a crowd of semi-trusted *annotators*. When sourcing labels from multiple semi-trusted contributors, sources will sometimes disagree on an object. In the case of humans, this can be due to many factors including: annotator bias, data ambiguity, lack of expertise, transcription errors, laziness, misunderstanding the task and malicious intent.

While majority voting can accurately determine the underlying *ground-truth* (or gold-standard) of the data (with sufficient labels under *reasonable* conditions (Snow et al., 2008)), greater accuracy with fewer labels can be had by building models of *annotator reliability* based on consistency of agreement with other annotators, and/or performance on *test* examples, e.g. (Raykar et al., 2010; Welinder & Perona, 2010; Karger et al., 2011). One practical use of these methods is to perform classification or regression from semi-trusted labels (Raykar et al., 2010; Venanzi et al., 2013). Alternatively, the ground-truth predictions may have value in themselves or more general use (Ipeirotis et al., 2010). Further, annotator reliability estimates can be used to more quickly acquire additional information, either by blacklisting the unreliable (Welinder & Perona, 2010), or with a more nuanced pairing of annotators with objects (Dickens & Lupu, 2014; Ipeirotis et al., 2010; Yan et al., 2011; Chen et al., 2013). This paper presents a new crowdsourcing model for inferring ground-truth and annotator expertise, for objects from a wide variety of different contexts (requiring different skills and knowledge), where information learnt in one context can *transfer* to others. Further, we give an approximate inference approach, using Variational-Bayes techniques, to efficiently infer model variables, including troublesome hyperparameter distributions.

Most existing crowdsourcing models place all labels within the same homogeneous context, where all labelling tasks are assumed to be similar, with consistent errors and biases, e.g. (Dawid & Skene, 1979; Raykar et al., 2010; Rzhetsky et al., 2009; Welinder & Perona, 2010; Karger et al., 2011). In contrast, crowdsourcing tasks are often grouped within similar contexts, both methodologically e.g. (Chilton et al., 2013; Bragg et al., 2013; Welinder et al., 2010), and in real world applications e.g. Amazon's MTurk and the Zooniverse project (Zooniverse). Moreover, annotators typically contribute across many such contexts. Under such circumstances, the above approaches would either have to: model each context separately, failing to share information about annotators across contexts; or pool all tasks together, losing differences between contexts.

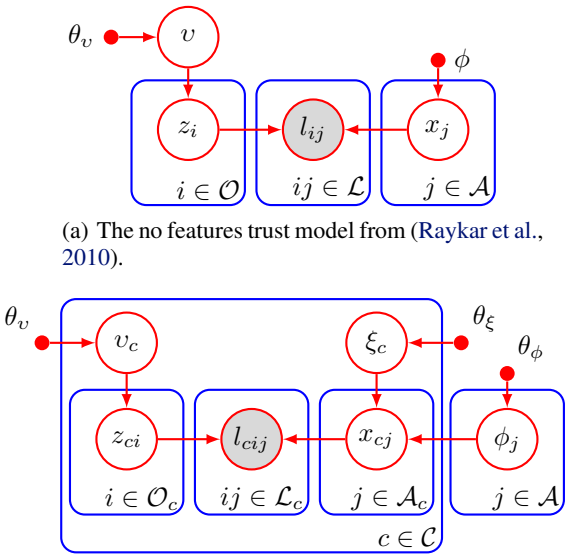
More sophisticated models can infer individual object difficulty (Whitehill et al., 2009) within a context, or more generally model various *confounding characteristics* for annotators and objects (Welinder et al., 2010). In a recent paper, (Mo et al., 2013) develop a cross-context model, which models confounding characteristics in a multi-context model for binary labels, and use Markov-chain Monte-carlo (MCMC) sampling to infer model parameters. We present a slightly simpler model that nonethe-

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

less captures inter- and intra-contextual information similarly, and which extends easily to categorical and real labels. Further, we develop a highly efficient and scalable inference method for this.

This paper makes the following contributions. We present a novel *context-aware* model, with a variational approximation method. In order to estimate problematic distributions in the context-aware model, we develop a novel estimation procedure, which can be applied to other probabilistic models with Dirichlet hyperparameter distributions. We also develop a novel variational-Bayes estimation approach to fit an influential single-context model from (Raykar et al., 2010). We evaluate our methods against competing technologies, and show that our single-context variational method competes with equivalent approaches on synthetic and real data, and our context-aware method outperforms single-context approaches in a number of ways.

In the remainder of this paper: Section 2 outlines an existing work which informs our approach; Section 3 presents our context aware model and estimation approach; Section 4 reports results of experiments to test the efficacy of our method; and Section 5 reflects on the material presented.



(a) The no features trust model from (Raykar et al., 2010).

(b) Our context-aware model, presented in Section 3.

Figure 1. (a) single-context and (b) context-aware probabilistic models for semi-trusted labels.

2. Background

Recent research on *crowdsourcing* has applied probabilistic models to large scale problems with semi-trusted meta-data, e.g. (Raykar et al., 2010; Welinder & Perona, 2010). Most such approaches focus on simple meta-data for discrete data items, e.g. binary, categorical, ordinal, real scalars and vectors, and simultaneously estimate the hidden ground-truth, and the expertise of the annotators (Dawid & Skene, 1979; Raykar et al., 2010; Rzhetsky et al., 2009; Welinder & Perona, 2010; Karger et al., 2011). These approaches exploit the principle that reliable annotators tend to agree with each other more often than those prone to random mistakes, bias, or error. Fig. 1(a) recreates the *no features* trust model used in (Raykar et al., 2010), but renames some variables for notational consistency. The model assumes a set of *objects* $\mathcal{O}$ (a meta-data requirement), and for each object $i \in \mathcal{O}$ an unknown *ground-truth* z_i (the required meta-data). Each *annotator*, $j \in \mathcal{A}$, is not entirely reliable, and when queried about z_i gives *label* l_{ij}, which we observe, and may (or may not) be equal to z_i. This uncertainty is modelled by j's *expertise* (alt. reliability) x_j, e.g., for binary labels, j's sensitivity and specificity. The set of labelled pairs is $\mathcal{L} \subseteq \mathcal{O} \times \mathcal{A}$. *Ground-truth parameter* υ controls the distribution over ground-truth values. ϕ and θ_υ are hyper-parameters. The earlier model from (Dawid & Skene, 1979) is the same except they fix υ and omit θ_υ.

