International Journal of Computer Trends and Technology
ISSN: 2231–2803 / https://doi.org/10.14445/22312803/IJCTT-V72I7P114

Volume 72 Issue 7, 108-112, July 2024

Original Article

Empirical Assessment of Nonconformity Score Functions for Classifiers in Conformal Prediction

Bhargava Kumar[1], Tejaswini Kumar[2], Swapna Nadakuditi[3]

[1]*Independent Researcher, Columbia University Alumni, NY, USA.*
[2] *Independent Researcher, Columbia University Alumni, NY, USA.*
[3]*Sr IT BSA, Florida Blue, Florida, USA.*

[1]*Corresponding Author : bhargava1409@gmail.com*

Received: 20 May 2024 Revised: 29 June 2024 Accepted: 19 July 2024 Published: 31 July 2024

Abstract *- Conformal prediction provides statistically guaranteed confidence measures for any machine learning model. This study investigates the effectiveness of three non-conformity score functions, namely Adaptive Prediction Sets (APS), Regularized Adaptive Prediction Sets (RAPS), and Sorted Adaptive Prediction Sets (SAPS), for sentiment analysis tasks. Expanding on past research that demonstrated the superiority of SAPS in classification tasks for image data, this study assesses whether this superiority extends to other domains, such as sentiment analysis. The study aims to evaluate these non-conformity score functions based on coverage and set sizes. The researchers conducted extensive experiments on a sentiment classification task using the GoEmotions dataset to gain insights into the versatility of SAPS and compared its performance with APS and RAPS. By examining the effectiveness of these non-conformity score functions, this study contributes to the understanding of the practicality of conformal prediction methods in real-world machine learning tasks beyond image classification.*

Keywords *- Adaptive Prediction Sets (APS), Conformal prediction, Non-Conformity Score Functions, Regularized Adaptive Prediction Sets (RAPS), Sorted Adaptive Prediction Sets (SAPS).*

1. Introduction

Conformal prediction [1] is a valuable statistical framework that delivers dependable measures of uncertainty in predictive modeling. Instead of generating single-point predictions, it offers a flexible approach by generating prediction sets, which account for uncertainty, ensuring that the true outcome falls within the prediction set with a predefined probability. This method is particularly useful in fields where precise uncertainty quantification is crucial, such as medical diagnosis, time series forecasting, medical diagnosis, and many others. [2,3,4]

In this study, three distinct non-conformity score functions used in conformal prediction are evaluated: Adaptive Prediction Sets (APS) [5], Regularized Adaptive Prediction Sets (RAPS) [6], and Sorted Adaptive Prediction Sets (SAPS) [7]. Each of these methods offers unique advantages in balancing prediction set size with coverage guarantees. APS effectively adjusts to intricate data distributions, ensuring adequate coverage through the use of a unique conformity score. RAPS bolsters stability by normalizing improbable class scores, resulting in smaller predictive sets with formal guarantees for coverage. SAPS reduces dependence on miscalculated probabilities, retaining only the highest softmax probability to formulate more compact yet informative prediction sets, all while upholding finite-sample coverage assurances.

Despite the demonstrated effectiveness of conformal prediction in various domains, there is a notable research gap in understanding how different nonconformity score functions perform in text-based classification tasks, particularly sentiment analysis. Previous research has primarily focused on image data, leaving a need to explore these methods in the context of sentiment analysis.

To address this gap, this study aims to examine the advantages and disadvantages of these nonconformity score functions by applying them to a sentiment classification task. A well-defined dataset is utilized, and an extensive analysis of each method's performance is conducted in terms of prediction set size and coverage probability. The findings reveal that while all three approaches achieve the desired coverage rate, RAPS consistently yields the smallest prediction sets, particularly at higher coverage rates, due to the limited number of labels considered [7]. This aligns with existing research indicating that RAPS is more effective in terms of prediction set size when the number of classes is small. In contrast, SAPS has been shown to perform better with larger numbers of classes in image data classification, but

this could not be directly tested as the dataset in this study had only a small number of classes.

