Conformal Classification with New Labels

Tianmin Xie TIANMINX@USC.EDU
University of Southern California, Los Angeles, CA, USA
Ziyi Liang LIANGZ25@UCI.EDU
University of California, Irvine, Irvine, CA, USA
Stefano Favaro STEFANO.FAVARO@UNITO.IT
University of Torino and Collegio Carlo Alberto, Torino, Italy
Matteo Sesia SESIA@MARSHALL.USC.EDU
University of Southern California, Los Angeles, CA, USA

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

Abstract

We propose Conformal Good-Turing Classification (CGTC), a novel conformal inference method for classification tasks where the true label space is unknown or potentially infinite. Traditional conformal classification methods rely on the assumption of a finite and fully known set of labels (Vovk et al., 2005; Romano et al., 2020; Angelopoulos et al., 2021). However, this assumption is sometimes violated in real-world applications, such as image classification tasks with dynamic datasets, where new classes can continually emerge (Bendale and Boult, 2015; Scheirer et al., 2012).

Consider, for example, a sequence $(X_1, Y_1), \ldots, (X_n, Y_n), (X_{n+1}, Y_{n+1})$ of facial images (X) with associated discrete identity labels (Y). In dynamic environments, new identities regularly appear, making it infeasible to predefine the entire label space. Consequently, applying traditional conformal methods in such contexts may lead to invalid coverage guarantees, as these methods fail to account for the possibility of encountering new, previously unseen labels.

The fundamental challenge in classification tasks with unknown and potentially infinite label spaces lies in accurately determining whether a new observation corresponds to an existing class or represents a new class. To address this, we integrate conformal inference with hypothesis testing for novelties. Specifically, we formalize two null hypotheses: the "new-label" hypothesis H_0^{new}, testing whether an observation represents a previously unseen label, and the "old-label" hypothesis H_0^{old}, testing whether it belongs to an existing class. These two hypotheses form a mutually exclusive and exhaustive partition—every test observation must either belong to a new class or to one of the existing classes, but not both.

Our decision rule operates as follows: If we reject H_0^{new}, we apply standard conformal classification methods and simply output the resulting prediction set. If we reject H_0^{old}, we conclude the observation belongs to an unseen label and output a special *catch-all* symbol indicating the presence of the new label explicitly. If we fail to reject both hypotheses at the chosen significance levels, we output the union of the standard prediction set and the *catch-all* symbol. This procedure maintains valid coverage guarantees.

To test these hypotheses, we introduce novel conformal p-values based on the classical Good-Turing frequency estimator (Good, 1953). Thus, we name our framework Conformal Good-Turing Classification. We develop multiple variants of these conformal p-values, including feature-enhanced versions that leverage additional covariate information beyond mere frequency counts.

Beyond the primary contributions, our framework addresses several practical challenges. First, we introduce a principled hyperparameter tuning strategy to optimally allocate the significance level between the classification and hypothesis testing components, enhancing prediction efficiency. Additionally, in settings with many rare labels appearing only once or a few times, random train-calibration splitting may result in some label classes appearing exclusively in the calibration set, rendering the conformal prediction set uninformative. We propose a selective splitting strategy to ensure each observed class is represented in the training set. While this breaks exchangeability, we design appropriate weights and prove that valid coverage is maintained.

Empirically, we evaluate our method using both synthetic experiments and the CelebA dataset. For synthetic data, we generate labels from a Dirichlet Process (DP) with a concentration parameter θ controlling the likelihood of new classes, and a uniform base distribution P_0 over $[0, 1]$. The Dirichlet Process is a canonical example of species sampling models, where labels are generated sequentially with predictive probabilities. This directly models our setting where new labels emerge dynamically. For each generated label $Y_i = y$, the corresponding feature vector X_i is sampled from a shifted multivariate normal distribution. For real-world evaluation, we use the CelebA dataset containing 202,599 face images of various celebrities representing 10,177 unique identities. We apply the MTCNN for face detection on the raw images and FaceNet to extract 128-dimensional feature embeddings. We then subsample a smaller group from the dataset and apply CGTC. Our experiments demonstrate that standard conformal classification methods fail to maintain the nominal coverage level when new labels appear. In contrast, CGTC successfully corrects the coverage to the target level while outputting more efficient prediction sets.

Keywords: conformal inference, classification, novelty detection, Good-Turing estimator, conformal p-values

Acknowledgments

T. X. and M. S. were partly supported by NSF grant DMS 2210637, by a Capital One CREDIF Research Award, and by a Google Scholar Award.

References

Anastasios Nikolas Angelopoulos, Stephen Bates, Michael Jordan, and Jitendra Malik. Uncertainty sets for image classifiers using conformal prediction. In *International Conference on Learning Representations*, 2021.

Abhijit Bendale and Terrance Boult. Towards open world recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1893–1902, 2015.

Irving J Good. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3-4):237–264, 1953.

Yaniv Romano, Matteo Sesia, and Emmanuel J. Candès. Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*, 33, 2020.

Walter J Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E Boult. Toward open set recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(7):1757–1772, 2012.

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