Proceedings of Machine Learning Research 95:192-207, 2018 ACML 2018

Distinguishing Question Subjectivity from Difficulty for Improved Crowdsourcing

Yuan Jin[1] YUAN.JIN@DEAKIN.EDU.AU
Mark Carman[2] MARK.CARMAN@MONASH.EDU
Ye Zhu[1] YE.ZHU@IEEE.ORG
Wray Buntine[2] WRAY.BUNTINE@MONASH.EDU
[1] *School of Information Technology, Deakin University, Melbourne, Australia*
[2] *Faculty of Information Technology, Monash University, Melbourne, Australia*

Editors: Jun Zhu and Ichiro Takeuchi

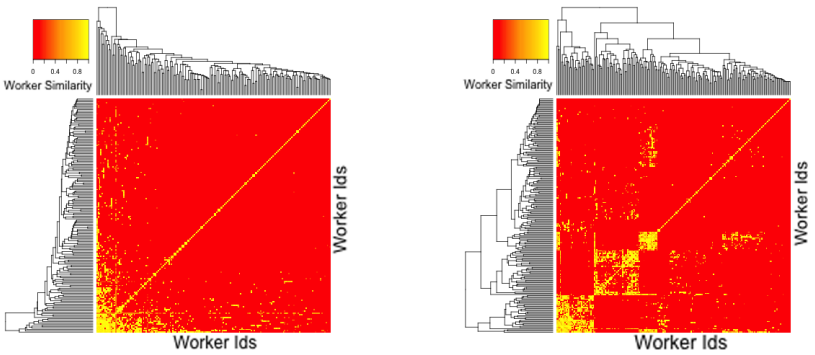
Abstract

The questions in a crowdsourcing task typically exhibit varying degrees of *difficulty* and *subjectivity*. Their joint effects give rise to the variation in responses to the same question by different crowd-workers. This variation is low when the question is easy to answer and objective, and high when it is difficult and subjective. Unfortunately, current quality control methods for crowdsourcing consider only the question difficulty to account for the variation. As a result, these methods cannot distinguish workers' *personal preferences* for different correct answers of a *partially subjective* question from their *ability* to avoid objectively incorrect answers for that question. To address this issue, we present a probabilistic model which (i) explicitly encodes question difficulty as a model parameter and (ii) implicitly encodes question subjectivity via latent *preference factors* for crowd-workers. We show that question subjectivity induces grouping of crowd-workers, revealed through clustering of their latent preferences. Moreover, we develop a quantitative measure for the question subjectivity. Experiments show that our model (1) improves both the question true answer prediction and the unseen worker response prediction, and (2) can potentially provide rankings of questions coherent with human assessment in terms of difficulty and subjectivity.

Keywords: Crowdsourcing, Subjectivity, Difficulty, Statistical modelling

1. Introduction

Outsourcing tasks to a flexible online workforce (aka crowdsourcing) has proven a successful paradigm for data collection in numerous fields due primarily to its overall lower costs and shorter turnaround time as compared to in-house expert-based data collection. The downside of online crowdsourcing is that the quality of the answers collected from crowd-workers is usually not guaranteed, even when multiple responses are collected and aggregated for each question, and workers are trained and vetted using gold-standard questions. To address this issue, quality control methods for the crowdsourced answers have been proposed. These methods rely on the assumptions that most crowd-workers are reliable when answering the questions and that a given worker is more likely to be reliable should she agree with the majority of her co-workers on the majority of their jointly answered questions. Thus, the methods have focused on modelling the *ability* of individual workers, assuming this to be correlated with the quality of their responses (Dawid and Skene, 1979; Venanzi



(a) (b)

Figure 1: Heatmaps showing inter-worker response similarity (% of response agreement) for two different tasks: (a) a *relatively objective* product matching task and (b) a *more subjective* fashion judging task, both involving binary worker responses. Hierarchical clustering was performed to order workers such that similar workers are close together. The three yellow blocks in (b) indicate three groups of response behaviour and higher subjectivity for task (b).

et al., 2014; Tian and Zhu, 2015). In recent years, it has become popular for quality control methods to also model the influence that individual questions exert on the quality of the responses (Whitehill et al., 2009; Welinder et al., 2010; Wang et al., 2017). Broadly speaking, the following two key properties of questions have drawn the modelling attention:

- **Difficulty.** The modelling of question difficulty is founded on the assumption that greater agreement among workers' responses to a question indicates less difficulty for them in determining the correct response. Quality control methods often encode this assumption using a function in which worker ability counteracts question difficulty for predicting the probability of a correct response. The probability is known as the *quality of the response*: the more difficult the question, the lower the quality of a response, and vice versa.

- **Subjectivity.** In crowdsourcing, there exist tasks that contain (*purely* or *partially*) subjective questions (Nguyen et al., 2016). Intuitively, the degree of subjectivity of a question depends on the number of response options that are correct. Being purely subjective means all of the options are correct, while being partially subjective means more than one but not all of them are correct. It is widely recognised that even expert assessors can disagree with each other on the correct responses to questions in crowdsourcing tasks, such as relevance judgement, which are considered to be partially subjective (Voorhees, 1998). In this case, the objectivity assumption on the questions does not hold and most of the quality control methods based on this assumption cannot distinguish the *ability* of workers from their *preferences* for the different response options.

Note that question subjectivity is different from the *domain* difficulty of *objective* questions which involves different groups of workers finding different types of objective questions more or less challenging than others. Handling such domain difficulty would require the modelling of *domain-level expertise*. For example, geography students might find physics questions more difficult than geography questions and vice versa. Modelling domain-level expertise and (objective) difficulty has been studied by Welinder et al. (2010). We did not study this

subject for this particular work but extending our model to account for domain expertise and difficulty is very straightforward.

When dealing with subjective questions, collaborative filtering for recommendation (Koren et al., 2009) considers users who share similar preferences to form *groups*. Those within the same group respond similarly towards subjective questions (e.g. rating movies) which share certain characteristics. This grouping effect can also be observed in crowdsourcing when crowd-workers respond to partially subjective questions. Figure 1 illustrates this effect by providing heat maps of pairwise worker similarity for two tasks: (a) a relatively objective task and (b) a more subjective one. The objective task required workers to judge whether a pair of products were the same based on their names, descriptions and prices, while the subjective task asked workers to judge whether an image contained "fashion related items"[1]. The similarity between pairs of workers is calculated as the percentage of agreement across the jointly answered questions[2] and hierarchical clustering was performed to group similar workers together. The three yellow boxes along the diagonal for the more subjective task (b) indicates the three distinct groups of worker response behaviour for this task, which was absent in the more objective task (a). Since the workers were mostly reliable on both tasks, we conjecture that the grouping of workers in the fashion judgement task reflects their tendencies for giving different correct responses to the same question (due to its subjectivity).

