UNIKASSEL VERSITÄT

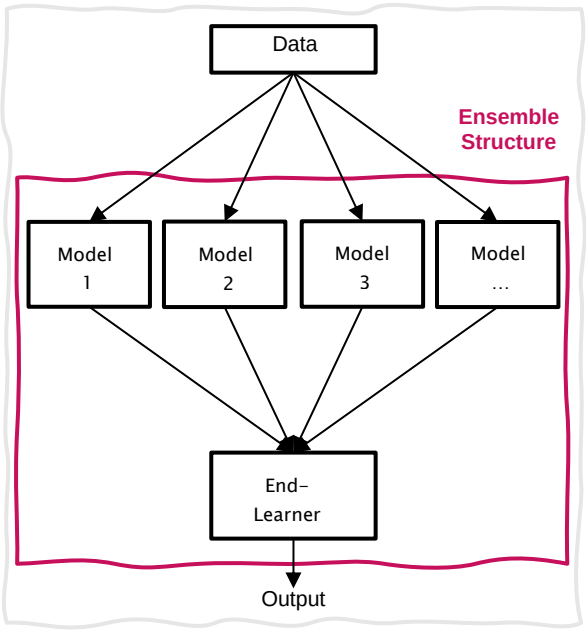
mrt
Mess- und Regelungstechnik
Univ.-Prof. Dr.-Ing. Andreas Kroll

Internship
Investigation of Ensemble Regression for small and sparse datasets

André Zanardi Creppe (ID: uk088114)

Since decades the data-driven modeling for real-time prediction of the behavior of complex systems in production and quality assurance is being investigated. The main problem in successful data-driven modeling of complex systems is the sparse and small amount of data for the modeling process. Machine learning (ML) algorithms are interesting alternatives in this respect, 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 this work, it is important to understand each of these methods in detail and consider them when modeling predictions. 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 bagging technique is applied to find the appropriate inputs for each model and also reduce dimensionality, which implies variance reduction [1].



The main purpose of this work is to find effective solutions for dealing with high-dimensional systems for which a small and sparse amount of data is available. This data can be quantitative or even qualitative in a static system. For this reason, ensemble learning methods such as feature bagging for dimension reduction and boosting method for dealing with small and sparse data can be appropriate here. This project aims to find out how the tradeoff between bias and variance in terms of the learning algorithm, the diversity of base learners, the type of algorithm, and the optimization framework in the ensemble structure can improve the performance of the algorithm.

The following steps are part of the task:

- Research and review the literature to identify different ensemble methods with respect to small and sparse datasets and also to find appropriate small and sparse datasets.
- Methodological comparison of selected algorithms and methods and investigation of their limitations, challenges, application areas and design procedures.
- Implementation (with available Python libraries) of an appropriate feature bagging method to reduce the dimension of the input space, and test it with selected datasets in the first step.
- Implementation (with available Python libraries) of a boosting method considering data augmentation technique to improve weakness prediction of baseline learners, and test it with selected datasets in the first step.
- The hyperparameters of the algorithms implemented in steps 3 and 4 should be optimized using the available AutoML, NSGA-II, and 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: 01.09.2022
End: 31.01.2023

References:

[1] SHAHHOSSEINI, Mohsen; HU, Guiping; PHAM, Hieu. Optimizing ensemble weights and hyperparameters of machine learning models for regression problems. *Machine Learning with Applications*, 2022, 7. Jg., S. 100251.

[2] SAGI, Omer; ROKACH, Lior. Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2018, 8. Jg., Nr. 4, S. e1249.

[3] DONG, Xibin, et al. A survey on ensemble learning. *Frontiers of Computer Science*, 2020, 14. Jg., Nr. 2, S. 241-258.

[4] PALANISWAMY, Sampath Kumar; VENKATESAN, R. Hyperparameters tuning of ensemble model for software effort estimation. *Journal of Ambient Intelligence and Humanized Computing*, 2021, 12. Jg., Nr. 6, S. 6579-6589.

[5] MINAEI-BIDGOLI, Behrouz; TOPCHY, Alexander P.; PUNCH, William F. A comparison of resampling methods for clustering ensembles. In: *IC-AI*. 2004. S. 939-945.

[6] LI, Yifeng; WU, Fang-Xiang; NGOM, Alioune. A review on machine learning principles for multi-view biological data integration. *Briefings in bioinformatics*, 2018, 19. Jg., Nr. 2, S. 325-340.

[7] LI, Yifeng; NGOM, Alioune. Data integration in machine learning. In: *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2015. S. 1665-1671.

[8] DIETTERICH, Thomas G. Ensemble methods in machine learning. In: *International workshop on multiple classifier systems*. Springer, Berlin, Heidelberg, 2000. S. 1-15.

[9] SENI, Giovanni; ELDER, John F. Ensemble methods in data mining: improving accuracy through combining predictions. *Synthesis lectures on data mining and knowledge discovery*, 2010, 2. Jg., Nr. 1, S. 1-126.

[10] MASOUDNIA, Saeed; EBRAHIMPOUR, Reza. Mixture of experts: a literature survey. *Artificial Intelligence Review*, 2014, 42. Jg., Nr. 2, S. 275-293.

[11] MINAEI-BIDGOLI, Behrouz, et al. Effects of resampling method and adaptation on clustering ensemble efficacy. *Artificial Intelligence Review*, 2014, 41. Jg., Nr. 1, S. 27-48.

[12] SHAHHOSSEINI, Mohsen; HU, Guiping; PHAM, Hieu. Optimizing ensemble weights for machine learning models: a case study for housing price prediction. In: *INFORMS international conference on service science*. Springer, Cham, 2019. S. 87-97.

[13] SRIVASTAVA, Nitish, et al. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 2014, 15. Jg., Nr. 1, S. 1929-1958.

[14] PURDY, David Gregory. *Sparse Models for Sparse Data: Methods, Limitations, Visualizations and Ensembles*. University of California, Berkeley, 2012.