Improving Model Selection by Employing the Test Data

Max Westphal [1] **Werner Brannath** [1]

Abstract

Model selection and evaluation are usually strictly separated by means of data splitting to enable an unbiased estimation and a simple statistical inference for the unknown generalization performance of the final prediction model. We investigate the properties of novel evaluation strategies, namely when the final model is selected based on empirical performances on the test data. To guard against selection induced overoptimism, we employ a parametric multiple test correction based on the approximate multivariate distribution of performance estimates. Our numerical experiments involve training common machine learning algorithms (EN, CART, SVM, XGB) on various artificial classification tasks. At its core, our proposed approach improves model selection in terms of the expected final model performance without introducing overoptimism. We furthermore observed a higher probability for a successful evaluation study, making it easier in practice to empirically demonstrate a sufficiently high predictive performance.

1. Motivation

Impressive progress has been made over the last years in a vast variety of supervised machine learning applications. End-to-end deep learning approaches have led to steadily improving results in traditional domains as well as in novel settings (LeCun et al., 2015; Jiang et al., 2017; Litjens et al., 2017; Miotto et al., 2017; Ching et al., 2018).

However, several challenges remain. In this work, we consider overfitting issues and overoptimistic claims concerning the predictive performance which are still common in applied machine learning (Boulesteix, 2009; Boulesteix & Strobl, 2009; Jelizarow et al., 2010; Boulesteix et al., 2013). This is in particular true when labelled data is expensive to acquire and datasets are thus only of modest size or the employed statistical methodology is inadequate for the particular study design. This issue is even more severe in critical applications like automated and assisted medical diagnosis where consequences could be ultimately life-threatening. In such scenarios, machine learning researches should conduct a rigorous model evaluation study and in fact may be even required to do so by regulators (Pepe, 2003; Knottnerus & Buntinx, 2009). While regulation is mandatory only in a few domains today, its introduction is heavily discussed in many environments (Cath et al., 2018; Gómez et al., 2018; Olhede & Wolfe, 2018; Pesapane et al., 2018; Reed, 2018).

From a statistical viewpoint, an evaluation study seeks to provide evidence that a novel prediction model has a sufficiently high performance compared to the reference standard (which provides the true labels). The true performance is of course unknown in practice and needs to be estimated based on data. Much can go wrong in this process which is why across the machine learning literature the following three-way data split is frequently recommended (Friedman et al., 2009; Japkowicz & Shah, 2011; Zheng, 2015; Goodfellow et al., 2016; Géron, 2017):

1. **Training**: used to train all initial candidate models.
2. **Validation**: used for algorithm / hyperparameter selection, the final model results from retraining this algorithm on training and validation data.
3. **Evaluation** (Test): used to access the performance of the final model.

A standard approach to address sampling variability of the empirical test performance, is to conduct (frequentist) statistical inference of the unknown generalization performance ϑ. This allows the construction of confidence intervals

$$\mathrm{CI}_{1-\alpha}(\vartheta) = (\vartheta^l, \vartheta^u) \tag{1}$$

based on the test data such that $\mathrm{CI}_{1-\alpha}(\vartheta)$ covers ϑ with probability $1-\alpha$. More formally, a test decision for the hypothesis problem

$$H_0 : \vartheta \leq \vartheta_0 \quad \text{vs.} \quad H_1 : \vartheta > \vartheta_0 \tag{2}$$

[1]Institute for Statistics, Faculty 3: Mathematics and Computer Science, University of Bremen, Bremen, Germany. Correspondence to: Max Westphal <mwestphal@uni-bremen.de>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

can be made. Hereby, it is required that the type 1 error rate, the probability of falsely rejecting the null, is (approximately) bounded by the significance level α, e.g. $\alpha = 0.05$. This approach may be motivated as follows: We assume that there is an existing comparator for the prediction task at hand, either in form of a threshold ϑ_0 or the performance of a reference model $\hat{f}_0$. For instance, an automated disease diagnosis system might only be deemed as useful when it can provide a classification accuracy of at least $\vartheta_0 = 90\%$ due to other advantages, e.g. decreased costs or lower invasiveness compared to the reference standard. Similarly, a company may decide to only replace their existing (and working) stock price forecasting model if enough evidence is available that the new candidate model performs better. It is sometimes criticized that the outcome of a statistical test is ultimately binary (reject or don't reject). We argue however that this binary outcome perfectly matches the pending decision in the given application: either we implement the candidate model in practice or we don't.

The statistical testing approach implicitly reflects that we hereby usually perceive type I errors (concluding a model performs well enough when it doesn't) to be more harmful than type II errors (failing to conclude superiority of a sufficiently good model). This view may be too rigorous for early-phase studies or algorithmic development attempts. It is however justified in critical applications and regulated environments.

Several works investigate and compare statistical methods for the evaluation of prediction models usable for this framework (García et al., 2010; Japkowicz & Shah, 2011; Raschka, 2018). Some of these methods are tailored towards the comparison of learning algorithms rather than models (Dietterich, 1998; Nadeau & Bengio, 2000; Hothorn et al., 2005; Eugster et al., 2008). In this case, each algorithm under investigation produces several models based on different datasets or partitions of the same dataset. The results of such a (benchmark) experiment are than based on a summary statistic of different performance estimates per algorithm, e.g. the average accuracy. We will however focus on model comparison and evaluation in this work. In the terminology of Dietterich (1998, figure 1), we primarily address research questions 1 and 3.