To facilitate the response quality control for the above tasks and generally, any crowdsourcing task that exhibits arbitrary degrees of question subjectivity and difficulty, we are motivated to develop a statistical model encoding both these properties. More specifically, question subjectivity causes groups of crowd-workers to emerge. A group specifies a particular correlation between the crowd-workers within it and the latent correct responses to different partially or purely subjective questions. We model such a correlation by factorising it into the latent preferences of the workers and the latent features of those questions. The assumption is that workers with similar latent preferences tend to perceive the same correct response when responding partially or purely subjective questions with similar characteristics. Specifically for the partially subjective questions, they possess certain degrees of difficulty. This difficulty corrupts the crowd-workers' perceptions as to (what tend to be) the correct responses to the question (determined by its subjectivity) to various extents depending on its level against the workers' levels of expertise. We model a greater extent of the corruption as a lower probability that the worker's response is equal to the subjective (worker-specific) correct response to the question, thereby the lower quality of the worker's response.

Our proposed model *considers both the subjective (i.e. worker-specific) truths regarding the correct answers to individual questions and also the difficulty-dependent probability that a worker's response to a question will equal her perceived subjective truth.* It encodes the question difficulty explicitly and the question subjectivity implicitly via latent variables for worker preferences and corresponding question features. Based on this model, we provide a Monte Carlo simulation approach for quantifying question subjectivity as the expected number of subjective truths perceived by different groupings of crowd-workers with respect to their preferences. Finally, we derive from the model a ranking of questions in terms of either difficulty or subjectivity which turns out to be coherent with human assessment.

1. The datasets for the two tasks have been listed in section 5.
2. Pairs of crowd-workers not sharing any questions had their similarity set to be .00001.

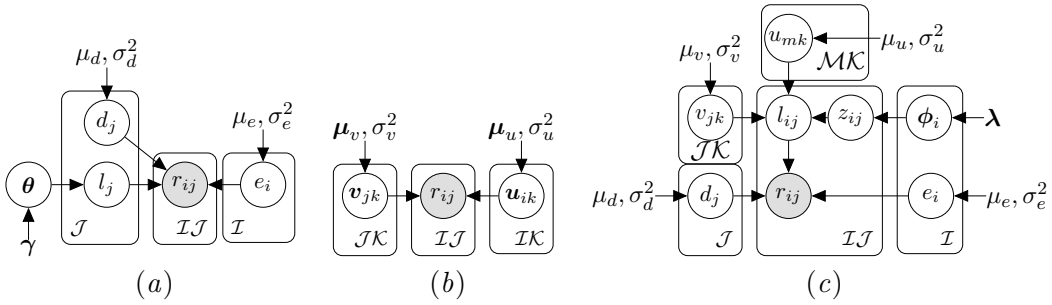


Figure 2: (a) shows GLAD with a latent variable l_j for each objective truth, (b) shows a collaborative filtering model without objective truths, and (c) is the proposed *subjectivity-and-difficulty response* (SDR) model for partially-subjective questions that is able to distinguish question difficulty from subjectivity.

2. Related Work

2.1. Latent Variable Modelling in Crowdsourcing

Most state-of-the-art response quality control methods in crowdsourcing have operated under the assumption that each question is purely objective. These methods are primarily based on statistical modelling of the interactions between crowd-workers and questions which determine either the marginal probabilities of the workers' responses equal to the corresponding correct answers (Whitehill et al., 2009; Welinder et al., 2010) or the conditional probabilities of the responses given the correct answers (Dawid and Skene, 1979; Venanzi et al., 2014). In comparison, the marginal probabilistic modelling is simpler than the conditional modelling, and also better at mitigating response sparsity problem in crowdsourcing. The basic marginal probabilistic modelling is GLAD from Whitehill et al. (2009), which models the correctness of each response as a logistic function where the question difficulty counteracts the expertise of the responding worker. Its graphical representation is shown by Figure 2(a) with the following generative scheme for a response r_{ij} of worker i given to question j: $\boldsymbol{\theta} \sim Dir(\boldsymbol{\gamma})$; $l_j \sim Discrete(\boldsymbol{\theta})$; $r_{ij} \sim Discrete(\boldsymbol{\pi}_{ij}|l_j)$. This means a true answer l_j is drawn for question j from a discrete distribution parametrised by $\boldsymbol{\theta}$, which was previously drawn from a Dirichlet distribution parametrised by $\boldsymbol{\gamma}$. Then, a response r_{ij} is drawn, conditioned on l_j, from a discrete distribution over the set of response options $\mathcal{K}$ with the k-th component of its parameters $\boldsymbol{\pi}_{ij}$ calculated as follows:

$$\pi_{ijk} = f(e_i, d_j) \text{ if } k = l_j \text{ else } \frac{(1 - f(e_i, d_j))}{K - 1}; \; f(e_i, d_j) = \frac{1}{1 + e^{-(e_i / \exp(d_j))}} \tag{1}$$

The function f takes in the expertise factor e_i of worker i and the difficulty factor d_j of question j. The output of the function is the probability of the response r_{ij} being correct. When $e_i \to +\infty$ or $d_j \to 0$, this probability grows, indicating a stronger positive correlation between r_{ij} and l_j. When $e_i \to 0$ or $d_j \to +\infty$ and the question has binary options, the probability approaches 0.5, which suggests r_{ij} is arbitrarily picked. When $e_i \to -\infty$, the probability decreases to 0, indicating a stronger negative correlation.

Few quality control methods have considered modelling the subjectivity of questions. One of the two papers that have made progress in this regard is by Tian and Zhu (2012). It assumes that a higher joint degree of difficulty and subjectivity for a crowdsourcing task can increase (decrease) the number of groups of responses given to the questions. The expected

size of each group becoming smaller (larger) indicates overall weaker (stronger) correlations of responses given to the questions. This paper makes no attempt in separating difficulty and subjectivity when only the difficulty accounts for the *quality of responses*. Moreover, this work requires every question to be answered by every worker, which is unrealistic in crowdsourcing. The other work by Nguyen et al. (2016) has focused on modelling partially subjective questions with just ordinal answers. It assumes each response to a question is generated by a univariate Gaussian the mean and the variance of which are linearly regressed over the observed features of the question. This means the model fits poorly any multi-modal distribution of responses (e.g. responses distributed only on rating 1 and 5).

