Estimating Consensus from Crowdsourced Continuous Annotations

Santosh Chapaneri
Dept. Electronics and Telecommunication Engg.
St. Francis Institute of Technology, University of Mumbai
Mumbai, India
santoshchapaneri@sfit.ac.in

Deepak Jayaswal
Dept. Electronics and Telecommunication Engg.
St. Francis Institute of Technology, University of Mumbai
Mumbai, India
djjayaswal@sfit.ac.in

***Abstract*—With the emergence of crowdsourcing services, the concept of the wisdom of crowds has gained immense popularity. To capture the subjectivity phenomena, multiple annotators are asked to give their responses using crowdsourcing tools. Unfortunately, inattentive and adversarial annotators pose a threat to the quality and trustworthiness of the consensus. In this work, we focus on crowd consensus estimation of continuous labels using a probabilistic approach. A lot of existing work is reported for annotator behavior modeling for the categorical case; however, there is limited work in the continuous case. We propose a maximum-likelihood solution to determine the estimated consensus while simultaneously modeling the behavior of various annotators. Further, to handle the long-tail phenomena commonly observed in crowdsourced datasets, a confidence-interval based estimated consensus is derived. The proposed technique is shown to perform better than using the average annotation values and existing work.**

***Keywords*—Crowdsourcing; Annotations; Consensus**

I. Introduction

Crowdsourcing is the concept of using the wisdom of the crowd for a specific task. Amazon Mechanical Turk (AMT) [1] and CrowdFlower [2] are examples of crowdsourcing platforms where the annotators submit multiple small tasks such as providing a response to specific target variables for a fee. Labeled datasets are considered to be valuable since they are expensive to produce. Supervised machine learning methods require labeled datasets for training the models. But providing the ground truth labels for large datasets is excessively time consuming and laborious. The labeling task is thus outsourced, especially for large datasets. Crowdsourcing is cost-effective as well as a fast method to solve this problem without the need of domain expertise. The process of collecting annotations from various annotators and using these annotations for estimating the consensus is known as crowd-labeling. Each person annotates a random subset of the dataset and every sample of the dataset is annotated by a subset of all annotators. This results in sparsely annotated data leading to the commonly observed long-tail phenomenon. The goal of crowd-labeling is to obtain the estimated consensus values from the sparsely annotated data.

Various benchmark datasets have been developed using crowdsourcing to handle the subjectivity issue, where multiple annotators are asked to provide their opinions for a random subset of data samples. The annotator responses can be inconsistent due to various personal and situational aspects such as personality, context, cultural background, etc. A typical approach is to estimate the ground truth of each data sample by averaging multiple annotations given to it. This approach implies that all annotators are equally reliable which may not be a valid assumption in practice since it often ignores the annotator errors (e.g. low-attention) and outliers (e.g. adversarial behavior) that can have a significant impact on the consensus. Thus, it is important to model the behavior and obtain the reliability of each annotator and consider this factor to determine the estimated consensus. Several studies can be found in the literature related to annotator behavior modeling and estimating the consensus for the categorical task, i.e. where the target variable is categorical. There is very limited existing work for estimating the consensus from continuous target responses and for modeling the annotator behaviors in this scenario.

Contributions: Extending the work of learning from crowdsourced data for a single-dimension target regression case [3], a maximum-likelihood solution is derived in this work to determine the multi-dimensional estimated consensus ($\mathbf{y} \in \mathbb{R}^D$) and simulataneously modeling the behavior of each annotator. Several annotator behaviors (spammer, adversarial, biased, competent, etc.) have been modeled in the literature for classification tasks (e.g. [4]), however, for the regression task, the annotator behavior modeling remains a problem to be solved. These behaviors can occur due to varying expertise, bad intent or low-attention of annotators. Inferring such behavior can be helpful to determine the annotation consensus. In this work, the behavior of adversarial, biased and reliable annotators is modeled for the multi-dimensional target task.

II. Related Work

A less reliable and an often overlooked assumption in most crowdsourcing techniques is that the ground truth can be obtained by taking the average of multiple annotator responses. To solve this problem, truth discovery analysis is required for finding the consensus among various annotators [3]–[8].

Raykar *et al* [3], [4] proposed methods to learn the annotator behaviors for the crowdsourced classification task. A survey on truth discovery analysis methods is presented in [5] focusing on crowdsourced opinions of multi-source data from categorical as well as continuous domains while considering the problem difficulty, spammer identification, constrained judgement, etc. In [6], an uncertainty-aware modeling approach is proposed to estimate the kernel density from multiple sources and learn trustworthy opinions. In [7], the joint distribution of annotators is considered to learn the estimated consensus.

978-1-7281-2339-4/20/$31.00 ©2020 IEEE

Authorized licensed use limited to: University of Exeter. Downloaded on July 16,2020 at 13:10:47 UTC from IEEE Xplore. Restrictions apply.

