An Uncertainty-Aware Pseudo-Label Selection Framework using Regularized Conformal Prediction

Matin Moezzi
matin.moezzi@gmail.com

Abstract

Consistency regularization based methods are prevalent semi-supervised learning (SSL) algorithms due to their exceptional performance. However, they mainly depend on domain-specific data augmentations, which are not usable in domains where data augmentations are less practicable. On the other hand, Pseudo-labeling (PL) is a general and domain-agnostic SSL approach that, unlike consistency regularization based methods, does not rely on the domain. PL underperforms due to the erroneous high confidence predictions from poorly calibrated models. This paper proposes an uncertainty-aware pseudo-label selection framework that employs uncertainty sets yielded by the conformal regularization algorithm to fix the poor calibration neural networks, reducing noisy training data. The codes of this work are available at: https://github.com/matinmoezzi/ups_conformal_classification

Index Terms

Semi-Supervised Learning, Pseudo-Labeling, Conformal Prediction, Uncertainty Quantification

I. Introduction

Much of the recent success in training large, deep neural networks is because of the existence of large labeled datasets. Yet, collecting labeled data is expensive for many learning tasks as it necessarily involves expert knowledge. Semi-supervised learning (SSL) seeks to alleviate the need for labeled data by allowing a model to leverage unlabeled data [Berthelot et al., 2019].

Consistency regularization based methods are a prevalent semi-supervised learning algorithm due to their exceptional performance. However, they mainly depend on domain-specific data augmentations, which are not usable in domains where data augmentations are less practicable. On the other hand, Pseudo-labeling (PL) is a general and domain-agnostic SSL approach that, unlike consistency regularization based methods, does not rely on the domain. PL underperforms due to the erroneous high confidence predictions from poorly calibrated models.

[Rizve et al., 2021] argues that conventional pseudo-labeling based methods achieve poor results because poor network calibration produces incorrectly pseudo-labeled samples, leading to noisy training and poor generalization. Since pseudo-labeling has been impactful due to its simplicity, generality, and ease of implementation, Rizve et al. propose an uncertainty-aware pseudo-label selection (UPS) framework, which attempts to maintain these benefits while addressing the calibration issue to improve PL performance drastically. PL leverages the prediction uncertainty to guide the pseudo-label selection procedure.

Recent advances in image classifiers (e.g., CNNs) have yielded high accuracy predictions. However, predicting the most likely label is not the only useful thing in consequential settings such as medical diagnostics. In particular, knowing the uncertainty of the model and considering other less probable labels is necessary for image classification applications where decision-making is crucial. For this purpose, Angelopoulos et al. propose an algorithm that modifies any classifier to output a predictive set, representing the quantified uncertainty of the classifier.

In this paper, we incorporate the uncertainty quantification idea into the UPS framework to reduce noisy training data in the process of pseudo-label selection.

II. Regularized Adaptive Prediction Set (RAPS)

In order to quantify the model uncertainty, Angelopoulos et al. proposed an algorithm called RAPS, which modifies the classifier to output a predictive set containing the true label with a user-specified probability. RAPS is based on a data-splitting version of the conformal prediction algorithm, a general algorithm to generate predictive sets and satisfies the coverage property for any predictors. The main RAPS contribution is to add a regularization term to the conformal score.

Formally, for an image classifier with a target label $Y \in \mathcal{Y} = \{1, \ldots, K\}$ and a feature vector $X \in \mathbb{R}^d$, let $\mathcal{C}(x, u, \tau) : \mathbb{R}^d \times [0,1] \times \mathbb{R} \to 2^{\mathcal{Y}}$ be a set function which generates a predictive set for input X. The τ parameter controls the size of the sets such that $\mathcal{C}(x, u, \tau_1) \subseteq \mathcal{C}(x, u, \tau_2)$ if $\tau_1 \leq \tau_2$.
First, RAPS selects the smallest τ, τ_{ccal}, that gives at least $1-\alpha$ coverage on the conformal calibration set ($U_i \sim U[0,1]$):

$$\hat{\tau}_{\text{ccal}} = \inf \left\{ \tau : \frac{|\{i : Y_i \in \mathcal{C}(X_i, U_i, \tau)\}|}{n} \geq \frac{\lceil (n+1)(1-\alpha) \rceil}{n} \right\} \tag{1}$$

Secondly, RAPS defines $C(x, u, \tau)$ such that:

$$\mathcal{C}^*(x, u, \tau) := \{y : \rho_x(y) + \hat{\pi}_x(y) \cdot u + \underbrace{\lambda \cdot (o_x(y) - k_{reg})^+}_{\text{regularization}} \leq \tau\} \tag{2}$$

Eventually, the predictive set for the input X is give by $\mathcal{C}^*(X, u, \hat{\tau}_{ccal})$[Eq. 2], which $\hat{\tau}_{ccal}$ is given by Eq. 1. Algorithm 1 [Angelopoulos, 2021] illustrates the pseudo code of RAPS.