Section 3 presents a context-aware model for semi-trusted labels. The dense connectivity in this model precludes exact inference (Bishop, 2007), and so some alternative approximate method will be required. Our approximate inference combines Variational-Bayes with a novel approach for hyperparameter distributions.

While *Expectation Maximisation* is a popular choice for inference in crowdsourcing models (Dawid & Skene, 1979; Raykar et al., 2010; Welinder & Perona, 2010). EM only predicts the mode of the distribution of interest, can be unstable when data is sparse, and inaccurate when distributions are not well behaved, e.g. multi-modal or skewed. To avoid these shortfalls (and those of other ML/MAP approaches, e.g. Laplace), we propose using variational approximations, which provide more robust approximations, with little (or no) computational overhead compared to EM. Two common variational methods are 1) *mean field approximations* as in (Parisi, 1988); and 2) *local Variational approximation* as in (Jaakkola & Jordan, 2000).

Mean field approximations are typically applied such that, for observed variables, X, and latent variables (and parameters), Z, the posterior distribution $p(Z|X) = p(X|Z)p(Z)/p(X)$, is approximated by $q(Z) \approx p(Z|X)$, where q assumes independence between groups of variables in Z. The best approximation, $q^\star$, is that which minimises the KL-divergence between q and p (Jaakkola & Jordan, 2000), e.g.

$$q^\star = \operatorname*{argmin}_q D_{\mathrm{KL}}(q(Z)||p(Z|X))$$

and this can be shown to occur when[1]

$$\ln q^\star(Z_i) = \mathbf{E}_{Z_{/i}}(\ln p(X, Z)) + \text{const} \quad (1)$$

where Z_i is a group of variables in Z, and $\mathbf{E}_{Z_{/i}}$ is the expectation over variables in Z not in Z_i. Repeatedly assigning each q from the expectation in Eqn. (1) leads to convergence on $q^\star$. Mean field approximations can lead to accurate approximations of the mean of $p(Z|X)$, but typically underestimate model uncertainty.

Local variational inference is used when estimation requires the expectation of a complex quantity (function), $f(x)$, over some distribution of interest, say $p(x)$ for $x \in \mathcal{X}$. This in turn, requires the evaluating the integral $\mathbf{E}_x f(x) = \int_{x\in\mathcal{X}} f(x)p(x)dx$. If this calculation is intractable, an approximation can be made by maximising an integral over lower bound functions $\hat{f}_\theta$, i.e. $\theta^\star = \text{argmax}_\theta \mathbf{E}_x \hat{f}_\theta(x)$. An equivalent approximation minimises an upper bound on $f(x)$.

A related variational approach to ours is Expectation propagation (EP) (Minka, 2001). This is typically more robust to local maxima, but can have convergence issues (Bishop, 2007). Complex distributions can also be fit to data using sampling methods, e.g. (Teh et al., 2004), e.g. Markov-chain Monte-carlo sampling. Appropriate sampling methods are guaranteed to converge on the true distribution, but tend to be very computationally demanding compared to variational approximations (Bishop, 2007), can have problems converging with complex models (Gelman & Hill, 2007), and must re-sample from scratch when new data is added, unlike EM and Variational approaches. We use elements of both mean field theory and local variational approximation for our inference.

3. A Context-Aware Model of Trust

This section outlines our context-aware model, and shows how to approximate model parameters using a combination of Variational-Bayes and our novel approximation technique for hyperparameter distributions. Fig. 1(b) shows the context-aware model for semi-trusted labels, which duplicates the single-context model from (Raykar et al., 2010) across multiple contexts, with additional information shared across contexts. More precisely, for each context $c \in \mathcal{C}$, ground-truth bias parameter υ_c controls the distribution over ground-truths z_{ci}, for all context objects $i \in \mathcal{O}_c$. Annotator j has *context expertise* x_{cj}, and provides observed noisy labels l_{cij} for some objects i. In addition, annotator j has *trust* ϕ_j, which captures context independent labelling characteristics for j, and each context c has *challenge* ξ_c, capturing annotator independent information about labelling uncertainty in c. Expertise x_{cj} depends on both ϕ_j and ξ_c. θ_υ, θ_ξ and θ_ϕ are fixed hyperparameters.

[1]The *const* variable captures terms independent of Z_i.

The joint probability for this model is

$$p(\mathbf{L}, \mathbf{Z}, \mathbf{X}, \Upsilon, \Xi, \Phi|\theta_\upsilon, \theta_\xi, \theta_\phi) = \prod_{j\in\mathcal{A}} p(\phi_j|\theta_\phi) \prod_{c\in\mathcal{C}} \Big(p(\upsilon_c|\theta_\upsilon)p(\xi_c|\theta_\xi) \prod_{i\in\mathcal{O}_c} p(z_{ci}|\upsilon_c) \prod_{j\in\mathcal{A}_c} p(x_{cj}|\phi_j, \xi_c) \prod_{ij\in\mathcal{L}_c} p(l_{cij}|z_{ci}, x_{cj}) \Big)$$

where $\mathbf{L}$, $\mathbf{Z}$, $\mathbf{X}$, Υ, Ξ, Φ are collections of all labels l_{ij}, ground-truths z_{ci}, expertise parameters x_{cj}, ground-parameters υ_c, challenge parameters ξ_c, and trust parameters ϕ_j respectively.

The topology of our model and the observed quantities guarantee certain independence properties (Bishop, 2007) – indicated below by $\stackrel{\Delta}{=}$, and for tractability we make additional independence assumptions – indicated by $\stackrel{\text{ass.}}{=}$, in:

$$q(\mathbf{Z}, \mathbf{X}, \Upsilon, \Xi, \Phi) \stackrel{\Delta}{=} q(\mathbf{Z}, \Upsilon)q(\mathbf{X}, \Xi, \Phi) \stackrel{\text{ass.}}{=} q(\mathbf{Z})q(\Upsilon)q(\mathbf{X})q(\Xi)q(\Phi) \quad (2)$$

Using Eqn. (1) and these assumptions (which we briefly discuss later in this section), we can write the optimal log variational distribution for each unobserved variable. For instance, the distribution of expertise parameters satisfies

$$\ln q^\star(x_{cj}) = \mathbf{E}_{\mathbf{Z}_c}\Big(\sum_{i\in\mathcal{O}_{cj}} \ln p(l_{cij}|z_{ci}, x_{cj}) \Big) + \mathbf{E}_{\phi_j,\xi_c}\Big(p(x_{cj}|\phi_j, \xi_c) \Big) + \text{const} \quad (3)$$

where collections with subscript c or j respectively contain only values corresponding to context c or annotator j, e.g. $\mathbf{Z}_c = \{z_{ci}|i \in \mathcal{O}_c\}$. Similar equalities exist for $\ln q^\star(z_{ci})$, $\ln q^\star(\upsilon_c)$, $\ln q^\star(\xi_c)$ and $\ln q^\star(\phi_j)$. In practice, we approach these optimal estimates by iteratively evaluating expectations on the right and assigning these to distributional forms on the left, with appropriate normalisation taking care of the constant terms (Bishop, 2007).