2.2. Latent Variable Modelling in Collaborative Filtering

In model-based collaborative-filtering (Koren et al., 2009), matrix factorization is applied to predicting ordinal ratings provided by users to items (e.g. movies). Its categorical version, shown in Figure 2(b), is less commonly applied but is important for the construction of our model for the quality control of crowdsourced categorical responses. It has a generative process: $r_{ij} \sim Discrete(\boldsymbol{\psi}_{ij})$ where $\boldsymbol{\psi}_{ij} = \{\psi_{ijk}\}_{k \in \mathcal{K}}$ with its k-th component calcuated as:

$$\psi_{ijk} = P(r_{ij} = k | \boldsymbol{U}_i, \boldsymbol{V}_j) = \exp(\boldsymbol{u}_{ik}^T \boldsymbol{v}_{jk}) / \sum_{k' \in \mathcal{K}} \exp(\boldsymbol{u}_{ik'}^T \boldsymbol{v}_{jk'}) \tag{2}$$

Here, $\boldsymbol{\psi}_{ij}$ is also called the *soft-max* function, $\boldsymbol{u}_{ik}$ and $\boldsymbol{v}_{jk}$ are respectively the latent preferences of worker i and the latent features of item j in relation to the k-th answer option. The inner product term $\boldsymbol{u}_{ik}^T \boldsymbol{v}_{jk}$ indicates how much tendency user i responds to item j with the k-th answer option.

3. Proposed Model

Our proposed model endeavours to combine the key characteristics of the latent variable models specified in sections 2.1 and 2.2. We call it *SDR* model (Subjectivity-and-Difficulty Response model), which comprises an *upstream module* which generates a *subjective* truth for a question based on the worker's perception of the correct answer, and a *downstream module* which imposes a *difficulty*-dependent corruption on the subjective truth for generating the actual response from the worker to the question. More specifically, in the upstream module, the latent subjective truth l_{ij} of question j as perceived by crowd-worker i is drawn from a soft-max function specified by Eq. (2) except that the original r_{ij} in the equation is now replaced by l_{ij}. This function explains how the worker's latent preferences interact with the question's latent features to generate the subjective truth behind her response to the question. In the downstream, conditioned on the latent subjective truth l_{ij}, the response r_{ij} actually given by worker i to question j is determined by the logistic function $f(e_i, d_j)$. It encodes how the worker expertise e_i counteracts the question difficulty d_j to corrupt the subjective truth into the response, which will be defined later in this section. Essentially, the above perception-corruption process is a generalisation of the corruption process of the correct answer signals from objective questions modelled in Welinder et al. (2010) by additionally considering the question subjectivity.

Unfortunately the upstream+downstream model described above suffers from an over-parameterisation issue whereby *both* the upstream component (which determines the worker-specific correct answer) and the downstream component (which determines the noise resulting

from worker inaccuracy) can *independently and adequately* explain the variance observed in worker responses to the same question. In other words, the varied responses from different workers to the same question could equally be due to different perceptions on what constitutes the correct answer to the question or to difficulty of the question causing low accuracy amongst the respondents. To remedy this situation we explicitly enforce a group structure over workers in order to limit the variation in the perceptions across workers. This is done by changing the upstream module to have *sparsity-inducing priors* over the latent preferences of crowd-workers. In this paper, we use the *Latent Dirichlet Allocation* (LDA) (Blei et al., 2003) as such priors. The final graphical representation of the SDR model is shown in Figure 2(c).The new upstream module of our model assigns a probability vector $\boldsymbol{\phi}_i$, which follows a Dirichlet with a concentration parameter $\boldsymbol{\lambda}$, to each worker i. Each component ϕ_{mi} of this probability vector reflects the worker's tendency of showing a particular preference m among the set of preferences $\mathcal{M}$ she possesses when answering any question. Then, a preference assignment z_{ij} is drawn from $\boldsymbol{\phi}_i$ for determining the specific preference worker i will show for answering question j. As for preference m, it has a weight u_{mk} for each answer option k to reflect how likely each option is to be selected given the preference m showed by any worker. In this paper, we fix the dimension of u_{mk} to be strictly 1. This weight is multiplied with the latent feature v_{jk} of question j and the result is input to a soft-max function for drawing the subjective truth l_{ij} behind the response r_{ij}. The above generative process can be formulated as: $\boldsymbol{\phi}_i \sim Dir(\boldsymbol{\lambda})$; $z_{ij} \sim Discrete(\boldsymbol{\phi}_i)$; $l_{ij} \sim Discrete(\boldsymbol{\psi}_{z_{ij}})$ with the k-th component of the soft-max function $\boldsymbol{\psi}_{z_{ij}}$ calculated as:

$$\psi_{z_{ij}k} = \exp(u_{z_{ij}k} v_{jk}) / \sum_{k' \in \mathcal{K}'} \exp(u_{z_{ij}k'} v_{jk'}) \tag{3}$$

Embodying the sparsity-inducing effect of LDA, the preference probabilities $\boldsymbol{\phi}_i$ are dedicated to revealing the underlying groups of crowd-workers while the soft-max specified by Eq. (3) governs the positive correlations between the latent correct answers to the same questions perceived by the workers within the same group. When the number of preferences in $\mathcal{M}$ = 1, the probability of the only preference ϕ_i is 1. This has a two-fold meaning that each question has one correct answer and every worker should perceive the correct answer of any question in the same way. When the size of $\mathcal{M}$ is greater than 1, this indicates certain numbers of underlying worker groups, which we can recover by applying K-means clustering to the estimated preference probabilities $\hat{\boldsymbol{\phi}}_i$ using the Elbow method to determine the right number of the groups.

The downstream module corrupts the correlations between the subjective truth l_{ij} and the response r_{ij}. It draws r_{ij} from a discrete probability distribution $\boldsymbol{\pi}_{ij}$ specified in Eq. (1) except the logistic function $f(e_i, d_j)$ has the following definition from Rasch (1960):

$$f(e_i, d_j) = 1/(1 + e^{-(e_i - d_j)}) \tag{4}$$

The term $(e_i - d_j)$ naturally explains the type of biases induced by *deceptive* questions when the difficulty d_j is much larger than the expertise e_i. This is not captured in Eq. (1) as the term $\exp(d_j)$ is never smaller than 0, meaning questions never bias workers to answer incorrectly due to their difficulty. Moreover, when the estimated values for this term are greater than zero for most responses, it means SDR deems them more likely to be correct. With more of them deemed correct, the number of inferred correct answers to any question tends to increase. As a result, the size of latent preference set $\mathcal{M}$ should grow, from the perspective of SDR, to fit the seemingly more diverse set of correlations between latent

correct answers across the questions. Thus, for our model to recover the right number of latent preferences for crowd-workers from their responses, the priors for e_i and d_j need to be set properly, which will be elaborated more in section 5.1.

