arXiv:2306.09335

Class-conditional conformal prediction with many classes

Tiffany Ding, Anastasios N. Angelopoulos, Stephen Bates, Michael I. Jordan, Ryan J. Tibshirani
University of California, Berkeley

Introduction

Standard conformal prediction (CP) methods are designed to take an input $X_{\text{test}} \in \mathcal{X}$ with unknown label $Y_{\text{test}} \in \mathcal{Y}$ (along with a labeled calibration set and a conformal score function $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$) and construct a prediction set $C(X_{\text{test}})$ that achieve *marginal coverage* for some small $\alpha > 0$:

$$P(Y_{\text{test}} \in C(X_{\text{test}})) \geq 1 - \alpha$$

However, in many settings we want the stronger guarantee of *class-conditional coverage*:

$$P(Y_{\text{test}} \in C(X_{\text{test}}) \mid Y_{\text{test}} = y) \geq 1 - \alpha \qquad \forall y \in \mathcal{Y} \quad (1)$$

Goal: Create a conformal prediction method that achieves good class-conditional coverage even in settings with *many classes* or *limited data*.

Methods

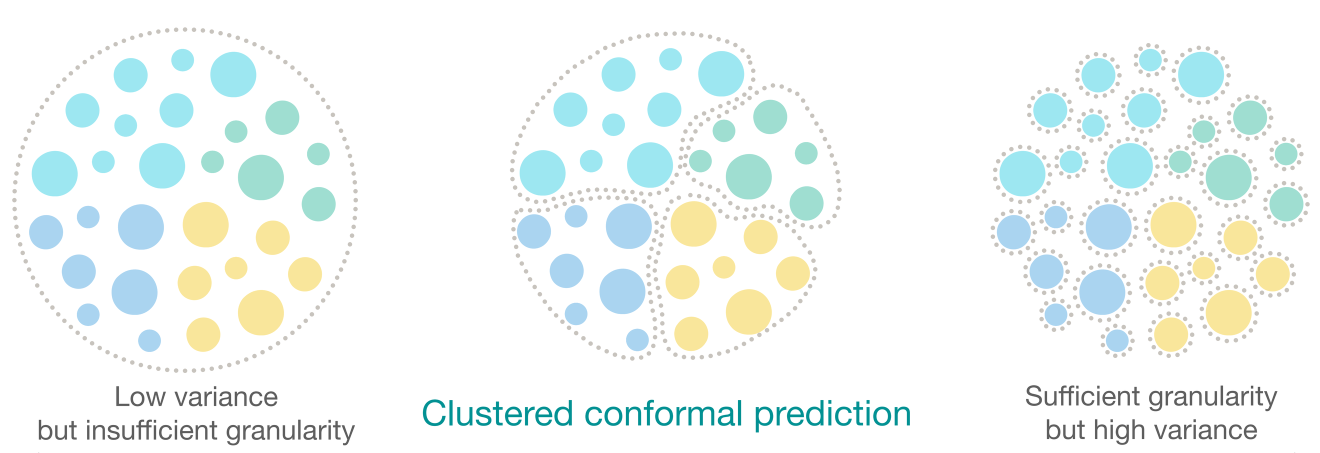


Fig 1. Existing methods either (1) do not split data but are only designed to achieve marginal coverage, or (2) are designed to achieve class-conditional coverage but use data inefficiently. Our method, *clustered conformal prediction*, achieves the best of both worlds by grouping together data from classes with similar conformal score distributions.

Standard CP: Estimated on *all* calibration data

$$C_{\text{STANDARD}}(X_{\text{test}}) = \{y : s(X_{\text{test}}, y) \leq \hat{q}^{\text{all}}\}$$

Classwise CP: Estimated using *only* data for class y

$$C_{\text{CLASSWISE}}(X_{\text{test}}) = \{y : s(X_{\text{test}}, y) \leq \hat{q}_y\}$$

Clustered CP (ours): Estimated using data in cluster that contains class y

$$C_{\text{CLUSTERED}}(X_{\text{test}}) = \{y : s(X_{\text{test}}, y) \leq \hat{q}(\hat{h}(y))\}$$

