Ordinal Programmatic Weak Supervision and Crowdsourcing for Estimating Cognitive States (Student Abstract)

Prakruthi Pradeep[1], Benedikt Boecking[1], Nicholas Gisolfi[1], Jacob R. Kintz[2], Torin K. Clark[2], Artur Dubrawski[1]

[1]Carnegie Mellon University
[2] The University of Colorado Boulder
prakruth@andrew.cmu.edu

Abstract

Crowdsourcing and weak supervision offer methods to efficiently label large datasets. Our work builds on existing weak supervision models to accommodate ordinal target classes, in an effort to recover ground truth from weak, external labels. We define a parameterized factor function and show that our approach improves over other baselines.

Introduction

To keep up with the labeled data demand of state-of-the-art machine learning models, crowdsourcing and weak supervision offer strategies to magnify the efficacy of humans tasked with annotating data. Existing methods show promise in classification tasks, but most methods tend to assume a non-ordinal nature of target classes (Ratner et al. 2016). Our work extends current weak supervision models to accommodate ordinal labeling, unlocking a new set of real-world application contexts that stand to benefit.

Our application context involves human-autonomy teaming in space habitats, where the goal is for AI and humans to cooperate to achieve mission goals. For successful teaming, it is important to consider how changes of human cognitive states (e.g., situation awareness) impact mission success, and how such states might be tracked. Specifically, we aim to predict cognitive states of astronauts in training, such as trust, mental workload, and situation awareness (TWSA), which are measured on ordinal scales. The ability to predict cognitive states from unobtrusive measures is vital since crew members cannot be queried while performing critical tasks. Current approaches for measuring cognitive states generally use subjective questionnaires, which are obtrusive to administer in operational contexts (Kintz et al. 2022) and expensive and time-consuming to collect. Thus, we consider using only weak, external cognitive state labels provided by secondary observer judgements or programmatic labeling functions.

We develop new methodology to handle the ordinal nature of the label space and to be able to learn from approximate, weak labels (Ratner et al. 2016)–as opposed to human crowd-sourced labels. Apart from data collection and adjudication, the core technical problem we address is to derive an estimate of the unobserved gold-standard labels purely based on weak, external labels.

Data Collection and Initial Analysis

We are interested in modeling cognitive workload, which is assessed using the modified Bedford scale, a standard measure in the field of human factors. It is a uni-dimensional rating scale designed to identify operators' spare mental capacity while completing a task. Using the Bedford scale, a user rates perceived workload at one of 10 different descriptive levels.

Our aim is ultimately to be able to predict cognitive states of astronauts in operational settings, without requiring them to complete questionnaires. This is done by having external judges or automated rules gauge the operators off-line, simulating a crew-member's cognitive state being assessed by judges on the ground, who have additional time compared to the strict schedules of astronauts. For evaluation, our algorithm predictions will be compared to the gold-standard (but obtrusive) questionnaire responses provided by participants. The initial experiment consisted of video and audio streams recorded of participants as they completed trials of a simulated spaceflight-relevant task. Recordings obtained from participants show where the participant was looking, their facial expressions, the current task display, and their body/posture. We use these recordings and provide them to 5 external judges. Judges met and discussed best practices and general strategies for assessing cognitive workload in this experiment. These judges were research personnel who were familiar with cognitive state estimation and the experiment setup. Judges then independently watched the recordings and completed the same subjective questionnaires that participants provided. The cognitive workload ratings provided by judges from these recordings are combined with data about participants' actions from the experiment and form our initial dataset.

To investigate the feasibility of using external observations to model unobserved ground-truth, we analyse the performance of individual judges. We computed various metrics over the dataset, including the RMSE of judges weak labels compared to the unobserved ground-truth. Initial analysis shows that external experts do in fact perform better than random at predicting cognitive workload, see Table 1, where we explored different notions of random predictions.

Label Estimate	RMSE
Uniform Random Prediction	3.675
Default model: predict global average	1.873
Worst Judge	2.041
Average Judge	1.929
Best Judge	1.732
Default label model: predict mean weak label	1.643

Table 1: Error associated with each method of estimating ground truth labels.

On average, our five expert judges have a mean RMSE of 1.929, with the best judge scoring an RMSE of 1.732 and the worst judge 2.255. *Even naive aggregations of the expert votes produce improved predictions of the unobserved ground truth*. When we use the mean weak label to predict workload, the RMSE drops to 1.643, better than the best expert. These results are promising for further improvements using label models which estimate the errors the external expert judges make in order to arrive at further improved estimates of the unobserved ground truth.