In a recent comprehensive overview article, Raschka (2018) presented different multiple tests which allow to conduct simultaneous inference regarding more than one models on a single test dataset while still controlling the probability of making a false positive claim. Raschka (2018) states that 'if we are honest and rigorous, the process of multiple hypothesis testing with appropriate corrections can be a useful aid in decision making.' We certainly agree with this statement. However, based on our observations, these methods are still not commonly employed in applied machine learning.

A novel perspective on model selection and evaluation which helps to justify the use of multiple testing methodology for model evaluation was presented by Westphal & Brannath (forthcoming 2019). In a nutshell, the authors propose to evaluate multiple models on purpose to allow the final model being selected on the test data. The resulting selection induced bias is countered by using a multiple test correction based on the approximate multivariate distribution of performance estimates. Westphal & Brannath (forthcoming 2019) used the so-called maxT-approach which is not well-known in the machine learning literature but has several appealing properties which we will highlight later. In numerical simulations of many repetitions over the complete learning-evaluation pipeline this approach resulted in final models with higher performance and also increased statistical power, i.e. the probability that the null hypothesis is correctly rejected. An additional motivation for this approach is given in case learning and evaluation data differ systematically, e.g. when the learning data is sampled from a perturbed data distribution. While this should generally be avoided, it is not uncommon in practice. For instance, diagnostic models in medical research are commonly trained on retrospectively collected data with potentially unrepresentative characteristics compared to the intended target population, e.g. on patients with less/more/other comorbidities or following different sample protocols. The evaluation data on the other hand is usually highly representative due to a prospective study design with restrictive inclusion criteria. If this is the case, it can be frustrating not to be allowed to use the test data for model selection for the sake of an unbiased estimation and valid statistical inference for the performance of the single prespecified model. We are not aware of other publications exploring a similar route, i.e. an intended model selection based on the test data.

The outline of this work is as follows: In the second section we will briefly introduce basic notation and the most important statistical aspects. In the third section we show results of new and extensive numerical simulations in which we compare the properties of different evaluation strategies. Finally, in the fourth section, we summarize our results, point out limitations of our work and give an outlook on future research.

2. Notation and Statistical Framework

2.1. The Default Data Splitting Approach

The goal of supervised machine learning is to learn a prediction model $\hat{f} : \boldsymbol{x} \mapsto \hat{y}$ which maps a P-dimensional feature vector $\boldsymbol{x}$ to a (scalar) label $\hat{y}$. This is achieved by a machine learning algorithm A which takes a learning set $\mathcal{L} = \{\boldsymbol{z}_i = (\boldsymbol{x}_i, y_i)\}_{i=1}^{n_\mathcal{L}}$ as input. Hereby, $\mathcal{L}$ is assumed to be sampled from the unknown joint data distribution $\mathfrak{D} = \mathfrak{D}_{(\boldsymbol{X},Y)}$, we write $\mathcal{L} \sim \mathfrak{D}^{n_\mathcal{L}}$ in short. Usually several

models are trained as it is rarely known beforehand which algorithm provides the best results for the problem at hand. This can be justified with the no free lunch theorem of machine learning (Shalev-Shwartz & Ben-David, 2014). We assume that that M algorithms A_m are used to train models $\hat{f}_m$, $m \in \mathcal{M} = \{1, \ldots, M\}$. A learning algorithm may depend on one or more so-called hyperparameters, e.g. the strength of a penalty term. In the following, we will treat the same algorithm with two different hyperparameter choices as two different algorithms A_m, A_k, $m \neq k$.

The true performance of the m-th model $\hat{f}_m$ is defined as

$$\vartheta_m = \vartheta(\hat{f}_m) = \mathbb{E}_{\mathfrak{D}}[s(\hat{f}_m(\boldsymbol{X}), Y)]. \tag{3}$$

Hereby, s is a similarity measure between prediction and label such as $s(\hat{y}, y) = \mathbb{1}(\hat{y} = y)$ for (binary) classification tasks (on which we will focus later). In practice, only sample estimates (empirical performances) $\hat{\vartheta}_m = \frac{1}{n}\sum_i s(\hat{y}_i, y_i)$ are available whereby the summation is meant to include all observations in the relevant dataset. It is well known that model selection based on the same data that was used for training generally favours (over)complex models. The usual recommendation to avoid overfitting the training data is to estimate performances on independent data (Friedman et al., 2009; Japkowicz & Shah, 2011; Zheng, 2015; Goodfellow et al., 2016; Géron, 2017). Hence we assume that the learning data is split into training data $\mathcal{T} \sim \mathfrak{D}^{n_\mathcal{T}}$ and validation data $\mathcal{V} \sim \mathfrak{D}^{n_\mathcal{V}}$. The performance of the resulting intermediate models $\hat{f}_m^- = A_m(\mathcal{T})$, $m \in \mathcal{M}$, (with training restricted to $\mathcal{T}$) are compared based on the validation performances. Usually, the empirical validation performance $\hat{\vartheta}^-$ is maximized to select the best algorithm $m^* = \arg\max_{m \in \mathcal{M}} \hat{\vartheta}_m^-$. The final prediction model $\hat{f}_{m^*}$ is the result of retraining the algorithm A_{m^*} on the whole learning data $\mathcal{L} = \mathcal{T} \cup \mathcal{V}$ as a non-decreasing generalization performance (on average) can be expected from any reasonable learning algorithm. We will call this approach the *default* selection rule hereafter.

