Technical University of Munich

Assessing Predictive Uncertainties in Remote Sensing Image Classification via Conformal Prediction

Christoph Koller [1,2], Protim Bhattacharjee [2], Peter Jung [2,3]

[1]: Technical University of Munich (TUM) [2]: German Aerospace Center (DLR) [3]: Technical University of Berlin (TUB)

Conformal Prediction in a Nutshell

- Conformal Prediction (**CP**) is a **post-hoc** calibration method with theoretical **coverage guarantees**
- Applied to a classification model, the CP framework yields so-termed **prediction sets** (subset of all available classes)
- After *conformalization*, the true class is supposed to lie in the prediction set with a prespecified probability

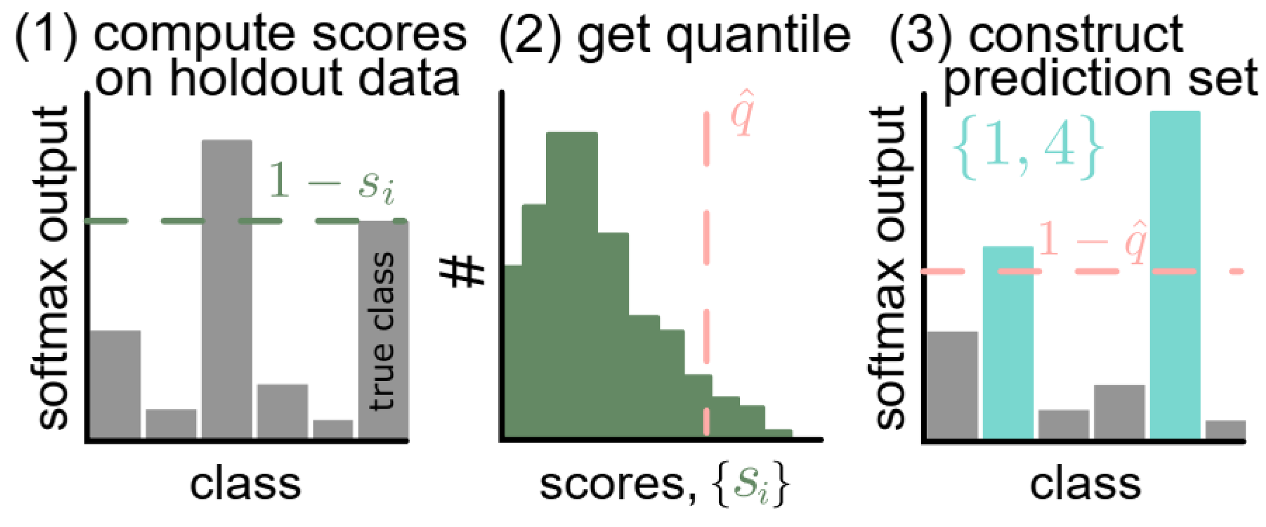
How it works:

- Choose an error rate $\alpha \in [0,1]$ and set aside a calibration dataset of size n_{calib}. The prediction set $\mathcal{C}(X_{test}) \subset \{1, \dots, K\}$ for a test data point X_{test} then should satisfy:

$$1 - \alpha \leq \mathbb{P}(Y_{\text{test}} \in \mathcal{C}(X_{\text{test}})) \leq 1 - \alpha + \frac{1}{n+1}$$

- Define the *conformal score* as 1 minus the softmax probability of the true class: $s_i = 1 - \hat{f}(X_i)_{Y_i}$
- Now set $\hat{q}$ as the $\lceil (n+1)(1-\alpha) \rceil$ / n quantile of $s_1, \dots, s_{n_{calib}}$
- Finally, create a prediction set for a new point as follows

$$\mathcal{C}(X_{\text{test}}) = \{y : \hat{f}(X_{\text{test}})_y \geq 1 - \hat{q}\}$$