3.1. Context-Aware Model for Binary Labels

Our context-aware model described can readily support binary, 1-of-K categorical, real scalar and real vector labels. For clarity and brevity, we consider the special case where labels and ground-truth values are binary, i.e. $z_{ci} \in \{0, 1\}$, and $l_{cij} \in \{0, 1\}$ for all relevant i, j and c (supporting K categories is a simple extension). We define the expertise of annotator j in context c, to be $x_{cj} = (x_{cj0}, x_{cj1})$ – the specificity and sensitivity of j in context c. Our choice for the probability of label l_{cij} reflects that in (Raykar et al., 2010) – a product of two Bernoulli distributions, i.e.

$$p(l_{cij}|z_{ci}, x_{cj}) = x_{cj0}^{(1-z_{ci})(1-l_{cij})}(1 - x_{cj0})^{(1-z_{ci})l_{cij}}(1 - x_{cj1})^{z_{ci}(1-l_{cij})}x_{cj1}^{z_{ci}l_{cij}} \quad (4)$$

This means that the total likelihood for labels within context c is a product of binomial distributions, i.e.

$p(\mathbf{L}_c|\mathbf{Z}_c, \mathbf{X}_c) = \prod_{ij} p(l_{cij}|z_{ci}, x_{cj})$, and is hence conjugate to a product of beta distributions (one for each specificity or sensitivity parameter). Therefore, the *variational conjugate* prior over expertise parameters is a joint beta distribution, parametrised by $\alpha_{cjmn}, m, n \in \{0, 1\}$, where

$$p(x_{cj}|\xi_c, \phi_j) = \prod_{m\in\{0,1\}} \text{Beta}(x_{cjm}|\alpha_{cjmm}, \alpha_{cjm\bar{m}}) \quad (5)$$

with $\bar{m}$ the complement of m. We interpret α_{cjmm} as the prior evidence for annotator j correctly labelling items of ground-truth m, with $\alpha_{cjm\bar{m}}$ for incorrect labelling. These prior weights depend on the challenge, ξ_c, and trust, ϕ_j, parameters. We assume a separate challenge and trust component for each α-parameter, and a simple summative model as this aids model interpretation[2], i.e.

$$\alpha_{cjmn} = \xi_{cmn} + \phi_{jmn} + 1 \quad (6)$$

Adding the 1 forces $\xi_{cmn} \geq 0$ – and can now be interpreted as evidence for the prototypical annotator in context c marking an object of ground-truth m with label n. Similarly, $\phi_{jmn} \geq 0$, and is the equivalent evidence for annotator j in a general context. This means $\xi_{cmm} = \xi_{cm\bar{m}} = \phi_{jmm} = \phi_{jm\bar{m}} = 0$, corresponds to zero information about x_{cjm}. Introducing Eqn. (6) into Eqn. (3) gives

$$\ln q^\star(x_{cj}) = \sum_{m\in\{0,1\}} \Big((\alpha'_{cjmm} - 1) \ln x_{cjm} + (\alpha'_{cjm\bar{m}} - 1) \ln(1 - x_{cjm}) \Big) + \text{const} \quad (7)$$

$$\alpha'_{cjmn} = \mathbf{E}\xi_{cmn} + \mathbf{E}\phi_{jmn} + \sum_{i\in\mathcal{O}_c} \Big((1 - l_{cij})^{(1-n)} l_{cij}{}^{n} (1 - \mathbf{E}z_{ci})^{(1-m)} \mathbf{E}z_{ci}{}^{m} \Big) \quad (8)$$

As anticipated, this is in the form of a joint-Beta distribution, and is conjugate with the prior. A similar reasoning process helps us determine distributions and parameters for z_{ci}, υ_c, so we just give the highlights here.

Ground-truth values By a similar inspection, we see that the log-variational over ground-truth values is the form of a Bernoulli distribution. Given a prior bias of υ_c, it follows that

$$p(z_{ci}|\upsilon_c) = (1 - \upsilon_c)^{(1-z_{ci})} \upsilon_c^{z_{ci}} \quad \text{and}$$
$$q^\star(z_{ci}) = (1 - \upsilon'_{ci})^{(1-z_{ci})} \upsilon'^{z_{ci}}_{ci} \quad (9)$$

Here, the updated bias, υ'_{ci}, is given by $\upsilon'_{ci} = \tilde{\upsilon}'_{ci1}/(\tilde{\upsilon}'_{ci0} + \tilde{\upsilon}'_{ci1})$ where for $m \in \{0, 1\}$

$$\ln \tilde{\upsilon}'_{cim} = \mathbf{E} \ln \left((1 - \upsilon_c)^{(1-m)} \upsilon_c^m \right) + \sum_j \Big((1 - l_{cij}) \mathbf{E} \ln \left(x_{cj0}^{(1-m)} (1 - x_{cj1})^m \right) + l_{cij} \mathbf{E} \ln \left((1 - x_{cj0})^{(1-m)} x_{cj1}^m \right) \Big) \quad (10)$$

[2]There is no conventional choice for this relationship, and an alternative might be multiplicative composition, e.g. $\alpha_{cjmn} = \xi_{cmn}\phi_{jmn} + 1$.

As noted, evaluation of the expertise parameters, Eqn. (7), requires the expected value of this distribution, namely $\mathbf{E}z_{ci} = \tilde{\upsilon}'_{ci1}/(\tilde{\upsilon}'_{ci0} + \tilde{\upsilon}'_{ci1})$.