A third and final dataset is eventually needed to allow for an unbiased performance estimation and, if needed, rigorous statistical inference regarding the true performance of the final model ϑ_{m^*}. This is due to the fact that the validation estimate for the final model m^* is overoptimistic due to selection induced bias. The magnitude of this bias is increasing in the number of models M that are compared (Jensen & Cohen, 2000). The final evaluation (or test) dataset will be denoted as $\mathcal{E} \sim \mathfrak{D}^{n_\mathcal{E}}$. This simple evaluation strategy works because of the complete separation of model selection (based on $\mathcal{V}$) and evaluation (based on $\mathcal{E}$). However, it might suffer if the selection results in a suboptimal model. From a strict statistical viewpoint (i.e. if we seek to control the type 1 error rate), we are not allowed to change our hypothesis which is implied by our model choice after we have investigated the evaluation data.

2.2. Simultaneous Evaluation of Multiple Models

The evaluation approach proposed by Westphal & Brannath (forthcoming 2019) can be formalized by allowing arbitrary selection rules $\boldsymbol{r} : \mathcal{L} \mapsto \mathcal{S} \subset \mathcal{M} = \{1, \ldots, M\}$. That is to say, a subset of models $\mathcal{S} \subset \mathcal{M}$ is selected for evaluation based on $\mathcal{L}$, particularly based on $\hat{\boldsymbol{\vartheta}}^-$. For simplicity, we will assume that the models are ordered such that the validation estimates form a decreasing sequence $\hat{\vartheta}_1^- \geq \ldots \geq \hat{\vartheta}_M^-$. In effect, we have $\mathcal{S} = \{1, \ldots, S\}$, $S \leq M$. In the evaluation phase we now have to simultaneously evaluate all selected models $m \in \mathcal{S}$. By evaluation we mean estimation of the unknown parameter vector $\boldsymbol{\vartheta} = (\vartheta_1, \ldots, \vartheta_S)$ and a test decision for the hypothesis system

$$\mathcal{H} = \{H_0^m : \vartheta_m \leq \vartheta_0,\ m \in \mathcal{S}\}. \tag{4}$$

The goal of the evaluation study is hence to reject at least one null hypothesis H_0^m, i.e. provide evidence that at least one model has a sufficiently high performance. We are particularly interested in the unknown performance ϑ_{m^*} of the final model. In the remainder of this work, we will focus on the most obvious way to choose the final model, namely the model with the highest evaluation performance $m^* = \text{argmax}_{m \in \mathcal{S}} \hat{\vartheta}_m(\mathcal{E})$. As a result model selection and evaluation now overlap (if $S > 1$) and statistical methods are needed to resolve bias introduced this way.

In general, a multiple test may be used to deal with hypothesis system (4). A multiple test is a mapping $\boldsymbol{\varphi} : \mathcal{E} \mapsto \{0, 1\}^S$ whereby hypothesis H_0^m gets rejected if and only if $\varphi_m = 1$. We will employ a particular test, the so-called maxT-approach which is also known as the projection method (Hothorn et al., 2008; Dickhaus, 2014). It is based on the approximate multivariate normal distribution of the sample estimates $\hat{\boldsymbol{\vartheta}} \dot{\sim} \mathcal{N}_S(\boldsymbol{\vartheta}, \hat{\boldsymbol{\Sigma}})$ where $\hat{\boldsymbol{\Sigma}}$ is a consistent estimate of the true covariance matrix $\boldsymbol{\Sigma}$ of $\hat{\boldsymbol{\vartheta}}$. Usually, when performance is defined as in (3), $\hat{\boldsymbol{\Sigma}}$ is the standard sample covariance of the similarity matrix

$$\boldsymbol{Q} = \left(s(\hat{f}_m(\boldsymbol{x}_i), y_i)\right)_{\substack{i=1,\ldots,n_\mathcal{E} \\ m=1,\ldots,S}} \tag{5}$$

scaled by the factor $1/n_\mathcal{E}$. To compute $\boldsymbol{Q}$, predictions for all evaluation observations need to be obtained for all selected models. In the following, we assume that the sample covariance matrix can be decomposed as $\hat{\boldsymbol{\Sigma}} = \hat{\boldsymbol{D}}^{-1/2}\hat{\boldsymbol{R}}\hat{\boldsymbol{D}}^{-1/2}$ whereby $\hat{\boldsymbol{D}}$ is the diagonal matrix of sample variances and $\hat{\boldsymbol{R}}$ the estimated correlation matrix. From this, a vector of test statistics can be constructed as

$$\boldsymbol{T} = \hat{\boldsymbol{D}}^{-1/2}(\boldsymbol{\vartheta} - \boldsymbol{\vartheta}_0). \tag{6}$$