There is limited existing work focusing on estimating consensus from continuous annotations. A non-parametric Gaussian process model is proposed in [8] to jointly learn the regression function as well as the behavior of annotators. An iterative probabilistic approach is used in [3] to determine the gold standard while also measuring the competence of the annotators. In [9], the focus is on ordinal labels where the continuous latent variables are discretized and treated as categorical variables. The crowd-labeling problem is treated as a bipartite graph in [10] with a belief propagation-based Bayesian iterative algorithm. However, the problem of annotator behavior modeling is largely an unsolved problem in the continuous annotation task.

III. Proposed Work

To measure the inter-annotator agreement, the consistency score J is evaluated given by Eq. (1) for N samples of the specific dataset.

$$J = \frac{1}{N}\sum_{i=1}^{N}\sqrt{\frac{\sum_{j\in\mathcal{A}_i}(y_i^j - y_{i_m})^2}{|\mathcal{A}_i|}} \tag{1}$$

For each i^{th} sample, its median annotation y_{i_m} is computed over all its annotated responses y_i^j and $\mathcal{A}_i$ is the set of workers that annotated the i^{th} sample. Intuitively, J measures the average deviation relative to the median, thus a lower value of J is better. For the crowdsourced datasets (discussed in Sec. IV), $J = 0.6807$ for the HeadPose dataset [11] and $J = 0.4639$ for the Age dataset [12]; these values indicate that not all annotators agree on the responses of the annotated samples. A simple approach to determine the consensus is to compute the average value, however, this assumes that all annotators are equally reliable which may not hold in practice.

Consider the dataset $\mathcal{D} = \{\mathbf{y}_i^1, \mathbf{y}_i^2, \ldots, \mathbf{y}_i^R\}_{i=1}^N$ consisting of multi-dimensional responses of N data samples by maximum R annotators, where each worker annotates only a subset of N samples through a crowdsourcing platform such as Amazon Mechanical Turk (AMT). To model the behavior of the j^{th} annotator, three parameters are considered:

a) **adversariness** a^j: if the annotator is adversarial, $a^j = -1$, else $a^j = 1$, thus $a^j \in \{-1, 1\}$ and $\left(a^j\right)^2 = 1$;
b) **bias** b^j: a Normal prior $\mathcal{N}(b^j|\mu_b, s_b)$ is used, where $\mu_b = 0$ to favor unbiased annotators and $s_b = 0.05$ to allow for some positive and negative bias;
c) **variability** α^j: this measures the variance of the j^{th} annotator, thus lower is better.

With $\mathbf{y}_i \in \mathbb{R}^D$ as the unknown D-dimensional ground truth for the i^{th} sample, a Gaussian distribution model $\mathcal{N}(\mathbf{y}_i^j|\mathbf{y}_i, \alpha^j, a^j, b^j)$ is assumed where $\mathbf{y}_i^j$ is the response of the j^{th} annotator. Assuming all samples are annotated independently by R annotators, the model parameters to be estimated are $\boldsymbol{\theta} = \{\mathbf{y}, \boldsymbol{\alpha}, \mathbf{a}, \mathbf{b}\}$ with the likelihood given by Eq. (2). The log-likelihood of Eq. (2) is given by Eq. (3) where the constant $C = \frac{-(N+1)R}{2}\ln(2\pi) - \frac{R}{2}\ln(s_b)$. Equating its gradients with respect to each parameter of $\boldsymbol{\theta}$ to 0, the updates are obtained given by Eq. (4) – (7) using the fact that $\|\mathbf{z}\|_2^2 = \mathbf{z}^\intercal\mathbf{z}$ and $\frac{d}{d\mathbf{z}}\mathbf{z}^\intercal\mathbf{z} = 2\mathbf{z}^\intercal$.

$$\begin{aligned} P(\mathcal{D}|\boldsymbol{\theta}) &= \prod_{i=1}^{N}\prod_{j=1}^{R}\mathcal{N}(\mathbf{y}_i^j|\mathbf{y}_i, \alpha^j, a^j, b^j) \times \prod_{j=1}^{R}\mathcal{N}(b^j|\mu_b, s_b) \\ &= \prod_{i=1}^{N}\prod_{j=1}^{R}\frac{1}{\sqrt{2\pi\alpha^j}}\exp\left[\frac{-1}{2\alpha^j}\|\mathbf{y}_i^j - a^j(\mathbf{y}_i + b^j\mathbf{1})\|_2^2\right] \\ &\quad \times \prod_{j=1}^{R}\frac{1}{\sqrt{2\pi s_b}}\exp\left[\frac{-1}{2s_b}\left(b^j - \mu_b\right)^2\right] \end{aligned} \tag{2}$$

$$\begin{aligned} \ln P(\mathcal{D}|\boldsymbol{\theta}) = C &- \frac{N}{2}\sum_{j=1}^{R}\ln(\alpha^j) \\ &- \sum_{i=1}^{N}\sum_{j=1}^{R}\frac{1}{2\alpha^j}\|\mathbf{y}_i^j - a^j(\mathbf{y}_i + b^j\mathbf{1})\|_2^2 \\ &- \sum_{j=1}^{R}\frac{1}{2s_b}\left(b^j - \mu_b\right)^2 \end{aligned} \tag{3}$$

