Improving Expert Predictions with Conformal Prediction - ICML 2023

Shin Yun Seop

February 25, 2025

Seoul national university, statistics, IDEA LAB

Topic

Improving Expert Predictions with Conformal Prediction

Eleni Straitouri [1] Lequn Wang [2] Nastaran Okati [1] Manuel Gomez Rodriguez [1]

Contents

Problem Formulation

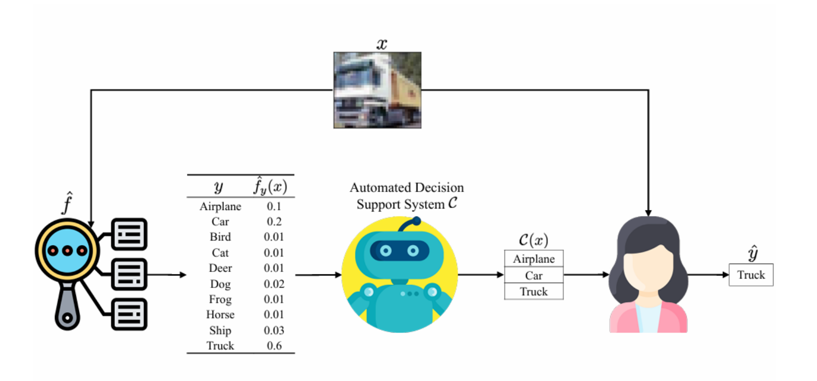

Figure 1: The human expert receives the recommended subset $\mathcal{C}(x)$, together with the sample, and predicts a label $\hat{y}$ from $\mathcal{C}(x)$ according to a policy $\pi(x, \mathcal{C}(x))$.

Problem Formulation: Notation

- $x \in \mathcal{X}$: feature vector with $x \sim P(X)$.
- $y \in \mathcal{Y} = \{1, \cdots, n\}$: label with $y \sim P(Y|X)$.
- $\mathcal{D}_{Prop}$: Proper training set having n data.
- $\mathcal{D}_{Cal}$: Calibration set having m data.
- $\mathcal{D}_{est}$: Estimation set having m data.
- $\hat{f} : \mathcal{X} \to [0, 1]^{|\mathcal{Y}|}$: trained classifier.

Problem Formulation: Notation

- $\mathcal{C} : \mathcal{X} \to 2^{\mathcal{Y}}$: Automated Decision Support System with $\mathcal{C}(x) \subset \mathcal{Y}$ using a trained classifier $\hat{f}$.
- $\Delta(\mathcal{Y})$: Probability simplex over the set of labels $\mathcal{Y}$.
- $\pi : \mathcal{X} \times 2^{\mathcal{Y}} \to \Delta(\mathcal{Y})$: expert's prediction policy.
- $\mathbb{P}[\hat{Y} = Y; \mathcal{C}]$: the expert's success probability if the human expert predicts a label $\hat{Y}$ among those in the subset $\mathcal{C}(x)$.

Problem Formulation: Goal

- Author want expert can only benefit from using the automated decision support system $\mathcal{C}$, i.e.,

$$\mathbb{P}[\hat{Y} = Y \mid \mathcal{C}] \geq \mathbb{P}[\hat{Y} = Y \mid \mathcal{Y}] \tag{1}$$

- Among those systems satisfying Eq (1), would like to find the system $\mathcal{C}^*$ that helps the experts achieve the highest success probability, i.e.,

$$\mathcal{C}^* = \arg\max_{\mathcal{C}} \mathbb{P}[\hat{Y} = Y \mid \mathcal{C}]. \tag{2}$$

- To address the design of such a system, we will look at the problem from the perspective of conformal prediction.

