UNI KASSEL VERSITÄT

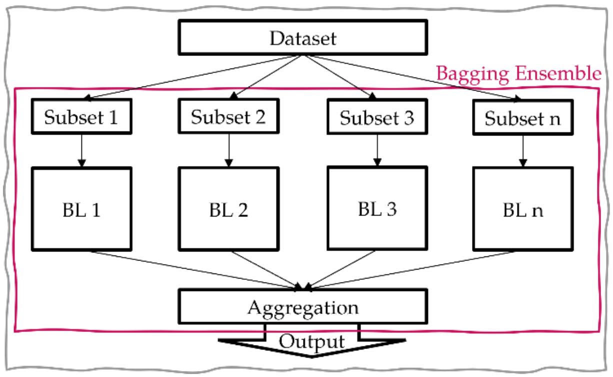
Mess- und Regelungstechnik
Univ.-Prof. Dr.-Ing. Andreas Kroll

Internship

Improving ensemble regression performance by optimally creating subsets of features from dataset for each base learner

Juan Andres Bueno Hortua

Since decades the data-driven modeling for real-time prediction of the behavior of complex systems in production and quality assurance is being investigated. Recently, several novel algorithms have been developed in the machine learning (ML) community, but training of models requires large amounts of data, which are not always available. Despite their promising empirical results and some initial successes, most ML approaches are still unable to extract interpretable information and knowledge from sparse and small datasets.



To overcome this problem, an ensemble structure is targeted here. In the ensemble structure, instead of using a single model to predict the system behavior, a combined structure with multiple algorithms is used to explore the data from different perspectives and achieve a better understanding of the patterns by combining their results into a final prediction. In this combined structure, the errors of a single algorithm are expected to be compensated by other algorithms, so that the resulting overall prediction performance of an ensemble is better than that of a single algorithm. The three main classes of ensemble learning methods are bagging, stacking, and boosting. In boosting technique, the main idea is to modify single weak learners into a strong learner by correcting prediction errors step by step, which leads to bias reduction, while the bagging technique is applied to find appropriate subsets of the dataset for each base learner (BL), which implies variance reduction [1]. Stacking (a popular meta-learning method) seeks to reduce bias and variance, or in other words, to find a balance in the bias-variance dilemma.

The main purpose of this work is to find effective solutions for dealing with high-dimensional systems for which only a small and sparse amount of data is available. This data can be quantitative or even qualitative from a static system. For this reason, ensemble learning methods such as feature sub-setting for dimension reduction for dealing with small and sparse data can be appropriate here. One criterion for subset formation can be the consideration of the monotonic constraints of the independent variables. Based on such categorization, on the one hand, dimensionality can be reduced and, on the other hand, only those inputs that show the same trend are treated in each BL. Both aspects are important for small and sparse data sets and can serve to reduce the variance dilemma.

The work includes the following tasks:

- Research and review the literature to identify different feature sub-setting ensemble methods with respect to small and sparse datasets.
- Methodological comparison of selected algorithms and methods and investigation of their limitations, challenges, application areas and design procedures.
- Investigate utilization of monotonic constraints in the available datasets for Hard Turning, Robot Arm function and Wing Weight function.
- Implement (with available Python libraries) appropriate data-driven methods for an ensemble structure.
- Create an optimal monotonic constraint-based framework for bagging the input features to reduce the dimensionality of datasets and improve the learning ability of algorithms in an ensemble structure and test the performance of the developed framework with the available datasets.
- Optimize the hyperparameters of the used algorithms using the available AutoML, NSGA-II, or Bayesian optimization libraries in Python.
- Documentation of work, technical report and colloquium presentation.

Supervisor: F. Rezazadeh M.Sc., Univ.-Prof. Dr.-Ing. A. Kroll
Start: 01.06.2023
End: 01.12.2023

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] 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.

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

[8] 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.

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

[10] 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.