Conformal Prediction without Nonconformity Scores

Abstract

Conformal prediction (CP) is an uncertainty quantification framework that allows for constructing statistically valid prediction sets. Key to the construction of these sets is the notion of nonconformity function, which assigns a real-valued score to individual data points: Only those (hypothetical) data points contribute to a prediction set that sufficiently conform to the data. The point of departure of this work is the observation that CP predictions are invariant against (strictly) monotone transformations of a nonconformity function. In other words, it is only the ordering of the scores that matters, not their quantitative values. Consequently, instead of scoring individual data points, a conformal predictor only needs to be able to compare pairs of data points, deciding which of them is the more conforming one. This suggests an interesting connection between CP and preference learning, in particular learning-to-rank methods, and makes CP amenable to training data in the form of (qualitative) preferences. Elaborating on this connection, we propose methods for learning (latent) nonconformity functions from data of that kind and show their usefulness in real-world classification tasks.

1 CONFORMAL PREDICITON

Alireza

Theorem 1.1 (equivalence)**.** *For any nonconformity function in conformal prediction, there exists a rank-equivalent preference relation that can be learned directly, provided that the appropriate data are available.*

Let us define a weak preference relation $\succsim$ to be a complete and transitive binary relation on a set $\mathcal{A}$ (which describes a decision makers ranking of all elements). We see that if $u : \mathcal{X} \longrightarrow \mathbb{R}$ is a utility function representing $\succsim$, $\succsim$ must be complete and transitive:

Table 1: Table of Notations.

Notation	Description
$x \in \mathcal{X}$	instance
$y \in \mathcal{Y}$	label
$\mathcal{C}(x)$	CP set for the instance x
$\pi^\star : \mathcal{X} \longrightarrow \mathbb{P}(\mathcal{Y})$	true label probability function
$\pi^\star(x)_y$	probability of class y for instance x
$\hat{\pi} : \mathcal{X} \longrightarrow \mathbb{P}(\mathcal{Y})$	an estimate of $\pi^\star$
$s \colon \mathcal{X} \times \mathcal{Y} \longrightarrow \mathbb{R}$	conformity score function
$\succ_s$	conformity (score) order relation
$\mathcal{D}_{\text{calib}}$	calibration data
$\mathcal{D}_{\text{train}}$	training data
α	CP error rate

Transitivity. Suppose $x \succsim y$ and $y \succsim z$. Since u represents $\succsim$, we have $u(x) \geq u(y)$ and $u(y) \geq u(z)$. By transitivity of $\geq$, $u(x) \geq u(z)$. Thus, $x \succsim z$. Hence, $\succsim$ is transitivity.

Completeness. For any $x, y \in \mathcal{X} \times \mathcal{Y}$, $u(x)$ and $u(y)$ are real numbers. Therefore, either $u(x) \geq u(y)$ or $u(y) \geq u(x)$. Since u represents $\succsim$, this implies $x \succsim y$ or $y \succsim x$. Hence, $\succsim$ is complete.

Following, the imposed preference relation can be learned by a ranker, for example by learning a binary predictor for every pair of items (and aggregating the individual predictions to a ranking at test time) or fitting a Placket-Luce model.

Theorem 1.2 (validity)**.** *If the data points in $\mathcal{D}_{calib} \cup (x_{new}, y_{new})$ are exchangeable, then*

$$\mathbb{P}(y_{new} \in \mathcal{C}(x)) \geq 1 - \alpha.$$

There are two major approaches towards modeling preference relations, which are binary preference predicates

Algorithm 1 Split conformal prediction without nonconformity score

Input: calibration data $\mathcal{D}_{\text{calib}}$, training data $\mathcal{D}_{\text{train}}$, error rate α, test instance x

Use preference data $\mathcal{D}_{\text{train}}$ to infer preference relation $\succ_s$

Sort $\mathcal{D}_{\text{calib}}$ according to $\succ_s$

Let (x_q, y_q) be the $\lceil (1-\alpha)(\frac{n}{2}+1) \rceil$ -th element in the sorted list