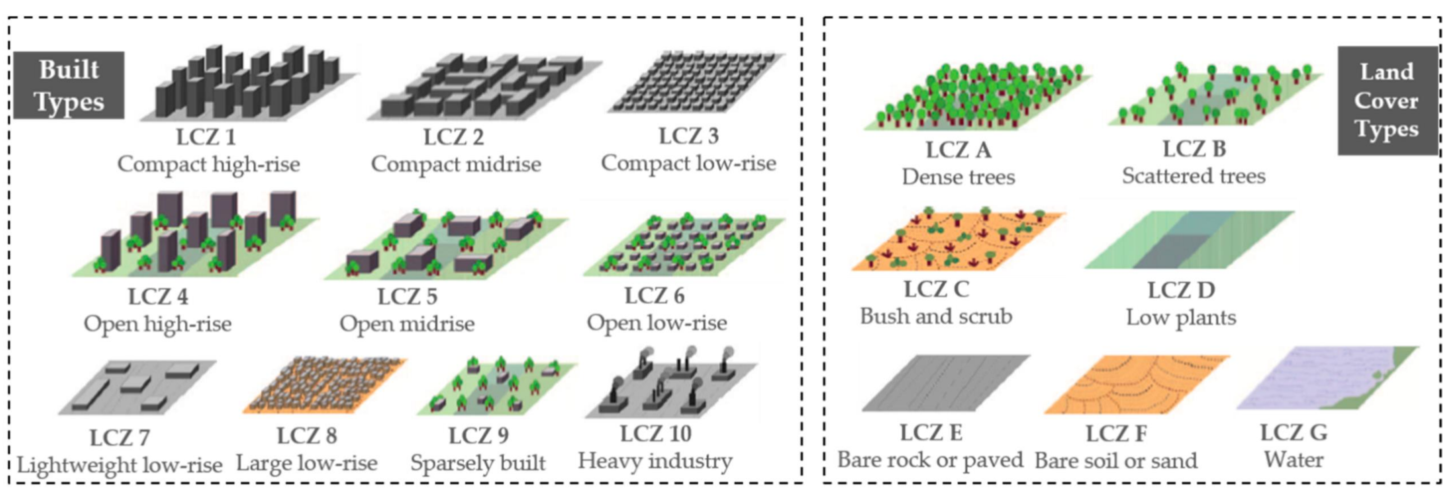


LAC Algorithm explained (Angelopoulos, Bates 2021)

Label Uncertainty in Remote Sensing

We study a subset of the *So2Sat LCZ42* (Zhu et al. 2020) dataset: **10 European cities** labeled by 10 remote sensing experts.

- Can CP help to derive suitable prediction sets with coverage guarantees being met?
- Are the prediction sets covering the human label uncertainty?



Local Climate Zone (LCZ) classification scheme (Zhao et al. 2019)

(Regularized Adaptive) Prediction Sets

We investigate multiple **CP** methods for the Urban classes (1-10):

- **Least Ambiguous set-valued Classifier (LAC):** Algorithm left
- **Naive approach**: Softmax values are ranked and summed up until the threshold is reached
- **Adaptive Prediction Sets (APS):** Softmax scores are summed up until true label is reached, last label can be in- or excluded or randomly decided (based on uniform sampling)
- **Regularized Adaptive Prediction Sets (RAPS):** APS with regularization hyperparameters based on *tuning dataset*
- **Top-k:** Fixed prediction set size based on rank of true label

