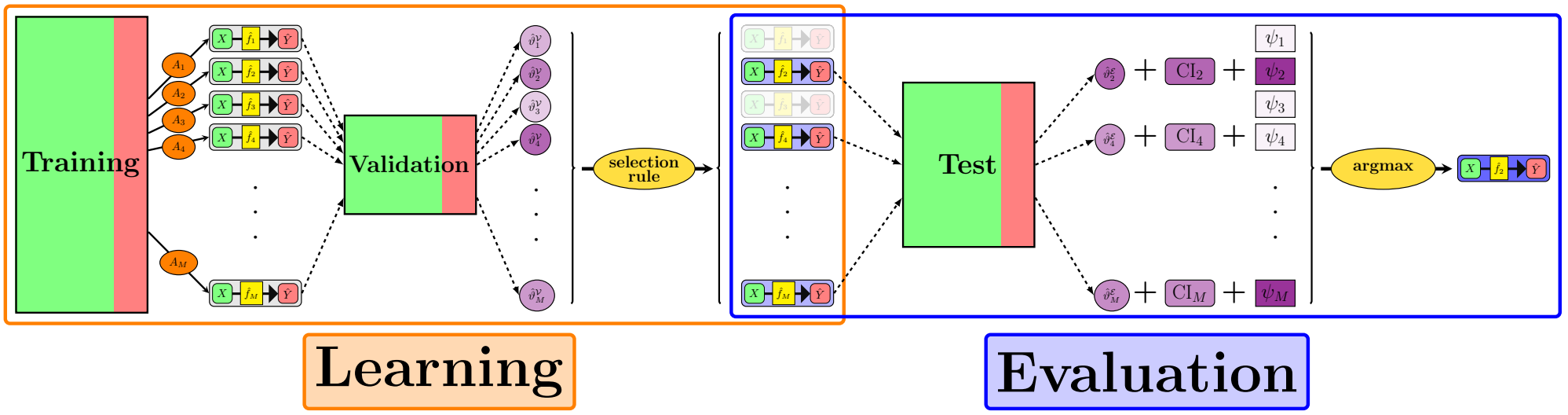
Model Selection and Evaluation in Supervised Machine Learning

