Can a Set of Calibrated Weak Learners Create a Single Calibrated Strong Learner?

Yash Shah, Minseok Bae*, Jason Park**

{ynshah,minseok,jpark26}@stanford.edu

*Equal contribution

Introduction

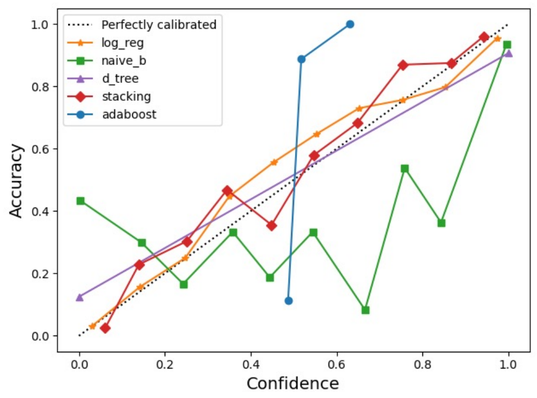
Calibration: $\mathbb{P}(\hat{y} = y|\hat{p} = p) = p, \quad \forall p \in (0,1)$
i.e., among the samples a model assigned a probability of 0.8, approximately 80% of them belong to the positive class.

Advantages

- ✓ Trustworthy, reliable, and robust models
- ✓ Can indicate when a model is likely to be incorrect and humans must be involved
- ✓ Application to safety-critical decision-making tasks:
  - ❑ autonomous driving
  - ❑ healthcare
  - ❑ time series forecasting, etc.

Problem Setup

We focus our attention toward the AdaBoost ensemble algorithm, known to be highly uncalibrated. We answer the question, “can a set of calibrated weak learners create a single calibrated strong learner?”.



Our Contributions

1. A novel algorithm using doubly- calibrated weak decision stumps to produce a single calibrated strong ensemble classifier using AdaBoost
2. Improvement for Thresholded Adaptive Calibration Error (TACE) over previous works
3. Ablation studies to verify utility and integrity of our proposed algorithm

Related Works

1) Platt Scaling [1]:

$$\mathbb{P}(y = 1|\hat{p}) = \sigma(a\hat{p} + b)$$

2) Isotonic Regression [2]:

$$\arg\min_{g} \sum_{i} (g(\hat{p}_i) - y_i)^2$$

3) Beta calibration [3]:

$$\mu_{beta}(s; a, b, c) = {}^{1}/_{1 + 1/(e^{c} s^{a}/(1-s)^{b})}$$

4) Reliability diagram: Plot of expected sample accuracy as a function of model confidence

Datasets

Name	Samples	Features	(-) class	(+) class
landsat	6435	36	{1,7}	{2,3,4,5,6}
abalone	4177	8	{1,...,9}	{10,...,29}
yeast	1484	8	{1,3}	{2,4,...,10}

From the UCI Machine Learning Repository [4]:

a) landsat: predict soil type from multi-spectral values of pixels in 3x3 tiles of a satellite image

b) abalone: predict age of abalone given sex, length, weight, rings, diameter, etc.

c) yeast: predict cellular localization sites of proteins given numeric scores and signals

We do not perform any pre-processing on these datasets.

Models

- ○ Logistic Regression: Solved using the quasi-Newton LBFGS [5] algorithm by computing estimates of the inverse Hessian matrix
- ○ Naive Bayes: A multinomial NB classifier
- ○ Decision Tree: Gini loss splits; max depth 25
- ○ Stacking: Stack of the above three classifiers; logistic regression on their outputs (5-fold CV)
- ○ AdaBoost: Sequence of 200 decision stumps fit using the SAMME.R [6] algorithm which uses soft probabilities for re-weighting

Future Work

Given that our algorithm currently is a little computation intensive, future efforts will focus on analyzing and optimizing the computational efficiency compared to post-hoc calibration techniques. Additionally, we aim to extend our research to multi-class classification tasks to broaden the utility of our findings. This will help ensure an effective balance between accuracy and computational practicality.

References

[1] Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, 1999.

[2] Zadrozny and Elkan. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *ICML*, 2001.

[3] Kull, et al. Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers. In *AISTATS*, 2017.

