Learning calibrated belief functions from conformal predictions
Vitor Martin Bordini, Sebastien Destercke and Benjamin Quost
CNRS, UMR 7253 Heudiasyc, UTC

Objectives
- Set $L_1$ learns Classifier $h_1$, Unlabeled data, confident data for $h_1$
- Set $L_2$ learns Classifier $h_2$, confident data for $h_2$

Problem
- Common problem on machine learning predictions: poor calibration.
- Calibration definition: The level of confidence actually reflects the chance that the associated output turns out to be true.
- Inductive Conformal Prediction (ICP) is a possible solution to this problem.
- What is the relation (if any) between ICPs and Belief Functions?

Possibility theory
- Possibility distribution: $\pi: \Omega \rightarrow [0, 1]$
- Necessity Measure $N$ and Possibility Measure $P$ (equivalent to Belief and Plausibility Functions, respectively): $P(B) = 1 - N(B^C) = \max_{\pi(x) \in B} \pi(x), B \subseteq \Omega$.
- Limitation: we can only extract imprecise probabilities from a possibility distribution.
- $\alpha$-cut: $\pi_{\alpha} = \{ x \in \Omega | \pi(x) \geq \alpha \}$
- Property: $P(\pi_{\alpha}) \geq 1 - \alpha$.
- In the example below, we can compute the mass functions as $m(y_1) = p_{\pi_{y_1}} - p_{\pi_0}$ and $m(y_2, y_3) = p_{\pi_{y_2}}$.

Experiments
- CIFAR-10 dataset.
- Classifier: Fitnet backbone + softmax layer.
- Regressor: Fitnet backbone + linear layer with activation function $\phi(x) = \min(x, 1)$.
- $\pi(\phi) = 0$.
- Training parameters: batch size 25, learning rate 0.001, momentum 0.9 and an Adam optimizer.

P-values comparison
- Comparison between ICP and the regressor outputs.
- Calibration dataset = 10% of the test dataset.
- The Mean Square Root (MSR) and the R² coefficient are 0.02 and 0.8, respectively.
- Output averaged by the number of classes.
- Results for 200 outputs. Ideal result = Blue line.

Calibration size influence
- Goal = Change the calibration set size from 10 to 1000 instances and check how the accuracy of ICP and our algorithm grows.
- Regressor performs slightly worse after 100 instances, at most 2.5% below than the ICP, but has a better performance with less data.

Conclusion
- Calibration techniques make model predictions statistically valid.
- The ICP is a popular calibration technique but it is slower and requires more data.
- Our algorithm decrease the dependence of ICP on the calibration dataset while also being less computationally expensive and having similar performance.
- However, it still requires a minimum amount of data and takes more time to learn.
- Future works may solve this problem using co-learning techniques.

References
[3] Dmytro Mishkin and Jiri Matas. All you need is a good init, 11 2015.