Crowd-Labeling for Continuous-Valued Annotations

PhD Thesis Defense

Yunus Emre Kara
Bogazici University, Department of Computer Engineering

01.06.2018

In 1906, statistician Francis Galton observed a contest held in a fair.

An ox was on display and 787 people answered the question:

What would the ox weigh after being slaughtered and dressed?

In 1906, statistician Francis Galton observed a contest held in a fair.

An ox was on display and 787 people answered the question:

What would the ox weigh after being slaughtered and dressed?

He calculated their median guess: 1207 pounds

In 1906, statistician Francis Galton observed a contest held in a fair.

An ox was on display and 787 people answered the question:

What would the ox weigh after being slaughtered and dressed?

He calculated their median guess: 1207 pounds

True answer: 1198 pounds

The Wisdom of Crowds

Collective opinion of a group is often superior to the opinion of any individual in the group.

Wisdom of Crowds for Machine Learning

Labeled datasets are invaluable

Machine learning era is here: virtual assistants, smart devices, ambient intelligence, self-driving cars, ...

Major challenges in dataset labeling

- Labeling large datasets is excessively time consuming.
  Solution: Outsourcing the labeling process
- Expert labelers are expensive.
  Solution: Wisdom of Crowds – Employing unskilled annotators through Amazon Mechanical Turk, CrowdFlower, etc.

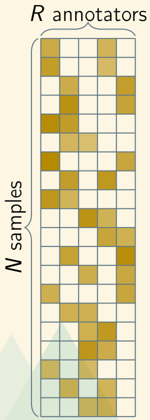
"Crowd-Labeling"

Crowd-Labeling (CL)

The process of collecting annotations from crowds and using them for estimating consensus values to be used as labels.

Problem definition

- Assume that we have N samples and R annotators
- An annotator annotates a subset of N samples
- Each sample is annotated by a group of annotators
- Goal: Obtaining a consensus label for each sample in a fast and cost-effective way
- Passive CL: Random sample-annotator pairing
- Active CL: Smart sample-annotator pairing



Literature on Binary and Categorical-Valued Labels

- **Binary-valued annotations**:

  Passive crowd-labeling: Carpenter (2008), Raykar et al. (2010), Rodrigues, Pereira, and Ribeiro (2013), Welinder, Branson, Belongie, and Perona (2010), Zhang and Obradovic (2012), Yan et al. (2010), Raykar and Yu (2012), Bi and Wang (2013)

  Active crowd-labeling: Sheng, Provost, and Ipeirotis (2008), Donmez and Carbonell (2008a, 2008b), Donmez, Carbonell, and Schneider (2009), Hsueh, Melville, and Sindhwani (2009), Welinder and Perona (2010), Yan, Rosales, Fung, and Dy (2011), Gao, Liu, Ooi, Wang, and Chen (2013), Lin, Mausam, and Weld (2016), Tran-Thanh, Venanzi, Rogers, and Jennings (2013), Tran-Thanh, Huynh, Rosenfeld, Ramchurn, and Jennings (2014), Fang, Yin, and Tao (2014), Raykar and Agrawal (2014), Mozafari, Sarkar, Franklin, Jordan, and Madden (2014), Nguyen, Wallace, and Lease (2015), Zhang, Wen, Tian, Gan, and Wang (2015), Zhuang and Young (2015), Zhu, Xu, and Yan (2015), Ho, Slivkins, and Vaughan (2016)

- **Categorical-valued annotations**:

  Passive crowd-labeling: Dawid and Skene (1979), Raykar and Yu (2011)

  Active crowd-labeling: Welinder and Perona (2010), Yan et al. (2011), Mozafari et al. (2014), Zhu et al. (2015), Kamar, Hacker, and Horvitz (2012), Kamar, Kapoor, and Horvitz (2013, 2015), Venanzi, Guiver, Kohli, and Jennings (2016)

Literature on Ordinal and Continuous-Valued Labels

	Acquisition	Input	Output	Other details
Raykar *et al.* (2010)	Passive	Continuous	Continuous	Can make use of features extracted from the data
Lakshmi. *et al.* (2013)	Passive	Ordinal	Ordinal	Task difficulty is incorporated to the discretization of continuous latent variables
Peng *et al.* (2013)	Passive	Categorical	Categorical	Protein structure prediction
Ok *et al.* (2017)	Passive	Continuous	Continuous	Object detection performance evaluation
Marcus *et al.* (2012)	Active	Binary	Counting	Use of gold standard labels to identify low-quality or spammer annotators
Guo *et al.* (2012)	Active	Binary	Maximum	Ordering by pairwise comparison aggregation
Welinder *et al.* (2010)	Active	Continuous	Continuous	Spammer detection and avoidance

Motivation and approach

- Focus: Sidelined issue of crowd-labeling for continuous-valued annotations
- We aim to target a wide range of labeling problems and thus specifically opt out of using:
  1. Gold standard labels (ground truth): Might be nonexistent or hard to acquire for some problems
  2. Data features: Limit the method to domain-specific problems
- Key characteristics of our approach
  - Input: Continuous, ordinal, or binary valued annotations
  - Output: Continuous consensus values to be used as labels
  - Outputs can be quantized to get ordinal/binarized labels
  - Unsupervised: No training required

Main contributions of this thesis

1. Developed a Bayesian approach for continuous-valued crowd consensus estimation
2. Introduced both passive and active crowd-labeling methods
3. Proposed novel metrics for
   - Assessing sample consensus quality
   - Measuring annotator quality
   - Preventing annotator domination in active crowd-labeling
4. Introduced early stopping in active crowd-labeling based on sample consensus quality
5. Introduced a multivariate crowd-labeling model for taking correlations among attributes into account
6. Collected annotations for datasets with known ground truth

Datasets

Age Annotations

Source of the images: The FGNet Aging Database
1002 pictures of 82 subjects in the age range 0–69
Question: Young/Old rating of the person in the picture
Answer: Annotations in 1–7 range

Crowdflower

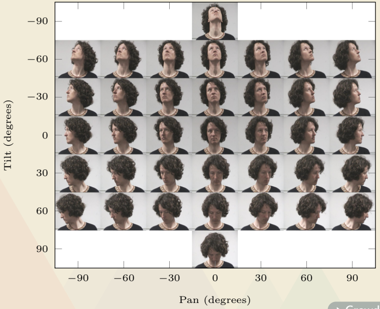
Affective Text Analysis[1]

Question: Annotate news headlines for six emotions (anger, disgust, fear, joy, sadness, and surprise)
Answer: Annotations in 0–100 range

Outcry at N Korea 'nuclear test'

Head Pose Annotations

Source of the images: The Head Pose Image Database
Subsampled 37 distinct head poses
Question: Up/Down and Left/Right orientation of the head
Answer: Tilt/Pan Annotations in 1–7 range

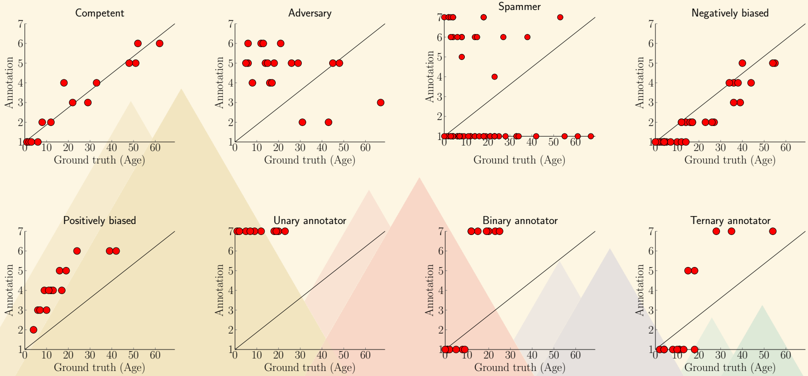


Crowdflower

[1] Snow, R., O'Connor, B., Jurafsky, D., and Ng, A. Y. (2008). 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* (pp. 254–263). Association for Computational Linguistics.

stats

Real annotator examples from the Age Annotations Dataset



Each graph presents all annotations of a single annotator

Details of the univariate models

Parameter	Name	Domain	Prior
Annotator Precision	λ_j	$\mathbb{R}_{>0}$	$\mathcal{G}(\lambda_j; \alpha_\lambda, \beta_\lambda)$
Adverseness	a_j	$\{-1, +1\}$	Flat
Opinion Scale	w_j	$\mathbb{R}_{>0}$	$\mathcal{G}(w_j; \beta_w + 1, \beta_w)$
Annotator Bias	b_j	$\mathbb{R}$	$\mathcal{N}(b_j; \mu_B, s_B^2)$

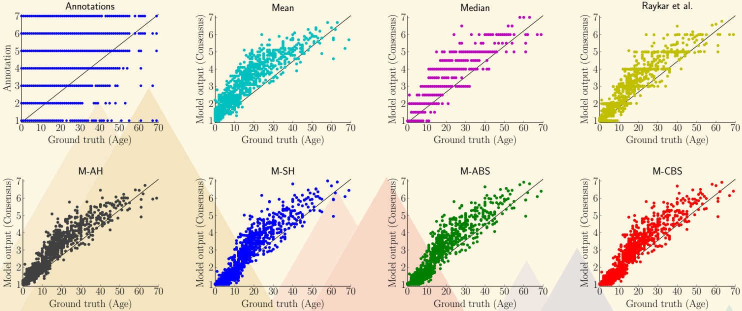
y_k: Value of the k^{th} annotation
x_i: Consensus value of the i^{th} sample
i_k: Sample index of the k^{th} annotation
j_k: Annotator index of the k^{th} annotation

$$p(y_k | x_{i_k}, \theta_{j_k}) = \mathcal{N}\left(y_k; \mu_{j_k}(x_{i_k}), \sigma_{j_k}^2\right)$$

Model	Name	$\mu_j(x)$	σ_j^2
M-AH	Adversary Handling	$a_j x$	$1/\lambda_j$
M-SH	Scale Handling	$a_j w_j x$	$1/\lambda_j$
M-ABS	Annotation Bias Sensitive	$a_j w_j x + b_j$	$1/\lambda_j$
M-CBS	Consensus Bias Sensitive	$a_j w_j (x + b_j)$	${w_j}^2/\lambda_j$

Update equations

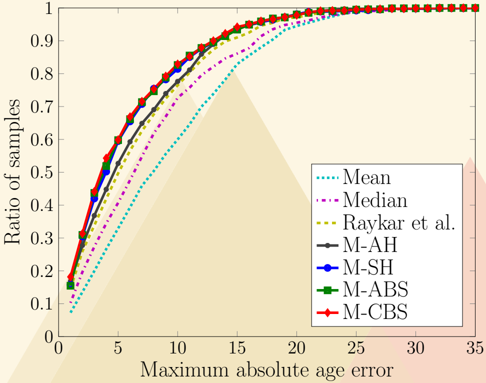
Ground truth estimation performance of models



Consensus vs. ground truth results for the models. In a good model, the dots should be located near the diagonal. (Top left: input annotation data)

Model	Mean	Median	Raykar	M-AH	M-SH	M-ABS	M-CBS
MAE	8.91	7.39	6.46	6.06	5.56	5.58	5.36

Cumulative match curves for the models



- The y coordinate of a point on the CMC is the ratio of the samples that have less error than the corresponding x coordinate.
- M-CBS is the best performer, followed closely by M-ABS and M-SH.