III. Uncertainty-Aware Pseudo-label Selection

Consider an SSL problem consists of a labeled dataset $D_L = \{(x_i, \mathbf{y}_i)\}_{i=1}^{N_L}$ with N_L samples where x_i is the input vector and $\mathbf{y}_i = [y_1^{(i)}, \ldots, y_C^{(i)}] \subseteq \{0,1\}^C$ is the corresponding label with C class categories, and an unlabeled dataset $D_U = \{x^{(i)}\}_{i=1}^{N_U}$ with N_U samples. For the unlabeled samples, pseudo-labels $\tilde{\mathbf{y}}^{(i)}$ are generated. Pseudo-labeling based SSL approaches involve learning a parameterized model f_θ on the dataset $\tilde{D} = \{(x^{(i)}, \tilde{\mathbf{y}}^{(i)})\}_{i=1}^{N_L+N_U}$, with $\tilde{\mathbf{y}}^{(i)} = \mathbf{y}^{(i)}$ for the N_L labeled samples.

In the conventional pseudo-labeling setting, the model is first trained with labeled data. Then, the model is used to predict labels for unlabeled data. The predicted labels (pseudo-labels) are target classes for unlabeled data as if they were true labels. Finally, the pre-trained model is trained in a supervised fashion with labeled and unlabeled data simultaneously [Lee, 2013].

In order to reduce noisy training data, a subset of pseudo-labels is intelligently selected, which are less noisy for training in each iteration. The high-confidence selection process is based on the network's output confidence probabilities and selects a subset of pseudo-labels by confidence thresholds. Since the poorly calibrated networks lead an incorrect label to have high confidence, the high-confidence based PL does not have sufficient accuracy.

Rizve et al. concluded that prediction uncertainties can reduce the effects of poor calibration. Thus, they propose an uncertainty-aware pseudo-label selection (UPS) process, which selects a more accurate subset of pseudo-labels used in training by employing both high confidence and uncertainty prediction (Eq 4). [Rizve et al., 2021] used the MC-Dropout sampling as the uncertainty estimation method.

UPS uses hard pseudo-labeling in which pseudo-labels are obtained directly from the network prediction. Let $\mathbf{p}^{(i)}$ be the probability outputs of a trained network on the sample $x(i)$, such that $p_c^{(i)}$ represents the probability of class c being present in the sample. the hard pseudo-label can be generated for $x^{(i)}$ as:

$$\tilde{y}_c^{(i)} = \mathbb{1}[p_c^{(i)} \geq \gamma] \tag{3}$$

Formally, Let $\mathbf{g}^{(i)} = [g_1^{(i)}, \ldots, g_C^{(i)}] \subseteq \{0,1\}^C$ be a binary vector representing the selected pseudo-labels in sample i, where $g_c^{(i)} = 1$ when $\tilde{y}_c^{(i)}$ is selected and $g_c^{(i)} = 0$ when $\tilde{y}_c^{(i)}$ is not selected. UPS improves the high-confidence selection process such that:

$$g_c^{(i)} = \mathbb{1}\left[u\left(p_c^{(i)}\right) \leq \kappa_p\right] \mathbb{1}\left[p_c^{(i)} \geq \tau_p\right] + \mathbb{1}\left[u\left(p_c^{(i)}\right) \leq \kappa_n\right] \mathbb{1}\left[p_c^{(i)} \leq \tau_n\right] \tag{4}$$

where $u(p)$ is the uncertainty of a prediction p, and κ_p and κ_n are the uncertainty thresholds, and τ_p and τ_n are the confidence thresholds for positive and negative labels.