The novelty of this work lies in its application to sentiment analysis, a domain that has not been thoroughly explored with these nonconformity score functions. By systematically comparing APS, RAPS, and SAPS in this new context, this research provides practitioners with practical insights into the trade-offs associated with each method, enabling more informed decisions when selecting a nonconformity score function for text-based classification tasks.

Ultimately, the study's results extend the utility of different conformal prediction approaches to sentiment analysis, enhancing our understanding of how different nonconformity score functions impact the performance and reliability of prediction sets in real-world applications. This novel application fills a significant gap in the current literature and offers a new perspective on the use of different conformal score functions in text-based classification tasks like sentiment classification.

2. Background Information

2.1. Adaptive Prediction Sets (APS)

The Adaptive Prediction Sets (APS) algorithm is a robust method for constructing prediction sets with guaranteed coverage for both categorical and unordered response labels in the field of conformal prediction. This algorithm is also applicable to regression tasks. The adaptive classification with the split-conformal technique begins by dividing the dataset into two subsets, one for calibrating the output of a black-box predictive model and the other for testing purposes. The model then generates standardized estimates of class probabilities for each data point in the calibration set, arranged in order from most likely to least likely. Conformity scores are calculated for each sample, quantifying the amount of probability mass needed to include the true class in the prediction set. The appropriate threshold for these conformity scores is then determined to create prediction sets with the desired coverage level. Finally, prediction sets are generated for unseen test data points using the calibrated conformity scores, ensuring that the sum of scores for the most probable classes exceeds the threshold. The APS algorithm provides a flexible and adaptable approach for constructing prediction sets, enabling reliable and interpretable predictions across various machine learning tasks, including multi-class classification. [5]

2.2. Regularized Adaptive Prediction Sets (RAPS)

The RAPS algorithm introduces a novel method for conformal prediction, focusing on constructing prediction sets with guaranteed coverage while also promoting smaller and more informative sets. A crucial aspect of RAPS is the concept of regularization, which penalizes the inclusion of unlikely classes, thus improving adaptiveness and reliability in uncertainty estimation. The parameter Kreg regulates the level of regularization, affecting the penalty imposed on less probable classes, while λ determines the extent of this penalty. Adjusting these parameters allows RAPS to strike a balance between set size and coverage, ensuring the inclusion of pertinent classes while minimizing unnecessary inclusions. In addition, RAPS employs randomization to address sudden changes in class probabilities, resulting in smooth transitions between inclusion and exclusion thresholds. This randomized term prevents multiple classes from being affected simultaneously, preserving the integrity of the prediction sets. Overall, these mechanisms enable RAPS to produce accurate and concise prediction sets, making it a valuable asset for uncertainty quantification in machine learning tasks. [6]

2.3. Sorted Adaptive Prediction Sets (SAPS)

Machine learning models often produce miscalibrated probability estimates, which can result in overly large prediction sets. To address this issue, the Sorted Adaptive Prediction Sets (SAPS) algorithm discards all probability values except for the maximum softmax probability. This approach helps to significantly reduce the size of the prediction sets while still effectively communicating the instance-wise uncertainty. The non-conformity score in SAPS, which is defined as a function of the maximum softmax probability, a uniform random variable, and a hyperparameter λ, regulates the weighting of the ranking information. By incorporating uncertainty through the maximum probability and mitigating the influence of tail probabilities with λ, SAPS presents a promising solution to the challenges posed by miscalibrated probability values in conformal prediction. Theoretical analyses have demonstrated the coverage guarantees and comparative advantages of SAPS over traditional methods like APS and RAPS, while experimental evaluations have confirmed its ability to generate smaller prediction sets on image classification tasks without compromising coverage accuracy. This leads to improved conditional coverage rates and better adaptation to varying levels of uncertainty. [7]

3. Experiment

3.1. Dataset

In this study, the GoEmotions dataset is utilized, which was introduced by Demszky et al. [8]. This dataset is comprised of 58,000 English Reddit comments that have been manually annotated for 27 emotion categories, including Neutral. This dataset stands out as the largest of its kind, offering a fine-grained typology that is adaptable to various downstream tasks. The high quality of its annotations has been validated through Principal Preserved Component Analysis, ensuring reliable labels.