Under the least favourable parameter configuration $\boldsymbol{\vartheta} = \boldsymbol{\vartheta}_0 = (\vartheta_0, \ldots, \vartheta_0) \in \mathbb{R}^S$ we have

$$\boldsymbol{T} \dot{\sim} \mathcal{N}_S(\boldsymbol{0}, \hat{\boldsymbol{R}}) \tag{7}$$

due to the multivariate central limit theorem. Standard packages allow the numerical calculation of a critical value c_α such that, given $\boldsymbol{\vartheta} = \boldsymbol{\vartheta}_0$,

$$\mathbb{P}(\max_{m \in \mathcal{S}} T_m \leq c_\alpha) \approx \int_{(-\infty, c_\alpha]^S} \phi_S(\boldsymbol{x}, \hat{\boldsymbol{R}}) d\boldsymbol{x} = 1 - \alpha. \tag{8}$$

Hereby, $\phi_S(\cdot, \hat{\boldsymbol{R}})$ is the density function of the $S-$dimensional standard normal distribution with correlation matrix $\hat{\boldsymbol{R}}$. Altogether, this leads to a simultaneous test procedure $\boldsymbol{\varphi}$ by defining $\varphi_m = 1 \Leftrightarrow T_m > c_\alpha$. This test controls the family wise error rate (FWER) in the strong sense asymptotically (as $n_\mathcal{E} \to \infty$). That is to say that the probability of any hypotheses in (4) being falsely rejected is bounded by the significance level α for all possible parameter configurations $\boldsymbol{\vartheta}$. If $S = 1$, the maxT-approach is equivalent to a standard Z-test. It is also possible to construct a simultaneous confidence region for the parameter vector $\boldsymbol{\vartheta}$ with (approximate) coverage probability $1 - \alpha$. Additionally, a corrected point estimator $\tilde{\boldsymbol{\vartheta}}$ may be defined via

$$\tilde{\vartheta}_m = \hat{\vartheta}_m - c_{0.5} \cdot \hat{\text{se}}(\hat{\vartheta}_m), \tag{9}$$

whereby $c_{0.5}$ fulfils equation (8) with $\alpha = 0.5$. This corrected estimator is median-conservative, i.e. the probability $\mathbb{P}(\cup_{m \in \mathcal{S}} \{\tilde{\vartheta}_m > \vartheta_m\})$ that any component $\tilde{\vartheta}_m$ overestimates the true performance ϑ_m is approximately (as $n_\mathcal{E} \to \infty$) bounded by 0.5. Further details are provided by Westphal & Brannath (forthcoming 2019).

The maxT-approach due to Hothorn et al. (2008) is not widespread in applied machine learning but has several appealing properties: (a) It is applicable to a wide variety of performance measures due to the employed normal approximation. (b) It takes into account the similarity of the models via the estimated correlation structure of the empirical performances. It is therefore less strict (smaller c_α) if similar models are evaluated. Simpler approaches such as the Bonferroni correction ignore the correlation structure completely (Dickhaus, 2014). (c) The framework allows corrected (median-conservative) performance estimates via (9) which is useful even if statistical testing is not of direct scientific interest. On the downside, the approximate nature of the procedure may result in too liberal test decision for small test sets. In the following section, the pros and cons of our multiple testing approach will be assessed systematically in an extensive simulation study.

3. Simulation Study

3.1. Goal

The goal of our simulation study is to compare different model evaluation strategies. We will combine different selection rules (based on the validation data) and the same statistical inference framework (on the test data). That is to say, the same multiple test we be employed in all cases, namely the maxT-approach described in the last section. Our main focus is how different selection rules impact the final model performance. Furthermore, we will compare estimation bias, statistical power and type 1 error rate.

The simulation study is inspired by previous work but was vastly extended regarding several aspects (Westphal & Brannath, forthcoming 2019). Most importantly, the diversity of learning algorithms (EN, RPART, SVM, XGB) was increased. The diversity of learning tasks was also increased as we now deal with unbalanced classification problems based on nonlinear risk scores. In addition, the number of candidate algorithms was increased to $200 = 4 \cdot 50$ and the hyperparameters are now sampled randomly instead of grid based. In total, we simulated 72,000 instances of the complete machine learning and evaluation pipeline and trained 200 models twice (pre and post validation) on each instance.

3.2. Software

All numerical experiments have been conducted in `R` (R Core Team, 2013). We used many existing packages, most importantly the `batchtools` (Lang et al., 2017) package for processing batch jobs and the `mvtnorm` (Genz et al., 2018) for computations concerning the multivariate normal distribution. For the machine learning part, we employed the `caret`[1] package as a wrapper for methods from `glmnet`, `rpart`, `LiblineaR`, and `xgboost`.

In addition, two newly developed packages were used: `SEPM`[2] (Statistical Evaluation of Prediction Models) provides the selection and statistical inference framework. `SEPM.MLE`[3] provides all functions used to conduct the numerical experiments presented in this work. To reproduce the simulations and analyses we recommend to follow the instructions provided in the public `R` project `SEPM.PUB`[4]

3.3. Setup

For the simulation study, we will focus entirely on binary classification and prediction accuracy. In the following, we will describe some of its main features.