In each iteration, the parameterized neural network f_θ is trained on labeled data D_L. Then, f_θ predicts the probability outputs for all unlabeled data D_U. Pseudo-labels are created by Eq. 3 and a subset of pseudo-labels are selected following Eq. 4. Next, f_θ is trained on a selected subset of pseudo-labels. Loss functions for positive and negative labels are calculated by cross-entropy loss which are given by

$$L_{\mathrm{NCE}}\left(\tilde{\boldsymbol{y}}^{(i)}, \hat{\boldsymbol{y}}^{(i)}, \boldsymbol{g}^{(i)}\right) = -\frac{1}{s^{(i)}} \sum_{c=1}^{C} g_c^{(i)} \left(1 - \tilde{y}_c^{(i)}\right) \log\left(1 - \hat{y}_c^{(i)}\right), \tag{5}$$

$$L_{\mathrm{BCE}}\left(\tilde{\boldsymbol{y}}^{(i)}, \hat{\boldsymbol{y}}^{(i)}, \boldsymbol{g}^{(i)}\right) = -\frac{1}{s^{(i)}} \sum_{c=1}^{C} g_c^{(i)} \left[\tilde{y}_c^{(i)} \log\left(\hat{y}_c^{(i)}\right) + \left(1 - \tilde{y}_c^{(i)}\right) \log\left(1 - \hat{y}_c^{(i)}\right)\right] \tag{6}$$

Where $\hat{\mathbf{y}}^{(i)} = f_\theta(x^{(i)})$ and $s^{(i)} = \sum_c g_c^{(i)}$ is the number of selected pseudo-labels for sample i. After calculating the losses of selected pseudo-labels, the back propagation step of f_θ is performed using the sum of labeled and selected pseudo-labels losses.

IV. Our Method

In this paper, we employ the RAPS framework in the UPS algorithm. RAPS takes an image classifier and outputs a predictive set for each input which represents the uncertainty of the classifier. We use RAPS as a wrapper around the UPS pre-trained model. To be specific, we use the predictive set yielded by RAPS instead of hard pseudo-labeling to predict labels for unlabeled data.

Formally, consider we have C class categories. For input $x^{(i)}$, RAPS produces a predictive set such that

$$\mathcal{C}^*(x^{(i)}, u, \hat{\tau}_{ccal}) \subseteq \mathcal{Y} = \{1, \ldots, C\}.$$

Which means for every c in $\mathcal{C}^*(x^{(i)})$, input $x^{(i)}$ belongs to the c class. Consider unlabeled input $\tilde{x}^{(i)}$, for multi-label case, the proposed method predicts the corresponding pseudo-label such that

$$\tilde{\mathbf{y}}^{(i)} = \{\tilde{y}_1^{(i)}, \ldots, \tilde{y}_C^{(i)}\} \tag{7}$$

$$\tilde{y}_c^{(i)} = \mathbb{1}[c \in \mathcal{C}^*(\tilde{x}^{(i)}, u, \hat{\tau}_{ccal})] \quad \forall c \in \mathcal{Y} \tag{8}$$

And for single-label case, the label with the highest conformal score is assigned as the pseudo-label such that:

$$\tilde{y}_c^{(i)} = \begin{cases} 1, & \text{if } c \in \mathcal{C}^*(\tilde{x}^{(i)}, u, \tau_{ccal}) \text{ and } \underset{j \in \mathcal{Y}}{\operatorname{argmax}}\, s_{i,j} = c \\ 0, & \text{Otherwise} \end{cases} \tag{9}$$

The conformal score of sample i and class j, $s_{i,j}$, is the softmax outputs of the network.

V. Learning Algorithm

First, a neural network $f_{\theta,0}$ is trained on the labeled dataset D_L. Once trained, RAPS takes this pre-trained network ($f_{\theta,0}$) and predicts the probabilities estimation beside predictive sets for all unlabeled data D_U. Next, pseudo-labels are created from predictive sets following Eq. 8 (Eq. 9 for the single-label case). Then, a subset of these pseudo-labels is selected using UPS (Eq. 4). Afterward, another network $f_{\theta,1}$ is trained on both labeled data D_L and selected pseudo-labels. This procedure is repeated iteratively until the number of selected pseudo-labels converges. A new network is initialized to prevent the error propagation issue in each iteration. This algorithm is illustrated in Algorithm 2.