4. Estimation

4.1. Model Parameter Estimation

We now provide equations used for parameter estimation, using the notation $\psi_{z_{ij}k}$ from Eq. (3) and $f(e_i, d_j) = f_{ij}$ from Eq. (4) to simplify the equations. The conditional probability for the preference assignment z_{ij} to worker i when answering question j is:

$$P(z_{ij}=m|e_i, d_j, \boldsymbol{u}_m, \boldsymbol{v}_j, \lambda) \propto \sum_{k\in\mathcal{K}} \psi_{mk} f_{ij}^{\delta_{ijk}} \left(\frac{1-f_{ij}}{K-1}\right)^{1-\delta_{ijk}} \frac{N^m_{i\neg j} + \lambda_m}{\sum_{m'\in\mathcal{M}} N^{m'}_{i\neg j} + \lambda_{m'}} \tag{5}$$

where $N^m_{i\neg j}$ denotes the number of questions excluding question j answered by worker i given her preference m. The joint probability of the other parameters given z_{ij} is:

$$\begin{aligned} Q &= p\Big(\{e_i\}_{i\in\mathcal{I}}, \{d_j, \boldsymbol{v}_j\}_{j\in\mathcal{J}}, \{\boldsymbol{u}_m\}_{m\in\mathcal{M}} | \{\boldsymbol{z}_{ij}\}_{i\in\mathcal{I}, j\in\mathcal{J}}, \mu_{\{e,d,u,v\}}, \sigma^2_{\{e,d,u,v\}}\Big) \\ &= -\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}} \log\left[\sum_{k\in\mathcal{K}} \left(\psi_{z_{ij}k} f_{ij}^{\delta_{ijk}} \left(\frac{1-f_{ij}}{K-1}\right)^{1-\delta_{ijk}}\right)\right] - \sum_{i\in\mathcal{I}} \log\left(p(e_i|\mu_e, \sigma^2_e)\right) - \\ &\quad \sum_{j\in\mathcal{J}} \log\left(p(d_j|\mu_d, \sigma^2_d)\right) - \sum_{k\in\mathcal{K}}\sum_{m\in\mathcal{M}} \log\left(p(u_{mk}|\mu_u, \sigma^2_u)\right) - \sum_{j\in\mathcal{J}} \log\left(p(v_{jk}|\mu_v, \sigma^2_v)\right) \end{aligned} \tag{6}$$

The partial derivatives for Q with respect to the other parameters:

$$\frac{\partial Q}{\partial u_{mk}} = -\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}} \zeta_{ijm} v_{jk} \sum_{k'\in\mathcal{K}} f_{ij}^{\delta_{ijk'}} \left(\frac{1-f_{ij}}{K-1}\right)^{1-\delta_{ijk'}} \omega_{ijkk'} + \frac{u_{mk} - \mu_u}{\sigma^2_u} \tag{7}$$

$$\frac{\partial Q}{\partial v_{jk}} = -\sum_{i\in\mathcal{I}} u_{z_{ij}k} \sum_{k'\in\mathcal{K}} f_{ij}^{\delta_{ijk'}} \left(\frac{1-f_{ij}}{K-1}\right)^{1-\delta_{ijk'}} \omega_{ijkk'} + \frac{v_{jk} - \mu_v}{\sigma^2_v} \tag{8}$$

$$\frac{\partial Q}{\partial e_i} = -\sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}} \left(\frac{-1}{K-1}\right)^{1-\delta_{ijk}} f_{ij}(1-f_{ij}) \psi_{z_{ij}k} + \frac{e_i - \mu_e}{\sigma^2_e} \tag{9}$$

$$\frac{\partial Q}{\partial d_j} = -\sum_{i\in\mathcal{I}}\sum_{k\in\mathcal{K}} (-1)^{\delta_{ijk}} \left(\frac{1}{K-1}\right)^{1-\delta_{ijk}} f_{ij}(1-f_{ij}) \psi_{z_{ij}k} + \frac{d_j - \mu_d}{\sigma^2_d} \tag{10}$$

Here, $\delta_{ijk} = \mathbb{1}\{r_{ij} = k\}, \zeta_{ijm} = \mathbb{1}\{z_{ij} = m\}$and $\omega_{ijkk'} = \psi_{z_{ij}k}\big(1-\psi_{z_{ij}k}\big)^{\mathbb{1}\{k=k'\}}\big(-\psi_{z_{ij}k'}\big)^{\mathbb{1}\{k\neq k'\}}$. The parameter estimation involves two alternating procedures: sample z_{ij} according to Eq. (5) and optimize Q in Eq. (6) using LBFGS based on Eq. (7), (8), (9) and (10).

4.2. True Answer Estimation

A worker-specific subjective truth l_{ij} (as perceived by worker i) fails to provide overall information about the correct answers to partially subjective question j. Thus, we should gather the l_{ij} values from all workers who answer the question. However, in practice, a question is usually assigned to only a limited number (usually 3 or 5) of workers, making the

estimate of the true answer distribution poor. Our solution to improving this estimate is to first find underlying clusters of workers (across all questions) by applying K-means with the Elbow method based on 10-fold cross validation to the posterior means $\hat{\mathbf{\Phi}} = \{\hat{\boldsymbol{\phi}}_i | i \in \mathcal{I}\}$ of the latent preference probabilities of all the workers. With the centroid $\hat{\boldsymbol{\phi}}_c$ of each resulting cluster c, we then calculate the probability that the true answer l_{cj} (as perceived by the workers in cluster c) takes the value k as follows:

$$P(l_{cj} = k|c) = \sum_{m \in \mathcal{M}} P(l_{cj} = k|m)P(m|c) = \sum_{m \in \mathcal{M}} \left(\frac{\exp(\hat{u}_{mk}\hat{v}_{jk})}{\sum_{k'=1}^{K} \exp(\hat{u}_{mk'}\hat{v}_{jk'})} \hat{\phi}_{mc} \right) \tag{11}$$

where $\hat{u}_{mk}$ and $\hat{v}_{jk}$ are the estimates of the weight u_{mk} for preference m and the latent feature v_{jk} of question j, both specific to option k. The best estimate regarding the correct answer l_{cj} according to the workers assigned to cluster c is then:

$$\hat{l}_{cj} = \arg\max_{k \in \mathcal{K}} P(l_{cj} = k|c) \tag{12}$$

Now we have a set of correct answer estimates $\hat{\mathcal{L}}_j = \{\hat{l}_{cj} | c \in \mathcal{C}\}$ for question j from all the worker clusters (with $\mathcal{C}$ being the set of the clusters). For the task of true answer prediction, we choose one from $\hat{\mathcal{L}}_j$ as the estimate of the true answer l_j by following certain strategies. Two simple strategies are to choose $\hat{l}_{cj}$ from the cluster c with the highest average expertise over its workers, or from the cluster with the largest proportion of workers assigned to it. The first strategy states that the answer perceived by on average the most expert group of workers is the most appropriate, while the second assumes it to be the one perceived by the largest group of workers which represents the mainstream school-of-thought. In this paper, we apply the second strategy because most crowdsourcing datasets used in the experiments correspond to relatively simple tasks for which the provided correct answers we believe are more likely to be mainstream opinions. As for the first strategy, it can be more useful for revealing a minority group of expert workers who show distinct preferences from the public.