Name	α Value	Coverage		No. of Null Sets		Avg. Pred. Set Size		Label Votes Cov.	
		One-Hot	Distr.	One-Hot	Distr.	One-Hot	Distr.	One-Hot	Distr.
Naive	$\alpha = 0.05$	77.5%	86.9%	0	0	**1.52**	2.20	56.8%	72.0%
	$\alpha = 0.1$	74.3%	82.0%	0	0	**1.30**	1.72	53.1%	64.4%
	$\alpha = 0.15$	72.8%	78.4%	0	0	**1.19**	1.47	50.9%	59.6%
	$\alpha = 0.2$	71.7%	76.1%	0	0	**1.12**	1.30	49.3%	55.8%
LAC	$\alpha = 0.05$	94.4%	95.0%	0	0	3.31	3.52	81.5%	86.0%
	$\alpha = 0.1$	89.1%	89.7%	0	0	2.37	2.43	69.0%	77.0%
	$\alpha = 0.15$	84.0%	84.5%	0	0	1.85	1.83	62.6%	67.6%
	$\alpha = 0.2$	79.4%	79.8%	0	2	1.48	1.41	57.7%	58.7%
APS w/ last label	$\alpha = 0.05$	95.6%	96.1%	0	0	3.75	3.83	86.3%	87.9%
	$\alpha = 0.1$	91.7%	**92.5%**	0	0	2.73	2.92	73.3%	**81.7%**
	$\alpha = 0.15$	87.7%	89.5%	0	0	2.22	2.39	67.2%	76.3%
	$\alpha = 0.2$	84.7%	**86.8%**	0	0	1.93	2.05	63.6%	**71.5%**
APS w/o last label	$\alpha = 0.05$	90.6%	91.6%	0	0	2.89	2.90	74.5%	80.1%
	$\alpha = 0.1$	82.8%	85.8%	0	0	1.97	2.07	62.9%	70.2%
	$\alpha = 0.15$	77.8%	80.6%	0	0	1.54	1.61	57.2%	62.4%
	$\alpha = 0.2$	75.0%	77.1%	0	0	1.33	1.38	53.8%	57.5%
APS w/ randomness	$\alpha = 0.05$	94.7%	94.5%	3	9	3.44	3.46	82.4%	85.1%
	$\alpha = 0.1$	89.3%	89.6%	20	55	2.47	2.57	69.8%	77.3%
	$\alpha = 0.15$	83.6%	84.5%	99	149	1.96	2.01	63.2%	69.5%
	$\alpha = 0.2$	78.5%	79.4%	236	296	1.66	1.69	58.3%	62.8%
RAPS	$\alpha = 0.05$	94.8%	94.9%	2	2	3.54	3.51	84.3%	86.0%
	$\alpha = 0.1$	89.4%	89.8%	2	22	2.59	2.56	70.9%	77.5%
	$\alpha = 0.15$	83.8%	85.1%	66	106	1.95	2.04	62.9%	70.5%
	$\alpha = 0.2$	78.9%	80.2%	106	151	1.57	1.66	57.7%	63.2%
Top-k	$\alpha = 0.05$	**96.6%**	95.6%	0	0	5	4	**93.1%**	88.8%
	$\alpha = 0.1$	90.5%	91.0%	0	0	3	3	75.6%	80.3%
	$\alpha = 0.15$	90.5%	**91.0%**	0	0	3	3	75.6%	**80.3%**
	$\alpha = 0.2$	83.4%	82.3%	0	0	2	2	64.3%	66.6%

Results for various methods on LCZ42 Evaluation Dataset (https://mediatum.ub.tum.de/1659039). Validation dataset was used for calibration. Coverage = True label in prediction set. Sen2LCZ (Qiu et al. 2020) was used as network classifier for the Urban classes. One-Hot = Training with single label (majority vote of experts), Distr. = Training with empirical distribution of label votes. Label Votes Cov. = Percentage of expert label votes covered by prediction sets.

Findings

- Naive approach overconfident; bad coverage despite small sets
- Strong results with APS, regularization seemingly without effect
- Randomization leads to large no. of null sets
- Top-k conformalization shines with seemingly great results, but comes with comparably large prediction sets
- Strong performance increase with distributional label approach

References:

- Angelopoulos, A. N., & Bates, S. (2021). A gentle introduction to conformal prediction and distribution-free uncertainty quantification. arXiv preprint arXiv:2107.07511.
- Qiu, C. et al. (2020). Multilevel feature fusion-based CNN for local climate zone classification from sentinel-2 images: Benchmark results on the So2Sat LCZ42 dataset. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 13.
- Zhao, N., Ma, A., Zhong, Y., Zhao, J., & Cao, L. (2019). Self-training classification framework with spatial-contextual information for local climate zones. Remote Sensing, 11(23).
- Zhu, X. X. et al. (2020). So2Sat LCZ42: A benchmark data set for the classification of global local climate zones. IEEE Geoscience and Remote Sensing Magazine, 8(3).

Technical University of Munich
TUM School of Engineering and Design
Chair of Data Science in Earth Observation

German Aerospace Center (DLR)
Remote Sensing Technology Institute
EO Data Science

This present contribution was supported by the Helmholtz Association under the Joint Research School Munich School for Data Science—MUDS.