Algorithm 1 RAPS

```
procedure RAPS(the calibration dataset, the model, the new image)
    calibrate: perform Platt scaling on the model using the calibration set.
    get conformal scores: For each image in the training set, define $E_j = \sum_{i=1}^{k'}(\hat{\pi}_{(i)}(x_j) + \lambda\mathbb{1}[j > k_{reg}])$ where $k'$ is the model's ranking of the true class $y_j$, and $\hat{\pi}_{(i)}(x_j)$ is the $i^{th}$ largest score for the $j^{th}$ image.
    find the threshold: assign $\hat{\tau}_{ccal}$ to the $1-\alpha$ quantile of the $E_j$.
    predict: Output the $k^*$ highest-score classes, where $\sum_{i=1}^{k^*}(\hat{\pi}_{(i)}(x_{n+1}) + \lambda\mathbb{1}[j > k_{reg}]) \geq \hat{\tau}_{ccal}$.
end procedure
```

Algorithm 2 UPS algorithm using RAPS's prediction sets

```
procedure UPS-RAPS(labeled dataset $D_L$, unlabeled dataset $D_U$, $\kappa_p$, $\kappa_n$, $\tau_p$, $\tau_n$, uncertainty estimator $u$)
    Train $f_{\theta,0}$ on $D_L$
    $\tilde{D} \leftarrow D_L$
    for $k = 1 \ldots MaxIterations$ do                                          ▷ Repeats until convergence
        for $\tilde{x}^{(i)}$ in $D_U$ do
            $\mathbf{p}^{(i)}$, predictive set $\mathcal{C}^* \leftarrow$ RAPS($\tilde{D}$, $f_{\theta,i-1}$, $\tilde{x}^{(i)}$)   ▷ RAPS from Algorithm 1
            for $c$ in $\{1, \ldots, C\}$ do                                      ▷ Equation 8
                if $c$ in $\mathcal{C}^*$ then
                    $\tilde{y}_c^{(i)} \leftarrow 1$
                else
                    $\tilde{y}_c^{(i)} \leftarrow 0$
                end if
                positive mask $\leftarrow \left(u(p_c^{(i)}) \leq \kappa_p \times p_c^{(i)} \geq \tau_p\right) \times 1$
                negative mask $\leftarrow \left(u(p_c^{(i)}) \leq \kappa_n \times p_c^{(i)} \geq \tau_n\right) \times 1$
                $g_c^{(i)} \leftarrow$ positive mask + negative mask                 ▷ Equation 4
            end for
            $\tilde{\mathbf{y}}^{(i)} \leftarrow [\tilde{y}_1^{(i)}, \ldots, \tilde{y}_C^{(i)}]$
            $\mathbf{g}^{(i)} \leftarrow [g_1^{(i)}, \ldots, g_C^{(i)}]$
        end for
        $D_{selected} \leftarrow \{(\tilde{x}^{(i)}, \tilde{\mathbf{y}}^{(i)}, \mathbf{g}^{(i)})\}_{i=1}^{N_U}$     ▷ Selected pseudo-labels
        $\tilde{D} \leftarrow D_L \cup D_{selected}$
        Initialize a new network $f_{\theta,i}$
        Train $f_{\theta,i}$ using samples from $\tilde{D}$                         ▷ Using loss functions in Eq. 5-6
        $f_\theta \leftarrow f_{\theta,i}$
    end for
    return $f_\theta$
end procedure
```

VI. Discussion

The hard pseudo-labeling method for generating pseudo-labels is based on the network output probabilities, which, as claimed in the original paper, the network suffers poor calibration. Hence, using conformal prediction and predictive sets rather than hard pseudo-labels improves the performance and alleviates the noisy pseudo-labels for training.

VII. Future Work

The RAPS algorithm and generally the conformal prediction method can adapt with uncertainty estimation methods and modify the conformal score. In particular, predictive sets can be obtained by conformal scores, which are based on variance-based uncertainty scalar or any uncertainty scalar $u(X)$ [Angelopoulos and Bates, 2021].

References

A. Angelopoulos. Uncertainty sets for image classifiers using conformal prediction. https://people.eecs.berkeley.edu/~angelopoulos/blog/posts/conformal-classification/, 2021.

A. Angelopoulos, S. Bates, J. Malik, and M. I. Jordan. Uncertainty sets for image classifiers using conformal prediction. *CoRR*, abs/2009.14193, 2020. URL https://arxiv.org/abs/2009.14193.

A. N. Angelopoulos and S. Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification, 2021. URL angelopoulos.ai/publications/downloads/gentle_intro_conformal_dfuq.pdf.

D. Berthelot, N. Carlini, I. J. Goodfellow, N. Papernot, A. Oliver, and C. Raffel. Mixmatch: A holistic approach to semi-supervised learning. *CoRR*, abs/1905.02249, 2019. URL http://arxiv.org/abs/1905.02249.

D.-H. Lee. Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks. 2013.

M. N. Rizve, K. Duarte, Y. S. Rawat, and M. Shah. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. *CoRR*, abs/2101.06329, 2021. URL https://arxiv.org/abs/2101.06329.