The processed dataset used in this study is based on the work of Jesse et al. [9], who illustrated the benefits of conformal prediction sets in enhancing human decision-making. During pre-processing, they focused on sentences

with a single label and excluded those with emojis due to compatibility issues with their analytical tools. To limit the number of classes for human experiments, the top 10 classes with the highest frequency were chosen - Love, Curiosity, Approval, Disapproval, Admiration, Gratitude, Neutral, Amusement, Annoyance, and Optimism. Stratified sampling was applied to both the calibration and test sets to maintain an equal distribution of samples across these classes, which yielded 2210 samples (calibration: 1180 samples; test: 1030 samples). Both datasets were processed similarly and treated as independent and identically distributed (i.i.d.) for conformal prediction, providing a solid basis for assessing the performance of various non-conformity score functions in sentiment analysis.

3.2. Models

For the model, a RoBERTa-Base model is selected [10], which is fine-tuned on the GoEmotions training set from HuggingFace [11]. RoBERTa (Robustly optimized BERT approach) is a transformer-based language model that builds upon the architecture of BERT (Bidirectional Encoder Representations from Transformers) and incorporates various modifications to enhance its performance. Fine-tuning involves adjusting the pre-trained RoBERTa model parameters using the specific dataset (in this case, the GoEmotions training set) to improve its performance on a particular task, such as sentiment analysis. By fine-tuning RoBERTa on the GoEmotions dataset, the researchers aimed to leverage its pre-trained knowledge while tailoring it to the nuances of emotion classification present in their specific domain. This approach allows the researchers to capitalize on the strengths of RoBERTa in capturing complex linguistic patterns and contextual information, ultimately enhancing the model's ability to classify sentiments expressed in text data accurately.

3.3. Experiment Setup

In the evaluations, the researchers employed the transformer model from HuggingFace to calculate softmax scores for various labels on the calibration set. These scores were subsequently utilized for calibration using the TorchCP library [12], a Python-based toolbox designed for conformal prediction research on deep learning models using PyTorch. After calibration, they employed the conformalized model to predict prediction sets for the test set. These prediction sets were then assessed based on marginal coverage and set size.

Marginal coverage allowed them to gauge the proportion of true labels captured within the prediction sets, while set size analysis provided insights into the uncertainty levels of the predictions, with smaller sizes indicating higher confidence. Through this extensive evaluation process, their objective was to determine the efficacy of APS, RAPS, and SAPS score functions in generating precise and informative prediction sets for sentiment analysis tasks on the GoEmotions dataset.

4. Results

The assessment of the three nonconformity score functions— APS, RAPS, and SAPS—reveals that all three methods attain the desired coverage rate, as seen in Figure 1 and Table 1. Although there are occasional minor discrepancies from the anticipated coverage, these deviations are insignificant and within acceptable parameters. This suggests that all three methods are reliable in maintaining the desired confidence level. [5,6,7]

It is worth noting that there is no clear winner in terms of which score function consistently achieves the highest coverage. The performance in this regard is highly dependent on the hyperparameters used to finetune each method. The variance in coverage is minor, indicating that all three approaches are sturdy and effective for sentiment classification tasks, provided they are properly calibrated.

