Generating Workers' Label Sets from Ground Truth

Step 1: **Probability Thresholds $\pi_a^{(k)}$**

To synthesize a multi-worker dataset from ground truth, we first use $U(0.4, 1)$ to obtain $\mathrm{M} \times \mathrm{K}$ probability thresholds.

$$\Pi = \left\{ \left\{ \pi_a^{(k)} \sim U(0.4, 1) \right\}_{k=1}^{K} \right\}_{a=1}^{M}$$

Step 2: **Generate Random $\rho^{(i)}$**

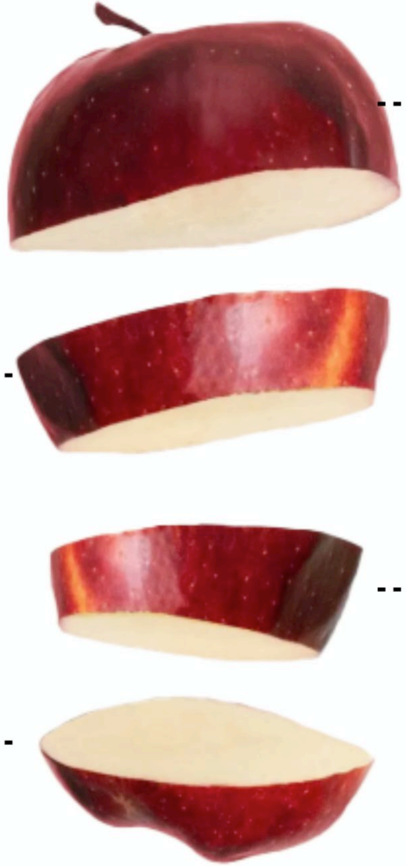
To generate the labels for each worker α, a random number $0 < \rho^{(i)} < 1$ is generated for each instance i in the dataset

Step 3: **Assigning Labels**

We assign the respective true label with probability $\pi_a^{(k)}$ and the opposite label with probability $\left(1 - \pi_a^{(k)}\right)$.

$$z_\alpha^{(i,k)} = \begin{cases} y^{(i,k)} & \text{if} \rho^{(i)} \leq \pi_\alpha^{(k)} \\ 1 - y^{(i,k)} & \text{if} \rho^{(i)} > \pi_\alpha^{(k)} \end{cases}$$

Weight Calculation Steps



4
Confidence Score Measurement

Calculating the confidence score corresponding to the aggregated label

3
Label Aggregation

Calculating the aggregated label

2
Weight Calculation

Calculating the weight corresponding to each annotator

1
Consistency Measurement

Calculating the consistency score corresponding to each annotator a

Uncertainty Measurement Techniques

$$-\sum_{g} p_{\alpha}^{(i,k),(g)} \log\left(p_{\alpha}^{(i,k),(g)}\right)$$

Entropy

$$\sqrt{\frac{1}{G-1}\sum_{g=1}^{G}\left(t_{\alpha}^{(i,k),(g)}-\mu\right)^{2}}$$

Standard Deviation

Monte Carlo Dropout

Committee-Based Methods

$$\frac{1}{G-1}\sum_{g=1}^{G}\left(p_{\alpha}^{(i,k),(g)}-\mu\right)^{2}$$

Predictive Interval

$$P\left(Q_L^k \le p_{\alpha}^{(i,k),(g)} \le Q_U^k\right) = \gamma$$

$$\Delta_{\alpha}^{(i,k)} = Q_L^k - Q_U^k$$

Conformal Prediction

- The reliability score is defined by averaging over consistency scores over all instances.

$$\psi_a^{(k)} = \frac{1}{N}\sum_{i=1}^{N} c_a^{(i,k)}$$

Weight Measurement

- Weight corresponding to annotator α will be as follows.

$$\omega_a^{(k)} = \frac{\psi_a^{(k)}}{\sum_{a=1}^{M} \psi_a^{(k)}}$$

Metrics

F1 Score

$$2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Accuracy

$$\frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} \delta \left(\nu^{(i,k)}, y^{(i,k)} \right)$$

AUC

Expected Calibration Error (ECE)

$$\sum_{b=1}^{B} \frac{|B_b|}{N} \left|\text{Accuracy}(B_b) - \text{Confidence-Score}(B_b)\right|$$

Brier Score

$$\frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} \left(F^{(i,k)} - y^{(i,k)} \right)^2$$

Weight Measurement Evaluation

Two strategies for classifier selection were investigated (Using different classifiers, vs. RF under different random seeds)

There was no considerable variations in the results.

Second strategy was adopted for faster processing time

For each worker , 10 RF were trained, each comprising 4 trees with a maximum depth of 4, under various random states.