Update solution for $\mathbf{y}_i$:

$$\frac{\partial \ln P(\mathcal{D}|\boldsymbol{\theta})}{\partial \mathbf{y}_i} = \mathbf{0} \implies$$

$$\sum_{j=1}^{R}\frac{1}{\hat{\alpha}^j}\left(\mathbf{y}_i^j - \hat{a}^j(\hat{\mathbf{y}}_i + \hat{b}^j\mathbf{1})\right)^\intercal \hat{a}^j = \mathbf{0}$$

Applying transpose on both sides,

$$\sum_{j=1}^{R}\frac{\hat{a}^j}{\hat{\alpha}^j}\mathbf{y}_i^j = \sum_{j=1}^{R}\frac{1}{\hat{\alpha}^j}\left(\hat{\mathbf{y}}_i + \hat{b}^j\mathbf{1}\right), \text{ since } \left(\hat{a}^j\right)^2 = 1,$$

$$\hat{\mathbf{y}}_i\sum_{j=1}^{R}\frac{1}{\hat{\alpha}^j} = \sum_{j=1}^{R}\frac{\left(\hat{a}^j\mathbf{y}_i^j - \hat{b}^j\mathbf{1}\right)}{\hat{\alpha}^j}$$

$$\therefore \boxed{\hat{\mathbf{y}}_i = \frac{1}{\sum_{j=1}^{R}\frac{1}{\hat{\alpha}^j}}\sum_{j=1}^{R}\frac{\left(\hat{a}^j\mathbf{y}_i^j - \hat{b}^j\mathbf{1}\right)}{\hat{\alpha}^j}} \tag{4}$$

Update solution for a^j:

$$\frac{\partial \ln P(\mathcal{D}|\boldsymbol{\theta})}{\partial a^j} = 0 \implies$$

$$-\frac{1}{\hat{\alpha}^j}\sum_{i=1}^{N}\left(\mathbf{y}_i^j - a^j(\mathbf{y}_i + b^j\mathbf{1})\right)^\intercal \times \left(\mathbf{y}_i + b^j\mathbf{1}\right) \times (-1) = 0,$$

$$\sum_{i=1}^{N}\mathbf{y}_i^{j\intercal}(\hat{\mathbf{y}}_i + \hat{b}^j\mathbf{1}) = \hat{a}^j\sum_{i=1}^{N}\|\hat{\mathbf{y}}_i + \hat{b}^j\mathbf{1}\|_2^2$$

Since $\|\hat{\mathbf{y}}_i + \hat{b}^j\mathbf{1}\|_2^2$ is always positive, we have

$$\therefore \boxed{\hat{a}^j = \operatorname{sgn}\left(\sum_{i=1}^{N}\mathbf{y}_i^{j\intercal}(\hat{\mathbf{y}}_i + \hat{b}^j\mathbf{1})\right)} \tag{5}$$

Authorized licensed use limited to: University of Exeter. Downloaded on July 16,2020 at 13:10:47 UTC from IEEE Xplore. Restrictions apply.

Update solution for b^j:

$$\frac{\partial \ln P(\mathcal{D}|\boldsymbol{\theta})}{\partial b^j} = 0 \implies$$

$$\sum_{i=1}^{N} \frac{1}{\hat{\alpha}^j}\left(\mathbf{y}_i^j - \hat{a}^j(\hat{\mathbf{y}}_i + \hat{b}^j\mathbf{1})\right)^\mathsf{T} \times (-\hat{a}^j\mathbf{1}) + \frac{1}{s_b}(\hat{b}^j - \mu_b) = 0$$

$$\therefore \sum_{i=1}^{N} \frac{\hat{a}^j}{\hat{\alpha}^j}\left(\mathbf{y}_i^j - \hat{a}^j(\hat{\mathbf{y}}_i + \hat{b}^j\mathbf{1})\right)^\mathsf{T}\mathbf{1} = \frac{1}{s_b}(\hat{b}^j - \mu_b)$$