Subset Selection using Conformal Prediction

- If the classifier $\hat{f}$ trained by $\mathcal{D}_{Prop}$, let $s_i = 1 - \hat{f}(x_i)$, $(x_i, y_i \in \mathcal{D}_{Cal},\ i = 1, \cdots, m)$ be conformal score.
- And $\hat{q}_\alpha$ is the $\frac{\lceil (m+1)(1-\alpha) \rceil}{m}$ empirical quantile of the conformal scores.
- Then, if construct the subsets $C_\alpha(X)$ for new data samples as follows:

$$\mathcal{C}_\alpha(X) = \{y \mid s(X, y) \leq \hat{q}_\alpha\}. \tag{3}$$

Subset Selection using Conformal Prediction

Theorem 1

For an automated decision support system $\mathcal{C}_\alpha$ that constructs the subsets $\mathcal{C}_\alpha(X)$ using Eq. (3), it holds that

$$1 - \alpha \leq \mathbb{P}[Y \in \mathcal{C}_\alpha(X)] \leq 1 - \alpha + \frac{1}{m+1},$$

where the probability is over the randomness in the sample it helps predicting and the calibration set used to compute the empirical quantile $\hat{q}_\alpha$.

Proof : Refer to Appendix D in Angelopoulos and Bates (2021).

Subset Selection using Conformal Prediction

- The problem is that, how to choose the α?
- If α is too small then, $\mathcal{C}_\alpha(X)$ is too large so that it is useless.
- But, if α is too large then, it is more probability $\mathcal{C}_\alpha(X)$ give wrong answer.
- So, author suggest the way to choose optimal α.

Subset Selection using Conformal Prediction

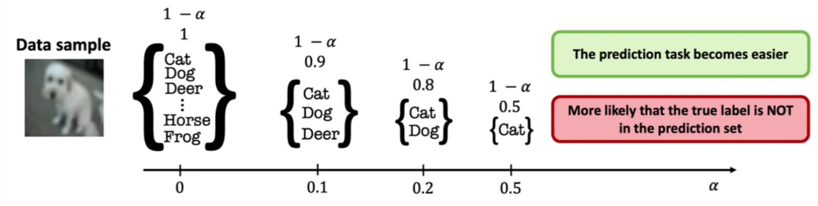

Figure 2: Relationship between α and $\mathcal{C}_\alpha(X)$.

Contents

Basic concept

- To find the optimal conformal predictor that maximizes the expert's success probability, we need to solve the following maximization problem:

$$\alpha^* = \arg\max_{\alpha \in \mathcal{A}} \mathbb{P}[\hat{Y} = Y \mid \mathcal{C}_\alpha], \tag{4}$$

where $\mathcal{A} = \{\alpha_i\}_{i \in [m]}$, with $\alpha_i = 1 - i/(m+1)$.

- To solve optimization problem (4) we need some assumption.
- Assume that author can access to (an estimation of) the confusion matrix **C** for the expert predictions in the (original) multiclass classification task i.e.,

$$\mathbf{C} = [C_{yy'}]_{y,y' \in \mathcal{Y}}, \quad \text{where } C_{yy'} = \mathbb{P}[\hat{Y} = y' \mid Y = y].$$

- Moreover, given a sample (x, y), we assume that the expert's conditional success probability for the subset $\mathcal{C}_\alpha(x)$ is given by

$$\mathbb{P}[\hat{Y} = y; \mathcal{C}_\alpha \mid y \in \mathcal{C}_\alpha(x)] = \frac{C_{yy}}{\sum_{y' \in \mathcal{C}_\alpha(x)} C_{yy'}}. \tag{5}$$

Estimation

- Then, we can compute a Monte-Carlo estimator $\hat{\mu}_\alpha$ of the expert's success probability $\mathbb{P}[\hat{Y} = Y; \mathcal{C}_\alpha]$ using the above conditional success probability $\mathbb{P}[\hat{Y} = y; \mathcal{C}_\alpha \mid y \in \mathcal{C}_\alpha(x)]$ and an estimation set $\mathcal{D}_{\text{est}} = \{(x_i, y_i)\}_{i \in [m]}$, i.e.,

$$\hat{\mu}_\alpha = \frac{1}{m} \sum_{i \in [m] \mid y_i \in \mathcal{C}_\alpha(x_i)} \mathbb{P}[\hat{Y} = y_i; \mathcal{C}_\alpha \mid y_i \in \mathcal{C}_\alpha(x_i)]. \tag{6}$$

- Then $\mathbb{E}(\hat{\mu}_\alpha) = \mathbb{P}[\hat{Y} = Y; \mathcal{C}_\alpha]$ and $\mathbb{P}[\hat{Y} = y_i; \mathcal{C}_\alpha \mid y_i \in \mathcal{C}_\alpha(x_i)]$ is in $[0, 1]$, we can apply Hoeffding's inequality.