Table 1. Coverage obtained empirically from various combinations of non-conformity score functions and error rates

	Error rate (%)		
	5 %	**10 %**	**25 %**
APS	94.1 %	90.9 %	75.0 %
SAPS	95.9 %	92.7 %	73.6 %
RAPS	93.5 %	91.4 %	74.0 %

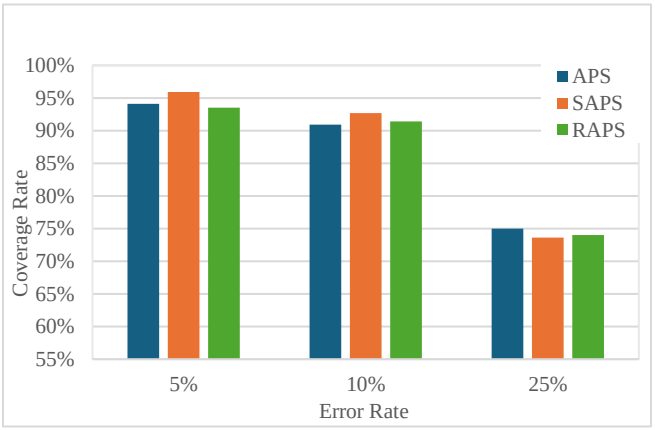

Fig. 1 Distribution of coverage across error rates for various non-conformity score functions

Regarding prediction set sizes, as seen in Figure 2 and Table 2, the Regularized Adaptive Prediction Sets (RAPS) method consistently yields the smallest average set size among the three score functions. This advantage is particularly beneficial because it suggests that RAPS can deliver more precise predictions while maintaining the necessary coverage, especially when the number of labels is small [7]. Furthermore, as the expected coverage rate diminishes, the average set sizes for all three methods also decrease, as expected. This trend aligns with the theoretical understanding that lower coverage requirements allow for more confident predictions.

Empirically, it is observed that the differences in set sizes between the methods become more pronounced as the target coverage rate increases (or equivalently, as the error rate diminishes). This indicates that RAPS's proficiency in maintaining smaller set sizes is particularly advantageous at higher coverage rates, where precision becomes increasingly critical.

Table 2. Average set size obtained from various combinations of non-conformity score functions and error rates

	Error rate (%)		
	5 %	**10 %**	**25 %**
APS	3.53	2.52	1.24
SAPS	3.32	2.39	1.23
RAPS	2.92	2.33	1.21

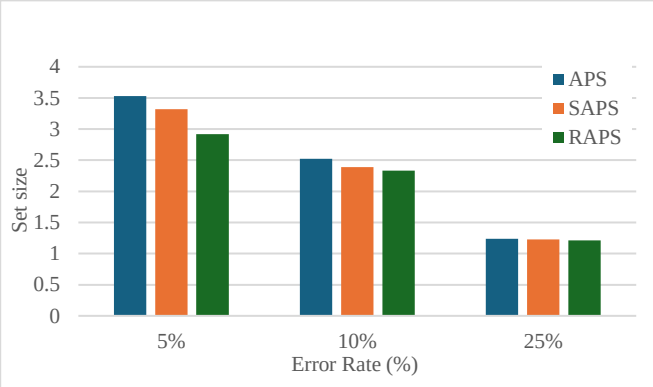

Fig. 2 Distribution of average set size across error rates for various non-conformity score functions

5. Discussion

Our empirical evaluation of APS, RAPS, and SAPS nonconformity score functions in sentiment classification reveals several key insights. All three methods effectively achieved the target coverage rates, demonstrating the robustness of conformal prediction techniques. The minor variations in coverage were primarily due to differences in hyperparameter settings, highlighting the importance of careful tuning for optimal performance.

Regarding prediction set sizes, RAPS consistently produced the smallest sets, underscoring its efficiency in providing precise predictions, especially in scenarios with fewer labels. This aligns with RAPS's design to penalize unlikely classes, making it particularly suitable for tasks with limited label sets. Conversely, SAPS exhibited strength in managing prediction set sizes when dealing with a larger number of labels, which corroborates previous findings in image classification tasks.

These results extend the applicability of conformal prediction methods to text-based tasks like sentiment analysis, offering new insights into their versatility. The study emphasizes the need to consider specific task requirements and label set characteristics when selecting a nonconformity score function, thereby aiding practitioners in optimizing conformal prediction methods for various machine learning applications.