Since $\left(\hat{a}^j\right)^2 = 1$,

$$\sum_{i=1}^{N}\left(\hat{a}^j\mathbf{y}_i^j - \hat{\mathbf{y}}_i\right)^\mathsf{T}\mathbf{1} - N\hat{b}^j = \frac{\hat{\alpha}^j}{s_b}(\hat{b}^j - \mu_b)$$

Rearranging,

$$\boxed{\hat{b}^j = \frac{1}{N + \frac{\hat{\alpha}^j}{s_b}}\left(\sum_{i=1}^{N}\left(\hat{a}^j\mathbf{y}_i^j - \hat{\mathbf{y}}_i\right)^\mathsf{T}\mathbf{1} + \frac{\hat{\alpha}^j\mu_b}{s_b}\right)} \qquad (6)$$

Update solution for α^j:

$$\frac{\partial \ln P(\mathcal{D}|\boldsymbol{\theta})}{\partial \alpha^j} = 0 \implies$$

$$\frac{-N}{2\hat{\alpha}^j} + \frac{1}{2(\hat{\alpha}^j)^2}\sum_{i=1}^{N}\|\mathbf{y}_i^j - \hat{a}^j(\hat{\mathbf{y}}_i + \hat{b}^j\mathbf{1})\|_2^2 = 0$$

$$\therefore \boxed{\hat{\alpha}^j = \frac{1}{N}\sum_{i=1}^{N}\|\mathbf{y}_i^j - \hat{a}^j(\hat{\mathbf{y}}_i + \hat{b}^j\mathbf{1})\|_2^2} \qquad (7)$$

Since not all instances will be annotated by each annotator in general, denote $\mathcal{A}_i$ as the set of annotators providing the response for the i^{th} sample and $\mathcal{R}_j$ as the set of songs for which the j^{th} annotator provided the response. The resulting solution given by Eq. (8) is equivalent to the EM (Expectation-Maximization) algorithm [13] where the E-step determines the estimated consensus $\hat{\mathbf{y}}_i$ and the M-step determines the annotator parameters (adversariness $\hat{a}^j$, bias $\hat{b}^j$ and variability $\hat{\alpha}^j$). These two steps are iterated till convergence (e.g. delta change of $\|\hat{\mathbf{y}}\| < 10^{-6}$). The EM algorithm is initialized with $\hat{b}^j = 0$, $\hat{\mathbf{y}}_i$ as the median of $\{\mathbf{y}_i^j\}_{j\in\mathcal{A}_i}$, $\hat{\alpha}^j$ as the variance of $\{\mathbf{y}_i^j\}_{j\in\mathcal{A}_i}$ and $\hat{a}^j$ as given in Eq. (8).

$$\begin{aligned}
\hat{\mathbf{y}}_i &= \frac{1}{\sum_{j\in\mathcal{A}_i}\frac{1}{\hat{\alpha}^j}}\sum_{j\in\mathcal{A}_i}\frac{1}{\hat{\alpha}^j}(\hat{a}^j\mathbf{y}_i^j - \hat{b}^j\mathbf{1}), \\
\hat{a}^j &= \operatorname{sgn}\left(\sum_{i\in\mathcal{R}_j}\mathbf{y}_i^{j\,\mathsf{T}}(\hat{\mathbf{y}}_i + \hat{b}^j\mathbf{1})\right), \\
\hat{b}^j &= \frac{1}{|\mathcal{R}_j| + \frac{\hat{\alpha}^j}{s_b}}\left(\sum_{i\in\mathcal{R}_j}\left(\hat{a}^j\mathbf{y}_i^j - \hat{\mathbf{y}}_i\right)^\mathsf{T}\mathbf{1} + \frac{\hat{\alpha}^j\mu_b}{s_b}\right), \\
\hat{\alpha}^j &= \frac{1}{|\mathcal{R}_j|}\sum_{i\in\mathcal{R}_j}\|\mathbf{y}_i^j - \hat{a}^j(\hat{\mathbf{y}}_i + \hat{b}^j\mathbf{1})\|_2^2
\end{aligned} \qquad (8)$$

Typically, the crowdsourced datasets have a long-tail problem, i.e. many annotators provide a response to few samples and few annotators provide a response to many samples. This phenomenon is illustrated in Fig. 1 for the HeadPose and Age datasets. In such a scenario, the solution of Eq. (8) will be over-optimistic. To overcome this problem, the $(1-\beta)$ confidence interval (CI) of annotator variability is considered where β is the significance value (e.g. 5%). Since the sum of squares of Gaussian random variables follows a χ^2 distribution, the $(1-\beta)$ confidence interval is obtained given by Eq. (9). If $|\mathcal{R}_j|$ is large, then the χ^2 value will be very close to $|\mathcal{R}_j|$, i.e. the algorithm will automatically adjust the weights for annotators with the different number of responses. From Table I, we observe that annotators with IDs 30172026 and 7837812 of the HeadPose dataset as well as IDs 4711962 and 22201476 of the Age dataset obtained similar variance $\hat{\alpha}^j$ with Eq. (8), however, the upper bound (UB) of confidence interval provides a realistic solution.