Algorithm 1: Subjectivity estimation for question j

Input: $\hat{v}_j$; $\{\hat{\boldsymbol{u}}_m\}_{m \in \mathcal{M}}$; $\hat{\mathbf{\Phi}}_c = \{\hat{\boldsymbol{\phi}}_c\}_{c \in \mathcal{C}}$; $T = 50,000$ (maximum number of iterations).
Output: $\mathbb{E}_{\mathcal{C}}[|\mathcal{L}_j|]$ /* Expected number of correct answers for question j */
$n_j \leftarrow 0$. /* Initialise number of correct answers for question j to zero */
for $t \leq T$ *and sample standard deviation of estimates for* $\mathbb{E}_{\mathcal{C}}[|\mathcal{L}_j|]$ *over* $(t-100, t] \geq 10^{-4}$ **do**
  $\hat{\mathcal{L}}_j \leftarrow \{\}$. /* Initialise set of correct answers to be sampled at iteration t. */
  for $c = 1...C$ **do**
    /*Sample group preference z_{cj} and correct answer $\hat{l}_{cj}$ perceived by worker cluster c.*/
    $z_{cj} \sim Discrete(\hat{\boldsymbol{\phi}}_c)$. $\hat{l}_{cj} \sim \boldsymbol{\psi}_{z_{cj}}$, where $\psi_{z_{cj}k} = \exp(\hat{u}_{z_{cj}k}\hat{v}_{jk}) / \sum_{k' \in \mathcal{K}} \exp(\hat{u}_{z_{cj}k'}\hat{v}_{jk'})$.
    $\hat{\mathcal{L}}_j \leftarrow \hat{\mathcal{L}}_j \cup \hat{l}_{cj}$ only if $\hat{l}_{cj} \notin \hat{\mathcal{L}}_j$ /* Add sampled $\hat{l}_{cj}$ to $\hat{\mathcal{L}}_j$ when it first appears. */
  end
  $n_j^{(t)} \leftarrow n_j^{(t-1)} + |\hat{\mathcal{L}}_j|$./* Increase n_j by number of distinct correct answers sampled at t.*/
  $\mathbb{E}_{\mathcal{C}}[|\mathcal{L}_j|] \approx \frac{n_j^{(t)}}{t}$. /* Divide n_j by t as estimate of question j's subjectivity at t. */
end

4.3. Subjectivity Estimation

Despite not being directly estimated in the model, question subjectivity can still be quantified and estimated after the model has been learned. This is achieved based on the reasonable assumption that the subjectivity of each question is proportional to the number of correct answers it has. Despite not knowing the actual number of correct answers $|\mathcal{L}_j|$ to question j, we can estimate the value by taking its expectation with respect to the clusters of workers derived in section 4.2. More precisely, the expected number of correct answers to question j with respect to worker clusters $\mathcal{C}$ is: $\mathbb{E}_{\mathcal{C}}[|\mathcal{L}_j|] = \sum_{n=1}^{|\mathcal{K}|} nP_{\mathcal{C}}(|\mathcal{L}_j| = n)$. In this equation, n iterates over the possible number of correct answers (from 1 to the size of $\mathcal{K}$). The probability $P_{\mathcal{C}}(|\mathcal{L}_j| = n)$ denotes how likely it is that the number of correct answers to question j equals n, with respect to the worker clusters $\mathcal{C}$. When $\mathcal{C}$ and $\mathcal{K}$ are large, it is difficult to calculate this probability due to a combinatorial explosion. Thus we apply Monte Carlo simulation to estimate (a measure of) the subjectivity of question j as $\mathbb{E}_{\mathcal{C}}[|\mathcal{L}_j|]$ using Algorithm 1. For testing the convergence of this algorithm, the *sample standard deviation* of estimates for $\mathbb{E}_{\mathcal{C}}[|\mathcal{L}_j|]$ over every 100 consecutive iterations (i.e. $(t-100, t]$) is monitored. The reason for using this quantity is that it is always large at the beginning of the sampling process. As the process proceeds, the sample standard deviation becomes smaller. When the process converges, it stabilises at a very small value ϵ. We observed $\epsilon = 10^{-4}$ to be a proper general value across the partially subjective datasets used in the experiments. Finally, the algorithm outputs $n_j^{(t)}/t$ as the estimate of $\mathbb{E}_{\mathcal{C}}[|\mathcal{L}_j|]$ where t is either the iteration that satisfies the 10^{-4} threshold or the maximum number of iterations T.

5. Experiments and Results

The evaluation of our proposed model consists of four parts. The first part is its sensitivity to various degrees of subjectivity in different crowdsourcing tasks. The second and the third parts are its performance of predicting respectively the provided correct answers of questions and the responses to be given by crowd-workers to unseen questions. The last part is its consistency with human assessors in assessing the difficulty and the subjectivity of questions. We have used 10 crowdsourcing datasets to evaluate the performance of our model across the four parts. Table 1 summarises these datasets as being either (primarily) objective or partially subjective. The identification tasks of *event time ordering* (Time), *dog* and *duck breeds*, and *identical products* (Product) concern objective factual knowledge. The judgement tasks of *image beauty* (Image), *document relevance* (Rel 1&2)[3], *fashionable items* (Fashion), *facial expression* (Face) and *adult content* (Adult) contain certain degrees of subjectivity.

5.1. SDR Hyper-parameter Setup

As discussed at the end of section 3, to find the right number of latent preferences, the hyper-parameters of the expertise e_i and the difficulty d_j in the SDR model need to be set properly. This is achieved by *held-out validation* which leverages *noise* within worker responses for detecting signs that SDR might be overfitting the responses by introducing more

3. The questions of relevance judgement task 2 come from the part of TREC 2011 crowdsourcing track (Lease and Kazai, 2011) that does not contain the questions of relevance judgement task 1. We collected crowdsourced judgements for the task 2 from CrowdFlower.

Table 1: The objective and the partially subjective datasets used in this paper.

Objective datasets	($\|\mathcal{I}\|$, $\|\mathcal{J}\|$, $\|\mathcal{R}\|$)	*Partially subjective datasets*	($\|\mathcal{I}\|$, $\|\mathcal{J}\|$, $\|\mathcal{R}\|$)
Time (Snow et al., 2008)	(76, 462, 4620)	**Image (Tian and Zhu, 2012)**	(402, 60, 24120)
Dog (Zhou et al., 2012)	(109, 807, 8070)	**Rel1 (Buckley et al., 2010)**	(642, 1787, 13310)
Duck (Welinder et al., 2010)	(53, 240, 9600)	**Rel2 (Lease and Kazai, 2011)**	(83, 585, 1755)
Product (Wang et al., 2012)	(176, 8315, 24945)	**Fashion (Loni et al., 2013)**	(199, 3837, 11511)
		Face (Mozafari et al., 2014)	(27, 584, 5242)
		Adult (adu)	(269, 333, 3324)

latent preferences than necessary. We construct a held-out validation dataset by randomly sampling a response from each worker. Thus, the size of such a dataset equals the number of workers participating in a task. Then, given a certain setting of the hyper-parameters, we learn our model based on the remaining responses and use the parameter estimates from the learned model to calculate the prediction accuracy: $1-MAE$ (Mean Absolute Error) over the held-out dataset. We repeat the model learning process with each hyper-parameter setting over the same 100 random held-out validation data subsets. We then obtain the average prediction accuracy for our model across these subsets for each hyper-parameter setting. Finally, we choose the setting (including the number for latent preferences) that yields the highest average prediction accuracy for use in the experiments.