Max Westphal, Institute for Statistics, University of Bremen, Bremen, Germany

RTG 2224



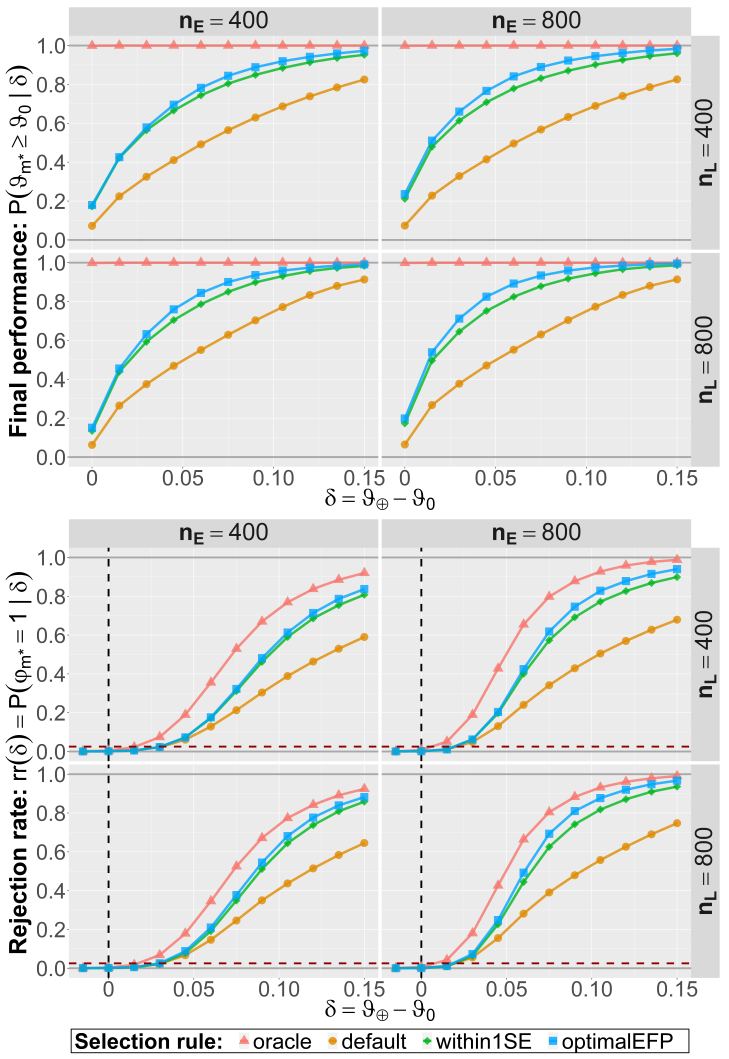
Motivation

The predominant recommendation in the machine learning literature is to strictly separate model development (training and selection) and evaluation by means of data splitting.
Following this *default* strategy to evaluate only a single final model on a distinct test dataset has two main advantages: (a) The generalization performance (or error) can be estimated without bias and (b) it is easy to test statistical hypotheses regarding the final performance. However, this approach is inflexible as one cannot alter the final model choice after the performance assessment without compromising the statistical inference (e.g. the type I error rate)
This work explores new model evaluation strategies where multiple models are deliberately assessed simultaneously to allow the test data to be used for the final model selection, hereby potentially correcting a flawed validation ranking.[1,2,3,4]

Methods

We consider evaluation strategies $\psi = (\kappa, \varphi)$ which consist of a selection rule κ based on the hold-out- or cross-validation ranking and a multiple test φ depending on the test data.
As a generic simultaneous test procedure φ, we employ the so-called maxT-approach which relies on a multivariate normal approximation and is applicable for most performance measures.[1,2] As an alternative Bayesian procedure for the analysis of multiple correlated proportions (e.g. classification accuracies), we propose a multivariate Beta-Binomial model.[3]
We compared different selection rules κ on various synthetic binary classification tasks for which $M = 200$ initial candidate models were trained via prominent learning algorithms (CART, EN, SVM, XGBoost). The results shown here concern the so-called co-primary endpoint analysis whereby sensitivity and specificity of a binary classifier or both required to be high.[4] We observed similar effects when the overall accuracy is assessed as a single endpoint.[1,2]
Besides the *default* approach (selection of single best validation model), we implemented the heuristic *within 1 SE* rule (selection of all models within 1 standard error of best validation model). Moreover, we assessed the *optimal EFP* rule which we derived in the framework of Bayesian decision theory.[4] Finally, the *oracle* rule cannot be implemented in practice but illustrates the best theoretically achievable performance.

Results & Conclusion



A disadvantage of our procedure is the introduction of a slight estimation bias which can however be corrected by means of a median-conservative point estimator.

We conclude that the test data can and should be used to improve model selection in supervised machine learning unless an unbiased estimation is deemed to be the main study goal. Hereby, statistical power can be increased while still controlling the type I error rate of false positive performance claims when a suitable adjustment for multiple comparisons is employed.

[1] Westphal, Max and Werner Brannath (2019a). "Evaluation of multiple prediction models: A novel view on model selection and performance assessment". In: Statistical Methods in Medical Research. Advance online publication. DOI: 10.1177/ 0962280219854487. URL: https://doi.org/10.1177/0962280219854487.
[2] Westphal, Max and Werner Brannath (2019b). "Improving Model Selection by Employing the Test Data". In: Proceedings of the 36th International Conference on Machine Learning. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. Long Beach, California, USA: PMLR, pp. 6747–6756. URL: http://proceedings. mlr.press/v97/westphal19a.html.
[3] Westphal, Max (2019). Simultaneous Inference for Multiple Proportions: A Multivariate Beta-Binomial Model. arXiv: 1911.00098 [stat.ME].
[4] Westphal, Max, Antonia Zapf, and Werner Brannath (2019). A multiple testing framework for diagnostic accuracy studies with co-primary endpoints. arXiv: 1911.02982 [stat.ME].

DFG Universität Bremen