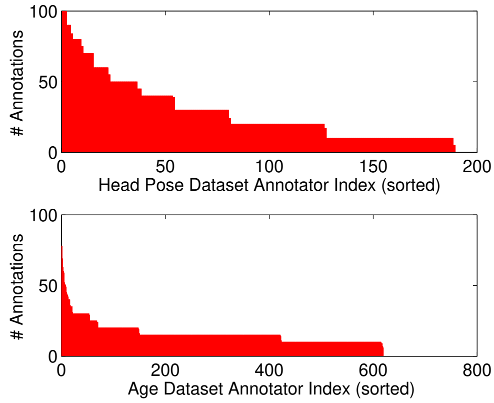


Fig. 1: Long tail phenomena in the Head Pose (above) and Age (below) datasets

$$\frac{1}{\alpha^j}\sum_{i\in\mathcal{R}_j}\|\mathbf{y}_i^j - \hat{a}^j(\hat{\mathbf{y}}_i + \hat{b}^j\mathbf{1})\|_2^2 = \frac{|\mathcal{R}_j|\hat{\alpha}^j}{\alpha^j} \sim \chi^2(|\mathcal{R}_j|)$$

$$P\left[\chi^2_{(1-\frac{\beta}{2},|\mathcal{R}_j|)} < \frac{|\mathcal{R}_j|\hat{\alpha}^j}{\alpha^j} < \chi^2_{(\frac{\beta}{2},|\mathcal{R}_j|)}\right] = 1-\beta$$

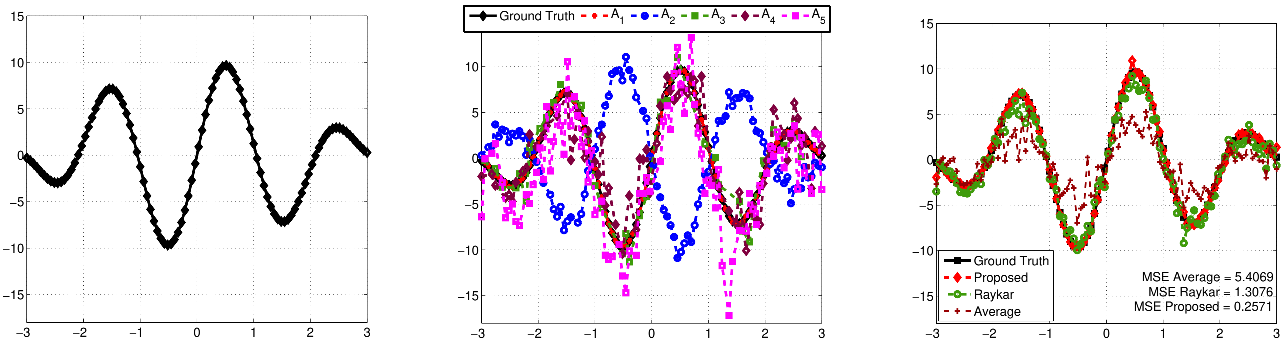
$$\therefore CI_{1-\beta} = \left\{\frac{|\mathcal{R}_j|\hat{\alpha}^j}{\chi^2_{(1-\frac{\beta}{2},|\mathcal{R}_j|)}}, \frac{|\mathcal{R}_j|\hat{\alpha}^j}{\chi^2_{(\frac{\beta}{2},|\mathcal{R}_j|)}}\right\} \qquad (9)$$

TABLE I: Confidence intervals of estimated annotator variabilities (top 4 rows refer to the HeadPose dataset and the bottom 4 rows refer to the Age dataset).

AnnotatorID	#Annotations	Variability ($\hat{\alpha}$)	95% CI
22540655	100	0.2491	(0.2178, 0.3217)
30507455	75	0.3657	(0.2152, 0.9174)
30172026	**30**	**1.4651**	**(0.7932, 3.2845)**
7837812	**5**	**1.6796**	**(0.2583, 6.8521)**
17525614	78	0.3615	(0.2718, 0.3217)
20730328	50	0.3529	(0.2436, 0.4173)
4711962	**20**	**0.5342**	**(0.2076, 2.6819)**
22201476	**6**	**0.5618**	**(0.1738, 4.6931)**

Outlier removal techniques such as Minimum Covariance Determinant (MCD) [14] etc. can be used to remove the outliers that may exist in the annotated data. However, we use the computationally inexpensive weighted median that is less sensitive to outliers compared to the weighted mean for the

Authorized licensed use limited to: University of Exeter. Downloaded on July 16,2020 at 13:10:47 UTC from IEEE Xplore. Restrictions apply.