Binarizing continuous labels

Model	Input	MCC	Accuracy	Sensitivity	Specificity
Welinder[2]	Binarized	0.427 ± 0.009	0.718 ± 0.009	0.686 ± 0.010	1.000 ± 0.002
Mean	Ordinal	0.521	0.814	0.796	0.980
Median	Ordinal	0.491	0.782	0.758	*1.000*
Raykar	Ordinal	0.614 ± 0.001	0.880 ± 0.000	0.871 ± 0.000	0.961 ± 0.001
M-AH	Ordinal	0.626 ± 0.000	0.884 ± 0.000	0.874 ± 0.000	0.971 ± 0.000
M-SH	Ordinal	0.644 ± 0.007	0.896 ± 0.003	0.888 ± 0.003	0.961 ± 0.005
M-ABS	Ordinal	0.642 ± 0.008	0.895 ± 0.003	0.887 ± 0.004	0.961 ± 0.005
M-CBS	Ordinal	0.648 ± 0.002	0.897 ± 0.001	0.890 ± 0.001	0.961 ± 0.000

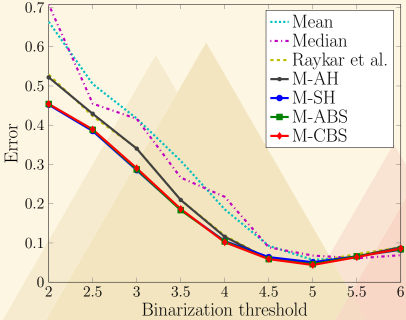
- For Welinder results, the input annotations are binarized
- The resulting consensus values are binarized for the continuous models

$$\text{Matthews correlation coefficient (MCC)} = \frac{\mathrm{TP} * \mathrm{TN} - \mathrm{FP} * \mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}$$

$$\text{Accuracy} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}, \quad \text{Sensitivity} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \quad \text{Specificity} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}$$

[2] Welinder, P., Branson, S., Belongie, S., and Perona, P. (2010). The Multidimensional Wisdom of Crowds. *Advances in Neural Information Processing Systems, 23*, 2424–2432.

Effect of binarization threshold



- Errors of all models are minimized around 5
- One would expect the optimal threshold to be around 4
- Confirms the global bias we observed previously
- Possible reason: Annotators are unaware of the upper limit of the subjects' age in the dataset.

So far, we have...

- Introduced the crowd-labeling problem
- Collected annotations for datasets with known ground truth
- Proposed Bayesian models for consensus estimation
- Adapted the methods to work with binary labeled data
- Outperformed frequently used methods in continuous-valued consensus estimation

What does active crowd-labeling achieve?

- Classical use of crowd-labeling
  - Careless shopper buys excessively
  - No proper planning
  - Purchase is thrown away when the product is of low quality or unneeded.
- Smart shopping (annotation collection):
  Choosing which item (annotation) to buy (incorporate into the annotation pool)
- Active crowd-labeling: The process of smart annotation collection using crowdsourcing
  - Meticulous shopper with limited time and money
  - What am I in need of purchasing?
  - Which vendor should I purchase it from?

Active Crowd-Labeling Algorithm

- Aim: Attaining high quality labels from continuous-valued annotations while reducing the cost of the annotation process.
- Approach: We develop an algorithm that
  - Decides which sample needs a new annotation and who should annotate it.
  - Does simultaneous annotator modeling and consensus estimation
- Iterative algorithm
  1. Choose a sample-annotator pair and request an annotation
  2. Add the new annotation to the annotations set
  3. Estimate consensus and relearn annotators

Which sample needs a new label?

- Identifying samples with low consensus quality is a good way to minimize annotation costs.
- Motivation: A sample's quality may roughly be assessed by the variance of the consensus posterior distribution: $p(x_i|Y,\theta)$
- The consensus quality score $S_S(i)$ of sample i is the inverse of the posterior's variance:

M-AH	M-SH	M-ABS	M-CBS
$\sum_{k:i_k=i} \lambda_{j_k}$	$\sum_{k:i_k=i} {w_{j_k}}^2 \lambda_{j_k}$	$\sum_{k:i_k=i} {w_{j_k}}^2 \lambda_{j_k}$	$\sum_{k:i_k=i} \lambda_{j_k}$

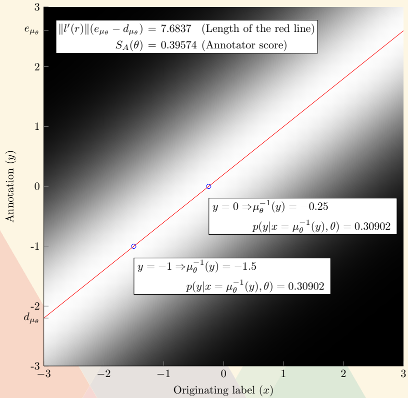
- Inferred form readily available annotator parameters.
- A sample's consensus quality is only as good as the annotators' precision and opinion scale that have annotated it.

Annotator Competence Scoring: Who annotates better?

- We need to distinguish between good and bad annotators for efficiently spending the budget.
- Annotator competence score is the path integral of the joint probability along the mode of the posterior $p(y|x,\theta)$:

$$S_A(\theta) = \int_l p(x,y|\theta)ds$$

- Annotator score is high when:
  - Opinion scale w_j is close to 1
  - Bias b_j is close to 0
  - Precision λ_j is high

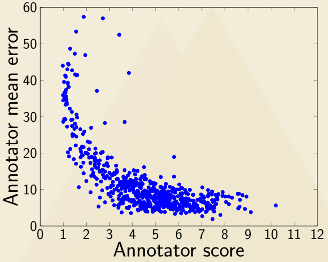


Grayscale values represent posterior probability of annotation value ($p(y|x,\theta)$); the higher the intensity, the higher the probability.

Annotator score formulas

Verification of annotator scoring

Annotator Score vs Annotator Mean Error for age labels



Our annotator scoring mechanism associates low-error annotators with a high score.

The error with top 50% performing annotators is half of the error with the bottom 50% performers.

Model	Top 50%	Bottom 50%
Mean	5.56	13.49
Median	6.19	12.63
Raykar	6.13 ± 0.037	12.44 ± 0.019
M-AH	5.65 ± 0.000	11.25 ± 0.072
M-SH	5.60 ± 0.075	10.06 ± 0.285
M-ABS	5.60 ± 0.078	10.12 ± 0.337
M-CBS	5.52 ± 0.000	10.18 ± 0.091

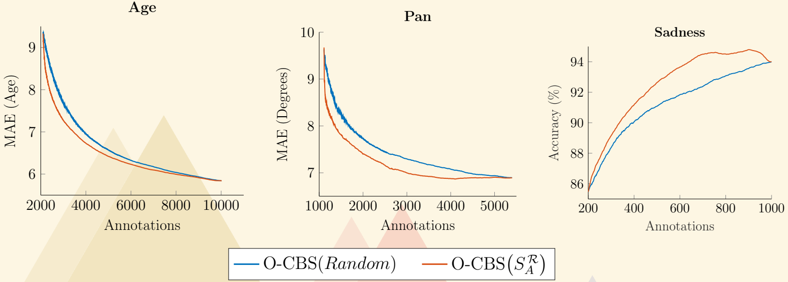
To wrap up, we...

- Outlined our active crowd-labeling algorithm
- Introduced a sample consensus quality score function (S_S)
- Introduced and verified an annotator competence score function (S_A)

Improving the Existing Consensus Using Active Crowd-Labeling

- Scenario: Assume that necessary annotations are already collected and labels are estimated for each sample
- O-CBS is designed to improve the consensus qualities using *the same annotators* by reconsulting them for the samples that they did not annotate beforehand
- Sample scoring function (S_S) is fixed as introduced before
- Denoted as: O-CBS(S_A), O-CBS($S_A^{\mathcal{K}}$), O-CBS(S_A^1), ..., O-CBS($Random$)

Effectiveness of S_S



- O-CBS($Random$): Both samples and annotators are selected randomly
- O-CBS($S_A^{\mathcal{R}}$): Annotators are selected randomly, samples are selected with S_S
- S_S is a favorable sample selection strategy in terms of MAE and accuracy
- Even without an annotator selection strategy, S_S by itself provides significant improvement

All results

Shortcomings of S_A

- In a small crowd, annotators with high workload have more influence on the system
- Spammers abuse the system for more money
  Thus, incompetent annotator domination is highly probable in the early phases of the crowd-labeling
- Incompetent annotator domination
  $\Longrightarrow$ Opinion of a competent annotator will be an outlier
  $\Longrightarrow$ Fewer annotations from the competent annotator
- For a more balanced system, we need to:
  - Prevent early annotator overloading
    Suppress S_A proportionally to the annotator workloads
  - Reduce this effect in time
    Employ competent annotators more frequently later on

Dominance suppression based annotator competence score

- *Dominance suppression factor:* $\left|\mathcal{K}^j\right|^{-\varphi\frac{|\mathcal{J}'|}{|\mathcal{K}|}}$
  1. $|\mathcal{K}^j|$: the number of annotations of annotator j

Dominance suppression based annotator competence score

- *Dominance suppression factor:* $\left|\mathcal{K}^j\right|^{-\varphi\frac{|\mathcal{J}'|}{|\mathcal{K}|}}$
  1. $|\mathcal{K}^j|$: the number of annotations of annotator j
  2. $\varphi > 0$: **the dominance suppression coefficient**

Dominance suppression based annotator competence score

- *Dominance suppression factor:* $\left|\mathcal{K}^j\right|^{-\varphi\frac{|\mathcal{J}'|}{|\mathcal{K}|}}$
  1. $|\mathcal{K}^j|$: the number of annotations of annotator j
  2. $\varphi > 0$: **the dominance suppression coefficient**
  3. $\frac{|\mathcal{J}'|}{|\mathcal{K}|}$: inverse of the average annotator workload
     - $|\mathcal{K}|$: current number of annotations
     - $|\mathcal{J}'|$: the number of active annotators

Dominance suppression based annotator competence score

- *Dominance suppression factor:* $\left|\mathcal{K}^j\right|^{-\varphi\frac{|\mathcal{J}'|}{|\mathcal{K}|}}$
  1. $|\mathcal{K}^j|$: the number of annotations of annotator j
  2. $\varphi > 0$: **the dominance suppression coefficient**
  3. $\frac{|\mathcal{J}'|}{|\mathcal{K}|}$: inverse of the average annotator workload
     - $|\mathcal{K}|$: current number of annotations
     - $|\mathcal{J}'|$: the number of active annotators
- The **updated annotator competence score** is

$$S_A^{\varphi}(j) = S_A(j)\left|\mathcal{K}^j\right|^{-\varphi\frac{|\mathcal{J}'|}{|\mathcal{K}|}}$$

Dominance suppression based annotator competence score

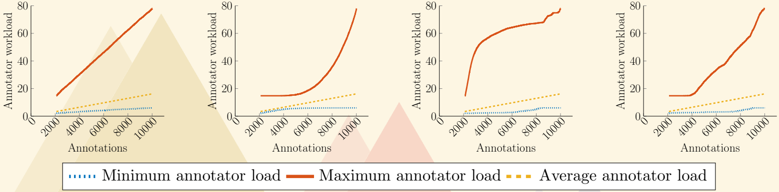
- ***Dominance suppression factor:*** $\left|\mathcal{K}^j\right|^{-\varphi\frac{|\mathcal{J}'|}{|\mathcal{K}|}}$
  1. $|\mathcal{K}^j|$: the number of annotations of annotator j
  2. $\varphi > 0$: **the dominance suppression coefficient**
  3. $\frac{|\mathcal{J}'|}{|\mathcal{K}|}$: inverse of the average annotator workload
     - $|\mathcal{K}|$: current number of annotations
     - $|\mathcal{J}'|$: the number of active annotators
- The **updated annotator competence score** is

$$S_A^{\varphi}(j) = S_A(j)\left|\mathcal{K}^j\right|^{-\varphi\frac{|\mathcal{J}'|}{|\mathcal{K}|}}$$

- As a **baseline**, simple annotator score based only on the **annotator's workload**: $S_A^{\mathcal{K}}(j) = |\mathcal{K}^j|^{-1}$

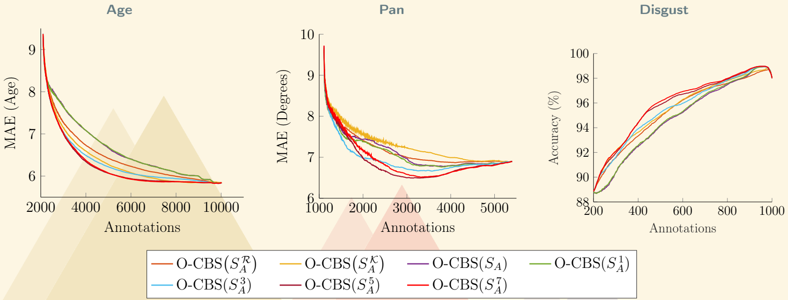
Effect of Dominance Suppression on Annotator Workloads

O-CBS$\left(S_A^{\mathcal{R}}\right)$	O-CBS$\left(S_A^{\mathcal{K}}\right)$	O-CBS(S_A)	O-CBS$\left(S_A^{5}\right)$
Random Selection	Selecting annotators inversely proportion to workload	Selecting highest ranking annotators at the time	Annotator selection with dominance suppression ($\varphi = 5$)



- The results are provided for the Age Annotations dataset
- If we use S_A, the system over-employs the high scoring annotators. This may be risky if the system starts out with ill-intentioned annotators.
- When we have dominance suppression, annotator workloads are distributed more evenly at the beginning.
- Later on the effect of dominance suppression diminishes, and the system focuses on high quality annotators.

