Crowdsourcing with Contextual Uncertainty

Viet-An Nguyen, Peipei Shi, Jagdish Ramakrishnan, Narjes Torabi, Nimar S. Arora, Udi Weinsberg, Michael Tingley

Meta

KDD 2022
Prevalence of violations

• Prevalence of violations: the rate at which policy violations occur
Estimating the prevalence

Population → Samples → Human labels → Prevalence Estimates

Sample → Send for human review → Estimate
### Two problems in practice

<table>
<thead>
<tr>
<th>Low prevalence</th>
<th>Labeling mistakes</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Uniform sampling will result in too few violating examples</td>
<td>• Human labelers make mistakes</td>
</tr>
<tr>
<td>• Solution:</td>
<td>• Solution:</td>
</tr>
<tr>
<td>• Sampling non-uniformly based on item’s features (i.e., context) which correlate with the likelihood of being a violation</td>
<td>• Send each item to multiple labelers and use statistical models to infer the true labels based on the observed labels</td>
</tr>
<tr>
<td>• In practice, there is a classifier which converts items’ features into a single score</td>
<td></td>
</tr>
</tbody>
</table>
Modeling crowdsourced labels

• The “classic” Dawid-Skene (D&S) model

Crowdsourced data
Modeling crowdsourced labels

- The "classic" Dawid-Skene (D&S) model

Per-labeler

**Sensitivity (TPR)** and **Specificity (TNR)**
for binary label

Crowdsourced data

Overall Prevalence
Theodon

Population

Samples

Human labels

- Sampling weights based on classifier scores
- Collecting multiple labels per item

Send for human review

Estimate

Theodon

[Graph showing prevalence, sensitivity, and specificity curves for Theodon scores]
Generative process: D&S model

- Overall Prevalence

- Sensitivity
- Specificity

- Beta prior

- Observed labels
- Observed labelers

- Item True Label

- Sensitivity: 0.95, 0.05
- Specificity: 0.2, 0.8

- Observed labels: 0.9, 0.1
- Overall: 0.9, 0.1

- Beta prior
Generative process: D&S model

Overall Prevalence

Item True Label

Classifier score

Observed labelers

Observed labels

Score

Prevalence

Sensitivity

Specificity

Score

Beta prior

Score

Prevalence

Sensitivity

Specificity

Observation labels

Model parameters

Data labels

Score

Beta prior
Generative process: Theodon

Gaussian process prior

Overall Prevalence

Score

Prevalence

Gaussian process prior

Item True Label

Classifier score

Observed labelers

Observed labels

Score

Sensitivity

Specificity

Score

Score

Score

Score

Score
Gaussian processes (GPs)

A GP is a stochastic process which defines a probability distribution over the function

\[ p(f \mid x) = \mathcal{GP}(m(x), K(x, x')) \]

- \( m(x) \): the mean function
- \( K(x, x') \): the covariance function

\[ K_{\alpha, \rho}(x, x') = \alpha^2 \exp\left(-\frac{(x - x')^2}{2\rho^2}\right) \]
## Related models

<table>
<thead>
<tr>
<th>Baseline</th>
<th>Prevalence</th>
<th>Sensitivity &amp; Specificity</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>FL-FL</td>
<td>Flat</td>
<td>Flat</td>
<td>Dawid &amp; Skene (D&amp;S) (1979)</td>
</tr>
<tr>
<td>LR-FL</td>
<td>Logistic Regression</td>
<td>Flat</td>
<td>Raykar et al. (ICML 2009, JMLR 2010)</td>
</tr>
<tr>
<td>GP-FL</td>
<td>Gaussian Process</td>
<td>Flat</td>
<td>Rodrigues et al. (ICML 2014)</td>
</tr>
<tr>
<td>LR-LR</td>
<td>Logistic Regression</td>
<td>Logistic Regression</td>
<td>Yan et al. (AISTATS 2010, MLJ 2014)</td>
</tr>
<tr>
<td>Theodon</td>
<td>Gaussian Process</td>
<td>Gaussian Process</td>
<td>Our work</td>
</tr>
</tbody>
</table>
Deployment at Meta
Empirical results

Applications
• Prevalence measurement
• Labeler performance measurement
• Classifier calibration
• Per-item label aggregation

Datasets
• Data generated from crowdsourcing applications at Meta
• Public crowdsourcing datasets: Music and Sentiment
Crowdsourced data for prevalence measurement at Meta