(a) Ground truth data

(b) Data from annotators

(c) Estimated Consensus

Fig. 2: Illustration of the proposed technique for obtaining estimated consensus

solution of $\hat{\mathbf{y}}_i$. The resulting equations for estimated consensus are given by Eq. (10), which are iterated till convergence.

$$\hat{\mathbf{y}}_i = \texttt{wMedian}\left(\hat{a}^j\mathbf{y}_i^j - \hat{b}^j\mathbf{1}, \frac{1}{\hat{\alpha}^j}\right),$$
$$\hat{\alpha}^j = \frac{1}{\chi^2_{(\frac{\beta}{2},|\mathcal{R}_j|)}} \sum_{i\in\mathcal{R}_j} \|\mathbf{y}_i^j - \hat{a}^j(\hat{\mathbf{y}}_i + \hat{b}^j\mathbf{1})\|_2^2 \qquad (10)$$

Note that the updates for $\hat{a}^j$ and $\hat{b}^j$ remain unchanged as in Eq. (8).

IV. Experimental Results

The performance metric for evaluation of the proposed technique is mean square error (MSE) given by Eq. (11), where $\hat{\mathbf{y}}_i$ is the multi-dimensional estimated consensus and $\mathbf{y}_i$ is the corresponding ground truth.

$$MSE = \frac{1}{N}\sum_{i=1}^{N} \|\hat{\mathbf{y}}_i - \mathbf{y}_i\|_2^2 \qquad (11)$$

The proposed technique for estimating the consensus from crowdsourced annotations is validated on one simulated, two synthetic as well as three benchmark datasets as follows:

(i) **Simulated**: For illustrative purposes, 5 annotators are simulated with varying behavior and synthetic data of 100 samples from $y_i^j = f(x_i) + \epsilon^j$ is used. The ground truth data is $f(x) = 10\sin(3x)\cos(\frac{1}{2}x)$ and $\epsilon^j \sim \mathcal{N}(0, \alpha^j)$. The variability levels of annotators are $\boldsymbol{\alpha} = \{0.1, 0.8, 1.5, 2.2, 3\}$ (lower is better), the number of samples annotated per worker is $|\mathcal{R}| = \{90, 95, 40, 70, 85\}$, the 2^{nd} annotator is assumed to be adversarial and the 5^{th} annotator is biased. The ground truth data is shown in Fig. 2(a), the annotations are shown in Fig. 2(b) and the estimated consensus is shown in Fig. 2(c), which is also compared with the average estimate as well as the method proposed by Raykar *et al* [3]. The proposed technique is effective in determining the estimated consensus as very close to the ground truth data with smaller mean square error (MSE = 5.4069 with average calculation, 1.3076 of [3] and 0.2571 of proposed algorithm), since it can identify the adversarial, biased as well as reliable annotators.

(ii) **Synthetic**: For the synthetic dataset, 200 data samples and 300 annotators having variability levels α^j, $j = 1, \ldots, 300$ are generated. $300 \times p$ annotators are considered as unreliable (out of which a fraction q are assumed to be adversarial and a fraction r are biased with $s_b \sim \mathcal{U}(0.01, 0.05)$), and the remaining $300 \times (1-p)$ annotators as reliable. For the j^{th} reliable annotator, we generate $\alpha^j \sim \mathcal{U}(0.01, 0.05)$ and for the unreliable annotator, α^j is generated from $\mathcal{U}(1, 5)$. For every i^{th} data sample, the number of annotators providing responses $|\mathcal{A}_i|$ is generated from a Poisson distribution $\mathcal{P}(\lambda)$, and $|\mathcal{A}_i|$ annotators are randomly selected to provide the responses. The ground truth is assumed to be $y_i = 1\ \forall\ i$ and the response y_i^j is generated from a Gaussian distribution $\mathcal{N}(y_i, \alpha^j)$. This implies that the unreliable annotators have significant variability and their responses are likely to be extreme values. The parameter p denotes the fraction of unreliable annotators, q denotes the fraction of adversarial annotators, r denotes the fraction of biased annotators, and λ denotes the average number of responses for each data sample. The proposed algorithm is tested with $p = 0.25, q = 0.2, r = 0.2, \lambda = 10$. To reduce random errors, 50 datasets are generated and the average MSE is reported.

(iii) **Housing**: Next, the benchmark Housing dataset from the UCI machine learning repository consisting of 506 data samples is considered with 'MEDV' as the ground truth target value y_i. In this dataset, 16 samples have the target value of 50.0 indicating either missing or censored values, hence these samples are discarded resulting in 490 samples. The annotators are simulated similar to the setup described in (ii) with 300 annotators.