Lemma 1 (Hoeffding's Inequality)

Let $Z_1, \ldots, Z_k$ be i.i.d., with $Z_i \in [a, b], i = 1, \ldots, k$, $a < b$ and $\hat{\mu}$ be the empirical estimate

$$\hat{\mu} = \frac{\sum_{i=1}^{k} Z_i}{k}$$

of $\mathbb{E}[Z] = \mathbb{E}[Z_i]$. Then:

$$\mathbb{P}[|\hat{\mu} - \mathbb{E}[Z] \geq \epsilon|] \leq 2 \exp\left(\frac{-2k\epsilon^2}{(b-a)^2}\right) \tag{7}$$

hold for all $\epsilon \geq 0$.

Estimation

Theorem 2

Under E.q. (6) following inequalities hold,

$$\mathbb{P}\left(\left|\hat{\mu}_\alpha - \mathbb{P}[\hat{Y} = Y; \mathcal{C}_\alpha]\right| \leq \epsilon_\delta\right) \geq 1 - \delta, \text{ for each } \alpha \in \mathcal{A} \tag{8}$$

and,

$$\mathbb{P}\left(\max_{\alpha \in \mathcal{A}}\left|\hat{\mu}_\alpha - \mathbb{P}[\hat{Y} = Y; \mathcal{C}_\alpha]\right| \leq \epsilon_{\delta/m}\right) \geq 1 - \delta \tag{9}$$

and, , where $\epsilon_\delta = \sqrt{\frac{\log \frac{1}{\delta}}{2m}}$

Proof: Use Hoeffding's Inequality to proof inequality (8) and Bonferroni correction technique to proof inequality (9).

Optimization

- Inequality (9) means that, with probability at least $1-\delta$, it holds that $\mathbb{P}[\hat{Y} = Y; \mathcal{C}_\alpha] \geq \hat{\mu}_\alpha - \epsilon_{\delta/m}$, $\forall \alpha \in \mathcal{A}$ simultaneously.
- *For any* $\delta \in (0,1)$, *consider an automated decision support system* $C_{\hat{\alpha}}$ *with*

$$\hat{\alpha} = \underset{\alpha \in \mathcal{A}}{\arg\max} \left(\hat{\mu}_\alpha - \epsilon_{\delta/m} \right). \tag{10}$$

- This paper does not explicitly mention why $\hat{\alpha}$ is determined in this way, but I think it means maximizing the minimum value of the probability $\mathbb{P}[\hat{Y} = Y; \mathcal{C}_\alpha]$ to be estimated.
- $\hat{\alpha}$ can be obtain we calculate $m'th$ $\hat{\mu}_\alpha - \epsilon_{\delta/m}$ for all $\alpha \in \mathcal{A}$.

Algorithm 1 Finding a near-optimal $\hat{\alpha}$

```
Require: f̂, D_est, D_cal, δ, m
Initialize: A = {}, α̂ ← 0, t ← 0
for i = 1, ..., m do
    α ← 1 − i/(m+1)
    A ← A ∪ {α}
end for
for α ∈ A do
    μ_α, ε_δ/m ← ESTIMATE(α, δ, D_est, D_cal, f̂)
    if t ≤ μ_α − ε_δ/m then
        t ← μ_α − ε_δ/m
        α̂ ← α
    end if
end for
return α̂
```

Experiment

	Classifier	Expert using $\mathcal{C}_{\hat{\alpha}}$
ResNet-110	0.928	0.981
PreResNet-110	0.944	0.983
DenseNet	0.964	0.987

Figure 3: Testing on the CIFAR-10H dataset.

Contents

Reference

1. Straitouri, Eleni, et al. "Improving expert predictions with conformal prediction." International Conference on Machine Learning. PMLR, 2023.
2. Angelopoulos, Anastasios N., and Stephen Bates. "A gentle introduction to conformal prediction and distribution-free uncertainty quantification." arXiv preprint arXiv:2107.07511 (2021).
3. Kerrigan, Gavin, Padhraic Smyth, and Mark Steyvers. "Combining human predictions with model probabilities via confusion matrices and calibration." Advances in Neural Information Processing Systems 34 (2021): 4421-4434.