3.3.1. Data generation

Training, validation and evaluation datasets $\mathcal{T}, \mathcal{V}$ and $\mathcal{E}$ are sampled from the same distribution $\mathfrak{D} = \mathfrak{D}_{(\boldsymbol{X}, Y)}$. In addition, a large population dataset $\mathcal{P}$ with 100,000 observations is generated. It is not used for selection or

[1] https://github.com/topepo/caret/
[2] https://github.com/maxwestphal/SEPM
[3] https://github.com/maxwestphal/SEPM.MLE
[4] https://github.com/maxwestphal/SEPM.PUB

evaluation but rather to calculate the ground truth (e.g. true classifier accuracies) with high precision. The standard error of calculated 'true' proportions is bounded by $\sqrt{0.5(1-0.5)/n_{\mathcal{P}}} \leq 0.0016$.

For each dataset, we firstly generate feature data $\boldsymbol{X} \in \mathbb{R}^P$ from a multivariate standard normal distribution with $P = 50$ features whereby only $P_{act} < P$ active features contribute to define the labels Y. In half of the cases the features are independent of each other. For the other half of instances features are redundant. That is to say, for each active feature, we have one 'partner' feature which is correlated ($\rho = 0.5$) to the active feature but does not contribute any information towards the true label.

True labels are then generated based on a risk score which is a function $g : \boldsymbol{x} \mapsto p = \mathbb{P}(Y = 1|\boldsymbol{X} = \boldsymbol{x}) \in (0, 1)$. We used six different risk scores. The first two (A, B) are linear in $\boldsymbol{X}$ and depend on $P_{act} = 5$ features (Westphal & Brannath, forthcoming 2019). They result in balanced learning problems, i.e. $\pi = \mathbb{P}(Y = 1) = 0.5$. The next two (C, D) are nonlinear risk scores also based on $P_{act} = 5$ features. The last two (E, F) are also nonlinear and based on $P_{act} = 9$ features. The risk scores for scenarios C through F were inspired by prediction tasks used by Friedman (1991, p. 35) and Breiman (1996, p. 139) and are tuned such that $\pi_C = \pi_E = 0.3$ and $\pi_D = \pi_F = 0.15$. Besides these realistic class balances we checked that all tasks have a realistic optimal performance ϑ_{opt} which is defined as the accuracy of the classifier resulting from thresholding the true (data-generating) risk score at 0.5, compare table 1.

Altogether we investigate $24 = 6 \cdot 2 \cdot 2$ different scenarios resulting from all combinations of risk score (A-F), feature distribution (independent [I] or redundant [R]) and size of the learning data $n_{\mathcal{L}}$ (either 400 or 800). The validation data size was set to $n_{\mathcal{V}} = n_{\mathcal{L}}/4$ in all cases. For each of the 24 distinct scenarios, we generated 3,000 data instances.

3.3.2. Machine learning

For every data instance, we train $M = 200$ models with randomly sampled hyperparameters on the training data $\mathcal{T}$ and again (with the same hyperparameters) on the learning data $\mathcal{L} = \mathcal{T} \cup \mathcal{V}$. We employed the following learning algorithms (in brackets: number of hyperparameters):

- EN: Elastic Net - Penalized Logistic Regression (2)
- CART: Cost-Sensitive Classification and Regression Trees (2)
- SVM: L2 Regularized Linear Support Vector Machines with Class Weights (3)
- XGB: eXtreme Gradient Boosting (7)

Each algorithm was used to train $M/4 = 50$ models on each data instance. These particular algorithms were chosen in order to achieve a good compromise between high performance and low training time. Details regarding the implementation and theoretical background are given in the `caret` documentation[1] and by Kuhn & Johnson (2013).

3.3.3. Selection rules

Our main focus is the comparison of the following two selection rules, motivated by previous research (Westphal & Brannath, forthcoming 2019).

(a) ***default***: evaluate the best the validation model only

(b) ***within 1 SE***: evaluate all models within one standard error of the best validation model

To reduce the computational costs for our simulation and potentially avoid a loss in statistical power we limited the number of selected model by imposing the condition $S \leq \lfloor \sqrt{n_{\mathcal{E}}} \rceil$. This can be motivated by noting that we have $\mathrm{tr}(\boldsymbol{\Sigma}) = \mathcal{O}(S n_{\mathcal{E}}^{-1/2})$ for the trace of the covariance matrix of $\hat{\boldsymbol{\vartheta}}$. Furthermore, we also included two oracle rules for comparison which cannot be implemented in practice. The *oracle [train]* rule selects the single best model based on true performances ϑ_m^- (before retraining) and the *oracle [learn]* rule is based on the true performances ϑ_m of the final models. For each of the 72,000 learning data sets, five test sets with different sizes $n_{\mathcal{E}} \in \{100, 200, 400, 800, 8000\}$ were sampled. All evaluation strategies are finally employed to the same 360,000 resulting combinations $(\mathcal{L}, \mathcal{E})$.