Effectiveness of Annotator Dominance Suppression



- Dominance suppression is beneficial in terms of MAE and accuracy
- Using S_A^5 and S_A^7 give the best results
- Small φ results in even lower performance than the baseline methods
- The trough shape in the *pan* dataset occurs due to distinguishing high quality annotators early on.

All results

To wrap up, we...

- Introduced O-CBS which improves the existing consensus obtained from a crowd
- Showed that the sample consensus quality score function S_S works well
- Introduced annotator competence scoring functions that prevent annotator dominance
- Showed that O-CBS and annotator dominance prevention works

Shortcomings of O-CBS

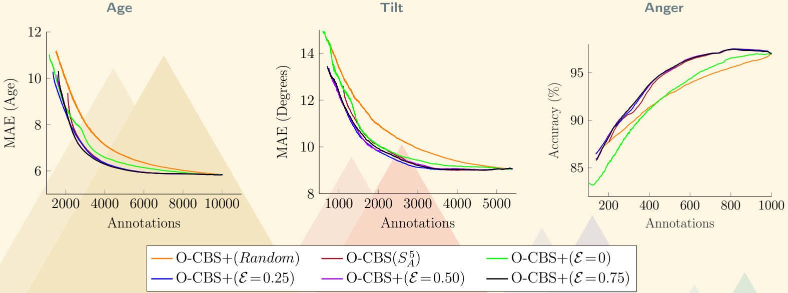
- O-CBS is designed for improving the existing consensus quality – it needs an annotated starting set
- More beneficial to identify the annotator quality A.S.A.P.
  Save money and time with fewer annotations
- Utilizes only current annotators; does not assess brand new annotators
- O-CBS is not designed for the addition of new samples.

O-CBS+: Online M-CBS from scratch

- O-CBS+ is designed for annotation collection from scratch
- Samples without any annotation have first priority
- Decide to exploit a known annotator or explore a new annotator with exploration parameter $\mathcal{E}$
- Exploring a new annotator:
  - Request two annotations
  - At least one of the annotations should be of an already annotated sample
- Baseline: O-CBS+$(Random)$, similar to O-CBS$(Random)$
- S_A^5 is fixed as the annotator competence scoring function

Algorithm

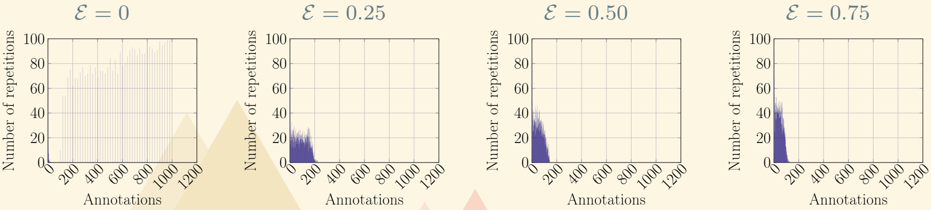
Effect of annotator exploration



- Annotator exploration is beneficial for reducing MAE or increasing accuracy
- Fast exploration is better (selecting a larger $\mathcal{E}$)

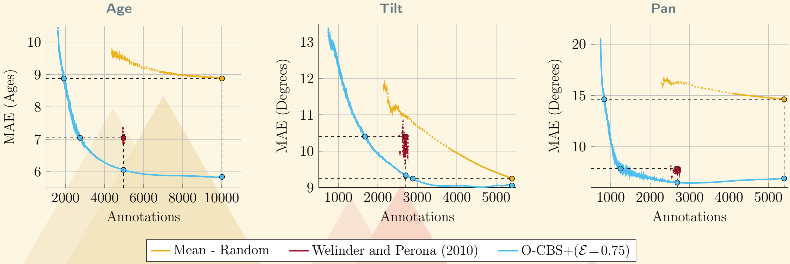
All results

Is it wise to take risks by incorporating new annotators?



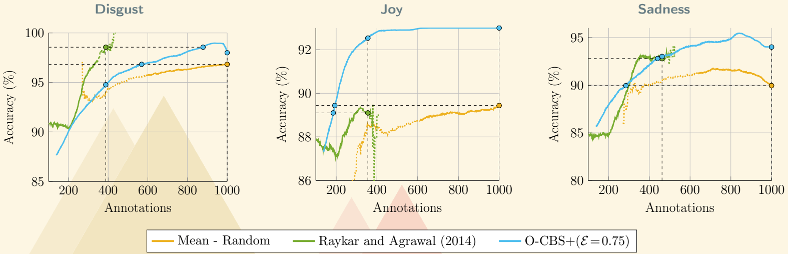
- Here, we observe when the new annotators are introduced into the system on the *anger* dataset
- Since the datasets contain a finite number of annotators, the system exhausts new annotators quickly.
- In limited annotator set: Assess all annotators quickly so that better annotators are utilized early on.
- In open-ended active crowd-labeling: Since annotators are virtually infinite, using a smaller $\mathcal{E}$ is advised.

O-CBS+ vs. Welinder and Perona (2010)



- We match the performance of the Mean-Random baseline method using 18%, 53%, and 15% of the annotations in the *age*, *tilt*, and *pan* datasets, respectively.
- Welinder and Perona's method acquires annotations sample by sample and disregards spammers. Thus, the red curve is very short.
- We match the performance of Welinder and Perona's method using 54%, 70%, and 52% of the annotations that their method requires in the *age*, *tilt*, and *pan* datasets, respectively.

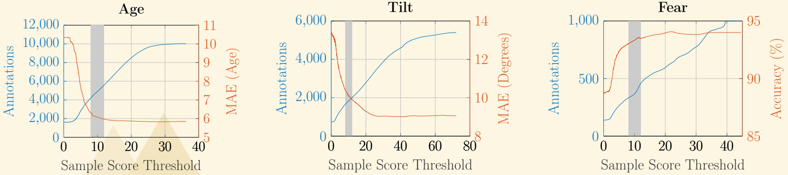
O-CBS+ vs. Raykar and Agrawal (2014)



- We present three interesting scenarios on the Affective Text Analysis datasets.
- On the *disgust* dataset, O-CBS+ matches the performance of Raykar and Agrawal with more annotations.
- O-CBS+ significantly outperforms both methods on the *joy* dataset both in terms of overall accuracy and using less annotations.
- On the sadness dataset, O-CBS+ matches the performance of Raykar and Agrawal with room for accuracy improvement if desired.

All results

Enforcing sample consensus quality induced stopping



- Up to this point, we stopped active crowd-labeling when a predefined number of annotations were collected. But, we may waste away the budget by collecting unnecessary annotations.
- We want to stop the annotation process upon attaining satisfactory sample consensus values.
- Our aim is keeping sample consensus posterior variances (inverse of S_S) below a threshold (*i.e.* we want $S_S(x_i) > \tau$ for all samples).
- Our motivation is simple: keep the sample consensus posterior variances below 0.1. This corresponds to τ=10.
- The gray band around τ=10 is a satisfactory point to stop: Error is significantly reduced at the time of termination.

All results

Comparative Performance of O-CBS+ with Existing Methods

	Welinder and Perona[3]		O-CBS+($\mathcal{E}=0.75$), $\tau=10$	
Dataset	Ann.	MAE	Ann.	MAE
Age	4969.77	7.02	4911.37	6.07
Tilt	2705.03	10.10	1836.39	10.11
Pan	2689.77	7.58	1721.22	7.13

	Raykar and Agrawal[4]		O-CBS+($\mathcal{E}=0.75$), $\tau=10$	
Dataset	Ann.	Acc.(%)	Ann.	Acc.(%)
Anger	415.86	96.07	386.20	94.58
Disgust	387.78	98.92	392.72	95.53
Fear	363.49	91.50	365.74	93.77
Joy	355.51	89.17	352.96	92.79
Sadness	462.34	93.31	390.84	92.72
Surprise	365.22	91.60	371.00	94.67

- Red values indicate that our method is significantly superior than the opponent method
- Yellow values indicate a tie
- Gray values indicate that the opponent method is better

[3] Welinder, P. and Perona, P. (2010). Online Crowdsourcing: Rating Annotators and Obtaining Cost-Effective Labels. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, 25–32.

[4] Raykar, V. C. and Agrawal, P. (2014). Sequential Crowdsourced Labeling as an Epsilon-Greedy Exploration in a Markov Decision Process. *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics (AISTATS-14), 33*, 832–840.

All results

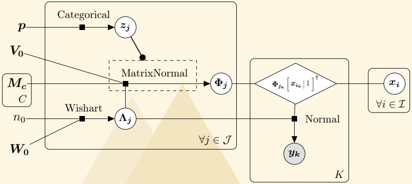
To wrap up, we...

- Introduced O-CBS+ which starts the active crowd-labeling from scratch
- Investigated explore-exploit trade-off for incorporating new annotators
- Showed that O-CBS+ indeed works well by comparing it with existing works
- Introduced a sample score related stopping criterion to active crowd-labeling process

Motivation and approach

- Some annotation problems have multiple attributes annotated by the same annotator
  - Head Pose Annotations (tilt, pan)
  - Affective Text Analysis (anger, disgust, fear, joy, sadness, surprise)
- Different attributes correlate with each other
- Taking these correlations into account may prove useful
- We introduce a multivariate model
- A variational Bayesian solution

The Model



- $\boldsymbol{y_k}$: Annotation ($d \times 1$ vector)
- $\boldsymbol{x_i}$: Consensus ($d \times 1$ vector)
- $\boldsymbol{\Lambda_j}$: Annotator precision ($d \times d$ positive definite matrix)
- $\boldsymbol{\Phi_j}$: Annotator opinion scale and bias ($d \times (d+1)$ matrix)
- $\boldsymbol{z_j}$: Annotator category (1-of-C binary vector)
- $\boldsymbol{M_c}$: Annotator category definition ($d \times (d+1)$ matrix, $\forall c \in \{1, \ldots, C\}$)
- $\boldsymbol{V_0}$: $(d+1) \times (d+1)$ positive definite matrix
- $\boldsymbol{W_0}$: $d \times d$ positive definite matrix
- $\boldsymbol{p} \in [0,1]^C$, such that $\sum_{c=1}^{C} \boldsymbol{p}_c = 1$
- $n_0 > d-1$ is a real scalar

$$p(Y|\Phi,\Lambda,X) = \prod_{k=1}^{K} \underbrace{\mathcal{N}_d\left(\boldsymbol{y_k}; \boldsymbol{\Phi_{j_k}} \begin{bmatrix} \boldsymbol{x_{i_k}} \\ \hdashline 1 \end{bmatrix}, \boldsymbol{\Lambda}_{\boldsymbol{j_k}}^{-1}\right)}_{p\left(y_k \middle| \boldsymbol{\Phi_{j_k}}, \boldsymbol{\Lambda_{j_k}}, \boldsymbol{x_{i_k}}\right)}$$

$$p(\Phi,\Lambda,Z,X) = \prod_{j=1}^{R} \underbrace{\mathcal{C}\left(\boldsymbol{z_j}; \boldsymbol{p}\right)}_{p\left(\boldsymbol{z_j}\right)} \underbrace{\mathcal{W}_d\left(\boldsymbol{\Lambda_j}; \boldsymbol{W_0}, n_0\right)}_{p\left(\boldsymbol{\Lambda_j}\right)} \underbrace{\prod_{c=1}^{C} \mathcal{MN}_{d,d+1}\left(\boldsymbol{\Phi_j}; \boldsymbol{M_c}, \boldsymbol{\Lambda}_{\boldsymbol{j}}^{-1}, \boldsymbol{V_0}\right)^{\boldsymbol{z_j}_c}}_{p\left(\boldsymbol{\Phi_j} \middle| \boldsymbol{\Lambda_j}, \boldsymbol{z_j}\right)}$$

