Master Thesis

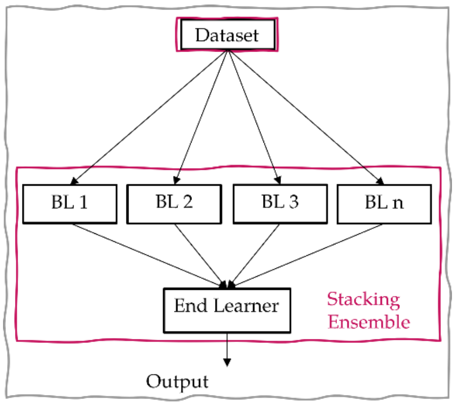
Improving Ensemble Regression Accuracy by Developing a New Metric for Base-Learner Selection and Enhancing Diversity During the Learning Process

Emad Olfatbakhsh

Data-driven modeling has long been employed to predict the behavior of complex systems in production and quality assurance. Although recent machine learning algorithms show promise, they struggle with small datasets. A stacking ensemble structure addresses this issue by employing multiple diverse algorithms to explore data from different perspectives, consolidating their results through a meta-learner (ML) for a final prediction. The challenge in stacking-based ensembles lies in achieving a trade-off between accuracy and diversity among base-learners (BLs), as well as effectively combining them using a ML.

In this master thesis, the student will first explore the concepts associated with diversity in the stacking ensemble structure and the challenges in BL selection by considering state-of-the-art associated metrics. Secondly, the student will investigate the effect of different MLs on the prediction accuracy of this structure in the process of combining selected BLs.

In the next step of the master thesis, the student will investigate the potential of enhancing ensemble learning by incorporating a new concept into the loss function of the ensemble structure. This will involve considering the ensemble structure as a whole and promoting diversity in the learning loop. To achieve this, the student will integrate the ensemble structure within an optimization loop, optimizing the hyperparameters of all BLs simultaneously. The goal is to address bias-variance and covariance issues in an ensemble structure during the learning process, with stacking aiming to find a balance in this dilemma to improve prediction accuracy. The objective is the errors of a single algorithm to compensate by other algorithms, resulting in a stacking ensemble structure with better overall prediction accuracy than a single algorithm.



The research will involve a comprehensive analysis of the effects of various BL and ML algorithms, including linear and non-linear models such as Linear Regression, Lasso, Kernel Ridge, Decision Tree, Random Forest, Gradient Boosting, Support Vector Machine, K-Nearest Neighbor, Gaussian Process Regression, and Multi-Layer Perceptron (available in Python libraries). The student will assess their performance across different complex problem domains using 10 test functions. The main purpose is to find effective stacking ensemble solutions for dealing with complex, high-dimensional problems with limited amount of data.

The work includes the following tasks:

- Familiarization with stacking ensemble learning and the concept of diversity in the context of regression.
- Methodological comparison of state-of-the-art BL selection metrics and investigation of their effects, as well as the impact of different MLs on the final performance of the ensemble structure with respect to accuracy. This aims to design a heterogeneous ensemble framework by optimally combining the selected diverse BLs.
- Develop a new BL selection metric based on acquired knowledge, which can address the weaknesses of considered metrics.
- Develop and integrate the proposed ensemble framework into the learning process, utilizing an optimization algorithm to improve BLs' accuracy and simultaneously increase their diversity via hyperparameter optimization, and combining them with an appropriate ML to enhance the overall prediction accuracy.
- Select 10 test functions and at least 2 evaluation criteria, and conduct comparative case studies using repeated cross-validation (10 runs) to ensure statistical validation.
- Result analysis and derivation of problem-specific recommendations.
- Documentation of work and colloquium presentation.

Supervisor: F. Rezazadeh M.Sc., Dr. rer. nat. H.J. Sommer, Univ.-Prof. Dr.-Ing. A. Kroll
Start: 01.02.2023
End: 30.09.2023

References:

[1] Shahhosseini, M., Hu, G., & Pham, H. „Optimizing ensemble weights and hyperparameters of machine learning models for regression problems“. In: Machine Learning with Applications, 7, 100251. 2022.

[2] Sagi, O., & Rokach, L. „Ensemble learning: A survey“. In: Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(4), e1249. 2018.

[3] Dong, X., Yu, Z., Cao, W., Shi, Y., & Ma, Q. „A survey on ensemble learning“. In: Frontiers of Computer Science, 14(2), 241-258. 2020.

[4] Palaniswamy, S. K., & Venkatesan, R. „Hyperparameters tuning of ensemble model for software effort estimation“. In: Journal of Ambient Intelligence and Humanized Computing, 12(6), 6579-6589. 2021.

[5] Purdy, D. G. „Sparse Models for Sparse Data: Methods, Limitations, Visualizations and Ensembles“. Dissertation: https://escholarship.org/uc/item/9qb472v2, University of California, Berkeley. 2012.

[6] Alizadeh, R., Jia, L., Nellippallil, A. B., Wang, G., Hao, J., Allen, J. K., & Mistree, F. „Ensemble of surrogates and cross-validation for rapid and accurate predictions using small data sets“. In: AI EDAM, 33(4), 484-501. 2019.

[7] Yu, J., Pan, R., & Zhao, Y. „High-dimensional, small-sample product quality prediction method based on mic-stacking ensemble learning“. In: Applied Sciences, 12(1), 23. 2022.

[8] Creppe, A., Rezazadeh, F., Kroll, A. „Investigation of diversity in ensemble regression for small and sparse data sets“. Internship, University of Kassel, Department of Measurement and Control. 2023.

[9] Rezazadeh, F., & Kroll, A. „Predicting the Compressive Strength of Concrete up to 28 Days-Ahead: Comparison of 16 Machine Learning Algorithms on Benchmark Datasets“. In: Proceedings 32. Workshop Computational Intelligence, 1, 52. 2022.

[10] Wood, D., Mu, T., Webb, A., Reeve, H., Lujan, M., & Brown, G. A. „Unified Theory of Diversity in Ensemble Learning“. arXiv preprint arXiv:2301.03962. 2023.

[11] Sahin, E. K., & Demir, S. „Greedy-AutoML: A novel greedy-based stacking ensemble learning framework for assessing soil liquefaction potential“. In: Engineering Applications of Artificial Intelligence, 119, 105732. 2023.