Ground-truth parameters Applying Eqn. (1) to the υ_c parameters, the log-variational again takes the form of a log Beta distribution. We therefore define the conjugate prior, $p(\upsilon_c|\theta_\upsilon) = \text{Beta}(\upsilon_c|\beta_1, \beta_0)$ with parameters $\theta_\upsilon = \{\beta_1, \beta_0\}$, and the variational posterior becomes

$$q(\upsilon_c) = \text{Beta}(\upsilon_c|\beta'_{c1}, \beta'_{c0}) \quad (11)$$
$$\beta'_{cm} = \beta_m + \sum_{i\in\mathcal{O}_c} (\mathbf{E}z_{ci})^m (1 - \mathbf{E}z_{ci})^{(1-m)}$$

We can now evaluate the required expectations for Eqn. (9), using Eqn.s (7) and (11), and a standard result for the expectation of a log-Beta variable (Bishop, 2007).

$$\mathbf{E} \ln \upsilon_c = \psi(\beta'_{c1}) - \psi(\beta'_{c1} + \beta'_{c0}) \quad (12)$$
$$\mathbf{E} \ln(1 - \upsilon_c) = \psi(\beta'_{c0}) - \psi(\beta'_{c1} + \beta'_{c0})$$
$$\mathbf{E} \ln x_{cjm} = \psi(\alpha'_{cjmm}) - \psi(\alpha'_{cjmm} + \alpha'_{cjm\bar{m}})$$
$$\mathbf{E} \ln(1 - x_{cjm}) = \psi(\alpha'_{cjm\bar{m}}) - \psi(\alpha'_{cjmm} + \alpha'_{cjm\bar{m}})$$

This uses the digamma function $\psi(x) = \frac{d}{dx} \ln \Gamma(x)$, which in turn depends the Gamma function, Γ, (Abramowitz & Stegun, 1964) (a generalisation of the factorial '!').

3.2. Challenge and Trust Parameters

Up to this point, we use a fairly conventional Variational Bayes approach (Bishop, 2007; Jaakkola & Jordan, 2000). Notable features being, a joint Bernoulli likelihood (Eqn. (4)), and our α_{cjmm} parameters (Eqn. (6)). However, the challenge parameters, ξ_c, and trust parameters, ϕ_j, give rise to problematic distributions, as we now show.

Within the binary label model, the challenge parameter ξ_{cmn} is the evidential weight given to a random annotator in context c, for assigning the label n to a random object with ground-truth m. Similarly, the trust parameter ξ_{jmn} is the evidential weight given to annotator j in a random context, for assigning the label n to a random object with ground-truth m. Both parameters types are similar, so we focus on challenge parameters, ξ_c. As with previous parameters, we begin by defining an approximate variational distribution using Eqn. (1) and independence assumptions from Eqn. (2), the resulting form factorises as: $q(\xi_c) = q(\xi_{c00}, \xi_{c01})q(\xi_{c10}, \xi_{c11})$. We now make two further independence assumptions for each context, namely

$$q(\xi_{cm0}, \xi_{cm1}) \overset{\text{ass.}}{=} q(\xi_{cm0})q(\xi_{cm1}) \quad \text{for } m \in \{0, 1\} \quad (13)$$

After some simplification this gives

$$\ln q^\star(\xi_{cmn}) = \mathbf{E}_{\xi_{cm\bar{n}}, \Phi_c} (G_{cmn}) + \ln p(\xi_{cmn}|\theta_\xi) + \xi_{cmn} \sum_{j\in\mathcal{A}_c} \mathbf{E}_{x_{cj0}} \ln \Big(p(l_{cij} = n|z_{ci} = m, x_{cj}) \Big) + \text{const} \quad (14)$$

which uses Eqn. (4) and

$$G_{cmn} = \sum_{j\in\mathcal{A}_c} \ln \frac{\Gamma(\xi_{cmn} + \phi_{jmn} + \xi_{cm\bar{n}} + \phi_{jm\bar{n}} + 2)}{\Gamma(\xi_{cmn} + \phi_{jmn} + 1)} \quad (15)$$

These distributions $q^\star(\xi_{cmn})$ do not have a standard form, and there is no natural choice for conjugate prior. However, we now show how to approximate $\ln q^\star(\xi_{cmn})$, by replacing the G_{cmn} term with a term of the form $a \ln \xi_{cmn}$ ($a > 0$). With this approximate form for the likelihood, and a conjugate Gamma prior, we can, in turn, approximate the variational posterior with a Gamma distribution, $\tilde{q}$, where

$$q^\star(\xi_{cmn}) \simeq \tilde{q}(\xi_{cmn}) \propto \xi_{cmn}^{(a'_{cmn}-1)} e^{(-b'_{cmn}\xi_{cmn})} \quad (16)$$

We argue that this is reasonable since: i) the exponential factor (far right) matches that in $q^\star$ precisely for all ξ_{cmn}, and this term will rapidly dominate for large ξ_{cmn}; and ii) the remaining factor in $q^\star$ scales in a *quasi*-power way. Identification of this candidate family of approximating distributions forms one of the contributions of our work. In the next section, we show how to approximate the target distribution by iteratively matching gradients of $q^\star$ and $\tilde{q}$ around the high probability region of $\tilde{q}$.

Reflexive Local-Variational Approximation Knowing our $\tilde{q}(\xi_{cmn})$ will be Gamma distributed, we define challenge and trust priors

$$p(\xi_{cmn}|\theta_\xi) = \text{Gamma}(\xi_{cmn}|a_{mn}, b_{mn})$$
$$p(\phi_{jmn}|\theta_\phi) = \text{Gamma}(\phi_{jmn}|d_{mn}, e_{mn})$$

with $\theta_\xi = \{a_{mn}, b_{mn} | m, n \in \{0, 1\}\}$ and $\theta_\phi = \{d_{mn}, e_{mn} | m, n \in \{0, 1\}\}$. The approximate posteriors of these four challenge and four trust parameters are then given by

$$p(\xi_{cmn}|\mathbf{X}_c, \theta_\xi) \simeq \tilde{q}(\xi_{cmn}) = \text{Gamma}(\xi_{cmn}|a'_{cmn}, b'_{cmn}) \quad (17)$$
$$p(\phi_{jmn}|\mathbf{X}_j, \theta_\phi) \simeq \tilde{q}(\phi_{jmn}) = \text{Gamma}(\phi_{jmn}|d'_{jmn}, e'_{jmn})$$

