Data Analytics and Machine Learning Group
Department of Computer Science
Technical University of Munich

Provably Reliable Conformal Prediction Sets in the Presence of Data Poisoning

Yan Scholten & Stephan Günnemann
Technical University of Munich

DAAD Zuse Schools
Konrad Zuse Schools of Excellence in Artificial Intelligence

tl;dr: Provably reliable conformal prediction sets (RPS)
- Pointwise reliability of conformal prediction sets under poisoning
- Adversaries can manipulate training and calibration data to alter prediction sets by (1) modifying, adding or deleting datapoints, and (2) by flipping labels
- We propose the first approach towards more reliable prediction sets and derive strong certificates that guarantee reliability under data poisoning

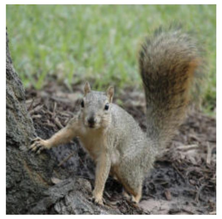
Context

- Conformal prediction provides prediction sets guaranteed to include the ground truth with any user-specified probability
- Machine learning models are susceptible to data poisoning attacks

Problem

Conformal prediction sets are not pointwise reliable under poisoning attacks, where adversaries manipulate both the training and calibration data by modifying, adding or deleting datapoints, or by flipping labels.

Test image x_{n+1}



Prediction set using clean data

$$\mathcal{C}(\mathrm{x}_{n+1}) = CP(D_{train}, D_{calib}, x_{n+1}) = \{squirrel\}$$

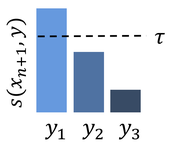
Prediction set using perturbed data

$$\widetilde{\mathcal{C}}(\mathrm{x}_{n+1}) = CP(\widetilde{D}_{train}, \widetilde{D}_{calib}, x_{n+1}) = \{marmot, dog\}$$

How can we make conformal prediction sets provably reliable in the presence of data poisoning?

Background: Conformal prediction

1. Train classifier $f: X \to Y$ on training set D_{train}
2. Compute conformal scores on the calibration set D_{calib} using a score function $s(x, y)$ to measure conformity
3. Compute empirical quantile of conformal scores S:
$$\tau = Quant(\alpha; S)$$
for user-specified significance level α
4. Given test image x_{n+1}, construct prediction set
$$\mathcal{C}(\mathrm{x}_{n+1}) = \{y \in \mathcal{Y} : s(x_{n+1}, y) \geq \tau\}$$



Marginal coverage guarantee

If $(x_{n+1}, y_{n+1}) \in D_{test}$ is exchangeable with D_{calib}, then

$$\Pr\left[y_{n+1} \in \mathcal{C}(\mathrm{x}_{n+1})\right] \geq 1 - \alpha$$

Majority prediction sets against calibration poisoning

1. Split the calibration data into k_c disjoint partitions
2. Compute conformal prediction sets $\mathcal{C}_i(\mathrm{x}_{n+1})$ using each calibration partition
3. Construct a majority prediction set $\mathcal{C}^M(x_{n+1})$ using quantile function $\hat{\tau}(\alpha)$ of the Binomial distribution $Bin(k_c, 1-\alpha)$

D_{calib}

P_1	$\mathcal{C}_1(\mathrm{x}_{n+1}) = \{y : s(x_{n+1}, y) \geq \tau_1\}$
P_2	$\mathcal{C}_2(\mathrm{x}_{n+1}) = \{y : s(x_{n+1}, y) \geq \tau_2\}$
$\vdots$	$\vdots$
P_{k_c}	$\mathcal{C}_{k_c}(\mathrm{x}_{n+1}) = \{y : s(x_{n+1}, y) \geq \tau_{k_c}\}$

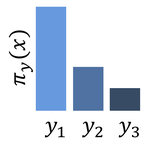
$$\mathcal{C}^M(x_{n+1}) = \{y : S(y) > \hat{\tau}(\alpha)\}$$

$$S(y) \triangleq \sum_{i=1}^{k_c} \mathbb{I}\{y \in \mathcal{C}_i(\mathrm{x}_{n+1})\}$$

For clean datasets of independent datapoints, the majority prediction set achieves marginal coverage: $\Pr\left[y_{n+1} \in \mathcal{C}^M(\mathrm{x}_{n+1})\right] \geq 1 - \alpha$

Smoothed score functions against training poisoning

1. Split the training data into k_t disjoint partitions
2. Train k_t classifiers $f^{(i)}$ separately on each partition
3. Construct a voting function $\pi_y(x) = \frac{1}{k_t}\sum_{i=1}^{k_t} \mathbb{I}\{f^{(i)}(x) = y\}$
4. Construct a score function by smoothing the voting function:
$$s(x, y) = e^{\pi_y(x)} / (\textstyle\sum_{i=1}^{K} e^{\pi_i(x)})$$



Additional softmax to resolve ties between scores deterministically

How to define reliability?

We are interested in certifying subset relationships and denote prediction sets
- coverage reliable, if we can guarantee $C(x_{n+1}) \subseteq \tilde{C}(x_{n+1})$,
- size reliable, if we can guarantee $C(x_{n+1}) \supseteq \tilde{C}(x_{n+1})$, and
- robust, if we can guarantee both "$\subseteq$" and "$\supseteq$".

How to certify reliability?

- Assume worst-case scenario: Each perturbed datapoint changes the prediction to the worst-case for at most one partition
- Since all votes are discrete, we can directly quantify the worst-case scores, worst-case quantiles, and worst-case counts in the majority prediction set $\mathcal{C}^M$

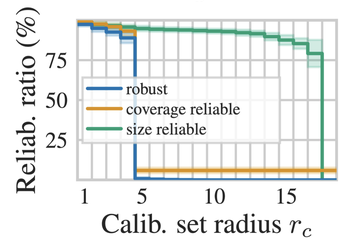
Our approach yields provably reliable prediction sets even under worst-case data poisoning and exchangeability violations described by our threat model

Experimental evaluation

Setting: ResNet18 on CIFAR10, $\alpha = 0.1$

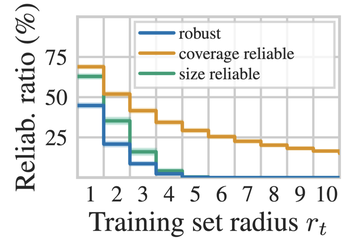
Reliability under calibration poisoning

Empirical coverage of 90.2% and average set size of 0.94 ($k_c = 22$)



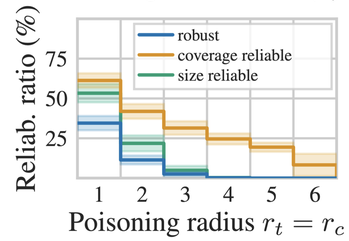
Reliability under training poisoning

Empirical coverage of 90.7% and average set size of 3.18 ($k_t = 100$)

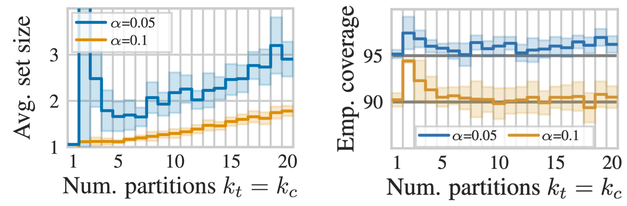


Reliability under training & calibration poisoning

Empirical coverage of 92% and avg. set size of 3.41 ($k_t = 100, k_c = 40$)



Average set size and empirical coverage



Paper, code, and more

SPONSORED BY THE
Federal Ministry of Education and Research