3.4. Main Results

The main results described here are averaged over all learning instances and stratified by relevant factors. A much more detailed supplementary report with several additional analyses is available online[5].

3.4.1. Prediction tasks

The 24 considered scenarios are summarized in table 1. The six different data distributions A-F are defined by the according risk score which are either linear (L) or nonlinear (N). The label depends either on $P_{act} = 5$ or 9 active features. The fourth column shows the prevalence of the positive class $\pi = \mathbb{P}(Y = 1)$. For each combination of data distribution A-F and learning sample size $n_{\mathcal{L}} \in \{400, 800\}$ the two rightmost column show the mean over all corresponding simulation instances of the maximum true performance $\vartheta_{max} = \max_{m \in \mathcal{M}} \vartheta_m$ for the case $M = 200$. These numbers are somewhat smaller when $n_{\mathcal{L}} = 400$ and when the

[5] https://maxwestphal.github.io/SEPM.PUB/MLE_SIM_ACC.html

Table 1. Description of the classification tasks investigated in the simulation study. Each task is defined by the its label distribution (A-F), number of learning samples (400 or 800) and an independent (I) or redundant (R) feature distribution.

$\mathfrak{D}$	TYPE	P_{act}	π	ϑ_{opt}	$n_{\mathcal{L}}$	ϑ_{max} [I]	ϑ_{max} [R]
A	L	5	.50	.885	400	.878	.877
					800	.882	.881
B	L	5	.50	.860	400	.850	.850
					800	.856	.855
C	N	5	.30	.951	400	.885	.853
					800	.867	.865
D	N	5	.15	.966	400	.892	.890
					800	.901	.899
E	N	9	.30	.950	400	.854	.852
					800	.865	.864
F	N	9	.15	.963	400	.891	.890
					800	.900	.900

features are independent (I) rather than redundant (R), as expected. The variation of ϑ_{max} across simulation instances is rather small, the largest standard deviation of 0.009 was observed for task (C, 400, I). The distance of ϑ_{max} to ϑ_{opt}, the optimal achievable performance, can be seen as a proxy of how well the considered algorithms are able to learn each task. Overall, the XGBoost algorithm is most likely to learn the best prediction model for most tasks. For tasks A and B however, the elastic net frequently provides the best models.

3.4.2. Final model performance

Our main goal is to investigate if our approach can be successfully used to improve model selection. Figure 1 shows the empirical distribution of performance gains Δ_ϑ as observed in the simulation study. Δ_ϑ is defined as the difference between the true model performance ϑ_{m^*} resulting from using the *within 1 SE* selection rule and the *default* selection rule. A positive Δ_ϑ means a higher performance for the former approach.

The mean performance gain is always greater than zero and this effect is increasing in $n_{\mathcal{E}}$ and decreasing in $n_{\mathcal{V}} = n_{\mathcal{L}}/4$. This is also true when stratifying the analysis further for learning task (see supplementary report). The magnitude of the mean effect varies between 0.8% and 0.9% accuracy if $n_{\mathcal{E}} = n_{\mathcal{V}}$. We furthermore observe that the distribution of Δ_ϑ is skewed as a large gain (5% or greater) is much more likely than an equally great loss in model performance. We also tested if the expected performance gain is significantly different from zero for the primary analysis ($M = 200$, $n_{\mathcal{V}} = n_{\mathcal{E}}$). The according null hypothesis $H_0 : \mathbb{E}\Delta_\vartheta = 0$ can be rejected in the overall analysis ($n_{sim} = 72000$) for the significance level $\alpha = 5\%$. This is also the case when conducting Bonferroni-corrected tests in all 24 encompassed scenarios ($n_{sim} = 3000$). That is to say, in all 24 individual scenarios, the multiplicity-adjusted lower confidence bounds for $\mathbb{E}\Delta_\vartheta$ are greater than zero.

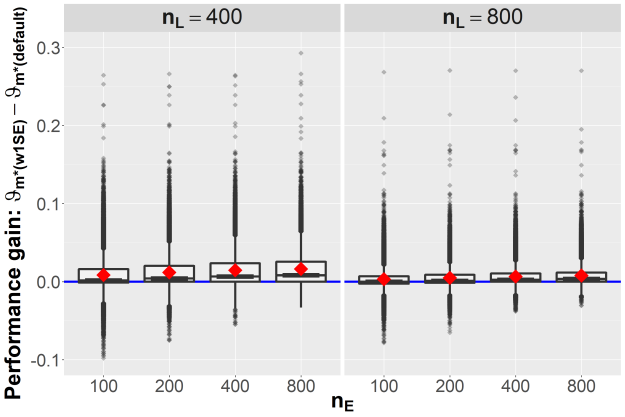

Figure 1. Gain in final model performance Δ_ϑ when the final selection is conducted on the test data depending on the number of learning observations $n_{\mathcal{L}}$ and the number of test observations $n_{\mathcal{E}}$.

3.4.3. Estimation

Evaluating only a single model has the undeniable advantage that an unbiased point estimate can be obtained from the test data. In contrast, evaluating multiple promising models and choosing the final model based on the test data is prone to selection induced bias. This is shown in the top row of figure 2: The mean relative bias of the naive estimator (sample mean) is around $+2\%$ for small test sets. This is the reason why the former approach is so popular and frequently recommended in the literature.

