UNIKASSEL VERSITÄT

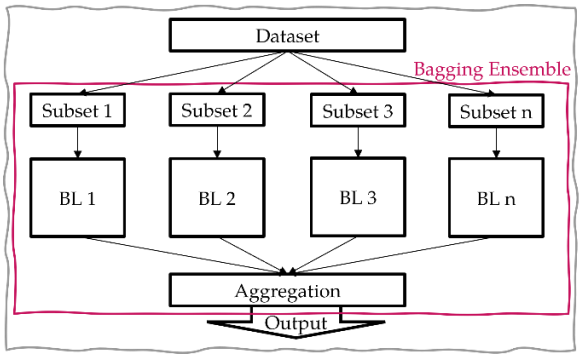
mrt
Mess- und Regelungstechnik
Univ.-Prof. Dr.-Ing. Andreas Kroll

Master- und Bachelorarbeit, Praktikum, BPS (auf Deutsch oder Englisch)

Improving Ensemble Regression Performance by Optimally Creating Specific Feature Subsets for Each Base Learner

NN

The rapid advancements in data-driven modeling have made machine learning (ML) essential for predicting complex systems in production and quality assurance. However, a key challenge arises when dealing with high-dimensional datasets that are also small in size. Such datasets can lead to overfitting, where models perform well on training data but fail to generalize to new data. Ensemble learning is a powerful approach to mitigate this issue by combining the predictions of multiple base learners (BLs), each exploring the data from different perspectives. This diversity allows the ensemble to achieve better overall performance than any single model could [1].



One of the main strategies to reduce overfitting is dimensionality reduction, which involves selecting a subset of features that capture the most relevant information from the dataset. However, the challenge becomes more complex in an ensemble structure [2], where each BL might benefit from a different subset of features, further enhancing the diversity of the model.

This internship project aims to design and implement a bagging-based ensemble framework that creates optimized feature subsets for each BL, applicable across various ML algorithms. The project will involve developing a method for feature selection that maximizes diversity among BLs while minimizing overfitting. A promising approach for this is the use of Genetic Algorithms as a multi-objective feature selection technique [3]. To achieve this, the well-developed *sing diversity metric* [4] will be employed to assess both prediction performance and the diversity of the BLs. These metrics will guide the multi-objective feature selection process to optimize the entire ensemble structure, ultimately leading to the identification of the optimal ensemble model. The intern will evaluate the performance of the proposed ensemble structure on high-dimensional, small-sized datasets and compare it to existing methods like Random Forest.

The ideal candidate should have a background in ML, proficiency in programming (Python), and a keen interest in ensemble learning. This project offers a unique opportunity to contribute to cutting-edge research in ML, with practical implications for predictive modeling in complex systems.

The work includes the following tasks:

- Literature Review and Methodological Comparison:
  - Conduct a thorough research and review of the literature to identify various feature subsetting methods specifically designed for ensemble learning in the context of small and sparse datasets.
  - Perform a methodological comparison of the selected algorithms and methods, focusing on their limitations, challenges, application areas, and design procedures.
- Design and Implementation of a Bagging Ensemble Structure:
  - Design a Bagging Ensemble structure using available sign diversity metrics as multi-objective criteria within a Genetic Algorithm optimization loop for feature selection.
  - Implement appropriate BLs using available Python libraries, such as scikit-learn.
  - Optimize the hyperparameters of the BLs using Python-based optimization algorithms.
- Application to Datasets:
  - Apply the developed methods to available high-dimensional datasets, including the UHPC dataset and at least two other test functions with high dimensionality, to evaluate the performance of the proposed ensemble structure.
- Documentation of work and colloquium presentation.

Supervisor: F. Rezazadeh M.Sc., Univ.-Prof. Dr.-Ing. A. Kroll

Start: …
End: …

References:

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

[2] 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.

[3] Rokach, L. “Genetic algorithm-based feature set partitioning for classification problems”. In: Pattern Recognition, 41(5), 1676-1700. 2008.

[4] Olfatbakhsh, E. “Improving Ensemble Regression Accuracy by Developing a New Base-Learner Selection Metric and Enhancing Diversity During the Learning Process”. Master’s thesis, University of Kassel. 2023.

[5] 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.

[6] 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.

[7] 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.

[8] Bryll, R., Gutierrez-Osuna, R., & Quek, F. “Attribute bagging: improving accuracy of classifier ensembles by using random feature subsets”. In: Pattern recognition, 36(6), 1291-1302. 2003.

[9] Błaszczyński, J., Stefanowski, J., & Słowiński, R. “Consistency driven feature subspace aggregating for ordinal classification”. In: International joint conference on rough sets, 580-589. 2016.

[10] Huang, J., Fang, H., & Fan, X. “Decision forest for classification of gene expression data”. In: Computers in biology and medicine, 40(8), 698-704. 2010.

[11] Kumar, V., & Minz, S. “Multi-view ensemble learning: an optimal feature set partitioning for high-dimensional data classification”. In: Knowledge and Information Systems, 49(1), 1-59. 2016.