A Variational Bayes Solution

- Approximates the posterior distribution of the latent variables Φ, Λ, Z, and X given the annotations Y
- For obtaining a tractable solution, factorize the latent variables:

$$q(\Phi, \Lambda, Z, X) = q(\Phi, \Lambda) q(Z, X) = \left(\prod_{j=1}^{R} q(\boldsymbol{\Phi_j} | \boldsymbol{\Lambda_j}) q(\boldsymbol{\Lambda_j}) \right) \left(\prod_{j=1}^{R} q(\boldsymbol{z_j}) \right) \left(\prod_{i=1}^{N} q(\boldsymbol{x_i}) \right)$$

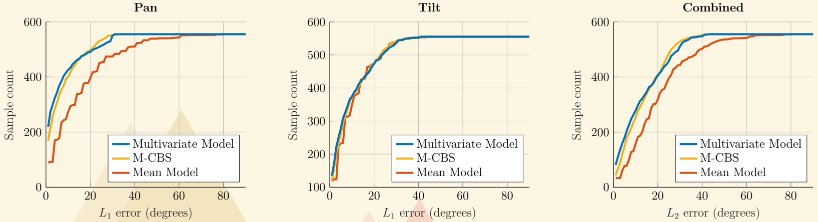
- Minimize the KL divergence, between $p(\Phi, \Lambda, Z, X|Y)$ and $q(\Phi, \Lambda, Z, X)$

$$q^*(\Phi, \Lambda, Z, X) = \left(\prod_{j=1}^{R} \underbrace{\mathcal{MN}_{d,d+1}\left(\boldsymbol{\Phi_j}; \boldsymbol{M_j}, \boldsymbol{\Lambda_j}^{-1}, \boldsymbol{V_j}\right)}_{q^*(\boldsymbol{\Phi_j}|\boldsymbol{\Lambda_j})} \underbrace{\mathcal{W}_d\left(\boldsymbol{\Lambda_j}; \boldsymbol{W_j}, n_0 + N_j\right)}_{q^*(\boldsymbol{\Lambda_j})} \right)$$

$$\left(\prod_{j=1}^{R} \underbrace{\mathcal{C}\left(\boldsymbol{z_j}; \boldsymbol{\rho_j}\right)}_{q^*(\boldsymbol{z_j})} \right) \left(\prod_{i=1}^{N} \underbrace{\mathcal{N}_d\left(\boldsymbol{x_i}; \boldsymbol{\mu_i}, \boldsymbol{\Sigma_i}^{-1}\right)}_{q^*(\boldsymbol{x_i})} \right)$$

▸ Update equations

Experiments on the Head Pose Annotations dataset



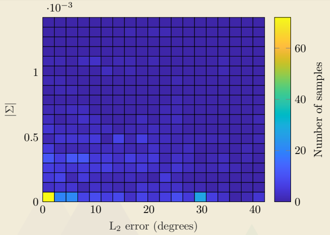
- Cumulative match curves: Comparison with two univariate models (mean model and M-CBS)
- Combined: Euclidean distance (L_2-norm) of a sample's 2-dimensional ground truth and its inferred consensus tuple
- The y coordinate of a point on the CMC is the number of samples that have less error than the corresponding x coordinate.

Lower bound

Observations on Annotators

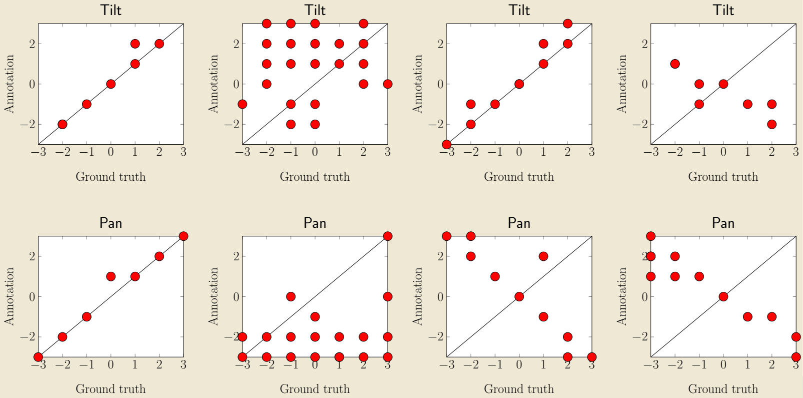
- Adverseness: Incorporate into the model using M_c matrices. Multivariate model categorized all annotators correctly:
  - One annotator is adverse in both *tilt* and *pan*,
  - 30 annotators are adverse in only *pan*,
  - Remaining 158 annotators are not adverse.
- Competent vs. spammer: We can rank the annotators using determinants of annotator precision matrices

Sample error - posterior sample variance relation



Preliminary Experiments

Competent annotator $(\max \mathbb{E}\left[\log \lvert\Lambda_j\rvert\right] \approx 23.2652)$	Spammer annotator $(\min \mathbb{E}\left[\log \lvert\Lambda_j\rvert\right] \approx -2.1495)$	Adversary annotator in *pan*	Adversary annotator in both *tilt* and *pan*



To wrap up, we...

- Introduced a multivariate annotation model that takes correlations of different attributes into account
- Gave a variational Bayes solution
- Performed preliminary experiments on the Head Pose Annotations dataset
- Showed that the proposed multivariate model holds potential for improvement

Observations and Discussion

- Hard to rate samples mislead most annotators, resulting in bad consensus values.
- After competent annotators are selected, simple models begin measuring up to complex models.
- Annotator dominance factor holds utmost importance in assessing the annotator quality correctly and timely.
- Timely exploration of new annotators prevents the initial annotators from biasing the model towards their opinion.
- Up to 80% reduction in the number of annotations (spending only one fifth of the original budget)
- Adverseness in different attributes are identified correctly with the multivariate model.

Conclusion

- Our focus in this thesis is on the rather sidelined issue of continuous-valued crowd-labeling.
- We propose Bayesian models that do not require gold standard labels or features extracted from data.
- We introduce active crowd-labeling methods
- We propose novel assessment ideas concerning sample consensus quality, annotator competence, and annotator dominance prevention.
- We introduce a sample score related stopping criterion to active crowd-labeling process
- Correlations of different attributes are taken into account in the multivariate annotation model.
- Proposed methods outperform contender methods in the literature.

Future directions

- Updating the models for better handling of binary annotations
- Incorporating a sample difficulty parameter
- Investigating different sample consensus quality and annotator competence scoring functions
- Addressing the issue of annotator competence fluctuation over time Distributing the tasks according to the recent performance
- Improving the multivariate model
    - Spammer avoidance: Disregard spam annotations completely
    - Applying active crowd-labeling

THANK YOU

¿Questions?

Publications related to this thesis:

- Kara, Y. E., Genc, G., Aran, O., & Akarun, L. (2015). Modeling Annotator Behaviors for Crowd Labeling. *Neurocomputing*, *160*, 141–156
- Kara, Y. E., Genc, G., Aran, O., & Akarun, L. (2018). Actively estimating crowd annotation consensus. *Journal of Artificial Intelligence Research*, *61*, 363–405
- “Multivariate Crowd-Labeling”, In preparation

APPENDICES

A. Passive CL (app.)

- Datasets
- Models

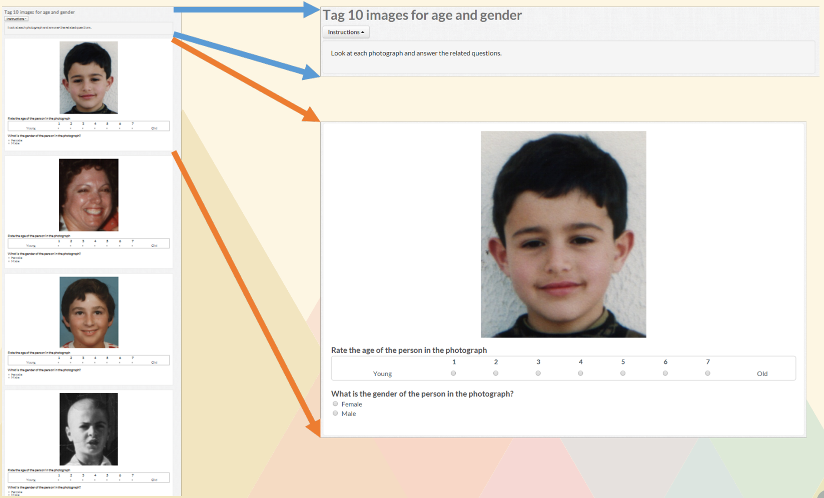
B. Active CL (app.)

C. O-CBS (app.)

D. O-CBS+ (app.)

E. Multivariate (app.)

Age Annotations Crowdflower Screen



Head Pose Annotations Crowdflower Screen

In the question below, the complete scale of 1 to 7 refers to:

1	2	3	4	5	6	7
Left	More towards left	Slightly left	Straight	Slightly right	More towards right	Right

Horizontal orientation of the head (pan)

	1	2	3	4	5	6	7	
Left	○	○	○	○	○	○	○	Right

Please annotate with respect to your own left and right

In the question below, the complete scale of 1 to 7 refers to:

1	2	3	4	5	6	7
Down	More towards down	Slightly down	Straight	Slightly up	More towards up	Up

Vertical orientation of the head (tilt)

	1	2	3	4	5	6	7	
Down	○	○	○	○	○	○	○	Up

Is the person wearing glasses?

○ Yes
○ No

Back

Datasets

Dataset	Annotations	Samples	Annotators	Ground Truth Range	Annotation Range
Age Annotations (introduced in this work)	10020	1002	619	$\{0,...,69\}$	$\{1,...,7\}$
Head Pose Annotations: *tilt*, *pan* (introduced in this work)	5399	555	189	$\{-90,...,90\}$	$\{1,...,7\}$
Affective Text Analysis[5]: *anger*, *disgust*, *fear*, *joy*, *sadness*, *surprise*	1000	100	38	$\{0,...,100\}$	$\{0,...,100\}$
ELEA Personality Impressions Data[6]: Big five personality traits	306	102	5	$\{1,...,7\}$	$\{1,...,7\}$

[5] Snow, R., O'Connor, B., Jurafsky, D., and Ng, A. Y. (2008). 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* (pp. 254–263). Association for Computational Linguistics.

[6] Aran, O. and Gatica-Perez, D. (2013). One of a Kind: Inferring Personality Impressions in Meetings. *Proceedings of the 15th acm on international conference on multimodal interaction* (pp. 11–18). ACM.