6. Conclusion

This research conducted an empirical evaluation of three nonconformity score functions— APS, RAPS, and SAPS—in the context of a sentiment classification task. The analysis focused on two main criteria: coverage and set size.

With respect to coverage, all three methods reliably achieved the target coverage rate, with only minor differences observed. There was no clear winner among the score functions in terms of coverage, as performance varied depending on the hyperparameters used for fine-tuning. This underscores the importance of carefully selecting hyperparameters to ensure optimal coverage and set size.

In terms of set size, RAPS consistently produced the smallest average prediction sets, demonstrating its efficiency and accuracy. Consistent with the results reported in [7], The findings indicate that SAPS tends to produce smaller set sizes when the number of labels is high, while RAPS yield lower set sizes when the number of labels is not excessive. This trend was particularly evident at higher coverage rates, where the differences between the methods' set sizes were more notable. The ability of RAPS to maintain smaller prediction sets without compromising coverage makes it a valuable tool for applications that require high precision.

7. Recommendations

For professionals and researchers utilizing conformal prediction in diverse classification tasks, several essential recommendations have been derived from this study. It is crucial to meticulously tune hyperparameters, as the performance of each score function substantially depends on this process. Properly tuned SAPS can yield smaller set sizes than APS, and a well-tuned RAPS can produce smaller set sizes than SAPS, emphasizing the significance of hyperparameter tuning for score functions.

In terms of method selection, this study's findings align with prior SAPS research. RAPS is recommended for situations with fewer labels due to its efficiency in maintaining compact prediction sets. In contrast, SAPS is beneficial for tasks with a substantial number of labels, as it can effectively manage prediction set sizes through appropriate weight adjustments.

By adhering to these recommendations, practitioners can leverage the strengths of conformal prediction methods to improve the reliability and accuracy of their classification models across a wide range of applications.

References

[1] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer, *Algorithmic Learning in a Random World*, Springer, pp. 1-476, 2022. [Google Scholar] [Publisher Link]

[2] Jonathan Alvarsson et al., "Predicting with Confidence: Using Conformal Prediction in Drug Discovery," *Journal of Pharmaceutical Sciences*, vol. 110, no. 1, pp. 42–49, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[3] Margaux Zaffran et al., "Adaptive Conformal Predictions for Time Series," *Proceedings of the 39th International Conference on Machine Learning*, pp. 25834-25866, 2022. [Google Scholar] [Publisher Link]

[4] Xianghao Zhan et al., "An Electronic Nose-based Assistive Diagnostic Prototype for Lung Cancer Detection with Conformal Prediction," *Measurement*, vol. 158, p. 107588, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[5] Yaniv Romano, Matteo Sesia, and Emmanuel J. Candès, "Classification with Valid and Adaptive Coverage," *arXiv*, 2020. [CrossRef] [Publisher Link]

[6] Anastasios Angelopoulos et al., "Uncertainty Sets for Image Classifiers using Conformal Prediction," *arXiv*, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[7] Jianguo Huang et al., "Conformal Prediction for Deep Classifier via Label Ranking," *arXiv*, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[8] Dorottya Demszky et al., "GoEmotions: A Dataset of Fine-Grained Emotions," *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4040-4054, 2020. [CrossRef] [Publisher Link]

[9] Jesse C. Cresswell et al., "Conformal Prediction Sets Improve Human Decision Making," *arXiv*, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[10] Yinhan Liu et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *arXiv*, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[11] Hugging Face, SamLowe/roberta-base-go_emotions, 2023. [Online]. Available: https://huggingface.co/SamLowe/roberta-base-go_emotions

[12] Hongxin Wei, and Jianguo Huang, "TorchCP: A Library for Conformal Prediction based on PyTorch," *arXiv*, 2024. [CrossRef] [Google Scholar] [Publisher Link]