We now describe how to find these approximate posteriors with a general challenge parameter, ξ_{cmn}. The corresponding rate parameter, b'_{cmn}, takes the linear terms from the Gamma prior and likelihood (Eqn. (14)) to give

$$b'_{cmn} = b_{mn} - \sum_{j \in \mathcal{A}_c} \mathbf{E}_{x_{cj}} \ln \left(p(l_{cij} = n | z_{ci} = m, x_{cj}) \right) \quad (18)$$

For the appropriate constant H_{cmn}, the shape parameter, a'_{cmn}, approximates remaining (non-constant) terms of the posterior with:

$$(a'_{cmn} - 1) \ln \xi_{cmn} + H_{cmn} \quad (19)$$

To approximate $q^\star(\xi_{cmn})$ around some given point $\bar{\xi}_{cmn}$, we take the log of both sides of Eqn. (17) and match gradients at $\bar{\xi}_{cmn}$ with

$$a'_{mn} = \bar{\xi}_{cmn} \frac{d}{d\xi_{cmn}} \left(\mathbf{E}_{\xi_{cm\bar{n}}, \Phi_c} G_{cmn} \right) \Big|_{\bar{\xi}_{cmn}} + a_{mn} \quad (20)$$

where

$$\frac{d}{d\xi_{cmn}} \left(\mathbf{E}_{\xi_{cm\bar{n}}, \Phi_c} G_{cmn} \right) \Big|_{\bar{\xi}_{cmn}} \simeq \quad (21)$$
$$\sum_{j \in \mathcal{A}_c} \left(\psi(\xi_{cmn} + \mathbf{E}\phi_{jmn} + \mathbf{E}\xi_{cm\bar{n}} + \mathbf{E}\phi_{jm\bar{n}} + 2) - \psi(\xi_{cmn} + \mathbf{E}\phi_{jmn} + 1) \right)$$

Eqn. (21) requires a Local-Variational approximation to push the expectation term into the G_{cmn} function from Eqn. (15) (see Section 2.1 in the supplementary materials). All that is now needed is a way to choose the point $\bar{\xi}_{cmn}$ at which to evalutate Eqn. (20). This approximation is exact at some point $\bar{\xi}_{cmn}$, a lower bound below $\bar{\xi}_{cmn}$ and a close approximation elsewhere (see Section 2.2 in the supplementary materials).

Reflexive Fitting As discussed in Section 2, conventionally local-variational methods approximate the expectation of a function with respect to a known distribution. We wish to choose a good value of $\bar{\xi}_{cmn}$ so that $\tilde{q}(\xi_{cmn})$ estimates $p(\xi_{cmn}|\mathbf{X}_c, \theta_\xi)$ well in regions of high density. However, a good choice of $\bar{\xi}_{cmn}$ depends itself on $p(\xi_{cmn}|\mathbf{X}_c, \theta_\xi)$. This is a catch 22 situation: we cannot (cheaply[3]) find the expectation in Eqn. (20) until we approximate $q^\star$, but we cannot approximate $q^\star$ until we perform the expectation. Instead, we make the following observation. The variational approximation needs to be most accurate *close* to the majority of the probability density of $p(\xi_{cmn}|\mathbf{X}_c, \theta_\xi)$. Therefore, a natural choice is to fit around the mean, i.e. $\bar{\xi}_{cmn} = \mathbf{E}\xi_{cmn}$. As the mean is not known in advance, we make an initial estimate for $\bar{\xi}_{cmn}$, then evaluate a'_{mn} using Eqn. (20). We can then estimate the mean with $\mathbf{E}\xi_{cmn} = a'_{mn}/b'_{mn}$, before repeating the process. More precisely, starting with initial estimate $\tilde{\xi}_{(0)}$ and $\tilde{a}_{(0)}$, respectively for $\mathbf{E}\xi_{cmn}$ and a'_{mn}, we apply the following two updates until convergence,

$$\tilde{a}_{(i+1)} \leftarrow \tilde{\xi}_{(i)} \left(G'_{cmn}(\tilde{\xi}_{(i)}) + \frac{a_{mn} - 1}{\tilde{\xi}_{(i)}} \right) + 1 \qquad \text{and}$$

$$\tilde{\xi}_{(i+1)} \leftarrow \frac{\tilde{a}_{(i+1)}}{b'_{cmn}}$$

At convergence, $i \to \infty$, we assign $a'_{mn} = \tilde{a}_{(\infty)}$ and $\bar{\xi}_{cmn} = \tilde{\xi}_{(\infty)}$, and the expected values are estimated as: $\mathbf{E}\xi_{cmn} \simeq \bar{\xi}_{cmn}$. Fig. 3.2 illustrates this approach. The trust parameters, ϕ_{jmn}, are found in a similar way.

Validity of Approach We make a number of approximations to make our approach tractable, and it is reasonable to ask whether these are sufficiently flexible to allow the model to learn appropriately. To justify our choices, we first note that independence assumptions and lower-bounds have led to very successful and accurate variational approximations elsewhere (Bishop, 2007; Jaakkola & Jordan, 2000). In the results section, we compare the predictive accuracy of our approach applied to single context models versus the more conventional EM approximation. For the multi-context model, we make similar assumptions to allow the hyperparameter distributions to be learnt. We show how well this performs as part of the full model.

[3] Quadrature could be used here but it is expensive, see Sec. 4.

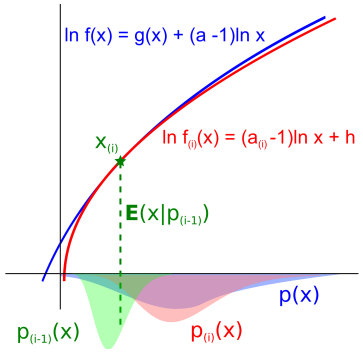


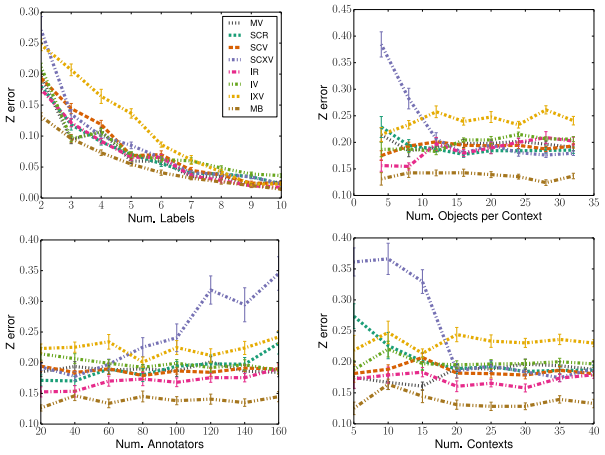
Figure 2. For parameters ξ_{cmn} and ϕ_{jmn} distributed as $p(x) = f(x)e^{-bx}$, we assume the form $p_{(i)}(x) \propto f_{(i)}(x)e^{-bx}$ and fit $\frac{d}{dx}\ln f_{(i+1)}\big|_{\bar{x}_i} = \frac{d}{dx}\ln f\big|_{\bar{x}_i}$ around the mean $\bar{x}_i = \mathbf{E}_{p_{(i)}}x$.

