Reliable Household Demographic Classification

Javier Carreno JAVIER.CARRENO.2023@LIVE.RHUL.AC.UK
Khuong An Nguyen KHUONG.NGUYEN@RHUL.AC.UK
Zhiyuan Luo ZHIYUAN.LUO@RHUL.AC.UK
Royal Holloway, University of London, Surrey TW20 0EX, United Kingdom
Andrew Fish ANDREW.FISH@LIVERPOOL.AC.UK
University of Liverpool, Liverpool L69 3BX, United Kingdom

Editor: Khuong An Nguyen, Zhiyuan Luo, Harris Papadopoulos, Tuwe Löfström, Lars Carlsson and Henrik Boström

Abstract

We propose the Hybrid Calibration Score (HCS), a new nonconformity measure for inductive conformal prediction. HCS combines instance-level scoring with global model calibration via Expected Calibration Error. On a real-world demographic classification task, HCS achieves 99% coverage with smaller prediction sets (APS = 1.55) and higher decisiveness (OneC = 55.19%) than standard measures, while preserving formal coverage guarantees.

Keywords: Conformal Prediction, Nonconformity Measure, Audience Modelling

1. Introduction

Accurate confidence estimation is essential for deploying machine learning in risk-sensitive settings. Conformal Prediction (CP) offers a principled way to generate prediction sets with formal coverage guarantees under minimal assumptions (Vovk et al., 2005). Its effectiveness hinges on the choice of nonconformity measure (NCM), which scores how atypical a prediction is (Papadopoulos et al., 2002; Aleksandrova and Chertov, 2021).

We propose the **Hybrid Calibration Score (HCS)**, a model-agnostic NCM that combines proper scoring rules (*Brier Score* and *Log Loss*) with a global Expected Calibration Error (*ECE*) element. This ensures both per-instance accuracy and global probabilistic trustworthiness, integrating seamlessly into any CP pipeline.

2. Hybrid Calibration Score (HCS)

Given a model's output distribution $p(y|x)$, HCS is defined as:

$$a(x, y) = \alpha \cdot \mathrm{Brier}(x, y) + \beta \cdot \mathrm{LogLoss}(x, y) + \gamma \cdot \mathrm{ECE}, \tag{1}$$

where $\alpha + \beta + \gamma = 1$, and weights are tuned using Bayesian optimisation.

The *Expected Calibration Error (ECE)* (Guo et al., 2017) quantifies the mismatch between predicted confidence and empirical accuracy across n confidence bins:

$$\mathrm{ECE} = \sum_{i=1}^{n} \frac{|B_i|}{N} \cdot |\mathrm{acc}(B_i) - \mathrm{conf}(B_i)| \tag{2}$$

where $|B_i|$ is the number of samples in bin i, N is the total number of samples, and $\mathrm{acc}(B_i), \mathrm{conf}(B_i)$ are the accuracy and mean confidence in bin B_i. ECE acts as a global regulariser, complementing instance-wise terms by enforcing population-level calibration.

3. Empirical Evaluation

We evaluate HCS on a real-world household classification task using 19,386 TV viewing instances across six categories (Carreno et al., 2024), with a Random Forest baseline. Data is split 60/30/10 (train/calibration/test) using stratified sampling. Experiments are implemented in Python with `scikit-learn`; a custom wrapper inspired by `mapie` enables flexible use of non-standard nonconformity measures. Results are averaged over five random seeds. Metrics follow Kato et al. (2023).

- **Coverage**: Fraction of prediction sets containing the true label.
- **APS**: Average Prediction Set size (lower is better).
- **OneC**: Rate of singleton prediction sets (higher implies greater decisiveness).

Table 1 summarises results at a significance level of $\alpha = 0.05$, corresponding to a 95% confidence level.

Table 1: Performance comparison of NCMs at $\alpha = 0.05$.

	Coverage (%)		**APS**		**OneC (%)**	
NCM	Mean	Std	Mean	Std	Mean	Std
Hinge Loss	95.80	0.20	1.22	0.02	78.91	1.43
Gap	95.61	0.22	1.28	0.03	81.09	0.88
Brier Score	95.58	0.25	1.22	0.01	80.57	0.99
HCS	94.84	0.25	**1.18**	0.01	**82.75**	1.27

Per-Class Analysis

To further assess the performance of the two strongest NCMs, we compare *Hinge Loss* and *HCS* across demographic categories at $\alpha = 0.05$ (95% confidence). Table 2 reports per-class coverage, APS, and OneC.

Table 2: Per-class comparison of *Hinge Loss* vs. *HCS* at $\alpha = 0.05$.

	Coverage (%)		**APS**		**OneC (%)**	
Category	Hinge	HCS	Hinge	HCS	Hinge	HCS
Couple w/ adult children	94.68 (1.37)	93.78 (1.21)	1.38 (0.04)	**1.31** (0.04)	62.86 (3.60)	**69.97** (3.10)
Couple w/ teenagers	85.30 (3.64)	80.79 (3.72)	1.51 (0.07)	**1.43** (0.05)	51.66 (5.98)	**58.94** (4.78)
Couple w/ young kids	77.14 (9.31)	70.00 (3.19)	2.03 (0.15)	**1.90** (0.08)	10.00 (8.14)	**18.57** (6.39)
Only middle-aged adults	98.03 (0.42)	97.40 (0.70)	1.12 (0.01)	**1.10** (0.01)	87.81 (1.54)	**89.79** (1.38)
Only young adults	99.66 (0.24)	99.49 (0.13)	1.20 (0.02)	**1.15** (0.02)	80.23 (1.61)	**84.90** (1.58)
Seniors	93.94 (1.89)	93.22 (1.55)	1.10 (0.01)	**1.08** (0.01)	90.25 (1.15)	**92.07** (1.02)

4. Conclusion

HCS delivers a compelling trade-off between reliability and efficiency. By incorporating global calibration into instance-wise scoring, it produces smaller, more decisive prediction sets without compromising coverage. Future work includes extensions to multi-label tasks and per-instance calibration.

References

Marharyta Aleksandrova and Oleg Chertov. Impact of model-agnostic nonconformity functions on efficiency of conformal classifiers: an extensive study. In *Conformal and Probabilistic Prediction and Applications*, pages 151–170. PMLR, 2021.

Javier Carreno, Khuong An Nguyen, Zhiyuan Luo, and Andrew Fish. Unlocking viewer insights in linear television: A machine learning approach. In *International Conference on Business Informatics Research*, pages 53–67. Springer, 2024.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

Yuko Kato, David MJ Tax, and Marco Loog. A review of nonconformity measures for conformal prediction in regression. *Conformal and Probabilistic Prediction with Applications*, pages 369–383, 2023.

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, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.