(iv) **Population**: The benchmark fact-finding Population dataset [15] obtained via crowdsourcing is used to validate the proposed technique. This dataset reflects the Wikipedia edit history regarding the city population for specific years. It consists of 1124 data samples (city names), 2344 annotators and 4008 responses obtained from the crowdsourced data. Pre-processing is done to retain only the latest claim (based on time-stamp) made by annotators for a specific city, and unreasonable claims such as 0 and 6.5979×10^{18} are discarded [6].

(v) **HeadPose**: The benchmark head pose dataset introduced in [11] consists of the head pose images of 15 people with

Authorized licensed use limited to: University of Exeter. Downloaded on July 16,2020 at 13:10:47 UTC from IEEE Xplore. Restrictions apply.

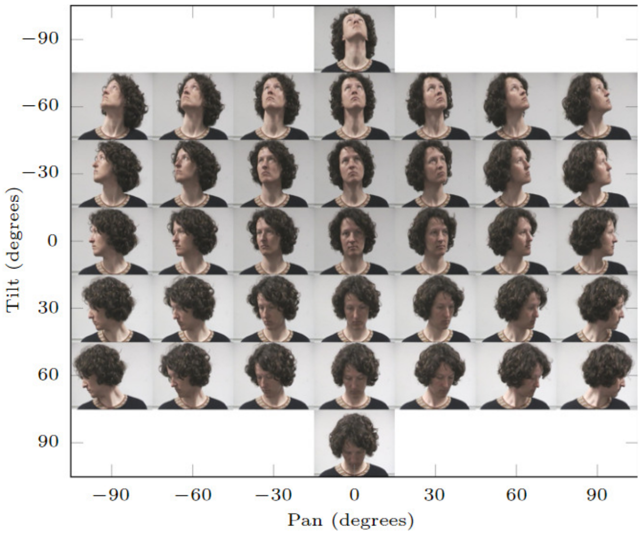


Fig. 3: Example of the head orientations in the HeadPose dataset

different tilt and pan orientations obtained from the Head Pose Image database [16]. An illustration of the various head (tilt and pan) orientations for a specific person is shown in Fig. 3. The annotations were collected using the CrowdFlower platform resulting in $5,399$ responses for 555 data samples annotated by 189 annotators. Each annotator was asked to specify their response for tilt and pan orientations by observing each head pose image. This dataset has two-dimensional responses for tilt and pan orientations both in the range $(-90, 90)$.

(vi) **Age**: The benchmark crowdsourced Age annotations dataset [12] consists of $10,020$ annotations of $1,002$ annotators for 619 data samples. The annotations were obtained using the CrowdFlower platform and the images were obtained from the FGNet Aging database [17]. An illustration of sample images of this dataset is shown in Fig. 4. Each data sample is an image of a person with known age in the range $(0, 69)$. Each annotator was shown several images and asked to rate the age of the person in each image.

Fig. 4: Example of images in the Age dataset

The annotator behaviors observed in the Age dataset is shown in Fig. 5. The competent annotator gives almost correct answers typically whereas the adversarial annotator intentionally gives incorrect answers. The biased annotators are biased either positively or negatively resulting in incorrect responses. Thus, it is crucial to identify the adversariness, biasedness, and variability of each annotator while determining the estimated consensus.

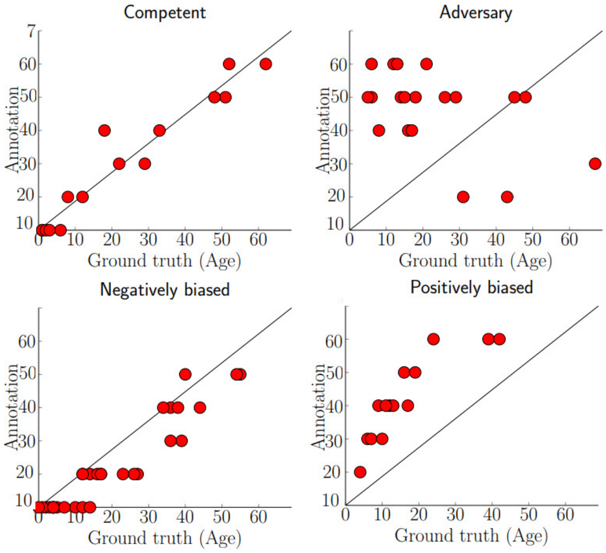


Fig. 5: Example of various annotator behaviors in the Age dataset

Table II shows the MSE results on the simulated, synthetic and benchmark datasets for the average consensus, consensus obtained with Raykar *et al* [3] and the proposed technique. For the synthetic and Housing datasets, the average MSE of 50 simulations is reported. Since the scale of target values for the Population, HeadPose and Age datasets is different, the range of MSE values varies accordingly. In all cases, the proposed technique achieves lower MSE due to its ability to identify adversarial as well as biased annotators.

