On the Utility of Prediction Sets in Human-AI Teams

Varun Babbar [1] Umang Bhatt [1] Adrian Weller [1,2]

[1]University of Cambridge [2]The Alan Turing Institute

Introduction

Machine learning models are increasingly being used in many real world settings involving high stakes decision making, such as medical diagnostics and computational drug discovery. In these settings, it is crucial for the human to be able to gauge and interpret the uncertainty of a model in order to facilitate robust decision making.

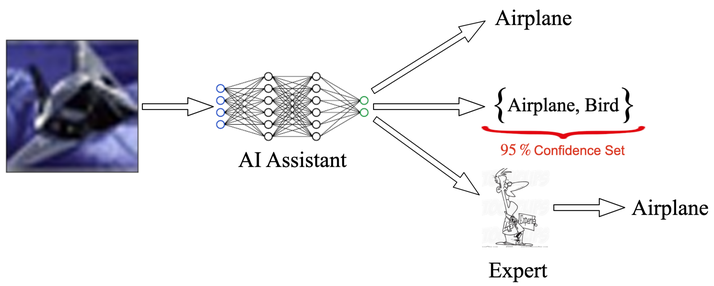

Figure 1. An AI assistant working alongside an expert can output one of three things: the most likely label, a set valued prediction with a predetermined error probability, or a deferral token indicating that the example should be labelled by the expert.

In this paper, we explore uncertainty in multi-class classification models from the perspective of prediction sets that provide theoretical guarantees on error tolerance. Specifically, we quantify how useful these sets are in human-AI teams and how we can generate even more useful sets.

Conformal Prediction (CP)

The goal of CP [5] is to construct predictive sets that are as small as possible for any user-defined error rate (or false negative rate) α. Formally, we construct sets of the following form:

$$1 - \alpha \geq P(Y \notin \Gamma(X)) \tag{1}$$

which holds in expectation for any datapoints (X, Y) that originate from the same distribution as the validation and training datasets. These predictive sets are generated in the following manner:

$$\Gamma(X) = \{y : \tau(X, y) \geq \tau_{cal}\} \tag{2}$$

where $\tau(X, y)$ is called a conformity score function and τ_{cal} is determined using a held out calibration dataset $\mathcal{D}_{cal} = \{(X_i, Y_i)\}_{i=1}^{N}$:

$$\tau_{cal} = \text{Quantile}(\alpha, \{\tau(X_i, Y_i)\}_{i=1}^{N}) \tag{3}$$

How useful are CP sets in human-AI teams?

Metric	Top-1	RAPS	p value	Effect Size
Accuracy	0.76 ± 0.05	0.76 ± 0.05	0.999	0.000
Reported Utility	5.43 ± 0.69	6.94 ± 0.69	**0.003**	1.160
Reported Confidence	7.21 ± 0.55	7.88 ± 0.29	0.082	0.674
Reported Trust	5.87 ± 0.81	8.00 ± 0.69	$<$ **0.001**	1.487

Table 1. Top-1 vs RAPS: All Examples

For our human subject experiments, we focus on one particular CP scheme called Regularised Adaptive Prediction Sets (RAPS) [1]. We split 30 participants in 2 groups and ask them to classify 15 CIFAR-100 images given their knowledge of Top-1 or RAPS and report other metrics such as utility, confidence, and trust on a scale of 10.

A scheme for providing more more useful CP sets: D-CP

- Our scheme is centered around two components: a deferral policy $\pi(x) : \mathcal{X} \to \{0, 1\}$ and a CP method.
- The deferral policy is based on our knowledge of the expert's strengths either acquired during training or a-priori. Using this black box policy, we first prune our calibration dataset, removing all examples where our deferral policy recommends deferral.
- After training a model and a suitable deferral policy, we perform conformal calibration on this pruned dataset of non-deferred examples.

In this procedure, for any predictive set $\Gamma(X_{test}, \tau_{cal})$ for an example X_{test} we can guarantee that:

$$1 - \alpha \leq P(Y \in \Gamma(X_{test}, \tau_{cal}) | \pi(X_{test}) = 0) \tag{4}$$

where 1 represents the action of deferral. From [1], when the conformity scores are known to be almost surely distinct and continuous, we can also guarantee:

$$P(Y \in \Gamma(X_{test}, \tau_{cal}) | \pi(X_{test}) = 0) \leq 1 - \alpha + \frac{1}{n+1} \tag{5}$$

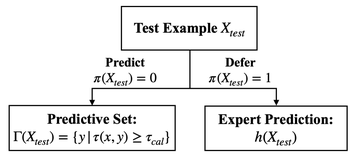

Figure 2. D-CP: Test Phase given a deferral policy $\pi(X)$

D-CP: Experiments with CIFAR-100 and CIFAR-10H

Deferral Rate	Team Accuracy	Predictive Set Size of Non-Deferred Examples		
		RAPS	APS	LAC
0	65.18 ± 0.30	3.75 ± 0.06	4.61 ± 0.08	3.26 ± 0.03
0.1	69.95 ± 0.31	2.81 ± 0.05	4.05 ± 0.06	2.13 ± 0.04
0.2	72.98 ± 0.30	2.36 ± 0.06	2.93 ± 0.10	2.07 ± 0.03

