Conformal prediction set for time-series

Chen Xu, Yao Xie
Industrial and Systems Engineering (ISyE)
Georgia Institute of Technology
cxu310@gatech.edu, yao.xie@isye.gatech.edu

ICML International Conference On Machine Learning

Workshop on Distribution-Free Uncertainty Quantification

1. Introduction

Goal and Challenges: Construct prediction sets for categorical time-series observations that arrive in sequence. The algorithm is primarily motivated by the **EnbPI** method [2]. In general, this is a challenging problem because both the conditional distribution $Y_t \,|\, X_t$ and the dependency among variables can be arbitrarily complex.

Contribution: Along the line of extending conformal prediction beyond exchangeable data, we develop *Ensemble Regularized Adaptive Prediction Set* (**ERAPS**). The benefits are

- Computationally and empirically, ERAPS allows arbitrary dependency among the features X_t and response Y_t. It can outperform non-ensemble-based ones based on numerical experiments.
- Theoretically bound coverage gaps and verify convergence of the estimated prediction set without assuming data exchangeability. Theoretical results hold for arbitrary definitions of non-conformity scores.

2. Problem Setup

Setup: Let $(X_t, Y_t), t \geq 1$ be a collection of random variables. $X_t \in \mathbb{R}^d, Y_t \in \{1,\ldots,K\}$. Denote $\pi := P_{Y|X}$ as the true conditional distribution function. Given T observations $\{(X_t, Y_t)\}_{t=1}^T$, one can construct an estimator $\hat{\pi} := \mathscr{A}(\{(X_t, Y_t)\}_{t=1}^T)$ using a classification algorithm $\mathscr{A}$ (e.g., neural network). Then, besides making a point prediction $\hat{Y}_t := \arg\max_{c\in[K]} \hat{\pi}(X_t, c)$, we want to construct a prediction set $C(X_t, \alpha)$ such that

$$\mathbb{P}(Y_t \in C(X_t, \alpha) \,|\, X_t) \geq 1 - \alpha, \tag{1}$$

which is stronger than the marginal coverage requirement

$$\mathbb{P}(Y_t \in C(X_t, \alpha)) \geq 1 - \alpha\,. \tag{2}$$

When observations are *exchangeable*, (2) always holds via existing conformal prediction techniques. However, (1) has been shown to be impossible in finite sample (unless trivially outputting $C(X_t, \alpha) = [K]$) without further assumptions.

3. ERAPS Procedure

In short, **ERAPS** first fits ensemble classifiers in an efficient “leave-one-out" (LOO) fashion and then compute the non-conformity scores introduced in [3]. During prediction on test data, it slides the past non-conformity scores forward to be more adaptive.

(Inputs: classification algorithm $\mathscr{A}$, training data $\{(X_t, Y_t)\}_{t=1}^T$, aggregation function ϕ (e.g., mean))

1. *[Bootstrap estimators]* Denote B as the total number of bootstrap models. Then, compute $\hat{\pi}^b$ as the b-th bootstrap estimator by applying $\mathscr{A}$ onto $\{(X_i, Y_i)\}_{i=b_1}^{b_T}$, where $S_b := \{b_1, \ldots, b_T\}$ denotes indices randomly sampled with replacement from $\{1,\ldots,T\}$.
2. *[LOO aggregation]* For each $t = 1,\ldots,T$, compute $\hat{\pi}_{-t} := \phi(\{\hat{\pi}^b : t \notin S_b\}_{b=1}^B$. In works, aggregate all bootstrap estimators whose training set does not contain (X_t, Y_t).
3. *[Non-conformity scores]* For each $t = 1,\ldots,T$, compute the non-conformity score $\hat{s}_t(Y_t)$ given penalty parameters $\{\lambda, k_{reg}\}$ as follows, which is proposed in [3]:

$$\hat{s}_t(Y_t) := m(Y_t) + \hat{\pi}_{-t}(X_t, Y_t) \cdot U_t + \lambda(r(Y_t) - k_{reg})^+ \tag{3}$$

$$m(c) := \sum_{c'=1}^{K} \hat{\pi}_{-t}(X_t, c')e_t(c'c), \qquad r(c) := |\sum_{c'=1}^{K} e_t(c', c)| + 1$$

$$e_t(c', c) := 1(\hat{\pi}_{-t}(X_t, c') > \hat{\pi}_{-t}(X_t, c))$$

In words, $m(c)$ denotes cumulative probabilities of classes more “likely" than c and $r(c)$ denotes the rank of c.