Return prediction set $\mathcal{C}(x) = \{y \in \mathcal{Y} \colon (x, y) \succ_s (x_q, y_q)\}$

and utility functions Fürnkranz and Hüllermeier [2011]. While theoretically both are applicable for our purpose, we employ the latter and learn a latent utility (or in this case conformity) function $f \colon \mathcal{X} \times \mathcal{Y} \longrightarrow \mathbb{R}$ in lieu of a nonconformity score. We consider pairwise preference data $\mathcal{D}_{\text{train}} = \{(x_{i_n}, y_{i_n}) \succ (x_{j_n}, y_{j_n})\}_{n=1}^{N}$ [1], where preference relation indicates that instance x_i with label y_i is more conformal than instance x_j with label y_j. Having access to training data of this kind, we can learn a conformity function via a generalized Bradley-Terry model Bradley and Terry [1952]. The probability of a pairwise preference is modeled as

rework notation

$$P(i \succ j) = \frac{\exp(f(i))}{\exp(f(i)) + \exp(f(j))} \tag{1}$$

Model parameters of f can then be learned via maximum likelihood estimation, where the negative log-likelihood function is given as

$$l(f) = \sum_{n=1}^{N} \ln(\exp(f(i) + \exp(f(j))) - f(i) \tag{2}$$

The negative log-likelihood function 2 can then be used as a loss for training models of f with gradient-based methods, such as deep neural networks. Due to its probabilistic nature, the Bradley-Terry model deals gracefully with noisy preference labels and is an appropriate choice for the task of learning a preference relation for conformal classification.

rework reference

1.1 FEEDBACK MODEL AND LEARNABLE SCORES

Jonas: I think it is also worth to discuss, which type of conformity function can be learned depending on the feedback model. As we discussed, with the "in-instance" comparisons, we can learn LAC but not APS. With "cross-instance" comparisons, any non-conformity score can be learned (as stated in Theorem 1.1). I think this has practical implications, depending on which types of comparisons can be obtained from a human annotator, there may be a constraint on the conformity functions that can be learned.

[1] For the ease of notation, we will refer to the instance-label-pairs with i and j in the remainder of this paper.

2 RELATED WORK

I am not sure whether a dedicated related work section is necessary, however, the following papers are based on classifiers that are then used for "ranking"

- Huang et al. [2024]
- Luo and Zhou [2024] (Preprint)

3 EXPERIMENTAL EVALUATION

3.1 RESEARCH QUESTIONS

- RQ2: How does a learned nonconformity score perform in comparison to established, pre-specified nonconformity scores on downstream tasks (image/text classification)
- ...

- Datasets, CV splits,
- Learn neural networks (loss function is Bradley-Terry NLL, Optimizer, Architecture, Epochs ...) both for classifiers as rankers
- Oracle annotators that mimic nonconformity scores and return preferences for pairs of observations
- Evaluation metrics: Coverage, worst-slice conditional coverage, y-conditional coverage, efficiency (beautiful boxplots)

3.2 REPLICATING NONCONFORMITY SCORES FROM ORACLE FEEDBACK

In the following, we will experimentally demonstrate that the aforementioned method can indeed be used to replicate existing nonconformity scores. To this end, we consider a synthetic setting in which the true label probability function $\pi^\star : \mathcal{X} \longrightarrow \mathbb{P}(\mathcal{Y})$ is known and an oracle annotator is simulated, that returns ordered pairs

$$(x_{i_n}, y_{i_n}) \succ (x_{j_n}, y_{j_n}) \iff s(x_{i_n}, y_{i_n}) > s(x_{j_n}, y_{j_n})$$

for a nonconformity score $s \colon \mathcal{X} \times \mathcal{Y} \longrightarrow \mathbb{R}$. In order to validate that the learned preference relation $\succ_s$ replicates s, we reserve observations $\mathcal{D}_{val}$ from the synthetic data

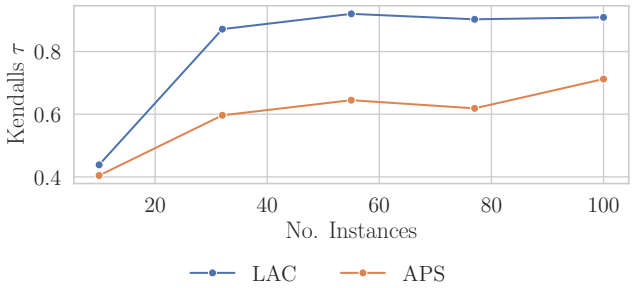

Figure 1: Rank correlation between $\mathcal{D}_{val}$ sorted according to the ground truth conformity score s and the preference relation $\succ_s$ inferred from pairwise annotations.

