Leave-One-Out Stable Conformal Prediction

Kiljae Lee, Yuan Zhang

The Ohio State University

Why Uncertainty Quantification?

ICLR International Conference On Learning Representations

The Ohio State University

In many machine learning applications, **point predictions** dominate.

Medical AI: *"This tumor is malignant"*

Autonomous Driving: *"A pedestrian is detected"*

But how it sure?

***Uncertainty quantification** is not optional. It is critical.*

Valid Prediction Set

Available data $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$

Test point $(X_{n+1}, \sout{Y_{n+1}})$

Our goal: Construct **a valid set** $C(X_{n+1})$ that contains Y_{n+1} with high probability ***without any distribution assumption***.

Definition (Validity)

*A set $C_\alpha(X_{n+1})$ is **valid** at a given level of α if $P\big(Y_{n+1} \in C_\alpha(X_{n+1})\big) \geq 1 - \alpha$ holds.*

Conformal Prediction

- **Conformal Prediction (CP)** provides valid prediction sets, relying solely on the ***i.i.d. assumption of data***.
- A core concept, **nonconformity scores**, quantify ***how well a given label Y_{n+1} conforms*** to a trained model.

(Examples)

Regression: $S(Y_i, \hat{f}(X_i)) = |Y_i - \hat{f}(X_i)|$

Classification: $S(Y_i, \hat{p}(X_i)) = 1 - \hat{p}_{Y_i}(X_i)$

(WARNING!) ***If some data points are favored*** in training, the validity breaks.

Full CP vs Split CP

There are **two** standard approaches ***to guarantee the equivalence (a.k.a. Exchangeability)*** across the S_i's.

Available data

$(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$

Test point

(X_{n+1}, y)

Full CP vs Split CP

Full CP (Vovk, 2005) computes nonconformity scores by ***fitting a model over all possible guesses*** of the test label.

Available data **Test point**

$(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ (X_{n+1}, y)

Train model $\hat{f}^y(\cdot)$ (for every y) & Calibrate scores $S(Y_i, \hat{f}^y(X_i))$

Pros: Allows the ***use of entire dataset*** for both training & calibration

Cons: Computational burden due to the ***model retraining***

ICLR International Conference On Learning Representations

Full CP vs Split CP

The Ohio State University

Split CP (Papadopoulos et al., 2002) ***divides the data*** into training and calibration sets.

Available data

$(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$

Test point

(X_{n+1}, y)

Training: $\hat{f}(\cdot)$

Calibration: $S(Y_i, \hat{f}(X_i))$

Pros: Only ***a single model fit*** is required

Cons: ***Less data*** is available for both training &calibration

Trade-off! ***Statistical Efficiency vs Computational Efficiency***

What has been done?

Returning to the Full CP, what if we ***replace*** $\boldsymbol{y}$ ***with an arbitrary guess*** $\tilde{y}$?

Available data | **Test point**

$(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ | $(X_{n+1}, \tilde{y})$

Train model $\hat{f}^{\tilde{y}}(\cdot)$ **& Calibrate scores** $S(Y_i, \hat{f}^{\tilde{y}}(X_i))$ **: No need to retrain**

Ndiaye (2022) proposed **Replace-One Stable CP (RO-StabCP)**:

"A single label replacement yields ***controllable changes***."

What has been done?

Returning to the Full CP, what if we ***replace*** $\boldsymbol{y}$ ***with an arbitrary guess*** $\tilde{y}$?

Available data **Test point**

$(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ $(X_{n+1}, \tilde{y})$

Train model $\hat{f}^{\tilde{y}}(\cdot)$ & Calibrate scores $S(Y_i, \hat{f}^{\tilde{y}}(X_i))$: No need to retrain

Definition (Algorithmic Stability: Replace-One)

*A model $\hat{f}$ is **replace-one stable**, if for all i, there exists τ^{RO} such that*

$$\sup_{z,y,\tilde{y}} \left| S\left(z, \hat{f}^{y}(X_i)\right) - S\left(z, \hat{f}^{\tilde{y}}(X_i)\right) \right| < \tau^{RO}.$$

ICLR International Conference On Learning Representations

Remaining Issue

The Ohio State University

However, in real-world scenarios, we often have ***more than one test point***.

Available data $\quad\quad$ **Multiple test points**

$$(X_1, Y_1), \dots, (X_n, Y_n) \quad (X_{n+1}, \cancel{Y_{n+1}}), \dots, (X_{n+m}, \cancel{Y_{n+m}})$$

Train model $\hat{f}_1^{\tilde{y}_1}(\cdot)$ & Calibrate scores $S(Y_i, \hat{f}_1^{\tilde{y}_1}(X_i))$: Only for Y_{n+1}

In the above case, **RO-StabCP** requires $\boldsymbol{m}$ model trainings since ***the model still includes*** $\boldsymbol{X_{n+j}}$.

Leave-One-Out Stable CP

"What if we train a model <u>without any test point information</u>?"

Available data — $(X_1, Y_1), \dots, (X_n, Y_n)$

Multiple test points — $(X_{n+1}, \cancel{Y_{n+1}}), \dots, (X_{n+m}, \cancel{Y_{n+m}})$

Train model $\hat{f}(\cdot)$

Calibrate scores $\boldsymbol{S}(\boldsymbol{Y}_i, \hat{\boldsymbol{f}}(\boldsymbol{X}_i))$

: For every $Y_{n+j},\ j = 1, \dots, m$

"**Excluding a single point** still yields *controllable changes*."

