Supplementary Material:
The Penalized Inverse Probability Measure for Conformal Classification

Paul Melki[*†], Lionel Bombrun[*‡], Boubacar Diallo[†], Jérôme Dias[†], Jean-Pierre Da Costa[*‡]
[*] IMS, CNRS, University of Bordeaux [†] EXXACT Robotics [‡] Bordeaux Sciences Agro

1. Some Mathematical Properties of PIP

The toy examples and the empirical results show that the proposed PIP nonconformity score behaves similarly to its baseline Hinge Loss (IP) measure in certain situations, while in other times its behavior resembles that of the Margin Score (MS). In fact, there is a direct relationship between these three scores, namely:

$$\Delta^{\text{PIP}}(y) = \underbrace{1 - \hat{p}^y}_{\Delta^{\text{IP}}(y)} + \sum_{r=1}^{R(y)-1} \frac{\hat{p}^{[r]}}{r} \mathbb{1}_{\{R(y)>1\}} \tag{1}$$

where $R(k)$ is the rank of class k after the estimated probabilities $p^1, ..., p^K$ have been sorted in decreasing order, and $\hat{p}^{[r]}$ the probability estimate of the class having rank r, such that $\hat{p}^k = \hat{p}^{[R(k)]}$. For any y such that $R(y) > 1$, we have:

$$\begin{aligned}
\Delta^{\text{PIP}}(y) &= 1 - \hat{p}^y + \hat{p}^{[1]} + \sum_{r=2}^{R(y)-1} \frac{\hat{p}^{[r]}}{r} \\
&= 1 \underbrace{- \hat{p}^y + \max_{k \neq y} \hat{p}^k}_{\Delta^{\text{MS}}(y)} + \sum_{r=2}^{R(y)-1} \frac{\hat{p}^{[r]}}{r} \\
&= 1 + \Delta^{\text{MS}}(y) + \sum_{r=2}^{R(y)-1} \frac{\hat{p}^{[r]}}{r} \qquad \blacksquare
\end{aligned} \tag{2}$$

From these relationships, we can study the behavior of $\Delta^{\text{PIP}}(y)$ in different possible scenarios and derive some upper and lower bounds:

1. Assume the case where the class of interest y is the most "certain" class. That is, $\hat{p}^y = 1$ and for all other classes $\hat{p}^k = 0,\ k \neq y$. In such a situation, the rank $R(y)$ of y will obviously be 1. In such an optimal scenario, y should be given the minimal possible score. Indeed:

   $$\Delta^{\text{PIP}}(y) = \Delta^{\text{IP}}(y) = 1 - \hat{p}^y = 0$$

   is a lower bound on Δ^{PIP}.

2. Assume the opposite scenario whereby a class $k \neq y$ is assigned the maximal probability $\hat{p}^k = 1$, which then means that $\hat{p}^y = 0$. Such a setting should be maximally penalized since the base classifier can be considered to have made a big mistake about class y:

   $$\begin{aligned}
   \Delta^{\text{PIP}}(y) &= 1 + \Delta^{\text{MS}}(y) + \underbrace{\sum_{r=2}^{R(y)-1} \frac{\hat{p}^{[r]}}{r}}_{=0} \\
   &= 1 + \underbrace{\max_{k \neq y} \hat{p}^k - \hat{p}^y}_{=1} \\
   &= 2
   \end{aligned} \tag{3}$$

   which is an upper bound on Δ^{PIP}.

3. Assume the theoretical scenario where the base classifier assigns the same probability estimate to all classes. That is, $p^k = 1/K,\ \forall k = 1, 2, ..., K$. In this case, the class of interest y will have the same probability estimate as all other classes and its PIP score will depend only on its rank $R(y)$, which can take any value in $1, 2, ..., K$:

   $$\begin{aligned}
   \Delta^{\text{PIP}}(y) &= 1 - \frac{1}{K} + \sum_{r=1}^{R(y)-1} \frac{\hat{p}^{[r]}}{r} \\
   &= 1 - \frac{1}{K} + \frac{1}{K} \left(\sum_{r=1}^{R(y)-1} \frac{1}{r} \right) \\
   &= 1 + \frac{1}{K} \left(\sum_{r=1}^{R(y)-1} \frac{1}{r} - 1 \right)
   \end{aligned} \tag{4}$$

   Based on this scenario, it is apparent that PIP generally guarantees assigning a different score to each class since even in the degenerate (and impossible) case when all the classes have the same probability estimates, the nonconformity score of each class will be different.

From these scenarios, we can further observe the "hybrid" behavior of PIP. Indeed, from scenario (1) it is apparent that when a class k has a high estimated probability $\hat{p}^k$ (close to 1), it is this probability estimate that plays the biggest role in the final score value. When this class k is the class of interest y, then $\hat{p}^y$ will play an important role in attenuating the final score as it will be used in the IP term, i.e. the first part in Equation 1. However, when k is different than the class of interest y it will play a role in increasing the score assigned to y as it will reside in the penalization component of the PIP score. When the base classifier is quite "ambivalent", assigning more or less the same scores to all classes, then the most important factor impacting the final score is the rank $R(y)$ of class y but which will have decreasing importance due to the inverse rank weighting in the penalization component.