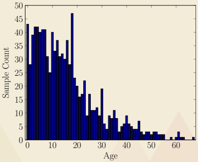
Back

Dataset statistics

Age Annotations

Annotator workload	Number of annotators Set 1	Set 2	Total
1	2	4	6
6	0	1	1
7	1	0	1
9	2	0	2
10	208	0	208
11	1	0	1
14	1	0	1
15	0	292	292
16	0	1	1
19	1	0	1
20	82	0	82
29	1	1	2
30	26	12	38
31	1	0	1
33	0	1	1
36	1	0	1
40	5	0	5
42	0	1	1
43	0	1	1
45	0	1	1
50	3	0	3
59	0	1	1

The FGNet Aging Database Age Histogram



ELEA Personality Impressions

Annotator Id	Workload
1	91
2	83
3	77
4	49
5	6

Head Pose Annotations

Annotator workload	Number of annotators
5	1
10	61
17	1
20	45
24	1
30	26
39	1
40	15
45	2
50	13
55	1
60	7
70	5
75	1
80	4
84	1
90	2
100	2

Head Pose Annotations

Sample annotation count	Number of samples
7	10
8	10
9	475
15	6
16	34
17	20

Back

M-AH update equations

$$x_i = \frac{\sum_{k:i_k=i} \lambda_{j_k} a_{j_k} y_k}{\sum_{k:i_k=i} \lambda_{j_k}}$$

$$\lambda_j = \frac{2(\alpha_\lambda - 1) + N_j}{2\beta_\lambda + \sum_{k:j_k=j} (y_k - a_j x_{i_k})^2}$$

$$a_j = \operatorname{sgn}\left(\sum_{k:j_k=j} y_k x_{i_k} \right)$$

Back

M-SH update equations

$$x_i = \frac{\sum_{k:i_k=i} \lambda_{j_k} w_{j_k} a_{j_k} y_k}{\sum_{k:i_k=i} \lambda_{j_k} w_{j_k}^2}$$

$$\lambda_j = \frac{2(\alpha_\lambda - 1) + N_j}{2\beta_\lambda + \sum_{k:j_k=j} (y_k - a_j w_j x_{i_k})^2}$$

$$a_j = \operatorname{sgn}\left(\sum_{k:j_k=j} y_k x_{i_k} \right)$$

$$w_j = \frac{a_j \sum_{k:j_k=j} y_k x_{i_k} - \frac{\beta_w}{\lambda_j}}{2 \sum_{k:j_k=j} x_{i_k}^2} + \sqrt{\left(\frac{a_j \sum_{k:j_k=j} y_k x_{i_k} - \frac{\beta_w}{\lambda_j}}{2 \sum_{k:j_k=j} x_{i_k}^2} \right)^2 + \frac{\beta_w}{\lambda_j \sum_{k:j_k=j} x_{i_k}^2}}$$

Back

M-ABS update equations

$$x_i = \frac{\sum_{k:i_k=i} \lambda_{j_k} w_{j_k} (a_{j_k} y_k - b_{j_k})}{\sum_{k:i_k=i} \lambda_{j_k} w_{j_k}^2}$$

$$\lambda_j = \frac{2(\alpha_\lambda - 1) + N_j}{2\beta_\lambda + \sum_{k:j_k=j} (y_k - a_j(w_j x_{i_k} + b_j))^2}$$

$$a_j = \operatorname{sgn}\left(\sum_{k:j_k=j} y_k (w_j x_{i_k} + b_j)\right)$$

$$b_j = \frac{a_j \sum_{k:j_k=j} y_k - w_j \sum_{k:j_k=j} x_{i_k} + \frac{\mu_B}{\lambda_j s_B^2}}{N_j + \frac{1}{\lambda_j s_B^2}}$$

$$w_j = \frac{\sum_{k:j_k=j} (a_j y_k - b_j) x_{i_k} - \frac{\beta_w}{\lambda_j}}{2 \sum_{k:j_k=j} x_{i_k}^2} + \sqrt{\left(\frac{\sum_{k:j_k=j} (a_j y_k - b_j) x_{i_k} - \frac{\beta_w}{\lambda_j}}{2 \sum_{k:j_k=j} x_{i_k}^2}\right)^2 + \frac{\beta_w}{\lambda_j \sum_{k:j_k=j} x_{i_k}^2}}$$

Back

M-CBS update equations

$$x_i = \frac{\sum_{k:i_k=i} \lambda_{j_k} \left(\frac{a_{j_k} y_k}{w_{j_k}} - b_{j_k} \right)}{\sum_{k:i_k=i} \lambda_{j_k}}$$

$$\lambda_j = \frac{2(\alpha_\lambda - 1) + N_j}{2\beta_\lambda + \sum_{k:j_k=j} \left(\frac{y_k}{w_j} - a_j(x_{i_k} + b_j) \right)^2}$$

$$a_j = \operatorname{sgn}\left(\sum_{k:j_k=j} y_k (x_{i_k} + b_j) \right)$$

$$b_j = \frac{\frac{a_j}{w_j} \sum_{k:j_k=j} y_k - \sum_{k:j_k=j} x_{i_k} + \frac{\mu_B}{\lambda_j s_B^2}}{N_j + \frac{1}{\lambda_j s_B^2}}$$

w_j is a root of the following equation:

$$V_3 \left(\frac{1}{w_j} \right)^3 + V_2 \left(\frac{1}{w_j} \right)^2 + V_1 \left(\frac{1}{w_j} \right) + V_0 = 0 \text{ where}$$

$$V_0 = -\beta_w,$$

$$V_1 = \beta_w - N_j,$$

$$V_2 = -\lambda_j a_j \sum_{k:j_k=j} y_k (x_{i_k} + b_j),$$

$$V_3 = \lambda_j \sum_{k:j_k=j} y_k^2.$$

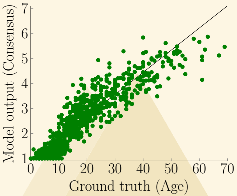
Back

Age label estimation errors on different sets

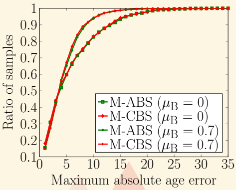
Model	Set 1	Set 2	Joint
Mean	9.68	8.95	8.91
Median	8.34	7.94	7.39
Raykar	7.20 ± 0.048	6.94 ± 0.062	6.46 ± 0.019
M-AH	6.59 ± 0.002	6.35 ± 0.001	6.06 ± 0.000
M-SH	6.06 ± 0.112	6.04 ± 0.098	5.56 ± 0.087
M-ABS	6.07 ± 0.116	6.04 ± 0.103	5.58 ± 0.083
M-CBS	5.91 ± 0.011	5.84 ± 0.006	5.36 ± 0.008

Removing global bias

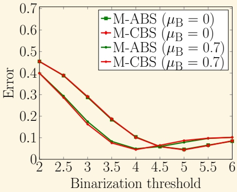
M-ABS



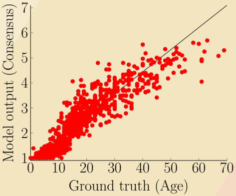
CMC



Binarization error



M-CBS



Models	MAE			RMSE		
	Set 1	Set 2	Joint	Set 1	Set 2	Joint
M-ABS	4.46	4.73	4.2	6.05	6.14	5.42
M-CBS	4.47	4.58	4.13	5.93	5.91	5.32
Original Results						
M-ABS	5.99	5.88	5.41	8.52	8.35	7.54
M-CBS	5.90	5.83	5.35	8.34	8.12	7.49

A. Passive CL (app.)

B. Active CL (app.)

- Algorithm
- Annotator score

C. O-CBS (app.)

D. O-CBS+ (app.)

E. Multivariate (app.)

Active Crowd-Labeling Algorithm

Input:

Set of all samples $\mathcal{I}$ to be annotated
Set of all annotators $\mathcal{J}$
Set of currently active annotators $\mathcal{J}'$
Set of current annotations $\mathcal{K}$

function ACL($\mathcal{I}, \mathcal{J}, \mathcal{K}$)
    EstimateLabels($\mathcal{I}, \mathcal{J}, \mathcal{K}$)
    repeat
        $k \leftarrow$ RequestAnnotation($\mathcal{I}, \mathcal{J}, \mathcal{J}', \mathcal{K}$)
        $\mathcal{K} \leftarrow \mathcal{K} \cup k$ ▷ Add the newly acquired annotation to the annotations set
        EstimateLabels($\mathcal{I}, \mathcal{J}, \mathcal{K}$) ▷ Estimate consensus and relearn annotators
    until Budget limit
end function

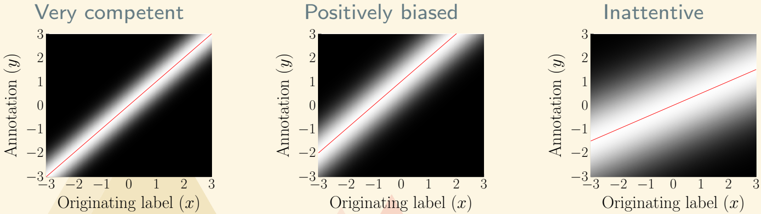
Annotator score formulas for the proposed models

	$\mu_\theta(x)$	$\|\|l'(y)\|\|$	$S_A(\theta)$
M-AH	ax	$\sqrt{2}$	$\sqrt{\frac{\lambda}{\pi}}(e_\theta - d_\theta)$
M-SH	awx	$\sqrt{1+\frac{1}{w^2}}$	$\frac{1}{w}\sqrt{\frac{\lambda(1+w^2)}{2\pi}}(e_\theta - d_\theta)$
M-ABS	$awx + b$	$\sqrt{1+\frac{1}{w^2}}$	$\frac{1}{w}\sqrt{\frac{\lambda(1+w^2)}{2\pi}}(e_\theta - d_\theta)$
M-CBS	$aw(x+b)$	$\sqrt{1+\frac{1}{w^2}}$	$\frac{1}{w^2}\sqrt{\frac{\lambda(1+w^2)}{2\pi}}(e_\theta - d_\theta)$

$$d_\theta = \min\{c, \max\{a_\theta \mu_\theta(-c), -c\}\}$$
$$e_\theta = \max\{-c, \min\{a_\theta \mu_\theta(c), c\}\}$$

Back

Some annotator types

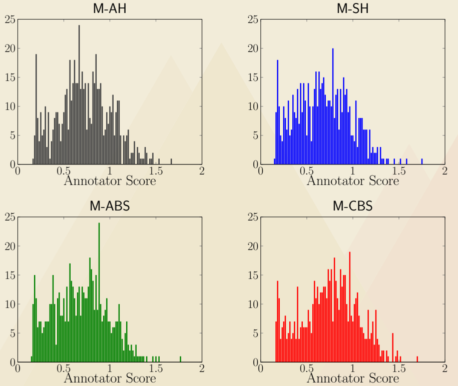


- Grayscale values represent posterior probability of annotation value ($p(y|x,\theta)$); the higher the intensity, the higher the probability.
- The red line is the peak of this distribution.
- For very competent annotators, opinion scale (w_j) is close to 1 and annotator bias (b_j) is close to 0. Additionally, they have high precision (λ_j) values resulting in a concentrated band of annotations around the peak.
- In contrast, inattentive annotators have lower λ_j values which result in more scattered annotations.

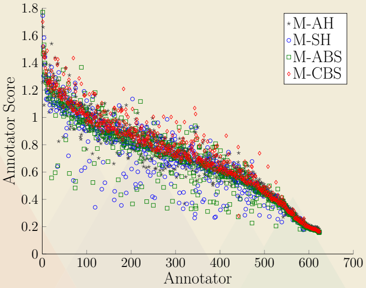
Back

Annotator score

Annotator score histograms for the proposed models

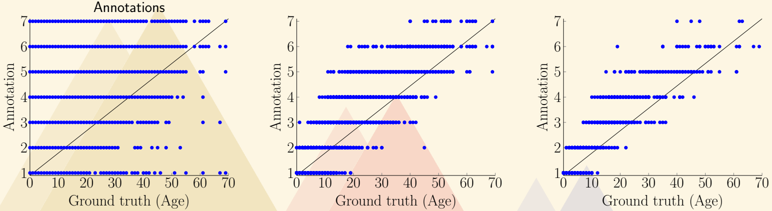


Annotator score comparison for the proposed models



Back

Annotations of only the best annotators



The annotations of (a)all (b)top 50% (c)top 10% scoring annotators.