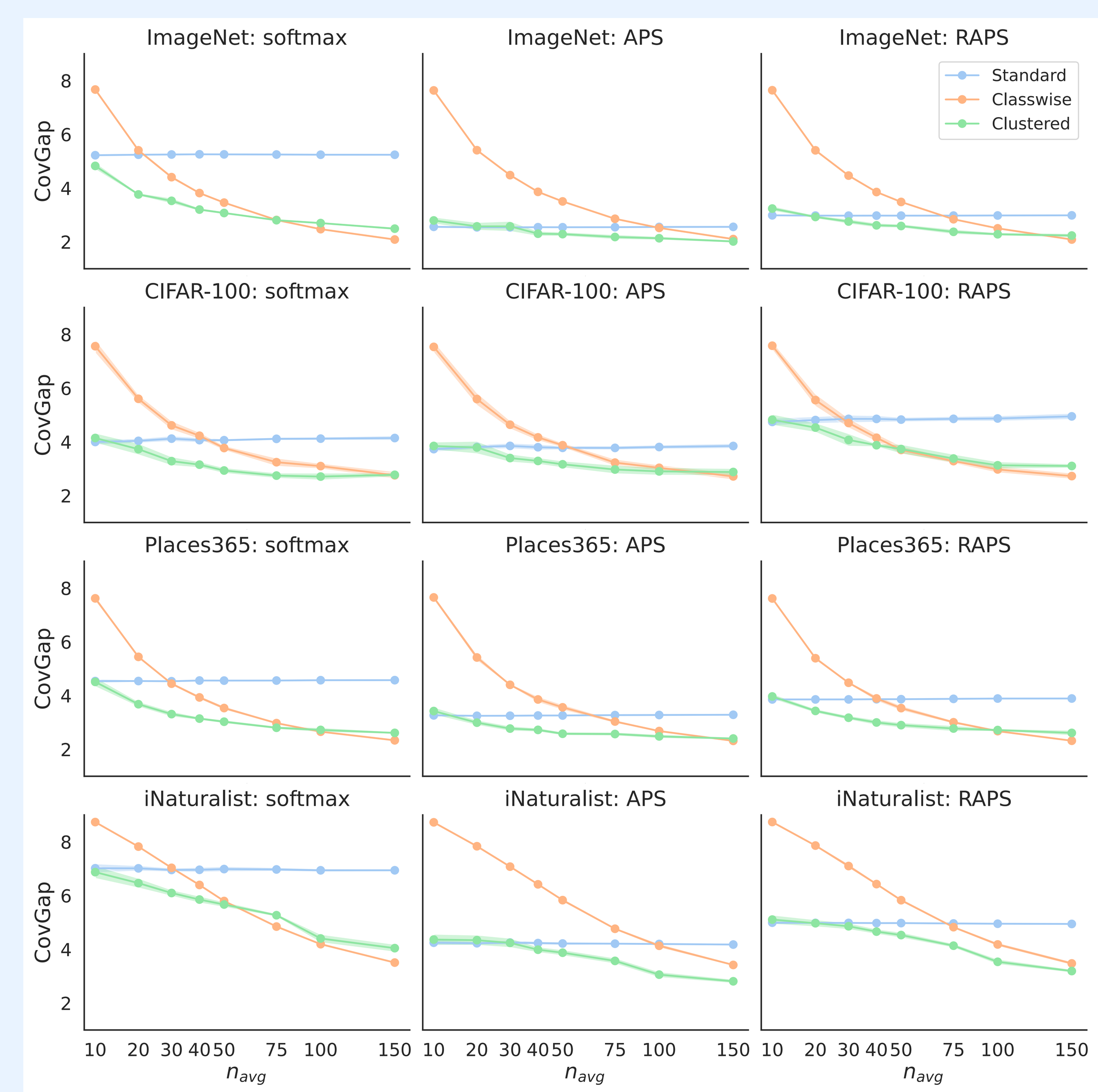
where $\hat{h} : \mathcal{Y} \to [M] \cup \text{null}$ is a clustering function learned by splitting off part of the calibration dataset, computing a quantile embedding for the data of each class, then performing k-means clustering.