4. *[Prediction]* Given new feature X_t, compute $\hat{\pi} := \phi(\{\hat{\pi}_{-t}\}_{t=1}^T)$. Then

$$C(X_t, \alpha) := \{c \in [K] : \hat{s}_t(c) \leq Q_{1-\alpha}(\{\hat{s}_j(Y_j)\}_{j=1}^T)$$

For the empirical quantile $Q_{1-\alpha}$. Slide and re-index the past T non-conformity scores using $\hat{s}_t(Y_t)$ after observing Y_t.

4. Theoretical Guarantee

The guarantee holds for arbitrary definition of the non-conformity score, but we illustrate the idea under the definition (3) at the first prediction index $t = T + 1$. Let s_t be defined in (3) by replacing $\hat{\pi}_{-t}$ with π, the true conditional distribution function of $Y_t \,|\, X_t$. Assume

1. [Estimation quality] There exists a real sequence γ_t such that $\sum_{t=1}^{T} (\hat{s}_t - s_t)^2/T \leq \gamma_t$
2. [Score dependency] $\{s_t\}_{t=1}^T$ are independent and identically distributed, with its CDF being Lipschitz.

Theorem (Coverage; Informal): $\mathbb{P}(Y_t \notin C(X_t, \alpha) \,|\, X_t) \in \mathcal{O}(\sqrt{\log(16T)/T} + \gamma_T^{2/3})$

Theorem (Width; Informal): Under stronger assumptions on estimation of s_t and the relationship between s_t and π, $C(X_t, \alpha)$ converges to the oracle prediction set as $T \to \infty$

5. Experiment

Result: We compare our **ERAPS** against **SRAPS** in [3], **SAPS** in [4], and Naive (i.e., top k most probable labels whose cumulative probabilities exceed $1 - \alpha$). We compare the marginal coverage in (3) and size of prediction sets under different α, where Table 1 shows that ERAPS maintains valid coverage at all times, with smaller set sizes than other methods on all but the PenDigits dataset. We also compare **ERAPS** against **SRAPS** upon conditioning on the label of Y_t, where Table 2 shows that **ERAPS** tends to maintain valid class-conditional coverage at most classes with similar set sizes, whereas **SRAPS** can be too conservative.

Table 1: Marginal coverage and set size on different datasets, where we choose $\alpha \in \{0.05, 0.075, 0.1, 0.15, 0.2\}$. We see that ERAPS almost always maintains valid coverage with smaller set sizes.

α	0.05		0.075		0.1		0.15		0.2	
Pedestrain	coverage	set size	coverage	set size	coverage	set size	coverage	set size	coverage	set size
ERAPS	0.94	1.69	0.92	1.18	0.90	1.04	0.85	0.96	0.81	0.91
SRAPS	0.95	4.09	0.94	3.25	0.92	3.00	0.89	2.02	0.82	1.17
SAPS	0.95	4.29	0.93	3.77	0.91	3.00	0.86	2.16	0.81	1.86
Naive	0.87	1.60	0.84	1.47	0.81	1.37	0.75	1.22	0.71	1.10
α	0.05		0.075		0.1		0.15		0.2	
Crop	coverage	set size	coverage	set size	coverage	set size	coverage	set size	coverage	set size
ERAPS	0.96	4.68	0.93	3.53	0.90	2.87	0.86	2.22	0.82	1.80
SRAPS	0.95	5.31	0.93	4.23	0.91	3.40	0.86	2.58	0.81	2.19
SAPS	0.95	4.51	0.93	3.72	0.90	3.32	0.86	2.79	0.82	2.35
Naive	0.96	4.65	0.94	3.97	0.92	3.50	0.88	2.88	0.83	2.46
α	0.05		0.075		0.1		0.15		0.2	
Pen digit	coverage	set size	coverage	set size	coverage	set size	coverage	set size	coverage	set size
ERAPS	0.94	0.96	0.91	0.93	0.89	0.91	0.84	0.86	0.79	0.81
SRAPS	0.92	0.95	0.90	0.93	0.88	0.91	0.83	0.86	0.78	0.81
SAPS	0.94	1.03	0.92	1.00	0.90	0.96	0.84	0.89	0.79	0.83
Naive	0.94	1.02	0.92	0.98	0.89	0.95	0.84	0.88	0.79	0.83