Back

A. Passive CL (app.)

B. Active CL (app.)

C. O-CBS (app.)

- The effect of S_S
- The effect of dominance suppression factor

D. O-CBS+ (app.)

E. Multivariate (app.)

O-CBS Algorithm

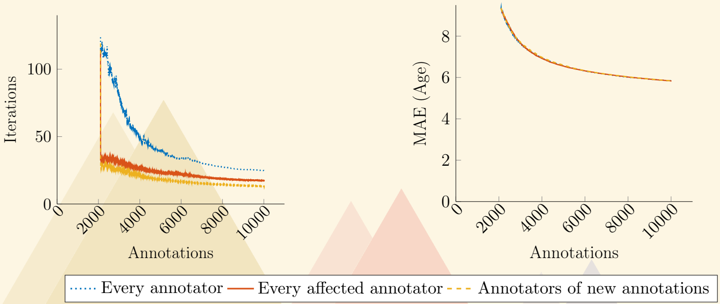
Input:
Set of all samples $\mathcal{I}$, all annotators $\mathcal{J}$, currently active annotators $\mathcal{J}'$, current annotations $\mathcal{K}$
i_k and j_k are the sample and annotator of annotation k, respectively
$S_S(\cdot)$ and $S_A(\cdot)$ are the sample consensus quality and annotator competence scoring functions, respectively
Output:
New annotation k

function RequestAnnotation($S_S(\cdot), S_A(\cdot), \mathcal{I}, \mathcal{J}, \mathcal{J}', \mathcal{K}$)
  for all $i \in \mathcal{I}$ **do**
    $\mathcal{K}_i \leftarrow \{k \in \mathcal{K} : i_k = i\}$ ▷ Annotations of sample i
    $\mathcal{J}_i \leftarrow \{j_k \in \mathcal{J} : k \in \mathcal{K}_i\}$ ▷ Annotators of sample i
  $i \leftarrow \underset{i' \in \mathcal{I} \text{ s.t. } \mathcal{J}' \setminus \mathcal{J}_{i'} \neq \emptyset}{\operatorname{argmin}} S_S(i')$ ▷ Select the sample with the worst consensus quality such that at least one of the currently active annotators has no annotations for that sample
  $j \leftarrow \underset{j' \in \mathcal{J}' \setminus \mathcal{J}_i}{\operatorname{argmax}} S_A(j')$ ▷ Select the most competent annotator from the set of active annotators who had not annotated sample i
  $k \leftarrow$ Request an annotation for sample i from annotator j
  return k

Details of the created subsets

Dataset	Subset size	Inter-Set Similarity (%) Min	Average	Max
Head Pose Annotations	1110	15.68	21.47 ± 1.17	26.13
Age Annotations	2100	18.43	21.13 ± 0.85	24.81
Affective Text Analysis	200	11	19.95 ± 2.63	29.5

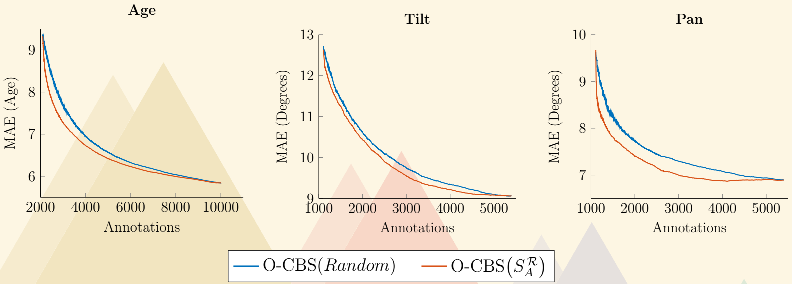
Speeding Up the Inference Process



The effect of three different random initialization approaches on the number of iterations for O-CBS$(S_A^{\mathcal{R}})$ (random annotation addition)

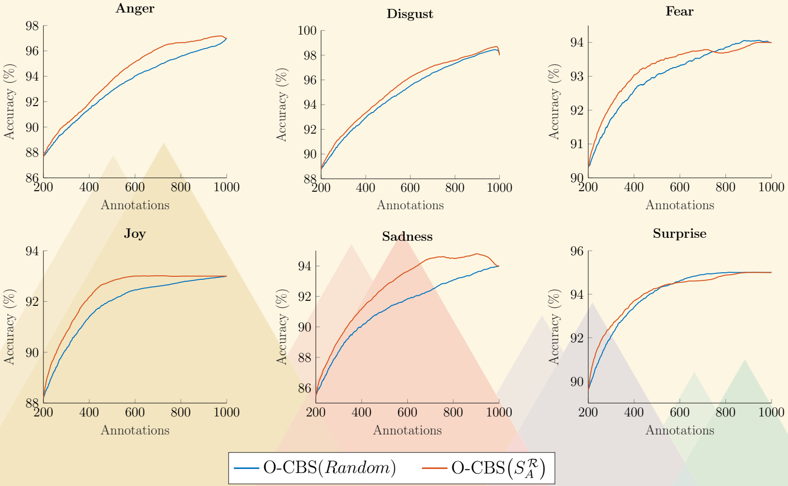
- The results are provided for the Kara Age Annotations dataset.
- Reinitializing the annotator parameters of only those providing new annotations results in much fewer iterations with the same MAE.

The effect of S_S



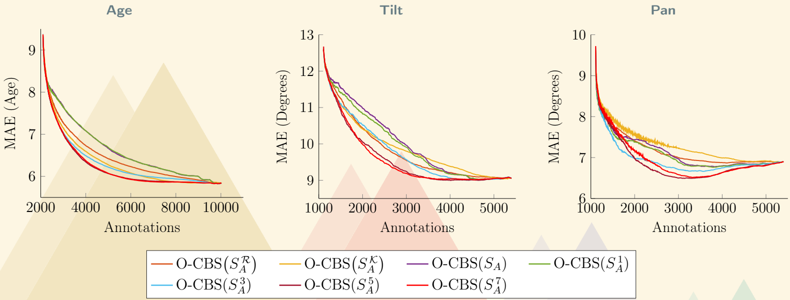
Back

The effect of S_S



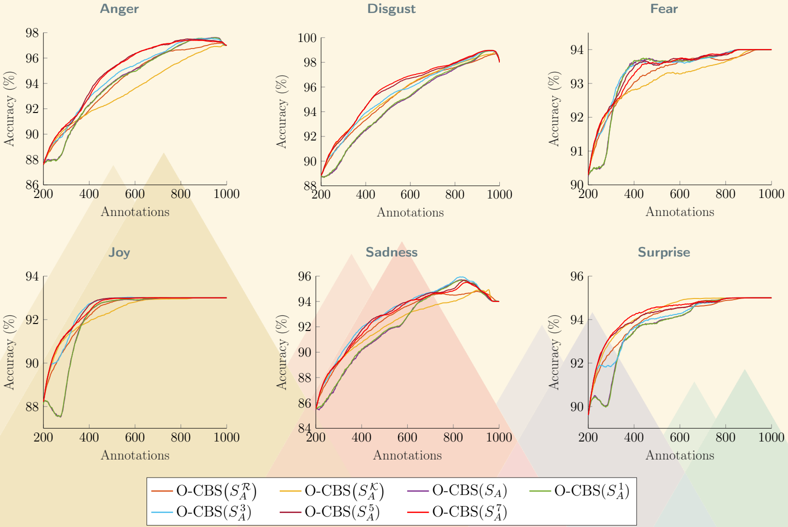
Back

The effect of dominance suppression factor



Back

The effect of dominance suppression factor



Back

A. Passive CL (app.)

B. Active CL (app.)

C. O-CBS (app.)

D. O-CBS+ (app.)

- The effect of annotator exploration parameter
- Comparative Performance of O-CBS+

E. Multivariate (app.)

Requesting annotations for smart label collection from scratch I

Input:

Set of all samples $\mathcal{I}$, all annotators $\mathcal{J}$, current annotations $\mathcal{K}$, currently active annotators $\mathcal{J}'$
i_k and j_k are the sample and the annotator of the annotation k, respectively
$S_S(i)$ is the consensus quality score of sample i, $S_A(j)$ is the competence score of annotator j
$\mathcal{E}$ defines the probability of exploring a new annotator

Output:

New annotation(s) $\{k\}$ or $\{k, k'\}$

function RequestAnnotationExp($S_S(\cdot), S_A(\cdot), \mathcal{E}, \mathcal{I}, \mathcal{J}, \mathcal{J}', \mathcal{K}$)
  for all $i \in \mathcal{I}$ **do**
    $\mathcal{K}_i \leftarrow \{k \in \mathcal{K} : i_k = i\}$ ▷ Annotations of sample i
    $\mathcal{J}_i \leftarrow \{j_k \in \mathcal{J} : k \in \mathcal{K}_i\}$ ▷ Annotators of sample i
  end for
  for all $j \in \mathcal{J}$ **do**
    $\mathcal{K}^j \leftarrow \{k \in \mathcal{K} : j_k = j\}$ ▷ Annotations of annotator j
  end for
  $\mathcal{U}_s \leftarrow \{i \in \mathcal{I} : |\mathcal{K}_i| = 0\}$ ▷ Samples without any annotation
  $\mathcal{U}_a \leftarrow \{j \in \mathcal{J} : |\mathcal{K}^j| = 0\}$ ▷ Annotators without any annotation
  if $|\mathcal{U}_s| > 0$ **then** ▷ If there is a sample without any annotation
    $i \leftarrow$ Randomly select from $\mathcal{U}_s$
  else
    $i \leftarrow \underset{i' \in \mathcal{I} \text{ s.t. } \mathcal{J}' \setminus \mathcal{J}_{i'} \neq \emptyset}{\operatorname{argmin}} S_S(i')$ ▷ Select the sample with the worst consensus quality such that at least one of the currently active annotators has no annotations for that sample

◂ Back

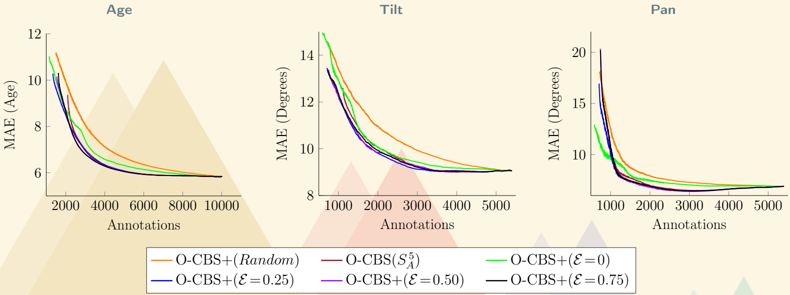
Requesting annotations for smart label collection from scratch II

end if
$\mathcal{R} \leftarrow \mathcal{U}_a \cap \mathcal{J}'$ ▷ Set of explorable annotators
$\mathcal{T} \leftarrow \mathcal{J}' \setminus (\mathcal{J}_i \cup \mathcal{U}_a)$ ▷ Set of exploitable annotators
if $|\mathcal{R}| > 0$ **and** $|\mathcal{T}| > 0$ **then** ▷ If there are both explorable and exploitable annotators
  explore $\leftarrow$ true with probability $\mathcal{E}$ ▷ Randomly decide whether to explore a new annotator or exploit an existing annotator
else if $|\mathcal{R}| > 0$ **then** ▷ If there are only explorable annotators
  explore $\leftarrow$ true
else if $|\mathcal{T}| > 0$ **then** ▷ If there are only exploitable annotators
  explore $\leftarrow$ false
end if
if explore **then**
  $j \leftarrow$ Randomly select from $\mathcal{R}$ ▷ Select an annotator from explorable annotators
  $i' \leftarrow$ Randomly select from $\mathcal{I} \setminus \mathcal{U}_s$ ▷ Select a sample from previously annotated samples
  $k' \leftarrow$ Request an annotation for a random sample i' from annotator j
else
  $j \leftarrow \underset{j' \in \mathcal{J}' \setminus \mathcal{J}_i}{\operatorname{argmax}} \; S_A(j')$ ▷ Select the most competent annotator from the set of active annotators who had not annotated sample i
end if
$k \leftarrow$ Request an annotation for the sample i from annotator j
if explore **then**
  return $\{k, k'\}$