Finally, while many existing variational techniques directly minimise the KL-divergence between the target and approximating distributions [4]. Our hyperparameter approximation does not do so directly. Instead, our approach matches exactly the order of decay of the target distribution for large variable values, and uses gradient matching at the mean to improve the approximation for smaller variable values. Hence the distribution is approximated for a range of values, not just at a single point, and we argue that this has similar characteristics to KL minimisation. Also, hyperparameters have a less direct, and consequently weaker influence on the accuracy of a model's predictions. Therefore, even imperfect approximations of hyperparameter distributions can improve predictive accuracy of a model versus no fitting, as we see in the next section.

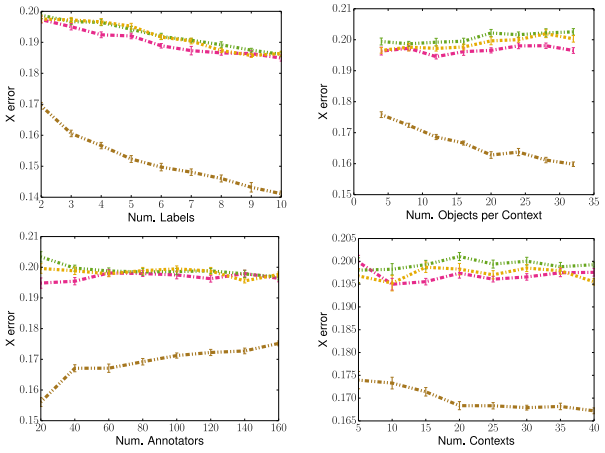
4. Results

In this section, we investigate the inferential and predictive accuracy on 1 synthetic dataset and 2 datasets from the literature: the headline affect score dataset from (Snow et al., 2008), and the Cub200 birds dataset from (Welinder & Perona, 2010). In all these experiments, method MV indicates a naive majority vote. Methods of the form SC* artificially *pool* all data into a single-context[5], and methods of the form I* fit each context independently. All SC* and I* methods are paired with three single-context methods: all *R fit the model shown in Fig. 1(a), using the EM method from (Raykar et al., 2010); all *XV use our novel variational fitting of the same model; and all *V use a novel variational fit of a simpler model, excluding ground-truth bias estimation (as in (Dawid & Skene, 1979)). Finally, MB is the context-aware approach outlined in Section 3.

In all experiments, probabilistic methods use non-informative ground-truth parameter priors of $(\beta_1, \beta_0) = (1., 1.)$, except *V which cannot, and instead assumes a prior-bias of $\upsilon_c = 0.5$ for all c. *R, *V and *XV methods assume weakly informative expertise priors of $(\alpha_{cjmm}, \alpha_{cjm\bar{m}}) = (2., 1.)$ for all c and j; this breaks model symmetry by assuming the average annotator is marginally more likely to be correct than incorrect. For the same reason, MB assumes all challenge and skill shape priors $a_{mn} = d_{mn} = 1$ for all m, n; but assumes rate priors on correct answers of $b_{mm} = e_{mm} = 0.2$, while rate priors on incorrect answers are $b_{m\bar{m}} = e_{m\bar{m}} = 0.5$.



(a) Ground-truth misclassification rate



(b) Expertise errors (RMSE)

Figure 3. Ground-truth and expertise estimation errors on synthetic data, while scaling #labels per object, #objects per context, #annotators, or #contexts. Defaults are #labels per object=2, #objects per context=10, #annotators=80, and #contexts=20.

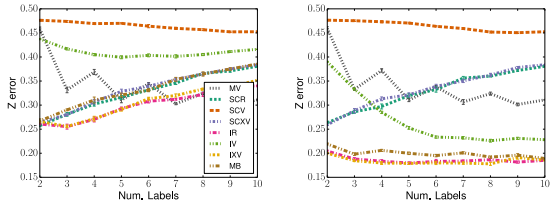
Figs. 6 (a) and (b) show performance on synthetic data, simulated by sampling from instantiations of the context-aware model, see Fig. 1(b); these use significantly stronger challenge and trust priors [6], than for the inference, to en-

[4] This is not necessarily implied by the term *variational*.

[5] Sometimes referred to as *data-pooling* (Mo et al., 2013)

[6] Synthetic data parameters are $a_{mm} = d_{mm} = 4$, $b_{mm} =$

courage high quality labels. Fig. 6 (a) shows that the ground-truth misclassification rate for MB is consistently equal to or better than all I* and SC* methods. I* methods suffer the most here, and this suggests that the advantage of pooling data (in SC* methods), outweighs uncertainty over ground-truth bias. The expertise errors in Fig. 6 (b), show MB substantially outperforms other methods, and directly reflects the advantages of challenge and trust parameters.



(a) Affect as context (b) Headline as context

Figure 4. Ground-truth errors on headline affect dataset, while scaling #labels.

We next investigate the headline affect data from (Snow et al., 2008), generated by a crowd of non-expert labellers scoring news headlines for anger, disgust, fear, joy, sadness, and surprise on a scale of 1 to 100. As in (Mo et al., 2013), we convert labels to binary *false* if less than 50 and *true* otherwise. Ground-truth relies on three *experts* labelling the same headlines, likewise converted to binary values. We can either model contexts as *question-type* with each headline as an object; or with headline as context, and each question as an object. In (Mo et al., 2013), they use question-type as context, and Fig. 4 shows performance as number of labels scales. This shows peculiar behaviour; methods SCR, SCXV, IR, IXV and MB all perform well with 2 or 3 labels per object, but method performance then degrades as labels are added, until worse than MV. This suggests that the models do not capture the data properties well and consequently overfit. IXV and IR are least affected by this suggesting little similarity between contexts. Finally, SCV and IV methods perform badly here suggesting extreme ground-truth bias. Using headline as context the performance is much better, in particular for IR, IXV and MB methods. The experiment from (Mo et al., 2013) first synthetically upsamples the data, by fitting a model (not described) to the total data, generating additional labels, then testing on a restricted set of upsampled data. They report average performance across all contexts of 0.84. This is marginally better than our IR, IXV and MB performance on headline-as-affect ($\sim$0.82 with as few as 3 labels per object).

