How do the performance of a Conformal Predictor and its underlying algorithm relate?

Giovanni Cherubin GCHERUBIN@MICROSOFT.COM
Microsoft, Cambridge, United Kingdom

Editor: Harris Papadopoulos, Khuong An Nguyen, Henrik Boström and Lars Carlsson

Abstract

Conformal Prediction (CP) offers a shift on the traditional supervised classification paradigm. Whereas in supervised learning one generally aims to optimize the error of a classifier at predicting the label correctly (*prediction error*), in CP one aims to optimize the size of a prediction set (*efficiency*), where this set is guaranteed to contain the true label with probability $\geq 1 - \varepsilon$, for a user-defined $\varepsilon \in [0, 1]$. CP works as a wrapper around a traditional learning model; yet, it is unclear how the prediction error of the underlying model affects the efficiency of the CP. In this note, we study a simple class of CPs whose *efficiency* is proportional to the *prediction error* of the underlying model.

Keywords: conformal prediction, efficiency, prediction error

Notation. Consider a sequence of IID random variables $\{(X_1, Y_1), ..., (X_N, Y_N), (X, Y)\} \in (\mathcal{X} \times \mathcal{Y})^{N+1}$, sampled from an unknown distribution, and let $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$; we assume that $\mathcal{Y}$ is finite. A Conformal Predictor (CP) predicts a set of labels for the test object X, given access to $\{(X_i, Y_i)\}_{i=1}^N$; for a user-defined $\varepsilon \in [0, 1]$, the prediction set output by CP is guaranteed to contain the true label Y with at least $1 - \varepsilon$ probability (Vovk et al., 2005).

A CP is defined for a function $A : \mathcal{Z} \times \mathcal{Z}^* \to \mathbb{R}_{\geq 0}$, the *nonconformity measure*, which indicates how "strange" an example $z = (x, y) \in \mathcal{Z}$ looks like w.r.t. a multiset of examples $\mathcal{D} \in \mathcal{Z}^*$; we call $\mathcal{D}$ the *training set*, and write $A(z; \mathcal{D})$ to indicate that the nonconformity measure is computed for z given $\mathcal{D}$. Informally, $A(z; \mathcal{D})$ takes a small value if z looks similar to the points in $\mathcal{D}$, and a large value otherwise. For simplicity, we develop our results on split CP (Papadopoulos et al., 2002). In split CP (henceforth, CP), the training set $\mathcal{D}$ of the nonconformity measure $A(\cdot; \mathcal{D})$ must be independent from (X, Y) and $\{(X_i, Y_i)\}_{i=1}^N$; typically, it is sampled from the same data distribution.

Coarse nonconformity measures. We study a special class of nonconformity measures, which we call *coarse*. They are defined on the basis of a classifier $g_{\mathcal{D}} : \mathcal{X} \to \mathcal{Y}$, trained on the training set $\mathcal{D}$. For a chosen classifier g, a coarse nonconformity measure is defined as:

$$A^g((x, y); \mathcal{D}) = I(g_{\mathcal{D}}(x) \neq y)\,; \tag{1}$$

I is the indicator function, which takes value 1 if $g_{\mathcal{D}}(x) \neq y$, and 0 otherwise. While not as informative as typical nonconformity measures, which incorporate the belief of the classifier on its prediction, coarse measures prove a useful tool to study CP.

P-values of CPs with coarse nonconformity measures. Consider a CP with coarse nonconformity measure A^g, tasked with making a prediction for the test object X. For every possible label $\hat{y} \in \mathcal{Y}$, the CP computes a p-value as follows:

$$P_{\hat{y}} = \frac{\#\{i \in [N] : A^g((X_i, Y_i); \mathcal{D}) \geq A^g((X, \hat{y}); \mathcal{D})\} + 1}{N + 1}$$

where # is the cardinality of a set. A label $\hat{y}$ is included in the prediction set if $P_{\hat{y}} > \varepsilon$.

We analyze the value of $P_{\hat{y}}$ for CPs with coarse nonconformity measures, and then state two simple corollaries on their performance (efficiency). Proofs are omitted for brevity.