Requesting annotations for smart label collection from scratch III

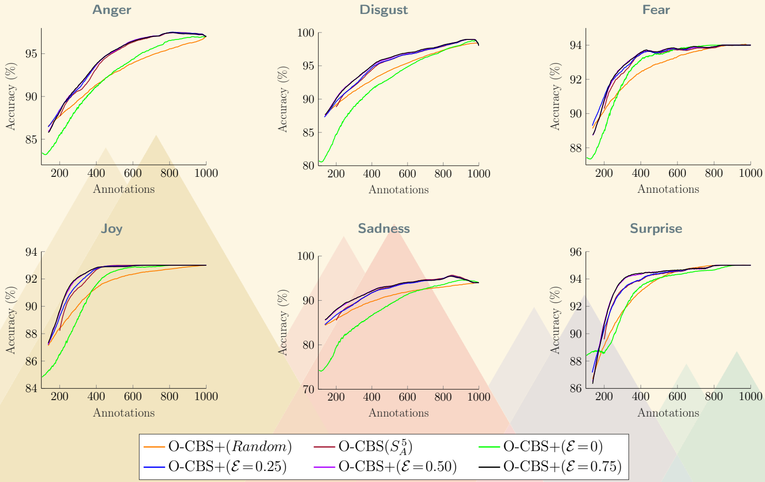
```
    else
        return {k}
    end if
end function
```

The effect of annotator exploration parameter



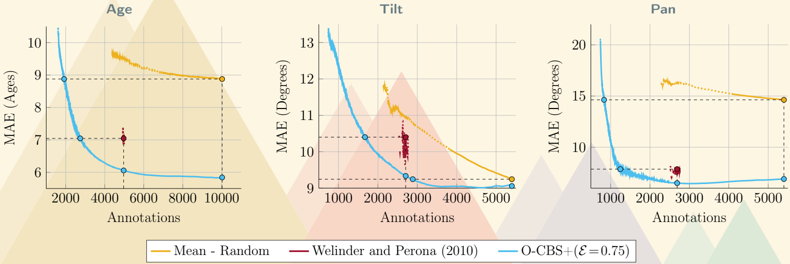
Back

The effect of annotator exploration parameter



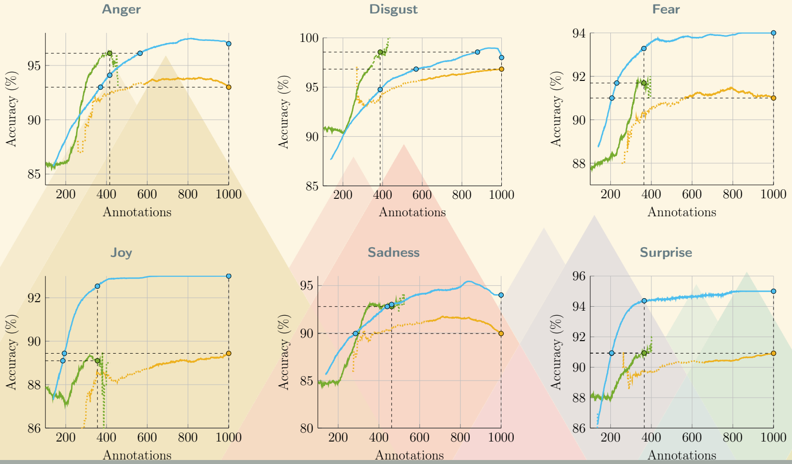
Back

The effect of enforcing annotation count or MAE limit



Back

The effect of enforcing annotation count or accuracy limit



Back

The effect of enforcing annotation count or MAE/accuracy limit

	Welinder and Perona[7]		O-CBS+($\mathcal{E} = 0.75$)	
Dataset	Annotations	MAE	MAE at target annotations	Required annotations to reach target MAE
Age	4969.77	7.02 ages	6.06 ages	2775.98
Tilt	2705.03	10.10 degrees	9.33 degrees	1892.16
Pan	2689.77	7.58 degrees	6.49 degrees	1387.88

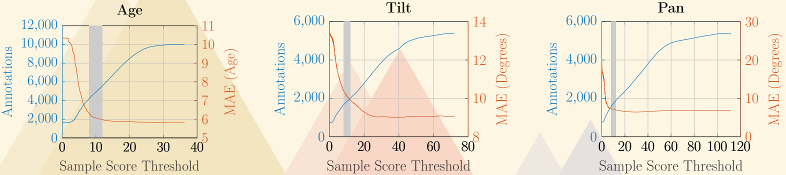
	Raykar and Agrawal[8]		O-CBS+($\mathcal{E} = 0.75$)	
Dataset	Annotations	MAE	MAE at target annotations	Required annotations to reach target MAE
Anger	415.86	96.07	94.11	535.81
Disgust	387.78	98.92	94.76	726.82
Fear	363.49	91.50	93.28	247.32
Joy	355.51	89.17	92.53	196.22
Sadness	462.34	93.31	93.01	522.80
Surprise	365.22	91.60	94.38	231.41

[7] Welinder, P. and Perona, P. (2010). Online Crowdsourcing: Rating Annotators and Obtaining Cost-Effective Labels. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, 25–32.

[8] Raykar, V. C. and Agrawal, P. (2014). Sequential Crowdsourced Labeling as an Epsilon-Greedy Exploration in a Markov Decision Process. *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics (AISTATS-14)*, *33*, 832–840.

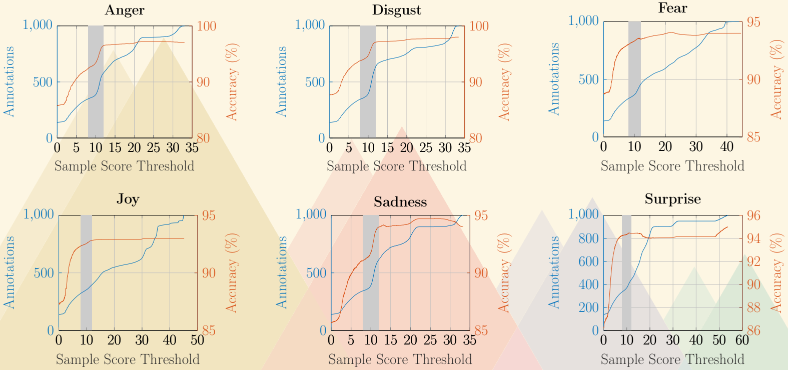
Back

The effect of enforcing sample consensus quality induced stopping criteria (1)



◂ Back

The effect of enforcing sample consensus quality induced stopping criteria (2)



Back

The effect of enforcing sample consensus quality induced stopping criteria (3)

	Welinder and Perona[9]		O-CBS+($\mathcal{E}=0.75$)					
			$\tau=8$		$\tau=10$		$\tau=12$	
Dataset	Ann.	MAE	Ann.	MAE	Ann.	MAE	Ann.	MAE
Age	4969.77	7.02	4189.93	6.33	4911.37	6.07	5607.13	5.97
Tilt	2705.03	10.10	1657.70	10.42	1836.39	10.11	2009.94	9.92
Pan	2689.77	7.58	1560.16	7.32	1721.22	7.13	1868.02	7.01

	Raykar and Agrawal[10]		O-CBS+($\mathcal{E}=0.75$)					
			$\tau=8$		$\tau=10$		$\tau=12$	
Dataset	Ann.	Acc.(%)	Ann.	Acc.(%)	Ann.	Acc.(%)	Ann.	Acc.(%)
Anger	415.86	96.07	347.83	93.38	386.20	94.58	564.59	97.24
Disgust	387.78	98.92	346.12	94.64	392.72	95.53	625.24	97.41
Fear	363.49	91.50	331.49	93.45	365.74	93.77	458.29	93.74
Joy	355.51	89.17	323.10	92.59	352.96	92.79	394.22	92.98
Sadness	462.34	93.31	343.58	91.96	390.84	92.72	603.89	94.50
Surprise	365.22	91.60	334.87	94.60	371.00	94.67	447.00	94.64

[9] Welinder, P. and Perona, P. (2010). Online Crowdsourcing: Rating Annotators and Obtaining Cost-Effective Labels. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, 25–32.

[10] Raykar, V. C. and Agrawal, P. (2014). Sequential Crowdsourced Labeling as an Epsilon-Greedy Exploration in a Markov Decision Process. *Proceedings of the Seventeenth International Conference on*

Back

A. Passive CL (app.)

B. Active CL (app.)

C. O-CBS (app.)

D. O-CBS+ (app.)

E. Multivariate (app.)

- Variational Bayes
- Experiments

Variational Bayes

$$\boldsymbol{V_j}^{-1} = \boldsymbol{V_0}^{-1} + \sum_{k:j_k=j} \left[\begin{array}{c:c} \boldsymbol{\Sigma_{i_k}} + \boldsymbol{\mu_{i_k}}\boldsymbol{\mu_{i_k}}^\top & \boldsymbol{\mu_{i_k}} \\ \hdashline \boldsymbol{\mu_{i_k}}^\top & 1 \end{array}\right] \quad \boldsymbol{M_j} = \left(\sum_{c=1}^{C} \rho_{jc}\boldsymbol{M_c}\boldsymbol{V_0}^{-1} + \sum_{k:j_k=j} \boldsymbol{y_k}\left[\begin{array}{c:c} \boldsymbol{\mu_{i_k}}^\top & 1\end{array}\right]\right)\boldsymbol{V_j}$$

$$\boldsymbol{W_j}^{-1} = \boldsymbol{W_0}^{-1} + \sum_{c=1}^{C}\rho_{jc}(\boldsymbol{M_c}-\boldsymbol{M_{j_z}})\boldsymbol{V_0}^{-1}(\boldsymbol{M_c}-\boldsymbol{M_{j_z}})^\top + (\boldsymbol{M_j}-\boldsymbol{M_{j_z}})\boldsymbol{V_0}^{-1}(\boldsymbol{M_j}-\boldsymbol{M_{j_z}})^\top$$
$$+ \sum_{k:j_k=j}\mathbb{E}\left[(\boldsymbol{y_k}-\boldsymbol{M_j}\boldsymbol{\chi_{i_k}})(\boldsymbol{y_k}-\boldsymbol{M_j}\boldsymbol{\chi_{i_k}})^\top\right]$$

$$\boldsymbol{\Sigma_i}^{-1} = \left[\begin{array}{c:c}\boldsymbol{I_d} & \boldsymbol{O_{d1}}\end{array}\right]\sum_{k:i_k=i}\left(d\boldsymbol{V_{j_k}} + (n_0+N_{j_k})\boldsymbol{M_{j_k}}^\top\boldsymbol{W_{j_k}}\boldsymbol{M_{j_k}}\right)\left[\begin{array}{c}\boldsymbol{I_d}\\ \hdashline \boldsymbol{O_{1d}}\end{array}\right]$$

$$\boldsymbol{\mu_i} = \boldsymbol{\Sigma_i}\left[\begin{array}{c:c}\boldsymbol{I_d} & \boldsymbol{O_{d1}}\end{array}\right]\sum_{k:i_k=i}\left((n_0+N_{j_k})\boldsymbol{M_{j_k}}^\top\boldsymbol{W_{j_k}}\left(\boldsymbol{y_k}-\boldsymbol{M_{j_k}}\left[\begin{array}{c}\boldsymbol{O_{d1}}\\ \hdashline 1\end{array}\right]\right) - d\boldsymbol{V_{j_k}}\left[\begin{array}{c}\boldsymbol{O_{d1}}\\ \hdashline 1\end{array}\right]\right)$$

$$\varrho_{jc} = \exp\left(\log \boldsymbol{p_c} + \frac{d+1}{2}\psi_d\left(\frac{n_0+N_j}{2}\right) + \frac{d+1}{2}\log|\boldsymbol{V_j}| - \frac{d(d+1)}{2}\log(\pi)\right.$$
$$-\frac{d}{2}\operatorname{Tr}\left(\boldsymbol{V_0}^{-1}\boldsymbol{V_j}\right) - \frac{(n_0+N_j)}{2}\operatorname{Tr}\left(\boldsymbol{V_0}^{-1}\boldsymbol{M_j}^\top\boldsymbol{W_j}\boldsymbol{M_j}\right) - \frac{d}{2}\log|\boldsymbol{V_0}|$$
$$\left.+(n_0+N_j)\operatorname{Tr}\left(\boldsymbol{V_0}^{-1}\boldsymbol{M_j}^\top\boldsymbol{W_j}\boldsymbol{M_c}\right) - \frac{(n_0+N_j)}{2}\operatorname{Tr}\left(\boldsymbol{V_0}^{-1}\boldsymbol{M_c}^\top\boldsymbol{W_j}\boldsymbol{M_c}\right)\right)$$

Back

- Incorporate adverseness behavior through $\boldsymbol{M_c}$ matrices

$$\boldsymbol{M_c} = \begin{bmatrix} \ddots & & 0 & 0 \\ & \boldsymbol{m}_c & & \vdots \\ 0 & & \ddots & 0 \end{bmatrix} \text{ where } \boldsymbol{m}_c \in \{-1, 1\}^d.$$

- No prior knowledge about the annotator behaviors occurring in the dataset: Flat prior over $\boldsymbol{p}$.
- $\boldsymbol{V_0} = 10^{-4}\boldsymbol{I_{d+1}}$, $\boldsymbol{W_0} = 10^4\boldsymbol{I_d}$, and $n_0 = 2$ for encouraging $|\boldsymbol{\Lambda_j}|$ to be large and assisting $\boldsymbol{\Phi_j}$ to somewhat resemble its mean $\boldsymbol{M_c}$.