5.2. Sensitivity Analysis

We first verify whether SDR is sensitive to various degrees of subjectivity in different crowdsourcing tasks. If a task is (almost entirely) objective, the optimal size of the latent preference set $\mathcal{M}$ should be 1, meaning that every worker perceives the correct answers in the same way. Consequently, the probabilities of latent preferences $\boldsymbol{\phi}_i$ for worker i collapse to $\phi_i = 1$, and the set of true answers $\mathcal{L}_j$ for question j collapses to a single true answer l_j. We conduct the held-out validation on our model across the objective datasets each with the 100 randomly sampled data subsets described in section 5.1. We expect that the average held-out prediction accuracy for SDR across these data subsets will decrease when the number of latent preferences it has increased from 1 to 2, since in this case the model starts to overfit by learning the noise in the training responses. If a task is sufficiently subjective, our model should uncover the right number of underlying groups of workers along with the right number of latent preferences. We conduct the experiment in the same way to see the difference in average prediction accuracy on held-out unseen responses with the number of preferences increasing from 1 to 3 over the partially subjective datasets. We expect the average prediction accuracy to be higher when the number of preferences is greater than 1. Since Tian and Zhu (2012) has provided us with the number of worker clusters emerging respectively from the five sub-tasks which constitute the *Image* data in Table 1, we thus compare the corresponding numbers of clusters derived from our model with theirs.

The results of the sensitivity analysis are shown in Tables 2(a) and 2(b). We can see from Table 2(a) that the average prediction accuracy of *SDR with 1 latent preference* is constantly higher than that of *SDR with 2 preferences* over all the objective datasets. This result indicates there is just one underlying group of workers for each of the tasks. This reflects that even though the difficulty-dependent corruption introduced noises to the objective truths to form the actual responses, SDR was still able to recover the right number of underlying

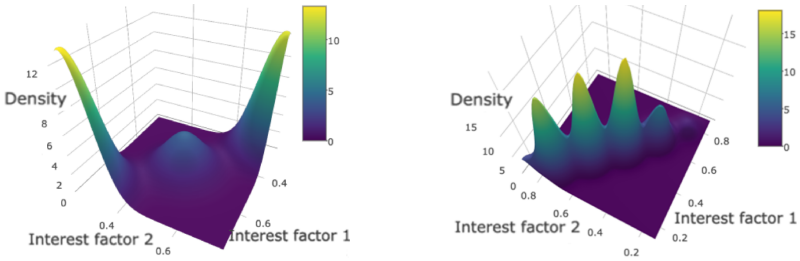
Table 2: (a) Average accuracy of our model with 1 and 2 latent preferences on predicting the held-out validation response of each worker over 4 objective tasks. (b) Average accuracy of our model with 1, 2 and 3 latent preferences on the held-out validation prediction over 10 partially subjective tasks the first 5 of which are sub-tasks of the *Image* task.

(*a*)

Dataset	The SDR model	
	m = 1	**m = 2**
Time	**0.8967**	0.8915
Dog	**0.6970**	0.6625
Duck	**0.8427**	0.8388
Product	**0.8396**	0.8291

(*b*)

Dataset	The SDR model		
	m = 1	**m = 2**	**m = 3**
Beauty 1	0.6736	**0.6944**	0.6924
Beauty 2	0.6914	**0.6998**	0.6937
Sky	0.8889	**0.8962**	0.8862
Building	0.8997	**0.9026**	0.9007
Computer	0.8098	**0.8117**	0.8074
Rel1	0.3956	**0.3985**	0.3983
Rel2	0.4426	**0.4481**	0.4481
Fashion	0.7517	**0.7589**	0.7522
Face	0.7181	**0.7203**	0.7123
Adult	0.7469	**0.7494**	0.7446



(*a*) (*b*)

Figure 3: (a) shows the 3 worker clusters on identifying sky from images and (b) shows the 4 worker clusters on judging beautiful images.

group of workers. From Table 2(b), *SDR with 2 preferences* clearly outperforms *itself with 1 preference* across all the partially subjective tasks. This means multiple groups of workers have emerged due to the sufficient subjectivity of these tasks. The table also shows that further increasing the number of latent preferences to 3 no longer improves the performance. This was most likely caused by over-fitting, and also suggests a two-dimensional latent space is accurate enough to explain the worker clustering effects emerged from these tasks. We now show the density of the workers' latent preference probabilities $\hat{\phi}_i$ estimated by SDR from the *Image* data. Due to a space limit, we only show two of them in Figure 3. According to Tian and Zhu (2012), the sub-task of judging whether images are beautiful is more subjective than the sub-task of identifying skies in images. This is re-confirmed by SDR with the inferred number of worker clusters for the former sub-task greater than that for the latter as shown in Figures 3(a) and 3(b).

5.3. Question True Answer Prediction

To verify the ability of the SDR model to predict the question true answers, we compare it with the Majority Vote (MV) and several frequently applied quality control methods including GLAD, *Multi-dimensional Wisdom of Crowds* (MdWC) (Welinder et al., 2010),

Table 3: Accuracy of all the models on predicting the true answers of the four partially subjective datasets (the results for the *Image* task are not included as the number of items in this task is too small to show any difference in the performance of different models).

Dataset	Question true answer prediction					
	SDR	**MV**	**GLAD**	**DS**	**CDS**	**MdWC**
Rel1	**0.4998**	0.4522	0.4457	0.4309	0.4697	0.4674
Rel2	**0.4752**	0.4544	0.4567	0.4512	0.4604	0.4586
Fashion	**0.8733**	0.8580	0.8689	0.8415	0.8463	0.8700
Face	**0.6423**	0.6404	0.6130	0.5924	0.5986	0.6079
Adult	**0.7598**	0.7568	0.7587	0.7534	0.7582	0.7556

Dawid&Skene (DS) (Dawid and Skene, 1979) and its variant *Community* DS (CDS) (Venanzi et al., 2014). All of these methods assume that each question has a single correct answer. The performance measure *true answer prediction accuracy* is calculated as: $\frac{1}{|\mathcal{J}|}\sum_{j\in\mathcal{J}}\mathbb{1}\{l_j=\hat{l}_j\}$, where $\hat{l}_j$ is inferred from the respective baselines. For SDR, it is: $\frac{1}{|\mathcal{J}|}\sum_{j\in\mathcal{J}}\mathbb{1}\{l_j=\hat{l}_{cj}\}$, where $c=\arg\max_{c\in\mathcal{C}} N_c$ with N_c the number of workers assigned to cluster c after Elbow K-means, and $\hat{l}_{cj}$ calculated by Eq. (12). The hyper-parameters for each baseline except MV are optimised using the held-out validation specified in section 5.1 on the exact same random held-out validation subsets of each dataset in Table 1.

