Class-Conditional Conformal Prediction with Many Classes

Ahn Youngmin

February 25, 2025

Seoul National University

Introduction

Main contribution

- Proposes an extension of standard conformal prediction, called **clustered conformal prediction**, for classification problems. *This method constructs a prediction set that satisfies class-conditional coverage instead of marginal coverage.*
- Demonstrates through empirical evaluation that the algorithm performs well even *when the number of classes is large (more than 100) and the data is limited.*

Outline

1. Background
2. Clustered Conformal Prediction
3. Experiments
4. Conclusion

Background

Standard Conformal Prediction

- Consider a calibration data set $\{(X_i, Y_i)\}_{i=1}^N$ and a test point (X_{test}, Y_{test}), where $X_i \in \mathcal{X}$, $Y_i \in \mathcal{Y}$, drawn from distribution F.
- Standard conformal prediction (STANDARD) creates a prediction set that satisfies marginal coverage:

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

for a coverage level $\alpha \in [0, 1]$.

Standard Conformal Prediction

Procedure of STANDARD

1. Given a pre-trained classifier f, define a nonconformity score function $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$.
   (A simple example is $s(x, y) = 1 - f_y(x)$, where $f_y(x)$ represents the yth entry of the softmax vector output by f for the input x.)

2. Evaluate the score of every calibration data point as $s_i = s(X_i, Y_i)$.

3. Compute the $(1 - \alpha)$ quantile of the empirical distribution defined by $\{s_i\}_{i=1}^N$:

$$\hat{q} = \text{Quantile}\left(\frac{\lceil (N-1)(1-\alpha) \rceil}{N}, \{s_i\}_{i=1}^N\right).$$

4. Construct prediction set:

$$C_{\text{STANDARD}}(X_{test}) = \{y \in \mathcal{Y} : s(X_{test}, y) \leq \hat{q}\}.$$

⇒ The core problem is that even though the average performance of the algorithm is good, the performance for some classes is quite poor.

Classwise Conformal Prediction

- Classwise conformal prediction (CLASSWISE) creates a prediction set that satisfies class-conditional coverage:

$$\mathbb{P}(Y_{test} \in C(X_{test}) \mid Y_{test} = y) \geq 1 - \alpha, \text{ for all } y \in \mathcal{Y},$$

for a coverage level $\alpha \in [0, 1]$.

- Use $(1 - \alpha)$ quantile of the empirical distribution of each class defined by $\{s_i\}_{i \in \mathcal{I}^y}$, where $\mathcal{I}^y = \{i \in \{1, ..., N\} : Y_i = y\}$:

$$\hat{q}^y = \text{Quantile}\left(\frac{\lceil (|\mathcal{I}^y| - 1)(1 - \alpha) \rceil}{|\mathcal{I}^y|}, \{s_i\}_{i \in \mathcal{I}^y}\right)$$

- The prediction set is constructed using a different threshold for each class:

$$C_{\text{CLASSWISE}}(X_{test}) = \{y \in \mathcal{Y} : s(X_{test}, y) \leq \hat{q}^y\}.$$

Clustered Conformal Prediction

Clusered Conformal Prediction

Clusered conformal prediction (CLUSTERED) clusters classes having similar conformal score distributions.

1. Randomly split the calibration data set into two parts: the clustering data set $D_1 = \{(X_i, Y_i) : i \in \mathcal{I}_1\}$ and a proper calibration data set $D_2 = \{(X_i, Y_i) : i \in \mathcal{I}_2\}$.

2. Apply a clustering algorithm to D_1 to obtain a clustering function $\hat{h} : \mathcal{Y} \to \{1, ..., M\} \cup \{null\}$.

3. Evaluate the score of the data points in D_2, and compute the $(1 - \alpha)$ quantile, $\hat{q}(\hat{h}(y))$, of the empirical distribution of each cluster.
   (For the *null* cluster, the quantile is computed using the entire data point in D_2, just as in STANDARD.)

4. The prediction set is constructed using a different threshold for each cluster:

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

Clusered Conformal Prediction

Proposition 1

The prediction sets $C = C_{\text{CLUSTERED}}$ from CLUSTERED achieve cluster-conditional coverage:

$$\mathbb{P}(Y_{test} \in C(X_{test} \mid \hat{h}(Y_{test} = m)) \geq 1 - \alpha, \text{ for all clusters } m = 1, ..., M.$$

Proposition 2

If $\hat{h} = h^*$, where h^* is an oracle clustering function that produces clusters of classes with the same score distribution, then the prediction sets from CLUSTERED satisfy class-conditional coverage for all classes y such that $h^* \neq null$.

$\Rightarrow$ Therefore, we need to find a clustering function that clusters classes that have similar score distributions.