Fig. 5 reports results on the Cub200 birds dataset from Caltech (Cub200) (Welinder & Perona, 2010), with responses

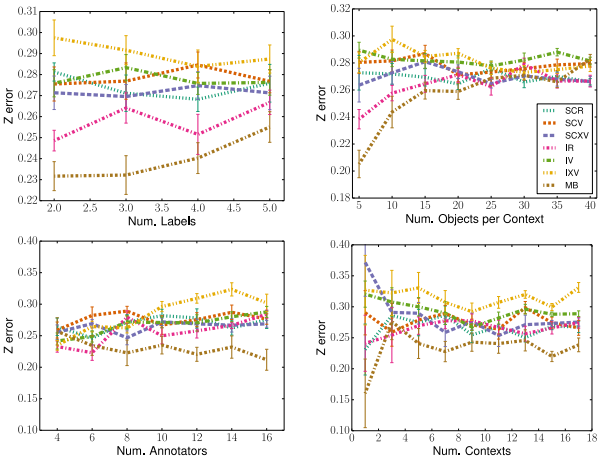


Figure 5. Ground-truth errors on Cub200 birds data, while scaling #objects per context, #annotators, or #contexts. The default #labels per object=3, #objects per context=10, #annotators=10, and #contexts=17.

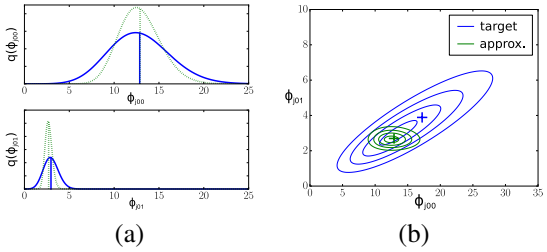
from crowdsourced `workers` (annotators), to questions on the presence/absence of `attributes`, e.g. curved beak shape, in `images` of birds. We define independent binary `attributes` from these to use as contexts, with randomly sampled (`attribute_id, image_id`) pairs as objects. Available objects and annotators are restricted to guarantee required label coverage. There is no objective ground-truth, so we use majority vote from all labels as a proxy (most objects had 5 labels in total). Notice that the accuracy degrades for more labels, and is most likely an artifact of choosing the majority vote as ground-truth. Nonetheless, our MB algorithm clearly outperforms both SC* and I* methods. This is particularly apparent when there are fewer labels per object or objects per context, or a large number of annotators overall. This suggests a possibly greater variation between contexts in the Cub200 data. However, I* methods are here less stable than SC*, so there is some value in pooling information across contexts here.

To illustrate the characteristics of our novel approximation approach, we show an example fit, see Figs. 4 (a) and (b). Fig. 4 (a) shows one dimensional conditional distributions for trust parameters ϕ_{j00} and ϕ_{j01} of a single annotator j. Each target distribution (blue) is of the form from Eqn. 3, with all other parameters as estimates from a 10 annotator 10 context model trained on Cub200 data. The reflexive approximation (Eqn. (17)) is shown in green. Means are indicated by vertical bars. Fig. 4 (b) shows the joint distribution (green) of the two approximations from the left figure, and the joint distribution (blue) that would pertain if assumptions of the form Eqn. (13) had not been made; the joint means are indicated by crosses.

It is the accuracy of the mean estimates from these reflex-

$e_{mm} = 1$, $a_{m\bar{m}} = d_{m\bar{m}} = 1$ and $b_{m\bar{m}} = e_{m\bar{m}} = 2$.

ive approximations that informs the rest of the inference (see Eqns. (8), (12), and (21)). The target distributions in Fig. 4 (a) are horizontal (ϕ_{j00}) and vertical (ϕ_{j01}) slices through the blue distribution from Fig. 4 (b) positioned at the *green* cross. These recover the means of the conditionals very accurately (this is a typical result). However, as the number of contexts increases, the distance between the means of the joint distributions increases (Fig. 4 (b)). This is due to the strong covariance between the two variables in the target distribution (again this is typical). Therefore, we conclude that while the reflexive fitting is remarkably accurate in recovering the conditional means, the assumptions from Eqn. (13) are the most likely cause for degraded performance with larger models. Note that a method to learn Dirichlet parameters is given in (Minka, 2000). However, it does not readily deal with our summative Dirichlet parameters from Eqn. 6, nor, as a maximum-likelihood method, would it accurately learn the distribution mean as desired.



(a) (b)

Figure 6. Example distributions for trust parameters ϕ_{jmn} from Cub200 data, using model predictions for other parameters. (a) Variational marginals, with targets (blue) and approximations (green). Means are vertical bars. (b) Variational joint target $q(\phi_{j00}, \phi_{j01})$ (blue) and approximation (green) is joint of independent approximations from (a). Means marked with $+$.

5. Discussion and Conclusion

This paper presents a context aware model for ground-truth inference with semi-trusted labels, and shows a significant accuracy improvement gained by using this approach. This would be of value to any investigator working with crowdsourced labels from multiple contexts, and is most beneficial with sparse data when information about annotators' reliability can be shared between contextual domains. Our approach also provides very high quality predictions about the expertise of particular annotators within a given context, and this information can be exploited by methods that efficiently harvest new labels, e.g. (Chen et al., 2013; Dickens & Lupu, 2014; Karger et al., 2011; Welinder & Perona, 2010). Interestingly, expertise estimates can be given for contexts without any existing labels, so this could be used to *transfer learning* to unseen contexts. Moreover, the newly emerging crowdsourcing market is one with few regulations and worker protections (Kittur et al., 2013). It is important that those of us making use of crowdsourced labour are not exploitative. The higher quality context free *trust* estimates provided by our model could be key in helping maintain good relationships with our contributors, for instance, using trust as part of performance related pay schemes.