The results of the question true answer prediction are listed in Table 3. Across all the partially subjective datasets except the Image data, the SDR model, based on the *largest-group* strategy for choosing the best worker clusters, is superior to the other 5 baselines. Especially, for the tasks of relevance judgement 1&2 and fashion judgement, SDR is able to outperform the best baselines by 3%, 1.5% and 0.3% with almost 54, 9 and 13 more correctly predicted question answers respectively. Since SDR is reduced to being very similar to GLAD when dealing with the objective datasets, it has achieved very similar results as GLAD did in true answer prediction over all the objective datasets except for the Duck data. In this task, SDR is superior to GLAD (0.69 versus 0.62 from GLAD) and very close to the accuracy achieved by MdWC. This suggests that SDR is at least as robust as GLAD when predicting true answers for objective tasks.

5.4. Worker Response Prediction

Predicting the responses to be given by crowd-workers to unseen questions is much more significant and common for (partially) subjective crowdsourcing tasks than it is for the objective ones. In this experiment, we evaluate the performance of all the models except MV on predicting the *next response* from each worker. We first sample one response from each worker to create a *held-out test* dataset, and then learn all the models from the rest of the data with their hyper-parameters optimised as described in section 5.1 using the exact same random validation data subsets. Finally, we evaluate the prediction performance of the models on the held-out test data using (1 - MAE). Due to the limitation of our computing power, in this experiment, we reduce the number of *held-out validation* iterations for each model to be 15 before a single iteration of *held-out test* is conducted. We performed 15 such random tests before the average performance of each model was elicited.

The results of the worker response prediction are shown in Table 4. We can see that SDR is not the best on 3 out of the 10 partially subjective datasets, topped by different baselines.

Table 4: Average accuracy of all the models on predicting the unseen held-out test response of each worker across all the partially subjective datasets.

Dataset	Unseen worker response prediction				
	SDR	GLAD	DS	CDS	MdWC
Beauty 1	**0.6974**	0.6884	0.6256	0.6927	0.6912
Beauty 2	0.7006	**0.7011**	0.6796	0.6842	0.6998
Sky	**0.9014**	0.8772	0.8801	0.8862	0.8903
Building	0.8987	0.8912	0.8956	**0.9006**	0.8976
Computer	0.8284	0.8139	0.8115	0.8196	**0.8336**
Rel1	**0.4067**	0.4035	0.3654	0.3972	0.3987
Rel2	**0.4386**	0.4312	0.4257	0.4304	0.4340
Fashion	**0.7659**	0.7593	0.6977	0.7621	0.7633
Face	**0.7224**	0.7193	0.6625	0.7081	0.7148
Adult	**0.7386**	0.7347	0.6767	0.7312	0.7354

Despite that, SDR has still performed adequately well (being second best on those datasets). We conjecture that this is because all these 3 datasets have binary answer options which intrinsically constrain the responding behaviour of workers. This results in overall weaker correlations both in the worker responses and in the underlying correct responses across the questions. For the other 7 datasets, 5 of them are with more than two answer options, thus containing stronger correct response correlations for SDR to exploit to achieve better performance. To see if the difference in the prediction accuracy between any two algorithms is significant, we conducted the Nemenyi post-hoc test (Demšar, 2006), with its parameter $\alpha = 0.1$, over performance ranks of models derived from Table 4. The result reveals that the performance difference between SDR and either CDS, GLAD or DS is statistically significant.

5.5. Subjectivity and Difficulty Estimate Consistency with Human Assessment

In this experiment, we investigate whether the estimates of the difficulty and the subjectivity of questions derived from the SDR model are *consistent* with the judgements of five human assessors. We focused on the object identification & image beauty task[4] from Tian and Zhu (2012) as the total number of its questions is 60, a manageable workload for the assessors to provide high-quality judgements. The assessors are either PhD or Master students, three of whom are avid photographers with adequate knowledge about what constitutes beautiful images, while the other two are novices who, during the group discussion, provided suggestions as to how novices might react to different images. We ask them to rank the images with respect to (i) difficulty and (ii) subjectivity. The respective instructions we gave to them are: *"rank all these images by how hard they are for crowd-workers to judge correctly by avoiding possible incorrect answers"* and *"rank them this time by how subjective they are for crowd-workers to judge"*. The assessors first independently came up with their two rankings without time limit. During the process, they could redo the two ranking tasks until they felt confident to submit. The assessors then worked together to merge their rankings into single rankings (for both difficulty and subjectivity) through group discussion and majority vote. The resulting rankings were then compared with the corresponding rankings based on the estimates from the learned SDR model. The assessors were also asked to categorise each image into one of the three levels of difficulty (i.e. *easy*, *medium*, and *hard*), and into one of