generating process, sort them according to $\succ_s$ and s and compute the Kendall's τ rank correlation coefficient Kendall [1938]. A detailed description of the experimental setup can be found in Appendix C. We consider the cases of the *least ambiguous set-valued classifier* (LAC) Sadinle et al. [2019] and the *adaptive prediction sets* (APS) Romano et al. [2020], Angelopoulos et al. [2020] nonconformity scores.

$$s_{\text{LAC}}(x, y) = 1 - \pi^{\star}(x)_y \tag{3}$$

$$s_{\text{APS}}(x, y) = \sum_{i=1}^{k} \pi^{\star}(x)_{y_i} \tag{4}$$

where $y = y_k$ and the probabilities are ranked from higher to lower.

Figure 1 shows the rank correlation between ground-truth nonconformity scores and the inferred preference relation $\succ_s$ on $\mathcal{D}_{val}$ for LAC and APS. We observe, that while both curves are rising, only the LAC curve is approaching 1 within the 100 instances. This is most likely due to the fact

4 LIMITATIONS AND FUTURE WORK

Limitations

- Calibration data still needs to consist of observations $(\boldsymbol{x}, y) \in \mathcal{X} \times \mathcal{Y}$
- Technically, the learned conformity function still returns a score
- So far, we only considered (multi-class) classification. Regression tasks cannot yet be accomplished by our method.

Future Work

- Dyad Ranking, allows for zero-shot predictions etc.

5 CONCLUSION

References

Anastasios Angelopoulos, Stephen Bates, Jitendra Malik, and Michael I Jordan. Uncertainty sets for image classifiers using conformal prediction. *arXiv preprint arXiv:2009.14193*, 2020.

Ralph Allan Bradley and Milton E. Terry. Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. *Biometrika*, 39(3/4):324, December 1952. ISSN 00063444. doi: 10.2307/2334029.

Johannes Fürnkranz and Eyke Hüllermeier. *Preference Learning*. SpringerLink Books. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011. ISBN 978-3-642-14124-9 978-3-642-14125-6. doi: 10.1007/978-3-642-14125-6.

Jianguo Huang, Huajun Xi, Linjun Zhang, Huaxiu Yao, Yue Qiu, and Hongxin Wei. Conformal Prediction for Deep Classifier via Label Ranking. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, 2024. URL `https://openreview.net/forum?id=b3pYoZfcoo`.

M. G. Kendall. A New Measure of Rank Correlation. *Biometrika*, 30(1/2):81–93, 1938. ISSN 00063444. doi: 10.2307/2332226. URL `http://www.jstor.org/stable/2332226`. Publisher: [Oxford University Press, Biometrika Trust].

Rui Luo and Zhixin Zhou. Trustworthy Classification through Rank-Based Conformal Prediction Sets. *CoRR*, abs/2407.04407, 2024. doi: 10.48550/ARXIV.2407.04407. URL `https://doi.org/10.48550/arXiv.2407.04407`. arXiv: 2407.04407.

Yaniv Romano, Matteo Sesia, and Emmanuel Candes. Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*, 2020.

Mauricio Sadinle, Jing Lei, and Larry Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 2019.

Title in Title Case
(Supplementary Material)

This Supplementary Material should be submitted together with the main paper.

A OMITTED PROOFS

YS: Please keep this section for now. I'll add notes here without disturbing the progress in the main paper.

Theorem A.1. *Let $s : \mathcal{Z} \to \mathbb{R}$ be any (pointwise) non-conformity score, and let $\rho : \mathcal{Z} \to \mathbb{R}$ be rank-equivalent to s. Then for every finite sample $\{z_1, \ldots, z_n\} \subset \mathcal{Z}$ and every new point z_{n+1} the conformal p-values coincide. Specifically, we have*

$$\pi_s(z_{n+t} \mid z_1, \ldots, z_n) = \pi_\rho(z_{n+t} \mid z_1, \ldots, z_n).$$

An immediate consequence of A.1 is that $\mathcal{C}_s(X_{n+1})$ and $\mathcal{C}_\rho(X_{n+1})$ coincide.