Leave-One-Out Stable CP

"What if we train a model <u>without any test point information</u>?"

Available data $\overbrace{(X_1, Y_1), \ldots, (X_n, Y_n)}$

Multiple test points $(X_{n+1}, \cancel{Y_{n+1}}), \ldots, (X_{n+m}, \cancel{Y_{n+m}})$

Train model $\hat{f}(\cdot)$

Calibrate scores $S(Y_i, \hat{f}(X_i))$

: For every $Y_{n+j},\ j = 1, \ldots, m$

Definition (Algorithmic Stability: Leave-One-Out)

*A model $\hat{f}$ is **leave-one-out stable**, if for all i, there exists τ^{LOO} such that*

$$\sup_{z,y} \left| S\left(z, \hat{f}^{y}(X_i)\right) - S\left(z, \hat{f}(X_i)\right) \right| < \tau^{LOO}.$$

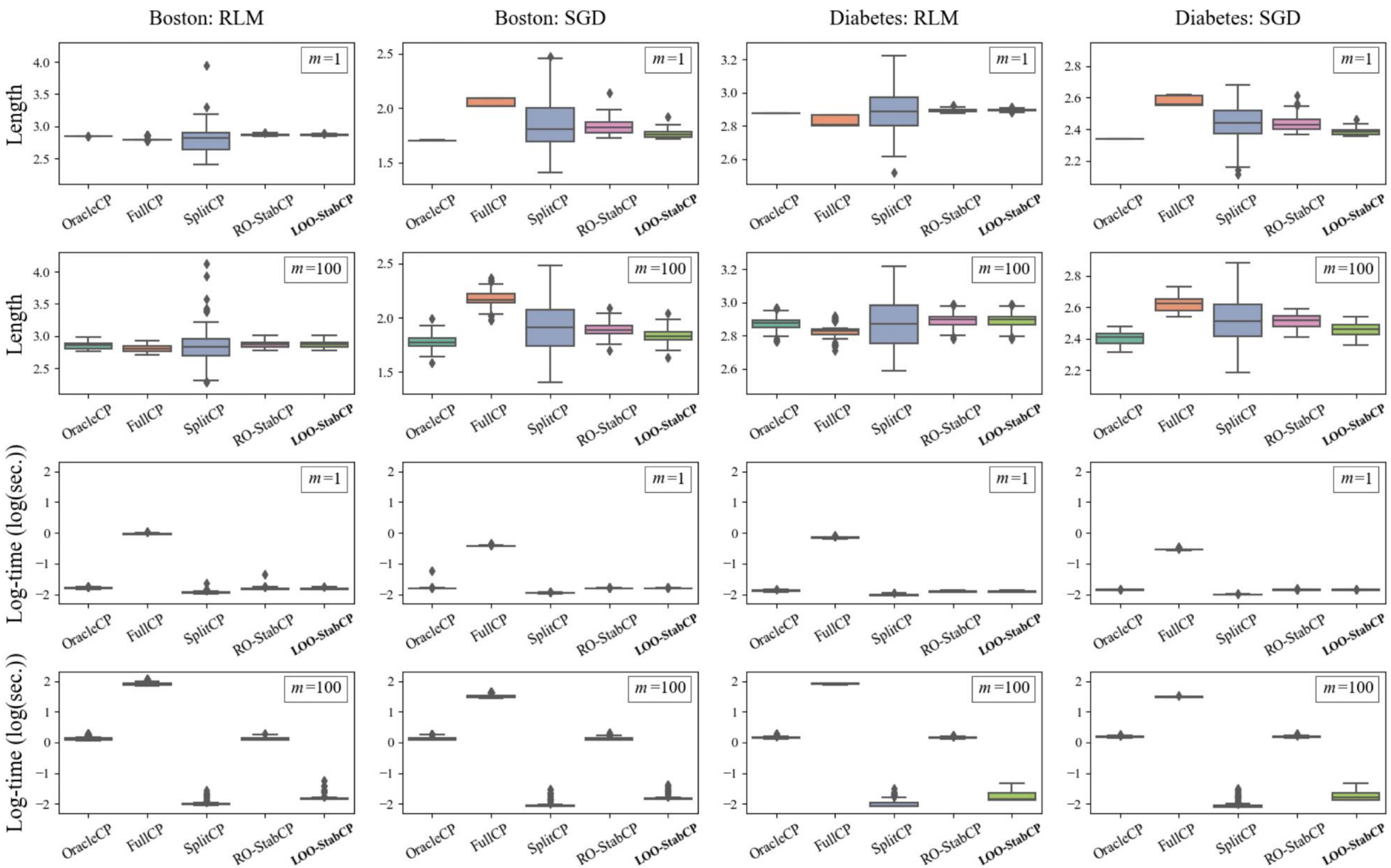
Computational Complexity

Method	# Model Fits	# Prediction Calls	Stable
FullCP	$\|\mathcal{Y}\| \cdot m$	$(n+1) \cdot \|\mathcal{Y}\| \cdot m$	✓
SplitCP	1	$n+m$	x
RO-StabCP	m	$(n+1) \cdot m$	✓
LOO-StabCP	**1**	$\boldsymbol{n+m}$	✓

ICLR International Conference On Learning Representations

Experimental Results

THE OHIO STATE UNIVERSITY



Summary

No Data-Split

LOO-StabCP

One Model

Multiple Tests

“One step closer to real-world deployment.”