4. Crowd-workers are asked whether an image is beautiful or not.

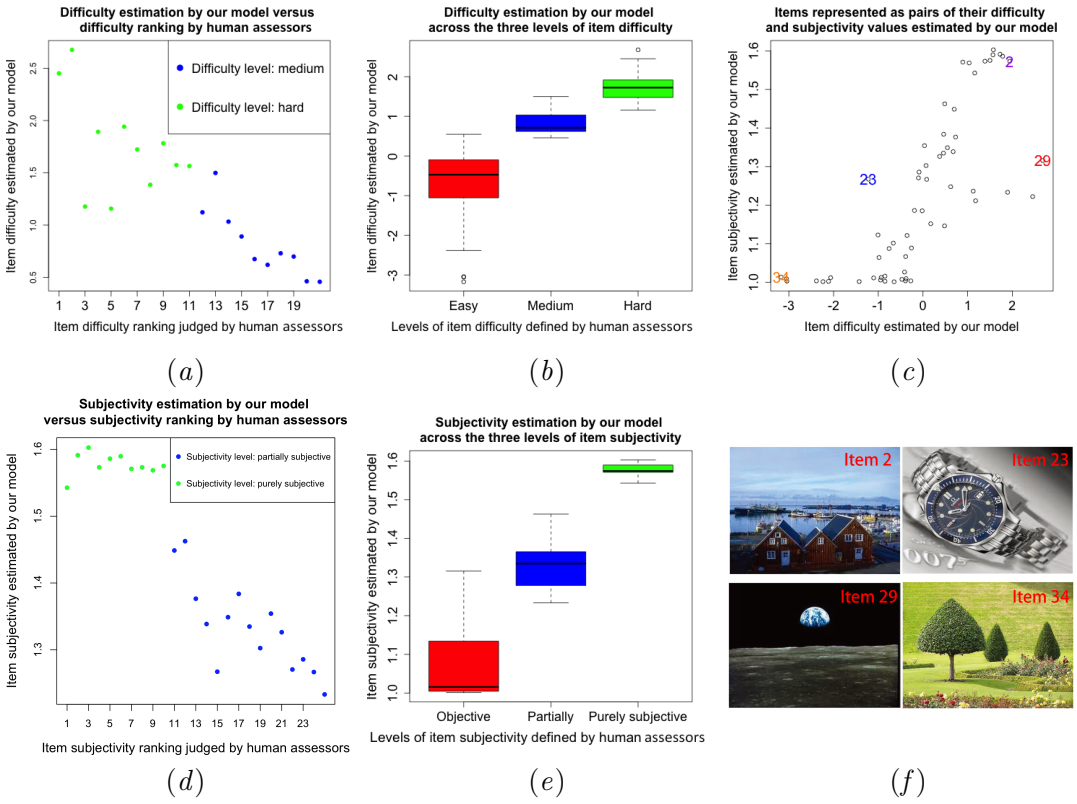


Figure 4: (a) and (d) show the correlations of both the difficulty and the subjectivity estimates with the corresponding rankings judged by human assessors, while (b) and (e) show the correlations respectively with the levels to which the images were categorized by the assessors. (c) shows the images as points with coordinates being the difficulty and the subjectivity estimates, and highlights some points, while (f) shows their corresponding images.

the three levels of subjectivity (i.e. *objective*, *partially subjective*, and *purely subjective*). We did this to see whether there existed any correlation between the difficulty or subjectivity levels to which images were categorised, and their corresponding estimates from SDR.

The results of the subjectivity and difficulty coherence evaluation have been summarised in Figure 4. Figures 4(a) and 4(d) show overall there is a strong negative correlation between the model estimates and the rankings judged by human assessors. The larger the estimate for either the difficulty or the subjectivity of an image, the higher it tends to be ranked by human assessors. Figures 4(b) and 4(e) show that there exist clear positive correlations between the levels of difficulty and subjectivity into which the images get categorised by the human assessors, and the estimated values of these two properties inferred by SDR. To support our argument about the efficacy of SDR in revealing the two key properties of images, we have selected four images highlighted in different colours in Figure 4(c) with their image ids. We can see that image 34 is inferred by our model to be both easy and objective as both of its estimates shown in Figure 4(c) are the smallest. This can be re-confirmed by visual inspection of the image in Figure 4(f). It is easy to see that there is no sky in the

image 34. Image 29 has been identified by SDR to be hard with low subjectivity according to its estimates shown in Figure 4(c). This is reasonable as the image indeed contains an extraterrestrial sky which is hard for novice workers to realise, while expert workers are able to realise and find the image objective. Images 2 and 23 both belong to the image beauty judgement task which requires workers to select 6 most beautiful images from 12 images. Our model has identified that image 2 is more subjective and harder to judge. This is probably because image 2 delivers a view of scenery which is more likely to resonate with workers while image 23 is merely showing an object. As a result, workers tend to show more different feelings and opinions towards image 2. On the other hand, image 23 does have better image quality and thus is easier for workers to make their decisions on whether it is beautiful or not.

6. Conclusion

In this paper, we have proposed the SDR (Subjectivity-and-Difficulty Response) model, a novel quality-control framework for crowdsourcing that is able to distinguish question subjectivity, which causes worker-specific truth for individual questions, from question difficulty, which determines the probability that a worker's response to each question equals her perceived subjective truth. Experiment results show that our model improves both the correct answer prediction for questions and the held-out unseen response prediction for crowd-workers compared to five baselines across numerous partially subjective crowdsourcing datasets. Moreover, our model shows robustness to both the objective and the partially subjective datasets by discovering the right numbers of underlying worker groups for them. Finally, our model is able to provide estimates of the difficulty and the subjectivity of questions that are consistent with the judgements from human assessors.

References

Adult datset. https://github.com/ipeirotis/Get-Another-Label/tree/master/data. Accessed: 2017-07-30.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022, 2003.

Chris Buckley, Matthew Lease, and Mark D Smucker. Overview of the TREC 2010 relevance feedback track (notebook). In *the 19th Text Retrieval Conference Notebook.*, 2010.

Alexander Philip Dawid and Allan M Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Applied Statistics*, pages 20–28, 1979.

Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8), 2009.

Matthew Lease and Gabriella Kazai. Overview of the TREC 2011 crowdsourcing track. In *Proceedings of the Text Retrieval Conference*, 2011.

Babak Loni, Maria Menendez, Mihai Georgescu, Luca Galli, Claudio Massari, Ismail Sengor Altingovde, Davide Martinenghi, Mark Melenhorst, Raynor Vliegendhart, and Martha Larson. Fashion-focused creative commons social dataset. In *Proceedings of the 4th ACM Multimedia Systems Conference*, pages 72–77, 2013.

Barzan Mozafari, Purna Sarkar, Michael Franklin, Michael Jordan, and Samuel Madden. Scaling up crowd-sourcing to very large datasets: a case for active learning. *Proceedings of the VLDB Endowment*, 8(2):125–136, 2014.

An Thanh Nguyen, Matthew Halpern, Byron C Wallace, and Matthew Lease. Probabilistic modeling for crowdsourcing partially-subjective ratings. In *Proceedings of the 4th AAAI Conference on Human Computation and Crowdsourcing*, pages 149–158, 2016.

Georg Rasch. Studies in mathematical psychology: Probabilistic models for some intelligence and attainment tests. 1960.

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. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2008.

Tian Tian and Jun Zhu. Uncovering the latent structures of crowd labeling. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 392–404. Springer, 2015.

Yuandong Tian and Jun Zhu. Learning from crowds in the presence of schools of thought. In *Proceedings of 14th SIGKDD Conference on Knowledge Discovery and Data Mining*, 2012.

Matteo Venanzi, John Guiver, Gabriella Kazai, Pushmeet Kohli, and Milad Shokouhi. Community-based bayesian aggregation models for crowdsourcing. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 155–164. ACM, 2014.

Ellen M Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. In *Proceedings of the 21st SIGIR Conference on Research and Development in Information Retrieval*, pages 315–323. ACM, 1998.

Jiannan Wang, Tim Kraska, Michael J Franklin, and Jianhua Feng. Crowder: Crowdsourcing entity resolution. *Proceedings of the VLDB Endowment*, 5(11):1483–1494, 2012.

Wei Wang, Xiang-Yu Guo, Shao-Yuan Li, Yuan Jiang, and Zhi-Hua Zhou. Obtaining high-quality label by distinguishing between easy and hard items in crowdsourcing. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 2017.

Peter Welinder, Steve Branson, Pietro Perona, and Serge J Belongie. The multidimensional wisdom of crowds. In *Advances in Neural Information Processing Systems*, 2010.

Jacob Whitehill, Ting-fan Wu, Jacob Bergsma, Javier R Movellan, and Paul L Ruvolo. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in Neural Information Processing Systems*, pages 2035–2043, 2009.

Denny Zhou, Sumit Basu, Yi Mao, and John C. Platt. Learning from the wisdom of crowds by minimax entropy. In *Advances in Neural Information Processing Systems*, 2012.