Table 2: MelbournePedestrian data with 10 classes: Class-conditional coverage (top two tables) and set size (bottom two tables) of ERAPS and SRAPS, where we condition on different road types. Both methods maintain valid coverage, but ERAPS produces much smaller sets than SRAPS at each class.

ERAPS: Conditional coverage for different pedestrian road types										
α	class 0	class 1	class 2	class 3	class 4	class 5	class 6	class 7	class 8	class 9
0.05	0.96	0.98	0.94	0.99	0.99	0.91	1.00	0.83	0.89	0.95
0.075	0.95	0.98	0.89	0.99	0.98	0.88	1.00	0.76	0.86	0.93
0.1	0.95	0.97	0.78	0.98	0.97	0.86	0.97	0.72	0.82	0.93
0.15	0.94	0.92	0.74	0.92	0.89	0.83	0.89	0.69	0.77	0.89
0.2	0.88	0.86	0.73	0.84	0.84	0.77	0.84	0.64	0.76	0.86
SRAPS: Conditional coverage for different pedestrian road types										
α	class 0	class 1	class 2	class 3	class 4	class 5	class 6	class 7	class 8	class 9
0.05	0.98	0.97	0.99	1	0.98	0.82	1	0.76	0.99	0.98
0.08	0.98	0.94	0.98	1	0.97	0.8	1	0.68	0.97	0.98
0.1	0.98	0.93	0.95	1	0.95	0.75	1	0.66	0.9	0.95
0.15	0.95	0.91	0.84	1	0.94	0.67	0.98	0.65	0.7	0.86
0.2	0.94	0.9	0.73	1	0.92	0.59	0.98	0.64	0.61	0.8
ERAPS: Conditional set size for different pedestrian road types										
α	class 0	class 1	class 2	class 3	class 4	class 5	class 6	class 7	class 8	class 9
0.05	2.03	2.14	2.17	1.93	1.07	1.22	1.00	1.35	2.04	2.01
0.075	1.15	1.21	1.66	1.07	1.01	1.07	1.00	1.14	1.52	1.11
0.1	1.05	1.05	1.16	1.02	0.99	1.00	0.97	1.00	1.19	1.05
0.15	1.00	0.96	1.00	0.95	0.91	0.98	0.89	0.98	1.00	0.97
0.2	0.93	0.90	0.99	0.87	0.86	0.91	0.84	0.92	0.99	0.93
SRAPS: Conditional set size for different pedestrian road types										
α	class 0	class 1	class 2	class 3	class 4	class 5	class 6	class 7	class 8	class 9
0.05	4	4.28	4.52	4	4.08	4.1	4	4.4	4.96	4
0.08	3.05	3.3	3.94	3	3.41	3.66	3.07	3.86	4	3.06
0.1	2.06	2.3	2.94	2	2.36	2.69	2.12	2.78	3	2.12
0.15	1.06	1.29	1.8	1.02	1.26	1.55	1.06	1.54	2	1.16
0.2	1	1.24	1.27	1.01	1.05	1.12	1.01	1.34	1.83	1

References

[1] Xu, Chen and Yao Xie. “Conformal prediction set for time-series” In: ICML 2022 DFUQ workshop

[2] Xu, Chen and Yao Xie. “Conformal prediction interval for dynamic time-series” Oral paper/long talk in ICML 2021

[3] Anastasios N. Angelopoulos, Stephen Bates, Jitendra Malik, & Michael I. Jordan. “Uncertainty Sets for Image Classifiers using Conformal Prediction” Spotlight in ICLR 2021

[4] Yaniv Romano, Matteo Sesia, and Emmanel Candes. “Classification with Valid and Adaptive Coverage” Spotlight in NeurIPS 2020