TABLE II: Evaluation results (MSE)

	Average	Raykar *et al* [3]	Proposed
Simulated	5.4069	1.3076	$\mathbf{0.2571}$
Synthetic	0.5631	0.3872	$\mathbf{0.1436}$
Housing	0.6548	0.4317	$\mathbf{0.2391}$
Population	$126,198$	$8,513$	$\mathbf{7,154}$
HeadPose	0.7082	0.4924	$\mathbf{0.2342}$
Age	21.6679	15.4278	$\mathbf{13.1816}$

V. Conclusion

A maximum-likelihood solution equivalent to the EM algorithm is proposed in this work to model the varying behavior of annotators and the confidence-interval based estimated consensus is derived for the continuous target task. The proposed work is especially useful in the case where the ground truth is not available and only the crowdsourced continuous annotations are available. The proposed technique can identify the adversariness, biasedness, and variability of each annotator through behavior modeling and simultaneously learn the unknown ground truth. For further work, the estimated consensus and annotator behavior modeling can be computed by also considering the features of the data samples besides the crowdsourced annotations.

Authorized licensed use limited to: University of Exeter. Downloaded on July 16,2020 at 13:10:47 UTC from IEEE Xplore. Restrictions apply.

References

[1] Amazon.com, Inc., *Amazon Mechanical Turk*, 2018, www.mturk.com

[2] Figure Eight, Inc., *CrowdFlower*, 2018, www.crowdflower.com

[3] V. Raykar, S. Yu, L. Zhao, G. Valadez, C. Florin, L. Bogoni and L. Moy, ”Learning from crowds”, *Journal of Machine Learning Research*, vol. 11, pp. 1297–1322, Apr 2010.

[4] V. Raykar and S. Yu, ”Eliminating spammers and ranking annotators for crowdsourced labeling tasks”, *Journal of Machine Learning Research*, vol. 13, pp. 491–518, Feb 2012.

[5] Y. Li, J. Gao, C. Meng, Q. Li, L. Su, B. Zhao, W. Fan and J. Han, ”A survey on truth discovery”, *ACM SIGKDD Explorations Newsletter*, vol. 17, no. 2, pp. 1–16, Feb 2016.

[6] M. Wan, X. Chen, L. Kaplan, J. Han, J. Gao and B. Zhao, ”From truth discovery to trustworthy opinion discovery: An uncertainty-aware quantitative modeling approach”, *Proc. of the ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*, pp. 1885–1894, Aug 2016.

[7] A. Ramakrishna, R. Gupta, R. Grossman and S. Narayanan, ”An Expectation Maximization approach to joint modeling of multidimensional ratings derived from multiple annotators”, *InterSpeech*, pp. 1555–1559, Sep 2016.

[8] H. Xiao, H. Xiao and C. Eckert, ”Learning from multiple observers with unknown expertise”, *Springer Proc. of the Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD)*, vol. 7818, pp. 595–606, Apr 2013.

[9] B. Lakshminarayanan and Y. Teh, ”Inferring ground truth from multi-annotator ordinal data: A probabilistic approach”, *arXiv preprint*, arXiv:1305.0015, pp. 1–19, Apr 2013.

[10] J. Ok, S. Oh, J. Shin, Y. Jang and Y. Yi, ”Iterative Bayesian learning for crowdsourced regression”, *arXiv preprint*, arXiv:1702.08840, pp. 1–22, Feb 2017.

[11] Y. Kara, G. Genc, O. Aran and L. Akarun, ”Actively estimating crowd annotation consensus”, *Journal of Artificial Intelligence Research*, vol. 61, pp. 363–405, Feb 2018.

[12] Y. Kara, G. Genc, O. Aran and L. Akarun, ”Modeling annotator behaviors for crowd labeling”, *Neurocomputing*, vol. 160, pp. 141–156, Jul 2015.

[13] C. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[14] P. Rousseeuw and K. Driessen, ”A fast algorithm for the minimum covariance determinant estimator”, *Technometrics*, vol. 41, no. 3, pp. 212–223, Aug 1999.

[15] J. Pasternack and D. Roth, ”Knowing what to believe (when you already know something)”, *Proc. of the 23rd International Conference on Computational Linguistics*, pp. 877–885, Aug 2010.

[16] N. Gourier, D. Hall and J. Crowley, *Head Pose Image Database*, www-prima.inrialpes.fr/perso/Gourier/Faces/HPDatabase.html, 2018.

[17] Face and Gesture Recognition Working group, *The FGNet Aging Database*, www-prima.inrialpes.fr/FGnet/html/benchmarks.html, 2018.

Authorized licensed use limited to: University of Exeter. Downloaded on July 16,2020 at 13:10:47 UTC from IEEE Xplore. Restrictions apply.