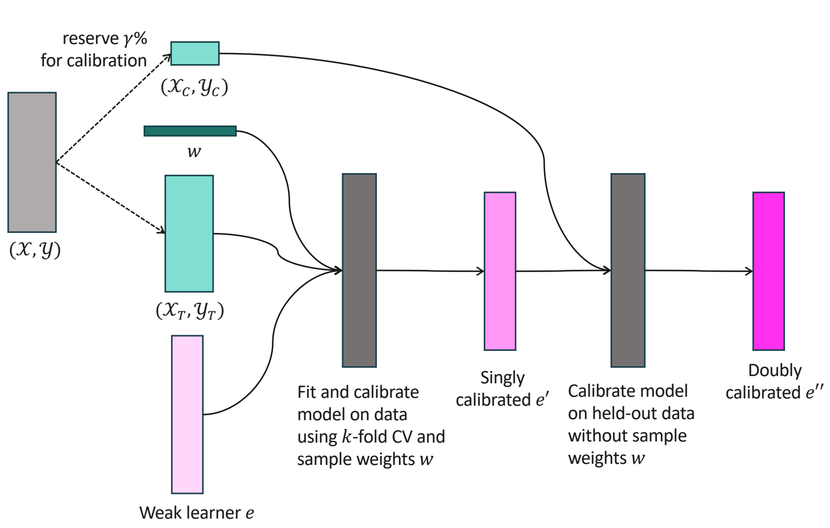
[4] Kelly, et al. The UCI Machine Learning Repository, https://archive.ics.uci.edu.

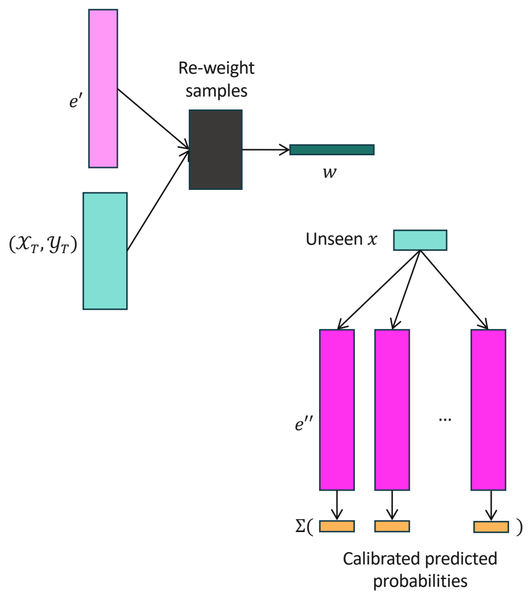
[5] Liu and Nocedal. On the limited memory bfgs method for large scale optimization. In *Mathematical Programming* 45, pp. 503–528, 1989.

[6] Zhu, et al. Multi-class adaboost. In *Statistics and its interface*, 2006.

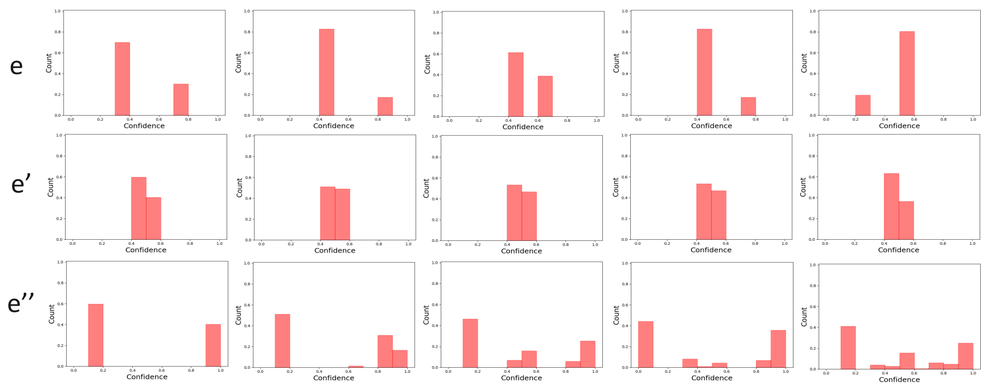
Methods

At each boosting step:





We propose a novel algorithm that uses doubly-calibrated weak decision stumps to produce a single calibrated strong ensemble classifier following the AdaBoost learning algorithm.



Results

dataset	metric	uncal	platt	isotonic	beta	ours
landsat	Train ACC (↑)	0.930	0.920	0.920	0.920	0.895
	Test ACC (↑)	0.900	0.900	0.897	0.902	0.882
	Train TACE (↓)	0.397	0.396	0.396	0.396	**0.036**
	Test TACE (↓)	0.365	0.047	0.043	0.044	**0.032**
abalone	Train ACC (↑)	0.800	0.798	0.798	0.798	0.754
	Test ACC (↑)	0.788	0.783	0.795	0.788	0.755
	Train TACE (↓)	0.282	0.283	0.283	0.283	**0.043**
	Test TACE (↓)	0.276	0.054	0.051	0.048	**0.048**
yeast	Train ACC (↑)	0.785	0.783	0.783	0.783	0.724
	Test ACC (↑)	0.734	0.714	0.720_3	0.714	0.731
	Train TACE (↓)	0.277	0.273	0.273	0.273	**0.045**
	Test TACE (↓)	0.224	0.066	0.063	0.061	**0.056**

Our experimental results are evaluated on two metrics: Accuracy (ACC) and Thresholded-Adaptive Calibration Error (TACE). We see that our model allows us to achieve the best of both worlds, outperforming traditional calibration methods and thus excelling in both accuracy and calibration.