Ensuring Class-Conditional Coverage for Pathological Workflows (Student Abstract)

Siddharth Narendra[1], Shubham Ojha[2], Aditya Narendra[2], Abhay Kshirsagar[3], Abhisek Mallick[4]

[1] Odisha University of Technology and Research, Bhubaneswar-751029, India
[2] Cincinnati Children's Hospital Medical Center
[3] University of Illinois Urbana-Champaign
[4] Northeastern University
{siddharthnarendra0708, shubham.ojha1000, adinarendra0108}@gmail.com
abhaysk2@illinois.edu, mallick.ab@northeastern.edu

Abstract

Conformal Prediction (CP) is an uncertainty quantification framework that provides prediction sets with a user-specified probability to include the true class in the prediction set. This guarantee on the user-specified probability is known as marginal coverage. Marginal coverage refers to the probability that the true label is included in the prediction set, averaged over all test samples. However, this can lead to inconsistent coverage guarantees across different classes, constraining its suitability for high-stakes applications such as pathological workflows. This study implements a *Classwise* CP method applied to two cancer datasets to achieve class-conditional coverage which ensures that each class has a user-specified probability of being included in the prediction set when it is the true label. Our results demonstrate the effectiveness of this approach through a significant reduction in the average class coverage gap compared to the *Baseline* CP method.

Introduction

Advancements in deep learning (DL) have facilitated its integration into various cancer screening methods within computational pathology workflows. However, the clinical deployment of these DL methods is limited because of their lack of uncertainty estimates and sole dependence on the maximum-likelihood estimates, which can undermine overall diagnostic performance. For example, even if the most likely diagnosis is a common throat infection, it is important for a screening method to rule out serious conditions like Tuberculosis (TB), COPD, or Lung cancer.

Conformal Prediction (CP) is a statistical framework that estimates uncertainty by creating prediction sets that guarantee the inclusion of the true label in the set with a user-specified confidence level. This guarantee known as marginal coverage, signifies that the probability of inclusion of the true class in the prediction set is averaged across all test examples.

However, the utility of this coverage guarantee can be limited in pathological workflows due to inconsistent coverage guarantees across various classes resulting in a higher risk of diagnostic errors and reduced clinical reliability. In practice, it is essential to have class conditional coverage where coverage guarantee is ensured across all classes leading to reliable prediction sets. Working upon it, this study explores a *Classwise* approach to achieve class-conditional coverage and evaluates its performance on the CRIC (Rezende et al. 2021) and BreakHis (Spanhol et al. 2016) datasets, comparing it to a baseline CP method that lacks class-conditional coverage.

CP Procedure & Working

The CP method used in this study is known as LAC (Sadinle, Lei, and Wasserman 2019) which uses a split-conformal prediction framework (Papadopoulos et al. 2002; Lei et al. 2018). It involves a trained model, a calibration data set, and a new test example $X_{\text{test}} \in \mathcal{X}$ with an unknown label $Y_{\text{test}} \in \mathcal{Y}$ and constructs prediction sets as follows:

First, we calculate the conformal score $s(x, y) \in \mathbb{R}$ for each calibration point (X_i, Y_i), defined as $s_i = 1 - f(X_i)_{Y_i}$, the $1-$softmax score for the true class. Next, we compute the threshold $\hat{q}$ as the $\frac{\lceil (n+1)(1-\alpha) \rceil}{n}$ quantile of the sorted conformal scores, where n is the number of examples in the calibration set, α is the error rate (e.g., $\alpha = 0.05$ for a 95% confidence level), and $\lceil \cdot \rceil$ denotes the ceiling function. Finally, for the new test example $(X_{\text{test}}, Y_{\text{test}})$, we construct the prediction set $C(X_{\text{test}}) = \{y : f(X_{\text{test}})_y \geq 1 - \hat{q}\}$, by including classes with scores meeting or exceeding $1 - \hat{q}$.

Let I_{test} be a test set and $|I_{\text{test}}|$ size of test set, then the empirical coverage $\hat{c}$ is defined in Eq. 1:

$$\hat{c} = \frac{1}{|I_{\text{test}}|} \sum_{i \in I_{\text{test}}} \delta[y_i \in C(x_i)] \tag{1}$$

where δ denotes an indicator function that is 1 when its argument is true and 0 otherwise.

Coverage Guarantee

In this section, we explain two coverage guarantees while considering (X_i, Y_i) and $(X_{\text{test}}, Y_{\text{test}})$ to be i.i.d.

Marginal Coverage Guarantee

The above mentioned CP method which is referred to as *Baseline* provides marginal coverage as defined by Eq. 2 :

$$P(Y_{\text{test}} \in C(X_{\text{test}})) \geq 1 - \alpha \quad (2)$$

This indicates that each test data point, regardless of its class, has at least the user-defined probability of having the true class included in the prediction set when averaged across all test points. This leads to an overall coverage guarantee with inconsistent coverage guarantees across various classes.