2. WE3DS Data Preparation

The experiments in this article are conducted on the public WE3DS dataset[1] published by Kitzler et al. [1]. Originally conceived as a dataset of densely annotated RGB-D images for semantic segmentation with 17 different plant classes and the *soil* class, it has been transformed through a simple procedure into a classification dataset.

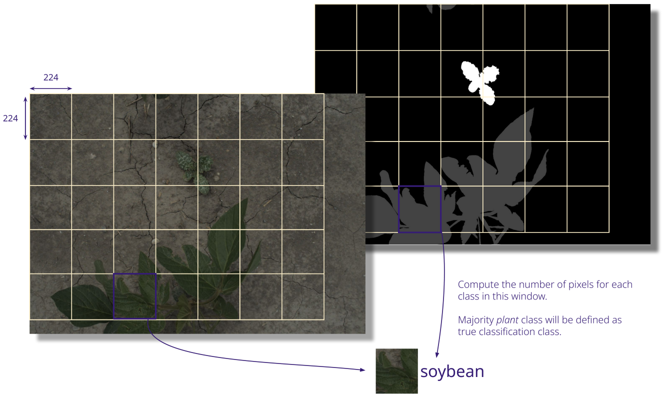
224
224
Compute the number of pixels for each class in this window.
Majority *plant* class will be defined as true classification class.
soybean

Figure 1. Demonstration of how the original WE3DS segmentation images are divided into smaller classification images.

As shown in Figure 1, after discarding the depth channel, each original RGB image of size 1600 × 1144 is divided into non-overlapping smaller images of size 224 × 224. The corner regions that do not align with the cropping grid (due to the original dimensions not being perfectly divisible by 224) are simply discarded. For each smaller image (like the one highlighted in blue in Figure 1), its corresponding area is considered in the ground-truth semantic mask. The number of pixels in each class is counted, then a decision is taken:

1. If the image contains only pixels of class *soil*, then *soil* is defined as the class label of the resulting classification image;

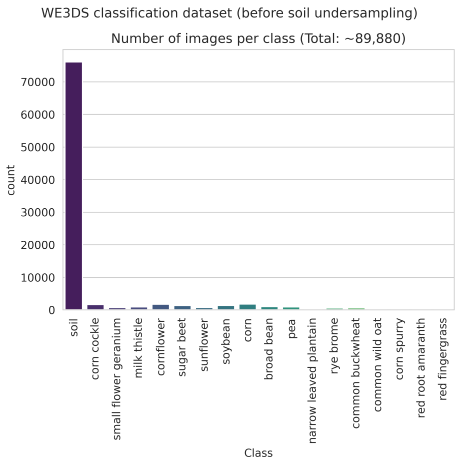

Figure 2. Count of images per class in the resulting classification dataset after window cropping from the original larger images. The *soil* class overshadows all other classes.

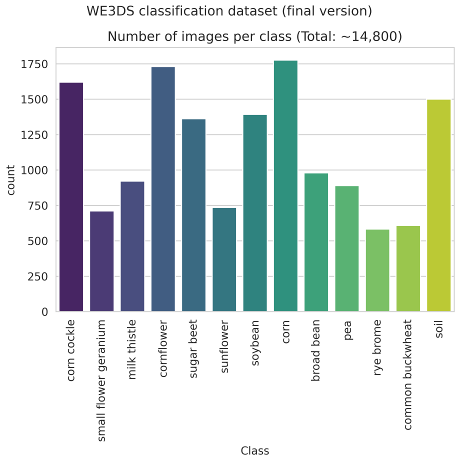

Figure 3. Count of images per class in the final classification dataset after *soil* random undersampling and the dropping of the five rarest classes.

2. If any other class exists in the image, the majority class is taken to be the true label.

This results in a classification dataset consisting of 89,880 images 18 different classes showing very high imbalance towards the heavily majoritarian *soil* class (Figure 2). In order to curb this imbalance problem, 1,500 images are randomly sampled from the *soil* class. Additionally, the five very rare classes (*corn spurry*, *narrow leaved plantain*, *common wild oat*, *red root amaranth* and *red fingergrass*) are removed, resulting in a dataset of 13 classes and 14,800 images as shown in Figure 3. This is the dataset used to conduct the experiments in the current work.

[1] The original dataset can be accessed and downloaded here: `https://zenodo.org/records/5645731`

References

[1] F. Kitzler, N. Barta, R. W. Neugschwandtner, A. Gronauer, and V. Motsch, “WE3DS: An RGB-D Image Dataset for Semantic Segmentation in Agriculture,” *Sensors*, vol. 23, no. 5, p. 2713, Mar. 2023. 2