• Simulate data based on crowdsourcing applications at Meta
Experimental setup

• Tasks
  • Prevalence measurement: estimating the prevalence function
  • Labeler performance measurement: estimating the sensitivity and specificity functions for each labeler

• Evaluation
  • Comparing the function estimate (mean $\hat{\theta}$ with 95%-CI $[\hat{\theta}_L, \hat{\theta}_U]$) with the true function $\theta^*$

Mean absolute error (MAE) of the mean $\hat{\theta}$:

$$\frac{1}{|S|} \sum_{s \in S} |\hat{\theta}(s) - \theta^*(s)|$$

Coverage rate of the confidence interval (CI):

$$\frac{1}{|S|} \sum_{s \in S} 1 \{\theta^*(s) \in [\hat{\theta}_L(s), \hat{\theta}_U(s)]\}$$
Results: prevalence & labeler performance measurement

Theodon consistently provides low Mean Absolute Error (MAE) while achieving high coverage rate compared to other baselines.
Public crowdsourcing datasets

• Two public datasets by Rodrigues et al.: *Music* and *Sentiment*
  • Each item has both *crowdsourced labels* and a *ground truth label*

• Label aggregation
  • Goal: inferring the ground truth label from crowdsourced labels
  • Metric: Area under the PR curve (AUC-PR)

• Classifier calibration
  • Goal: transforming the raw classifier scores into the true correctness probabilities using crowdsourced labels
  • Metric: Expected calibration error (ECE)
### Results: label aggregation & classifier calibration

<table>
<thead>
<tr>
<th>Metric</th>
<th>l</th>
<th>Base</th>
<th>MV</th>
<th>FV</th>
<th>Snorkel</th>
<th>Fl-Fl</th>
<th>Lr-Fl</th>
<th>Gp-Fl</th>
<th>Lr-Lr</th>
<th>Theodon</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sentiment</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ECE</td>
<td>3</td>
<td>0.2304</td>
<td>0.1360</td>
<td>0.0754</td>
<td>0.0989</td>
<td>0.0952</td>
<td>0.0855</td>
<td>0.0814</td>
<td>0.0729</td>
<td><strong>0.0661</strong></td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>0.2394</td>
<td>0.1116</td>
<td>0.0887</td>
<td>0.0765</td>
<td>0.0842</td>
<td>0.0725</td>
<td>0.0726</td>
<td>0.0704*</td>
<td><strong>0.0656</strong></td>
</tr>
<tr>
<td>AUC-PR</td>
<td>3</td>
<td>0.9210</td>
<td>0.8931</td>
<td>0.9169</td>
<td>0.9440</td>
<td>0.9283</td>
<td>0.9578</td>
<td>0.9604*</td>
<td>0.9641*</td>
<td><strong>0.9649</strong></td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>0.9345</td>
<td>0.9151</td>
<td>0.9498</td>
<td>0.9694</td>
<td>0.9536</td>
<td>0.9701</td>
<td>0.9716</td>
<td>0.9718</td>
<td><strong>0.9771</strong></td>
</tr>
<tr>
<td><strong>Music</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ECE</td>
<td>3</td>
<td>0.1816</td>
<td>0.0580</td>
<td>0.0514</td>
<td>0.0649</td>
<td>0.0487</td>
<td>0.0424*</td>
<td><strong>0.0403</strong></td>
<td>0.0470</td>
<td>0.0413*</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>0.1835</td>
<td>0.0570</td>
<td>0.0635</td>
<td>0.0582</td>
<td>0.0448</td>
<td>0.0420*</td>
<td><strong>0.0413</strong></td>
<td>0.0443</td>
<td>0.0423*</td>
</tr>
<tr>
<td>AUC-PR</td>
<td>3</td>
<td>0.7245</td>
<td>0.7139</td>
<td>0.7094</td>
<td>0.7808</td>
<td>0.7719</td>
<td>0.8276</td>
<td>0.8513*</td>
<td>0.8276</td>
<td><strong>0.8619</strong></td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>0.7660</td>
<td>0.7261</td>
<td>0.7399</td>
<td>0.8132</td>
<td>0.7906</td>
<td>0.8378*</td>
<td>0.8474*</td>
<td>0.8329*</td>
<td><strong>0.8515</strong></td>
</tr>
</tbody>
</table>
Conclusion

• Theodon: a system developed and deployed at Meta to model crowdsourced labels by capturing the dependencies of label’s prevalence and labelers’ performance on the input classifier score using Gaussian Processes

• Extensive empirical results on Meta’s and public datasets
Thank you!