Experiments

Setup: We compare the performance of Standard, Classwise, and Clustered CP using the softmax, APS, and RAPS conformal score functions for various amounts of calibration data. n_{avg} is the average # of calibration examples per class.

Datasets:

ImageNet	CIFAR-100	Places365	iNaturalist
1000 classes	100 classes	365 classes	663 classes & highly imbalanced



How close are we to the desired coverage level of $1 - \alpha$?

Fig 2. Class-coverage gap (CovGap), defined as

$$\text{CovGap} = 100 \times \frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{Y}} |\hat{c}_y - (1-\alpha)|$$

where $\hat{c}_y$ is the coverage of class y, as computed on our validation dataset.

How useful are the prediction sets?

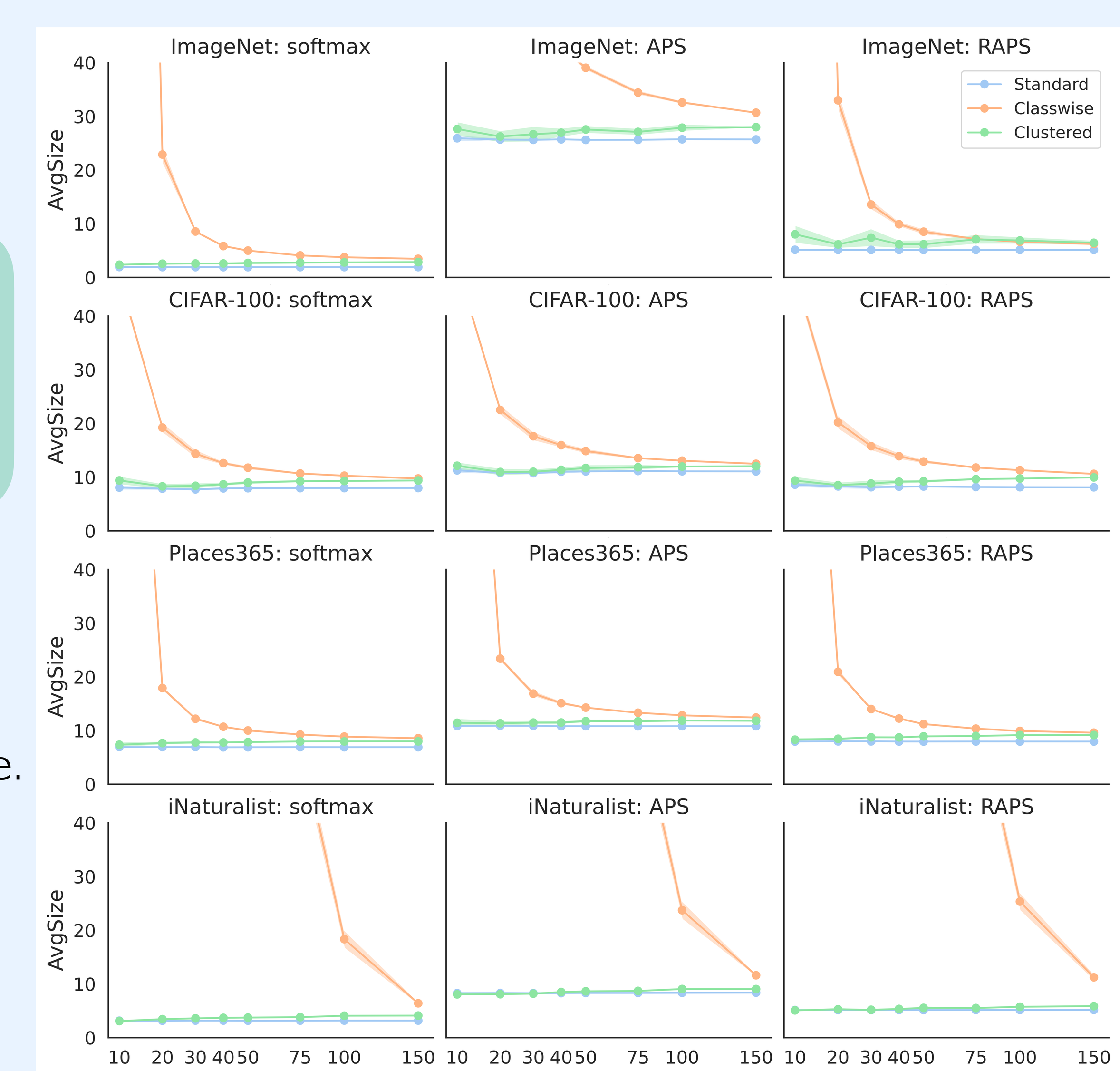


Fig 3. Average set size.

