Investigating Uncertainty in Ensemble Methods

Gurman Bhullar
Department of Computer Science
University of Toronto
gbhullar@cs.toronto.edu

Rajesh Marudhachalam
Department of Computer Science
University of Toronto
rajesh1804@cs.toronto.edu

Sarah Hindawi
Department of Computer Science
University of Toronto
shindawi@cs.toronto.edu

Sumant Bagri
Department of Computer Science
University of Toronto
sbagri@cs.toronto.edu

1 Motivation

In many high-risk applications, it is critical to estimate the uncertainty in a model's predictions in order to take safer actions, and human intervention when needed to prevent losses. In addition, we can also perform active learning, in uncertainty sampling, the active learner sequentially queries the label of those instances for which its current prediction is maximally uncertain [1]. Different types of uncertainties like data uncertainty and knowledge uncertainty can be solved using these methodologies. Gradient boosting gives a prediction model in the form of an ensemble of weak prediction models, which are typically decision trees. Hence enhancing the results to estimate the uncertainty in the predictions made by an ensemble of models can be really useful in handling abrupt and unexpected losses.

2 Related Work

Coulston et al. [2], use a Monte Carlo approach to quantify prediction uncertainty for random forest regression models. They test the approach by simulating maps of dependent and independent variables with known characteristics and compare actual errors with prediction errors. Since this approach is data driven, prediction intervals were either too wide or too narrow in sparse parts of the prediction distribution. This approach provides reasonable estimates of prediction uncertainty for random forest regression models.

The distinction between aleatoric and epistemic uncertainty has recently received a lot of attention in machine learning. Shaker and Hüllermeier [3], show how two general approaches for measuring the learner's aleatoric and epistemic uncertainty in a prediction can be instantiated with decision trees and random forests as learning algorithms in a classification setting. The first approach is based on entropy measure and the second is a measure based on relative likelihood. Both the approaches provide reasonable reasons for a learner to abstain from a prediction.

Stochastic Gradient Langevin Boosting [4] is based on a special form of the Langevin diffusion equation specifically designed for gradient boosting and theoretically guarantees the global convergence even for multimodal loss functions, while standard gradient boosting algorithms can guarantee only local optimum.

3 Project Overview and Final Goals

Our main objective is to explore the different types of uncertainties involved during training an ensemble model as well as those involved during predictions. The idea is to establish a comprehensive set of methodologies that track both the aleatoric (data uncertainty) and the epistemic (knowledge uncertainty). We start with a theoretical analysis of bayesian inference on an ensemble of probabilistic models, $\{P(y|x;\theta)\}_{m=1}^{M}$ for both regression and classification [5] establishing the mathematical definitions to compute the aleatoric and epistemic uncertainties. Next we generate boosted ensembles of GBDT (Gradient Boosted Decision Trees) using two techniques: SGB (Stochastic Gradient Boosting) and SGLB(Stochastic Gradient Langevin Boosting) as well as bagged ensemble (specifically, Random Forests). Subsequently, we will analyze the models using synthetically generated data to empirically understand the different attributes of the ensembles for estimating data and knowledge uncertainties. Finally, we will evaluate the plausibility of using the uncertainty measures to detect errors (Prediction-Rejection Ratio - PRR [6]) and out-of-domain inputs (AUC-ROC) on real-world data (such as prediction uncertainty for credit card fraud, uncertainty in prediction of medical diagnosis, etc).

As an extension, we want to investigate how the uncertainty can be affected when using boosting techniques to prune a bagged ensemble by retaining only classifiers that are essential for the classification. As suggested by [7], pruned bagging ensembles outperform full bagging in all classification tasks investigated, and they also outperform boosting in noisy classification tasks. Other mentioned potential benefits of using pruned bagging ensembles include an increase in the classification speed, lower memory requirements, and possibly improved classification performance.

References

[1] V.-L. Nguyen, M. H. Shaker, and E. Hullermeier. How to measure uncertainty in uncertainty sampling for active learning. *Machine Learning*, 111(1):89–122, 2022.

[2] J. W. Coulston, C. E. Blinn, V. A. Thomas, and R. H. Wynne. Approximating prediction uncertainty for random forest regression models. *Photogrammetric Engineering & Remote Sensing*, 82(3):189–197, 2016.

[3] M. H. Shaker and E. Hüllermeier. Aleatoric and epistemic uncertainty with random forests. In *International Symposium on Intelligent Data Analysis*, pages 444–456. Springer, 2020.

[4] A. Ustimenko and L. Prokhorenkova. Sglb: Stochastic gradient langevin boosting. In *International Conference on Machine Learning*, pages 10487–10496. PMLR, 2021.

[5] A. Ustimenko, L. Prokhorenkova, and A. Malinin. Uncertainty in gradient boosting via ensembles. *CoRR*, abs/2006.10562, 2020. URL https://arxiv.org/abs/2006.10562.

[6] A. Malinin, B. Mlodozeniec, and M. Gales. Ensemble distribution distillation. 2019.

[7] Using boosting to prune bagging ensembles. *Pattern Recognition Letters*, 28(1):156–165, 2007. ISSN 0167-8655.