Deferral Rate	Team Accuracy	Predictive Set Size of Non-Deferred Examples		
		RAPS	APS	LAC
0	82.02± 0.55	1.91 ± 0.03	2.83 ± 0.05	2.47 ± 0.03
0.1	86.53 ± 0.68	1.73 ± 0.08	2.56 ± 0.07	1.90 ± 0.04
0.2	89.43 ± 0.64	1.49 ± 0.06	2.13 ± 0.13	1.51 ± 0.03

Figure 3. Set size and overall team accuracy on the CIFAR-100 (top) and CIFAR-10H (bottom) datasets for the deferral scheme in [2] with $\alpha = 0.1$. Even with low deferral rates, we not only obtain smaller set sizes, but also benefit from increased human-AI team accuracy compared to not deferring. CIFAR-100: Synthetic Human Expert with 70% accuracy. CIFAR-10H: Real human annotations with 95% accuracy

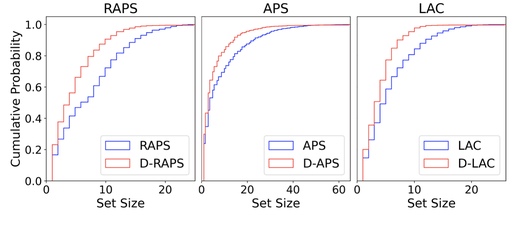

Figure 4. Cumulative CP and D-CP Set Size Distribution of Non-Deferred Examples in the CIFAR-100 dataset ($\alpha = 0.05$, deferral rate $b = 0.2$, Single Expert) for 3 different CP Schemes [4, 3]

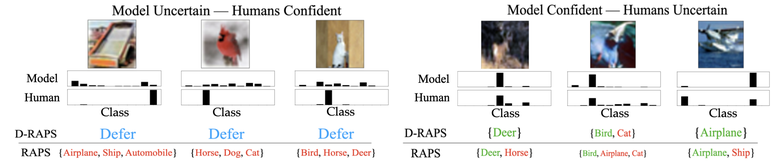

Figure 5. D-RAPS vs RAPS on CIFAR-10H examples ($\alpha = 0.05$, $b = 0.2$). Deferring whenever experts are more confident than the model yields smaller sets on examples where the model is more confident than the expert. Thus, we leverage both the model and the expert's strengths

D-CP: Human Subject Analysis

- We choose another set of 15 examples from the CIFAR-100 test set for which we generate RAPS prediction sets with error rate $\alpha = 0.1$ and D-RAPS prediction sets with deferral rate 0.2 and error rate $\alpha = 0.1$.
- We select 12 non-deferred examples at random wherein the D-RAPS predictive set is smaller than the RAPS predictive set, but the ground truth labels are contained in both sets.
- We choose the remaining 3 deferred examples where the model is underconfident, i.e. RAPS provides misleading predictions because the ground truth label is not in the set.

Metric	D-RAPS	RAPS	p value	Effect Size
Accuracy	0.76 ± 0.08	0.67 ± 0.05	**0.002**	0.832
Reported Utility	7.93 ± 0.39	6.32 ± 0.60	$<$ **0.001**	1.138
Reported Confidence	7.31 ± 0.29	7.28 ± 0.29	0.862	0.046
Reported Trust	8.00 ± 0.45	6.87 ± 0.61	**0.006**	0.754

Table 2. D-RAPS vs RAPS: All Examples

Metric	D-RAPS	RAPS	p value	Effect Size
Accuracy	0.88 ± 0.05	0.81 ± 0.04	0.058	0.508
Reported Utility	7.93 ± 0.39	6.19 ± 0.62	$<$ **0.001**	1.211
Reported Confidence	7.78 ± 0.33	7.31 ± 0.34	0.059	0.507

Table 3. D-RAPS vs RAPS: Non-Deferred Examples

- We define the **bias** toward incorrect labels as the proportion of examples where an incorrect prediction made by an expert is found in the predictive set output by the model averaged across all subjects.
- That is, given experts h, examples x, the associated label $y(x)$, and the CP set $\Gamma(x)$:

$$\text{Bias} = \mathbb{E}_{h,x}\left[\mathcal{I}_{h(x) \in \Gamma(x)} \mathcal{I}_{h(x) \neq y(x)}\right] \tag{6}$$

Metric	D-RAPS	RAPS Non-Deferred Examples	RAPS Deferred Examples s
Bias	0.063 ± 0.035	0.189 ± 0.046	0.933 ± 0.069

Table 4. D-RAPS vs RAPS: Bias towards incorrect or misleading labels. Comparing just the non-deferred examples we see that experts are much more biased towards incorrect predictions in RAPS sets than in D-RAPS sets.

References

[1] Anastasios Nikolas Angelopoulos, Stephen Bates, Michael Jordan, and Jitendra Malik. Uncertainty Sets for Image Classifiers Using Conformal Prediction. In *International Conference on Learning Representations*, 2020.

[2] Hussein Mozannar and David Sontag. Consistent Estimators for Learning to Defer to an Expert. In *International Conference on Machine Learning*, pages 7076–7087. PMLR, 2020.

[3] Yaniv Romano, Matteo Sesia, and Emmanuel Candes. Classification with Valid and Adaptive Coverage. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 3581–3591. Curran Associates, Inc., 2020.

[4] Mauricio Sadinle, Jing Lei, and Larry Wasserman. Least Ambiguous Set-Valued Classifiers with Bounded Error Levels. *Journal of the American Statistical Association*, 114:223–234, 9 2016.

[5] Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 01 2005.