Back

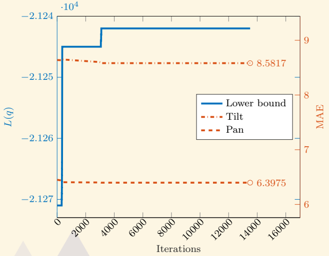
Experiments

$$\log p(Y) = \mathcal{L}(q) + \mathsf{KL}(q\|p)$$

where

$$\mathcal{L}(q) = \sum_Z \iiint q(\Phi, \Lambda, Z, X) \log \frac{p(Y, \Phi, \Lambda, Z, X)}{q(\Phi, \Lambda, Z, X)} d\Phi d\Lambda dX,$$

$$\mathsf{KL}(q\|p) = -\sum_Z \iiint q(\Phi, \Lambda, Z, X) \log \frac{p(\Phi, \Lambda, Z, X|Y)}{q(\Phi, \Lambda, Z, X)} d\Phi d\Lambda dX.$$



Change of the lower bound value ($L(q)$) and attribute error while fitting the model

◂ Back

References I

Aran, O. & Gatica-Perez, D. (2013). One of a Kind: Inferring Personality Impressions in Meetings. In *Proceedings of the 15th acm on international conference on multimodal interaction* (pp. 11–18). ACM.

Bi, J. & Wang, X. (2013). Min-Max Optimization for Multiple Instance Learning from Multiple Data Annotators. *KDD'13 August*.

Carpenter, B. (2008). Multilevel Bayesian Models of Categorical Data Annotation. *Unpublished manuscript*.

Dawid, A. P. & Skene, A. M. (1979). Maximum likelihood estimation of observer error-rates using the em algorithm. *Applied statistics*, 20–28.

Donmez, P. & Carbonell, J. G. (2008a). Paired-Sampling in Density-Sensitive Active Learning. In *The international symposium on artificial intelligence and mathematics*.

Donmez, P. & Carbonell, J. G. (2008b). Proactive Learning : Cost-Sensitive Active Learning with Multiple Imperfect Oracles. In *Proceedings of the 17th acm conference on information and knowledge management* (pp. 619–628).

Donmez, P., Carbonell, J. G., & Schneider, J. (2009). Efficiently learning the accuracy of labeling sources for selective sampling. *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '09*, 259–268.

Fang, M., Yin, J., & Tao, D. (2014). Active Learning for Crowdsourcing Using Knowledge Transfer. *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, 1809–1815.

References II

Gao, J., Liu, X., Ooi, B. C., Wang, H., & Chen, G. (2013). An Online Cost Sensitive Decision-Making Method in Crowdsourcing Systems. *Proceedings of the 2013 international conference on Management of data - SIGMOD '13*, 217–228.

Gourier, N., Hall, D., & Crowley, J. L. (2004). Estimating Face Orientation From Robust Detection of Salient Facial Structures. In *Fg net workshop on visual observation of deictic gestures* (pp. 1–9). FGnet (IST–2000–26434) Cambridge, UK.

Guo, S., Parameswaran, A., & Garcia-Molina, H. (2012). So Who Won?: Dynamic Max Discovery with the Crowd. In *Proceedings of the 2012 acm sigmod international conference on management of data* (pp. 385–396). ACM.

Ho, C.-J., Slivkins, A., & Vaughan, J. W. (2016). Adaptive Contract Design for Crowdsourcing Markets: Bandit Algorithms for Repeated Principal-Agent Problems. *Journal of Artificial Intelligence Research*, *55*, 317–359.

Hsueh, P.-Y., Melville, P., & Sindhwani, V. (2009). Data Quality from Crowdsourcing: A Study of Annotation Selection Criteria. In *Proceedings of the naacl hlt 2009 workshop on active learning for natural language processing* (June, pp. 27–35).

Kamar, E., Hacker, S., & Horvitz, E. (2012). Combining Human and Machine Intelligence in Large-Scale Crowdsourcing. *AAMAS '12 Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems - Volume 1*, 467–474.

Kamar, E., Kapoor, A., & Horvitz, E. (2013). Lifelong Learning for Acquiring the Wisdom of the Crowd. *International Joint Conference on Artificial Intelligence (IJCAI)*, *13*, 2313–2320.

References III

Kamar, E., Kapoor, A., & Horvitz, E. (2015). Identifying and Accounting for Task-Dependent Bias in Crowdsourcing. In *Proceedings, the third aaai conference on human computation and crowdsourcing (hcomp-15)* (pp. 92–101).

Kara, Y. E., Genc, G., Aran, O., & Akarun, L. (2015). Modeling Annotator Behaviors for Crowd Labeling. *Neurocomputing*, *160*, 141–156.

Kara, Y. E., Genc, G., Aran, O., & Akarun, L. (2018). Actively estimating crowd annotation consensus. *Journal of Artificial Intelligence Research*, *61*, 363–405.

Lakshminarayanan, B. & Teh, Y. (2013). Inferring Ground Truth From Multi-Annotator Ordinal Data: A Probabilistic Approach. *arXiv preprint arXiv:1305.0015*, 1–19. arXiv: arXiv:1305.0015v1

Lin, C. H., Mausam, & Weld, D. S. (2016). Re-active Learning : Active Learning with Relabeling. *Thirtieth AAAI Conference on Artificial Intelligence*, 1845–1852.

Marcus, A., Karger, D., Madden, S., Miller, R., & Oh, S. (2012). Counting with the Crowd. *Proceedings of the VLDB Endowment*, *6*(2), 109–120.

Mozafari, B., Sarkar, P., Franklin, M., Jordan, M., & Madden, S. (2014). Scaling Up Crowd-Sourcing to Very Large Datasets: A Case for Active Learning. *Proceedings of the VLDB Endowment*, *8*(2), 125–136.

Nguyen, A. T., Wallace, B. C., & Lease, M. (2015). Combining Crowd and Expert Labels using Decision Theoretic Active Learning. In *Proceedings of the 3rd aaai conference on human computation (hcomp)* (pp. 120–129).

Ok, J., Oh, S., Shin, J., Jang, Y., & Yi, Y. (2017). Iterative Bayesian Learning for Crowdsourced Regression. *arXiv:1702.08840 [cs.LG]*, 1–22. arXiv: 1702.08840

References IV

Peng, J., Liu, Q., Ihler, A., & Berger, B. (2013). Crowdsourcing for Structured Labeling With Applications to Protein Folding. *ICML Workshop on Machine Learning Meets Crowdsourcing*, 2008–2012.

Raykar, V. & Yu, S. (2012). Eliminating Spammers and Ranking Annotators for Crowdsourced Labeling Tasks. *Journal of Machine Learning Research*, *13*, 491–518.

Raykar, V. C. & Agrawal, P. (2014). Sequential Crowdsourced Labeling as an Epsilon-Greedy Exploration in a Markov Decision Process. *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics (AISTATS-14)*, *33*, 832–840.

Raykar, V. C. & Yu, S. (2011). Annotation Models for Crowdsourced Ordinal Data. *NIPS Workshop on Computational Social Science and the Wisdom of Crowds*.

Raykar, V. C., Yu, S., Zhao, L. H., Valadez, G. H., Florin, C., Bogoni, L., & Moy, L. (2010). Learning from crowds. *Journal of Machine Learning Research*, *99*, 1297–1322.

Rodrigues, F., Pereira, F., & Ribeiro, B. (2013). Learning from Multiple Annotators: Distinguishing Good from Random Labelers. *Pattern Recognition Letters*, *34*(12), 1428–1436.

Sheng, V. S., Provost, F., & Ipeirotis, P. G. (2008). Get Another Label? Improving Data Quality and Data Mining Using Multiple, Noisy Labelers. In *Proceedings of the 14th acm sigkdd international conference on knowledge discovery and data mining* (pp. 614–622).

Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. Y. (2008). 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* (pp. 254–263). Association for Computational Linguistics.

The FGNet Aging Database. (n.d.).

References V

Tran-Thanh, L., Huynh, T. D., Rosenfeld, A., Ramchurn, S., & Jennings, N. R. (2014). BudgetFix : Budget Limited Crowdsourcing for Interdependent Task Allocation With Quality Guarantees. *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-agent Systems*, 477–484.

Tran-Thanh, L., Venanzi, M., Rogers, A., & Jennings, N. R. (2013). Efficient Budget Allocation With Accuracy Guarantees for Crowdsourcing Classification Tasks. In *Proceedings of the 12th international conference on autonomous agents and multiagent systems* (pp. 6–10).

Venanzi, M., Guiver, J., Kohli, P., & Jennings, N. R. (2016). Time-sensitive bayesian information aggregation for crowdsourcing systems. *Journal of Artificial Intelligence Research*, *56*, 517–545.

Welinder, P., Branson, S., Belongie, S., & Perona, P. (2010). The Multidimensional Wisdom of Crowds. *Advances in Neural Information Processing Systems*, *23*, 2424–2432.

Welinder, P. & Perona, P. (2010). Online Crowdsourcing: Rating Annotators and Obtaining Cost-Effective Labels. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, 25–32.

Yan, Y., Rosales, R., Fung, G., & Dy, J. G. (2011). Active Learning from Crowds. In *Proceedings of the 28th international conference on machine learning* (pp. 1161–1168).

Yan, Y., Rosales, R., Fung, G., Schmidt, M., Hermosillo, G., Bogoni, L., . . . Dy, J. (2010). Modeling Annotator Expertise: Learning When Everybody Knows a Bit of Something. *International Conference on Artificial Intelligence and Statistics*, *9*, 932–939.

Zhang, P. & Obradovic, Z. (2012). Integration of Multiple Annotators by Aggregating Experts and Filtering Novices. *2012 IEEE International Conference on Bioinformatics and Biomedicine*, 1–6.

References VI

Zhang, Q., Wen, Y., Tian, X., Gan, X., & Wang, X. (2015). Incentivize Crowd Labeling Under Budget Constraint. *2015 IEEE Conference on Computer Communications (INFOCOM)*, 2812–2820.

Zhu, C., Xu, H., & Yan, S. (2015). Online Crowdsourcing. *arXiv:1512.02393 [cs]*. arXiv: arXiv:1512.02393v1

Zhuang, H. & Young, J. (2015). Leveraging In-Batch Annotation Bias for Crowdsourced Active Learning. *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (WSDM'15)*, 243–252.