Quantile-based clustering

Procedure of clustering

1. Denote by $\mathcal{I}_1^y = \{i \in \mathcal{I}_1 : Y_i = y\}$ the indices of examples in D_1 with label y.
2. Compute quantiles of the scores $\{s_i\}_{i \in \mathcal{I}_1^y}$ from class y at the levels

$$\mathcal{T} = \left\{ \frac{\lceil (|\mathcal{I}_1^y| + 1)\tau \rceil}{|\mathcal{I}_1^y|} : \tau \in \{0.5, 0.6, 0.7, 0.8, 0.9\} \cup \{1 - \alpha\} \right\},$$

   and collect them into an embedding vector $z^y \in \mathbb{R}^{|\mathcal{T}|}$.
3. If $|\mathcal{I}_1^y| < (1/\min\{\alpha, 0.1\}) - 1$, then the uppermost quantile in z^y will not be finite, so assign y to the *null* cluster.
4. For a pre-chosen number of clusters M, run k-means clustering with $k = M$ on the data $\{z^y\}_{y \in \mathcal{Y} \setminus \mathcal{Y}_{null}}$, where $\mathcal{Y}_{null}$ denotes the set of labels assigned to the *null* cluster.

Quantile-based clustering

Proposition 3

Let S^y denote a random variable sampled from the score distribution for class y, and assume that the clustering map $\hat{h}$ satisfies

$$D_{KS}(S^y, S^{y'}) \leq \epsilon, \text{ for all } y, y' \text{ such that } \hat{h}(y) = \hat{h}(y') \neq null,$$

where $D_{KS}(X, Y)$ is the Kolmogorov-Smirnov distance, which is defined between random variables X and Y as
$D_{KS}(X, Y) = \sup_{\lambda \in \mathbb{R}} |\mathbb{P}(X \leq \lambda) - \mathbb{P}(Y \leq \lambda)|$.
Then, for $C = C_{\text{CLUSTERED}}$ and for all classes y such that $\hat{h}(y) \neq null$,

$$\mathbb{P}(Y_{test} \in C(X_{test}) \mid Y_{test} = y) \geq 1 - \alpha - \epsilon.$$

$\Rightarrow$ If the score distributions for the classes that $\hat{h}$ assigns to the same cluster are similar enough, then we can provide an approximate class-conditional coverage guarantee.

Experiments

Settings

- Data sets

Data set	ImageNet	CIFAR-100	Places365	iNaturalist
Number of classes	1000	100	365	663*
Class balance	0.79	0.90	0.77	0.12
Example classes	mitten triceratops guacamole	orchid forest bicycle	beach sushi bar catacomb	salamander legume common fern

*The number of classes in the iNaturalist data set can be adjusted by selecting which taxonomy level (e.g., species, genus, family) to use as the class labels. We use the species family as our label and then filter out any classes with < 250 examples in order to have sufficient examples to properly perform evaluation.

- Base classifier: ResNet-50 fine-tuned on a small subset D_{fine} of the original data.
- Score function: softmax (1- softmax output of the base classifier), APS (a score designed to improve X-conditional coverage)
- Coverage level: $\alpha = 0.1$
- Size of calibration set D_{cal}: $n_{avg} \times \mathcal{Y}, n_{avg} = \{10, 20, 30, 40, 50, 75, 100, 150\}$
- Splitting D_1 and D_2: (1) Consider the size of D_1 to grow with the number of clusters, M. (2) Consider the size of D_2 to have at least 150 points per cluster on average.

Evaluation

Evaluation metric: Average class coverage gap (CovGap)

$\Rightarrow$ Measures how far the class-conditional coverage is from the desired coverage level of $1 - \alpha$ in terms of the ℓ_1 distance across all classes.

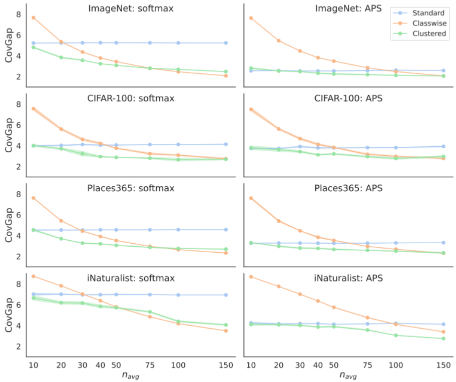

Figure 1: CovGap for ImageNet, CIFAR-100, Places365, and iNaturalist, for the softmax (left) and APS (right) scores, as we vary n_{avg}.

Conclusion

Summary

- Proposes **clustered conformal prediction** outperforms standard and classwise conformal, *in terms of class-conditional coverage, when there is limited calibration data available per class.*

Guidelines for Class-conditional Coverage

examples/class	< 10	$20 \sim 70$	$75 \sim 100$	$100 <$
Method	STANDARD	CLUSTERED	CLUSTERED +CLASSWISE	CLASSWISE

References

- Ding, Tiffany & Angelopoulos, Anastasios & Bates, Stephen & Jordan, Michael & Tibshirani, Ryan. (2023). Class-Conditional Conformal Prediction With Many Classes. 10.48550/arXiv.2306.09335.