Theoretical guarantees

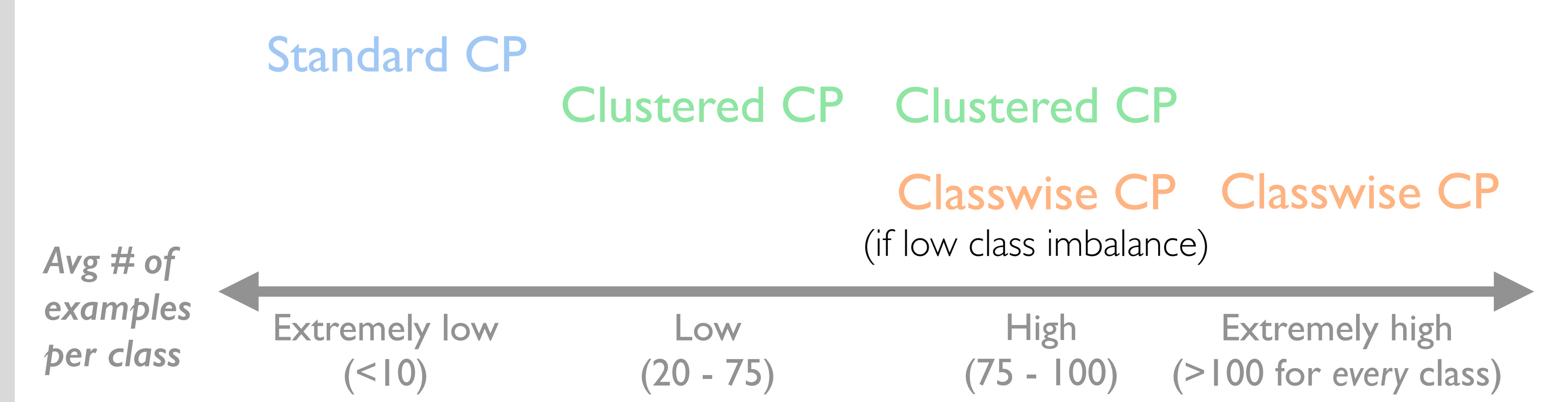
Proposition 1: (Under perfect clustering) Let h^* be an oracle clustering function such that all classes assigned to the same cluster have scores that are exchangeable. If $\hat{h} = h^*$, then $C_{\text{CLUSTERED}}$ will satisfy (1).

Proposition 2: (Under imperfect clustering) Suppose that the classes that $\hat{h}$ assigns to the same cluster are *almost* exchangeable (formally, let S^y denote a random variable sampled from the score distribution for class y, and assume $D_{\text{KS}}(S^y, S^{y'}) \leq \epsilon$ for all y, y' s.t. $h(y) = h(y')$), then $C_{\text{CLUSTERED}}$ will satisfy

$$P(Y_{\text{test}} \in C(X_{\text{test}}) \mid Y_{\text{test}} = y) \geq 1 - \alpha - \epsilon, \; \forall y \in \mathcal{Y}.$$

Practical takeaways

For a given problem setting, what is the best way to produce prediction sets that have good class-conditional coverage but are not too large to be useful?



Conclusion

Summary

1. Marginal coverage is not enough. In many settings, we want to have class-conditional coverage.
2. Class-conditional coverage is hard to achieve when there are many classes and limited data per class.
3. Clustering classes with similar score distributions allows us to share data between classes in a way that will achieve good class-conditional coverage

Future directions? Generalizing our clustering approach to achieve group-conditional coverage for any grouping.

tiffany_ding@berkeley.edu