Towards Well-Calibrated AutoML: A Theoretical Analysis based on Ensemble Diversity

Automated Machine Learning (AutoML) is a transformative method in ML that seeks to automate tasks such as model selection, hyperparameter tuning, and pipeline optimization. This democratization of advanced analytical tools has aided in making ML accessible to a broader range of users, including non-experts. However, most AutoML frameworks lay emphasizes on predictive accuracy while neglecting probability calibration. This is crucial for reliable uncertainty quantification in high-stake domains such as medical diagnosis, autonomous driving, and financial risk management. Poor calibration may result in misleading uncertainty estimates and harmful decision-making outcomes.

This work investigates the potential of ensemble-based AutoML to address this limitation. We focus on the widely used AUTO-SKLEARN framework (1). Drawing on recent theoretical findings, it is argued that diversity within ensembles plays a crucial role in improving probability calibration. This paper proposes a theoretical framework with ensemble diversity to reductions in log-loss, a proper scoring rule for probabilistic predictions, and examines how different phases of AUTO-SKLEARN contribute to this process. Specifically, this work analyzes the roles of Base Learners, Meta-Learning, Bayesian Optimization, and Caruana Ensemble construction in shaping both diversity and calibration quality using the AUTO-SKLEARN framework.

Empirical evaluations were conducted using 20 benchmark datasets across diverse application domains.The results shows that while Base Learners and Meta-Learning introduce a certain amount of diversity, they can likely yield poorly calibrated probabilities. Bayesian Optimization effectively reduces the log-loss but decreases diversity, whereas Caruana's ensemble method restores diversity in its construction, leading to robust and well-calibrated probability estimates. Our findings confirm that AUTO-SKLEARN, although not explicitly designed for calibration, inherently produces better uncertainty estimates through its ensemble-building process (2).

In conclusion, by uncovering the theoretical mechanisms through which ensemble-based AutoML enhances probability calibration, this work provides a foundation for the development of novel AutoML methodologies that prioritize not only predictive accuracy but also the reliability of uncertainty estimates. A key insight from our analysis is that diversity plays a crucial role in achieving well-calibrated probability estimates. Specifically, our findings show that any AutoML framework aiming to produce reliable probabilistic predictions should not only seek individual models with low log-loss but also ensure sufficient diversity among them. This balance between accuracy and diversity is essential for high-stakes applications, such as medical diagnosis, autonomous decision-making, and financial risk assessment, where miscalibrated probabilities can lead to suboptimal and potentially harmful outcomes. By leveraging these insights, future AutoML frameworks can be designed to produce models that offer both strong predictive performance and trustworthy probabilistic outputs, ultimately improving decision-making processes in critical real-world scenarios.applications.

References

[1] FEURER, M., KLEIN, A., EGGENSPERGER, K., SPRINGENBERG, J. T., BLUM, M., AND HUTTER, F. Efficient and robust automated machine learning. In *Advances in neural information processing systems* (2015), vol. 28, pp. 2962–2970.

[2] MASEGOSA, A. R. Learning under model misspecification: Applications to variational and ensemble methods. *Advances in Neural Information Processing Systems 33* (2020), 5479–5491.