In contrast, the corrected estimator $\tilde{\vartheta}_{m^*}$ introduced in equation (9) is rather biased downward by a similar margin (bottom row of figure 2). This bias also vanishes asymptotically. This conservatism makes the corrected estimate preferable for evaluation studies. Besides the bias, we also analysed the mean squared error and mean absolute deviation of $\hat{\vartheta}_{m^*}$ and $\tilde{\vartheta}_{m^*}$. Regarding these characteristics, the two estimators perform very similarly in our simulation.

3.4.4. Test decisions

In a rigorous model evaluation study, we may require that rejection of the null hypothesis $H_0^{m^*} : \vartheta_{m^*} \leq \vartheta_0$ is necessary to conclude superiority of the final model $\hat{f}_{m^*}$ over the benchmark ϑ_0. Figure 3 shows exactly this rejection rate $\pi(\delta) = \mathbb{P}(\varphi_{m^*} = 1|\delta)$ given $\delta = \vartheta_{max} - \vartheta_0$. Hereby ϑ_{max} is the true performance of the best candidate model $m \in \mathcal{M}$, compare table 1.

When $\delta \leq 0$, the global null is true, i.e. all models are worse than the benchmark ϑ_0. In this case $\pi(\delta)$ coincides with the type 1 error rate of the statistical test and hence should be bounded by α. The (one-sided) significance level $\alpha = 2.5\%$ is indicated as the red dashed line in figure 3. Apparently,

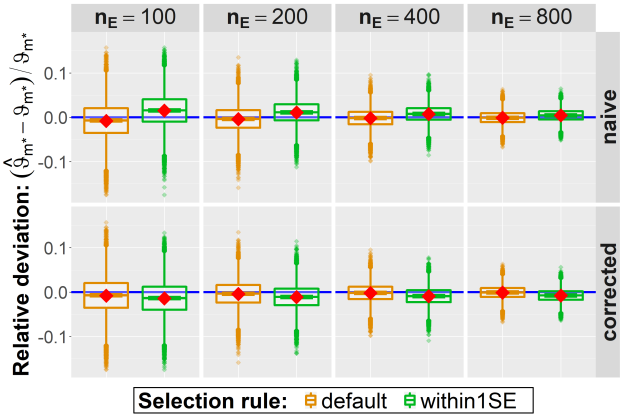

Figure 2. Relative deviation between estimated and true final model performance for the naive (top) and corrected (bottom) estimator. In this figure, only instances with $n_{\mathcal{L}} = 400$ are shown.

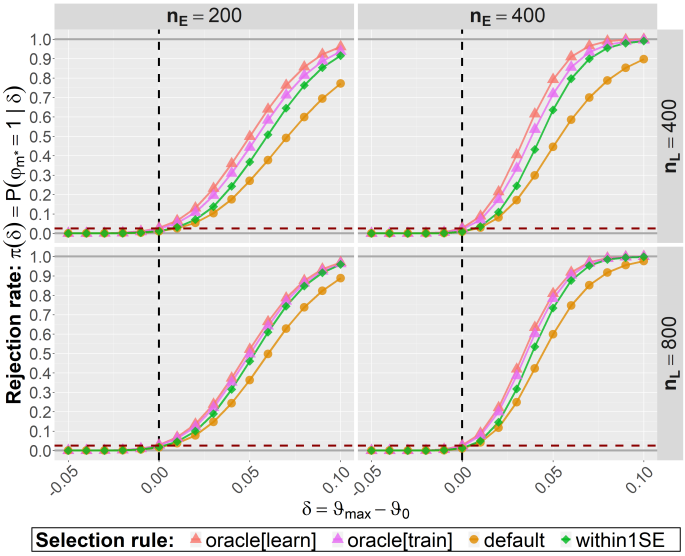

Figure 3. Overall rejection rate, i.e. probability to falsely ($\delta \leq 0$) or correctly ($\delta > 0$) identify a sufficiently good prediction model.

control of the type 1 error rate is possible for all employed selection rules. Only when model selection is perfect (*oracle [learn]* rule), the observed false positive rate of the test is slightly above α. This can be explained by the asymptotic nature of the employed statistical test. We remark here, that the test decision for all models $m \notin \mathcal{S}$ which have not been selected is to retain the null hypothesis. This definition is obvious but necessary to make selection rules comparable (Westphal & Brannath, forthcoming 2019).

In the more interesting case $\delta > 0$, the global null is no longer true, i.e. there are models with sufficiently high performance $\vartheta_m > \vartheta_0$. The rejection rate $\pi(\delta)$ then almost coincides with statistical power, i.e. the probability to correctly identify such a model. To be precise, $\pi(\delta)$ may also contain false positive test decisions for f_{m^*}, the empirically best model on the test data, when $\delta > 0$. We found in separate analyses that this difference is minor and only noteworthy at all for small positive δ. The curves showing the raw statistical power thus look almost identical to those shown in figure 3 for $\delta > 0$ and are provided in the supplementary report. Either way, evaluating multiple promising models increases statistical power uniformly (for all $\delta > 0$). For most of the investigated range of δ, the gain in power is substantial as it lies between 10 and 20 percent. This gain is decreasing in the number of validation samples $n_{\mathcal{V}} = n_{\mathcal{L}}/4$.