Proposition 1 *For any classifier $g : \mathcal{X} \to \mathcal{Y}$, the CP defined by the coarse nonconformity measure A^g outputs a p-value $P_{\hat{y}}$ for test object X such that*

$$P_{\hat{y}} = \begin{cases} 1 & \text{if} \quad g(X) = \hat{y} \\ \frac{N\hat{R}_g+1}{N+1} & \text{otherwise} \end{cases}$$

where $\hat{R}_g = \frac{1}{N}\sum_{i=1}^{N} I(g(X_i) \neq Y_i)$ is the empirical error of classifier g on $\{(X_i, Y_i)\}_{i=1}^{N}$.

Remark. This shows a direct correspondence between the error of the underlying classifier and that of the CP. In particular, for a large N, the p-value will be either $P_{\hat{y}} = 1$ or $P_{\hat{y}} \to R_g$, where R_g is the expected error of the classifier on the underlying distribution.

Efficiency of CPs with coarse nonconformity measures. Since a CP's error is guaranteed, the main parameter for judging its performance is the "tightness" of its prediction set; this is referred to as the *efficiency* of a CP. Various efficiency criteria have been considered in the past (Vovk et al., 2016). We study two: i) the *sum of p-values*, $E_{\mathcal{S}} = \sum_{\hat{y}\in\mathcal{Y}} P_{\hat{y}}$ (Corollary 2, Fig. 1), and ii) the *prediction set size*, $E_{\mathcal{N}}^{\varepsilon} = \#\{\hat{y} \in \mathcal{Y} : P_{\hat{y}} > \varepsilon\}$ (Corollary 3).

Corollary 2 (Sum of p-values) *Let $L = |\mathcal{Y}|$. The sum of p-values of a CP with coarse nonconformity measure A^g is $E_{\mathcal{S}} = 1 + (L-1)\frac{NR_g+1}{N+1}$.*

Corollary 3 (Prediction set size) *The prediction set of a CP with coarse nonconformity measure has size 1 for $\varepsilon \geq \frac{N\hat{R}_g+1}{N+1}$, and size L otherwise.*

Observe that the worst-case error is achieved when the classifier guesses the label uniformly at random; hence, $\hat{R}_g \leq U = \frac{L-1}{L}$ for all g. Combined with Corollary 3, we conclude that a CP with coarse nonconformity measure has perfect efficiency (i.e., $E_{\mathcal{N}}^{\varepsilon} = 1$) for $\varepsilon \geq \frac{NU+1}{N+1}$, regardless of the underlying classifier.

Conclusions and future work. We showed that the performance (efficiency) of CP is correlated with the performance (error rate) of its underlying classifier for the class of *coarse* nonconformity measures. A natural next step is to extend this analysis to more general (and informative) nonconformity measures, such as those deriving from the confidence of the classifier on its prediction; this can be done by replacing the 0-1 loss in Equation (1) with a loss that uses the confidence of the classifier. Obtaining tight bounds may be harder in this case, but we expect the same behavior to hold. Ultimately, we hope this line of work can lead to i) the ability to predict the efficiency of a CP before deployment, ii) a better understanding of what constitutes a good efficiency criterion for CP (Vovk et al., 2016), and in general iii) a deeper understanding of the connection between traditional learning and CP.

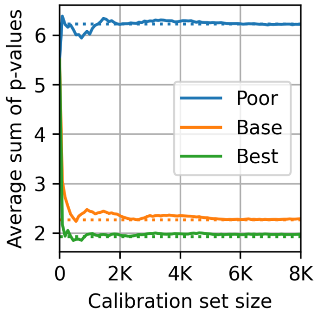

Figure 1: Sum of p-values ($E_{\mathcal{S}}$) on the MNIST dataset; average across 100 test points. Dotted line: asymptotic behavior implied by Corollary 2: $E_{\mathcal{S}} \to 1 + (L-1)R_g$. Results are for 3 coarse nonconformity measures (`scikit-learn` logistic regressors, default parameters), resp. trained on: 20 examples ("Poor"), 1K examples ("Base"), and on the entire training, calibration, and test data to observe near-optimal behavior ("Best").

References

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *Machine Learning: ECML 2002: 13th European Conference on Machine Learning Helsinki, Finland, August 19–23, 2002 Proceedings 13*, pages 345–356. Springer, 2002.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

Vladimir Vovk, Valentina Fedorova, Ilia Nouretdinov, and Alexander Gammerman. Criteria of efficiency for conformal prediction. In *Conformal and Probabilistic Prediction with Applications: 5th International Symposium, COPA 2016, Madrid, Spain, April 20-22, 2016, Proceedings 5*, pages 23–39. Springer, 2016.