Suggestion for the introduction:

B WORK IN PROGRESS

Theorem B.1. *For any nonconformity function in conformal prediction, an equivalent ranking problem can be learned directly, provided that the appropriate data are available.*

Let us define a weak preference relation $\succsim$ to be a complete and transitive binary relation on a set $\mathcal{A}$ (which describes a decision makers ranking of all elements). We see that if $u : \mathcal{X} \longrightarrow \mathbb{R}$ is a utility function representing $\succsim$, $\succsim$ must be complete and transitive:

Transitivity. Suppose $x \succsim y$ and $y \succsim z$. Since u represents $\succsim$, we have $u(x) \geq u(y)$ and $u(y) \geq u(z)$. By transitivity of $\geq$, $u(x) \geq u(z)$. Thus, $x \succsim z$. Hence, $\succsim$ is transitivity.

Completeness. For any $x, y \in \mathcal{X} \times \mathcal{Y}$, $u(x)$ and $u(y)$ are real numbers. Therefore, either $u(x) \geq u(y)$ or $u(y) \geq u(x)$. Since u represents $\succsim$, this implies $x \succsim y$ or $y \succsim x$. Hence, $\succsim$ is complete.

Following, the imposed preference relation can be learned by a ranker, for example by learning a binary predictor for every pair of items (and aggregating the individual predictions to a ranking at test time) or fitting a Placket-Luce model. See after 1.2.

Definition B.2 (Basic Spaces)**.** Let (X, Y) be a measurable space where:

- X is the feature space
- Y is the label space
- $Z = X \times Y$ is the example space

Definition B.3 (Nonconformity Function)**.** Let $\alpha : Z \longrightarrow \mathbb{R}$ be a nonconformity function where:

- For any $z \in Z_{test}$, $\alpha(z)$ measures the nonconformity of z with respect to the calibration set Z^*

Proof. **Part 1: Construction of the Ranking Function**

Definition B.4. For any nonconformity function α, define the ranking function $r_\alpha : Z \times Z \longrightarrow \{-1, 1\}$ as:

$$r_\alpha(z_1, z_2) = \text{sign}(\alpha(z_2) - \alpha(z_1))$$

Part 2: Equivalence in Conformal Prediction

Theorem B.5. *For any $z \in Z$, the p-value computed using α is equivalent to that computed using r_α.*

Proof. The p-value using α is defined as:

$$p_\alpha(z) = \frac{|\{z' \in Z : \alpha(z') \geq \alpha(z)\}|}{|Z| + 1}$$

Using r_α, we can express the same set:

$$\{z' \in Z : \alpha(z) \geq \alpha(z)\} = \{z' \in Z : r_\alpha(z', z) \leq 0\}$$

Therefore:

$$p_\alpha(z) = \frac{|\{z' \in Z : r_\alpha(z', z) \leq 0\}|}{|Z| + 1} = p_r(z)$$

□

Part 3: Learnability □

C ADDITIONAL EXPERIMENTS

C.1 REPLICATING NONCONFORMITY SCORES FROM ORACLE FEEDBACK

In the following, we will describe the experimental details for replicating nonconformity scores from Oracle feedback. We consider a multiclass scenario with three features and $K = 3$ classes. The classification instances follow a multivariate normal distribution $x \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and the conditional class distribution is modeled as a multinomial logistic regression $P(y = k \mid x) = \frac{\exp(x^T \beta_k)}{\sum_{j=1}^{K} \exp(x^T \beta_j)}$.

We draw n classification instances from the $P(x)$ and create all $n \cdot K$ possible observations (x, y). The oracle annotator proceeds to compute the nonconformity scores for all observations and returns all $\binom{n \cdot K}{2}$ ordered pairs as preference data $\mathcal{D}_{train}$ for the ranking model. The ranking model is a simple feed-forward neural network with three hidden layers of width three and sigmoid activation functions. We train it for 300 epochs at a learning rate of 0.01 with the Adam optimizer to infer a preference relation $\succ_s$.

Afterwards, we sample another 100 instances from $P(x)$ and again create all possible observations. These observations are then being sorted with respect to the ground-truth nonconformity score s and according to $\succ_s$. We compute the Kendalls τ rank correlation coefficient between these two rankings in order to validate whether $\succ_s$ indeed replicates s.