A Factor Graph Label Model for Ordinal Data

To obtain improved estimates of the unobserved ground truth, we sought to expand on existing approaches to modelling weakly labeled data. Ratner et al. (2016, 2020) present a factor-graph based method that focuses on modeling and integrating noisy signals provided by a set of labeling functions. The approach defines a factor graph which encodes labelling propensity, accuracy, and pairwise correlations of labelling functions. Inspired by this method, our algorithm encodes the generative model $p_w(\Lambda, Y)$, using the labeling accuracy of the experts. Let $\Lambda \in \{1, \ldots, 10\}^{m \times n}$ denote the matrix of weak labels for m samples annotated by n external judges. Given the label matrix, for a given data point x_i, expert j, and unobserved gold-standard label y_i, the labelling accuracy factor for a classification task is defined as follows:

$$\phi_{i,j}^{Acc}(\Lambda, Y) = \mathbb{1}\{\Lambda_{i,j} = y_i\}$$

For a given data point x_i, we define the vector of this factor for all n experts as $\phi_i(\Lambda, Y)$, and the corresponding parameter vector $w \in R^n$. This defines our model:

$$p_w(\Lambda, Y) = Z_w^{-1} \exp\left(\sum_{i=1}^{m} w^\top \phi_i(\Lambda, y_i)\right). \qquad (1)$$

To fit this model without access to the gold-standard labels Y, we minimize the negative log marginal likelihood given the observed label matrix Λ. We optimize this objective by interleaving stochastic gradient descent steps with Gibbs sampling.

Ordinal Labels In our ordinal data setting, labels that are closer together along the ordinal scale are more likely to be confused. Therefore, we want to weight weak labels that are off by small magnitudes similarly to correct weak labels, which we achieve by introducing an alternative factor function. We define the following, parameterized factor function,

Best Judge	Default Label Model	Snorkel	Ours
1.732	1.643	3.183	1.582

Table 2: Mental workload prediction performance (RMSE) of label models, computed by comparison to ground-truth.

which replaces the accuracy factor function:

$$\phi_{i,j}^{Ord}(\Lambda, Y) = \frac{1}{1 + e^{|\Lambda_{i,j} - Y_i| - \delta_j}}$$

The error parameter δ is a vector, much like the w vector in our objective function, and each entry is associated with the error we allow for expert j. This parameter shifts the inverted sigmoid function along the x-axis, where greater shifts or greater δ_j values allow for greater errors to be made by expert j. The δ parameter can either be determined through domain knowledge about what constitutes a good and an acceptable weak label, or it can be learned based on Λ, since ϕ^{Ord} is continuous, smooth, and differentiable. We do note that, if one aims to learn δ, the objective is no longer convex. Therefore, in many cases it might be better to have domain knowledge inform the values δ, i.e., acceptable deviations from the ground-truth.

Results

To identify ideal setting for our algorithm (learning rate, number of iterations), we choose a learning rate that fits the weak labels well and achieves a low negative marginal log likelihood during training, and stop training early as the likelihood curve flattens out.

Our approach's RMSE value of 1.582 represents an improvement over the best expert, who had an RMSE of 1.732, over the data programmming method Snorkel (Ratner et al. 2020)–designed for classification–with an RMSE of 3.183, and over the default label model predicting the mean weak label with an RMSE of 1.643.

Acknowledgements

This work was partially supported by a Space Technology Research Institutes grant from NASA's Space Technology Research Grants Program and by the U.S. Army Research Office and the U.S. Army Futures Command under Contract No. W911NF-20-D-0002.

References

Kintz, J. R.; Banerjee, N. T.; Zhang, J. Y.; Anderson, A. P.; and Clark, T. K. 2022. Estimation of Subjectively Reported Trust, Mental Workload, and Situation Awareness Using Unobtrusive Measures. *Human Factors (forthcoming)*. Doi: 10.1177/00187208221129371.

Ratner, A.; Bach, S. H.; Ehrenberg, H.; Fries, J.; Wu, S.; and Ré, C. 2020. Snorkel: Rapid training data creation with weak supervision. *The VLDB Journal*, 29(2): 709–730.

Ratner, A. J.; De Sa, C. M.; Wu, S.; Selsam, D.; and Ré, C. 2016. Data Programming: Creating Large Training Sets, Quickly. In *Advances in Neural Information Processing Systems*, 3567–3575.