Some probabilistic crowdsourcing models use latent variables to capture hidden characteristics, such as one or more dimensions of annotator skill and question difficulty (Welinder et al., 2010; Whitehill et al., 2009); this includes the context-aware model from (Mo et al., 2013). Extensions to our model could incorporate these ideas, but the increased complexity of the model may require sampling based estimation (as with (Mo et al., 2013)). We argue that models supporting approximate inference are greatly preferred, as these scale much more favourably, and can be incrementally updated where new data is continually added.

The paper also identifies a family of distributions that are good candidates for approximating Dirichlet hyperparameter distributions. Further, we present a novel self referential local variational technique to approximate these distribution within our model. We do not present a formal proof of convergence for the procedure here, but we do show that it can estimate challenge and trust parameters sufficiently well to outperform models that do not have such parameters. Directions for future work includes a formal proof of convergence, and possibly developing improved methods, for instance to capture the covariance between hyperparameters. Such approaches could be used to develop other similar context-aware methods based on different single-context models, e.g. those in (Welinder et al., 2010; Whitehill et al., 2009; Venanzi et al., 2013), or with any probabilistic models that have hyper-parameters on Dirichlet distributions, e.g. topic modelling with Dirichlet processes (Teh et al., 2004), where there could be big computational savings over sampling approaches. Multi-level models are also common elsewhere to test co-dependence, causal relationships and hidden factors in data (Gelman & Hill, 2007). Our general approach could be applied there too, perhaps with other troublesome distributions as well, such as the hyperparameters for Gamma distributions.

To conclude, while single context crowdsourcing models have had great success improving the quality of information extracted versus simple majority voting, there is additional informational value to be had where contributions are collected across multiple contexts. Moreover, this appears of most benefit when considering contributors' performance across contexts. Sampling approaches, while flexible, require costly retraining as new data is added, and this is not well suited to applications with a continuous input stream of contributions. Approximate inference may provide a way to overcome these limitations, but more advanced approximation techniques may yet be needed.

REFERENCES

Abramowitz, M. and Stegun, I.A. *Handbook of Mathematical Functions*. Dover, New York, fifth edition, 1964.

Bishop, Christopher M. *Pattern Recognition and Mach. Learning (Information Science and Statistics)*. Springer, October 2007.

Bragg, J, Mausam, and Weld, DS. Crowdsourcing multi-label classification for taxonomy creation. In *Conf. on Human Computation & Crowdsourcing (HCOMP)*, 2013.

Chen, X., Lin, Q., and Zhou, D. Optimistic knowledge gradient policy for optimal budget allocation in crowdsourcing. In *ICML*, pp. 64–72, 2013.

Chilton, Lydia B., Little, Greg, Edge, Darren, Weld, Daniel S., and Landay, James A. Cascade: crowdsourcing taxonomy creation. In *CHI*, pp. 1999–2008, 2013.

Dawid, A. P. and Skene, A. M. Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm. *Jnl. of the Royal Stats. Soc. Series C*, 28(1):20–28, 1979.

Dickens, L. and Lupu, E. On efficient meta-data collection for crowdsensing. In *Crowdsensing Workshop at PerCom*, 2014.

Gelman, A. and Hill, J. *Data Analysis using Regression and Multilevel/Hierarchical Models*. Analytical Methods for Social Research. Cambridge University Press, 2007. ISBN 9780521686891.

Ipeirotis, Panagiotis G., Provost, Foster, and Wang, Jing. Quality management on amazon mechanical turk. In *Proc. of ACM SIGKDD Workshop on Human Computation*, HCOMP '10, pp. 64–67, New York, NY, USA, 2010. ACM.

Jaakkola, T. S. and Jordan, M. I. Bayesian parameter estimation via variational methods. *Statistics and Computing*, 10:25–37, 2000.

Karger, David R., Oh, Sewoong, and Shah, Devavrat. Budget-optimal crowdsourcing using low-rank matrix approximations. In *Allerton*, pp. 284–291. IEEE, 2011. ISBN 978-1-4577-1817-5.

Kittur, A., Nickerson, J., Bernstein, M., Gerber, E., Shaw, A., Zimmerman, J., Lease, M., and Horton, J. The Future of Crowd Work. In *CSCW*, pp. 1301–1318, 2013.

Minka, T. P. Expectation propagation for approximate bayesian inference. In *UAI*, pp. 362–369, 2001.

Minka, Thomas P. Estimating a dirichlet distribution. Technical report, Microsoft Research, 2000. (revised 2003, 2009, 2012).

Mo, Kaixiang, Zhong, Erheng, and Yang, Qiang. Cross-task crowdsourcing. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, pp. 677–685, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2174-7.

Parisi, G. *Statistical Field Theory*. Addison-Wesley, Redwood City, CA, 1988.

Raykar, V. C., Yu, S., Zhao, L. H., Valadez, G. H., Florin, C., Bogoni, L., and Moy, L. Learning from crowds. *JMLR*, 11:1297–1322, 2010.

Rzhetsky, Andrey, Shatkay, Hagit, and Wilbur, W. John. How to get the most out of your curation effort. *PLoS Comp. Bio.*, 5(5):e1000391, May 2009. doi: 10.1371/journal.pcbi.1000391. URL http://dx.doi.org/10.1371/journal.pcbi.1000391.

Snow, R., Jurafsky, D., and Ng, A. Y. Cheap and Fast — But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. In *EMNLP*, pp. 254–263, 2008.

Teh, Yee Whye, Jordan, Michael I., Beal, Matthew J., and Blei, David M. Sharing clusters among related groups: Hierarchical dirichlet processes. In *NIPS*, 2004.

Venanzi, M., Rogers, A., and Jennings, N. R. Trust-based fusion of untrustworthy information in crowdsourcing applications. In *AAMAS*, pp. 829–836, 2013.

Welinder, P. and Perona, P. Online crowdsourcing: rating annotators and obtaining cost-effective labels. In *CVPR*, pp. 2262–2269, 2010.

Welinder, P., Branson, S., Belongie, S., and Perona, P. The multidimensional wisdom of crowds. In *NIPS*, pp. 2424–2432, 2010.

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*, pp. 2035–2043, 2009.

Yan, Yan, Rosales, Rómer, Fung, Glenn, and Dy, Jennifer G. Active learning from crowds. In *ICML*, pp. 1161–1168. Omnipress, 2011.

Zooniverse. Zooniverse: Real Science Online. Web Site/Service. URL www.zooniverse.org. accessed on: 24/10/2013.