Class Conditional Coverage Guarantee

Class conditional coverage is defined by Eq. 3 :

$$P(Y_{\text{test}} \in C(X_{\text{test}}) \mid Y_{\text{test}} = y) \geq 1 - \alpha, \text{ for all } y \in \mathcal{Y} \quad (3)$$

This indicates that every class y has at least the user-defined probability of being included in the prediction set when it is the true label. To achieve it, we employ a *Classwise* approach based on Mondrian Conformal Prediction (MCP) (Vovk, Gammerman, and Shafer 2005) which follows a similar workflow as the above-mentioned CP method but calibrates within each class separately. This leads to an overall coverage guarantee with consistent coverage guarantees across various classes.

Evaluating Coverage Guarantee

In this section, we provide the detailed evaluation methodology for the *Baseline* approach which ensures marginal coverage and the *Classwise* approach which aims to provide class-conditional coverage.

Evaluating Marginal Coverage

First, split the calibration and test data points irrespective of class. Calculate the threshold $\hat{q}$ upon all the calibration data points. For each test data point, generate prediction sets $C(x) = \{y : s(x, y) \geq 1-\hat{q}\}$ and then evaluate the coverage guarantee across the entire test dataset.

Evaluating Class Conditional Coverage

First, split the calibration and test data points by class. Within each class, calculate a separate threshold $\hat{q}^k$ for the calibration data. For each test data point, generate prediction sets $C(x)_k = \{y : s(x, y) \geq 1-\hat{q}^k\}$ for each class based on their respective thresholds $\hat{q}^k$. To obtain the final prediction set for each test data point, take the union of the prediction sets across all classes: $C(x) = \bigcup_{k \in \mathcal{K}} C(x)_k$, where $\mathcal{K}$ denotes all classes. Finally, evaluate the class-wise coverage guarantee for each class.

Experimental Setup

In this study, each dataset was divided into 70% training/-validation and 30% calibration/test sets, with 5-fold cross-validation. Multiple models were evaluated on CRIC and BreakHis datasets and DieT was selected as the underlying model for CP for its superior performance on both datasets. For training, Adam optimizer was used with a learning rate of 0.0002, a weight decay of 5e-3, and a batch size of 512.

For this study, we evaluate the class conditional coverage by comparing both *Baseline* and *Classwise* CP approaches, using the average class coverage gap (CovGap) metric, which is defined in Eq. 4 :

$$\text{CovGap} = 100 \times \frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{Y}} \max\left(0, (1 - \alpha) - \hat{c}_k\right) \quad (4)$$

where $\hat{c}_k$ is the empirical class-conditional coverage of class k.

Results

Dataset	Method	Cov = 90%	Cov = 95%	Cov = 99%
		CovGap	CovGap	CovGap
CRIC	*Baseline*	4.59 (4.16)	2.44 (2.30)	1.09 (1.18)
	Classwise	0.0 (0.0)	0.0 (0.0)	0.79 (0.99)
BreakHis	*Baseline*	4.44 (4.07)	1.82 (1.88)	1.30 (1.41)
	Classwise	0.0 (0.0)	0.0 (0.0)	0.66 (0.92)

Table 1: Average class coverage gap for Baseline and Classwise methods applied to CRIC and BreakHis datasets across various coverages

Table 1 demonstrates that the *Classwise* approach effectively minimizes the average class coverage gap (CovGap), with reductions ranging from 27.52% to 100% across various datasets and coverages.

Conclusion

This work highlights the importance of class-conditional coverage, particularly in pathological workflows, and explores a *Classwise* approach to minimize the average coverage gap. Utilizing class-conditional methods can enhance diagnostic precision and contribute to improved patient care outcomes.

References

Lei, J.; G'Sell, M.; Rinaldo, A.; Tibshirani, R. J.; and Wasserman, L. 2018. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523): 1094–1111.

Papadopoulos, H.; Proedrou, K.; Vovk, V.; and Gammerman, A. 2002. Inductive confidence machines for regression. In *Machine learning: ECML 2002: 13th European conference on machine learning Helsinki, Finland, August 19–23, 2002 proceedings 13*, 345–356. Springer.

Rezende, M. T.; Silva, R.; Bernardo, F. d. O.; Tobias, A. H.; Oliveira, P. H.; Machado, T. M.; Costa, C. S.; Medeiros, F. N.; Ushizima, D. M.; Carneiro, C. M.; et al. 2021. Cric searchable image database as a public platform for conventional pap smear cytology data. *Scientific data*, 8(1): 151.

Sadinle, M.; Lei, J.; and Wasserman, L. 2019. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525): 223–234.

Spanhol, F. A.; Oliveira, L. S.; Petitjean, C.; and Heutte, L. 2016. A Dataset for Breast Cancer Histopathological Image Classification. *IEEE Transactions on Biomedical Engineering*, 63(7): 1455–1462.

Vovk, V.; Gammerman, A.; and Shafer, G. 2005. *Algorithmic learning in a random world*, volume 29. Springer.