3.4.5. Sensitivity Analyses

We conducted several sensitivity analyses to investigate if the results change under modified conditions. They are described very briefly below. Details can be found in the supplementary report.

Most importantly, we stratified all analyses further by learning task scenario, i.e. label and feature distribution. In all cases the results were qualitatively similar to the those in the main (global) analysis. In particular we observed an increased final model performance and improved statistical power when multiple models are evaluated simultaneously.

The case of fewer than 200 initial candidate models was also emulated. For this purpose, we repeated the complete analysis after randomly sampling 40 or 100 from the 200 available models for each simulation instance. Hereby, we enforced that each of the four learning algorithms is represented equally in all sets of candidate models $\mathcal{M}$, i.e. each $\mathcal{M}$ (still) contains $M/4$ models of each type. In summary, none of the previously reported results is particularly sensitive to this change, neither qualitatively nor quantitatively.

We also investigated how our multiple testing approach compares against selection based on model complexity, at least for the subset of elastic net models ($M = 50$). A traditional selection rule in this case is to minimize the number of non-zero model coefficients under the side condition that the validation performance is within one standard error of the best validation model (Friedman et al., 2009, p. 61, p. 244). Somewhat surprisingly, this selection rule didn't even outperform the *default* approach regarding model performance and quality of test decisions. Moreover, it may be very difficult in general to find a criterion which measures model complexity for all candidate models when those are the result of a diverse set of learning algorithms.

Finally, we examined the effect of using the well-known Bonferroni correction instead of the advertised maxT-approach in conjuction with the *within 1 SE* selection rule.

This multiple test is based on dividing the global significance level α by the number of selected models S to obtain a local significance level $\alpha_{loc} = \alpha/S$. As expected, the Bonferroni correction results in a less powerful test procedure as it ignores the dependency structure of test statistics. The rejection rate was decreased by up to 10%, depending on the benchmark ϑ_0. The employed statistical test has no influence on model selection, as we would still pick the model with the highest empirical performance on the test data as the final model.

4. Conclusion

4.1. Summary and Interpretation

'The test data should not guide model selection and instead only be used for the assessment of a single (independently selected) model.' The main contribution of the present work is the demonstration that this omnipresent recommendation in machine learning should not be considered irrefutable. A simultaneous evaluation of multiple models is possible when suitable statistical methods are employed to counter the induced overoptimism. Moreover, our proposed evaluation strategy has two major advantages.

Firstly, in all scenarios investigated in our simulation, the expected model performance was increased compared to the *default* approach. This can be explained by the fact that effectively more data is used to inform the model selection, namely the test data which is otherwise only used for an assessment of a single prespecified prediction model. In other words, our approach enables the researcher to correct mistakes made due to the imperfect model ranking in the validation stage. The expected net benefit of this approach in our simulation study, just under one percent classification accuracy when $n_{\mathcal{V}} = n_{\mathcal{E}}$, may seem small. It may however also be seen as a relative decrease in classification error of (roughly) $10\% = \frac{0.01}{1-0.9}$.

The second benefit is the increased statistical power. In practice, the researcher has in effect a largely increased chance to show that at least one of the candidate models performs sufficiently well. The increase in power was substantial, between 10% and 20% in most relevant situations. In other words, less observational units are needed to achieve the same power. This is an important factor as conducting evaluation studies in regulated environments such as medical diagnosis can be very costly.

The main drawback of our approach is the introduced upward bias of the final performance estimate in the evaluation study. However, we showed that it is possible to adjust for this bias with a corrected, median-conservative estimator which is arguably preferable in the context of model evaluation.

4.2. Limitations and Outlook

Our conclusions are mainly based on an extensive simulation study. As a consequence, the results cannot be extrapolated naively to all possible machine learning scenarios. However, our conclusions are not limited to the employed learning algorithms and data distributions. We rather see those as reasonable to generate the truly relevant characteristics, namely the distribution of (intermediate) performances ϑ^-, ϑ and their dependence structures. These characteristics are partially analysed in table 1 and in the supplementary report. The range of the best learned model accuracy (85%-90%) in relation to the theoretically optimal accuracy (86%-97%) seems very realistic to us.

The biggest restriction of our simulation study may be the investigated learning sample sizes $n_{\mathcal{L}}$. We feel that 400 to 800 learning samples are realistic for applications in assisted disease diagnosis - our main target application. In 'big data' scenarios however, the performance gain will eventually vanish as all standard errors in the validation phase tend to zero asymptotically. In this regard, our *within 1 SE* rule 'converges' to the *default* selection rule as $n_{\mathcal{V}} \to \infty$.

The extension to other predictions tasks other than binary classification or other performance measures is easily possible. All that is required is the asymptotic multivariate normality of the performance estimate $\hat{\boldsymbol{\vartheta}}$. This is the case when performance (or error) estimates are computed as a sample average over (dis)similarities between predictions and labels and in many other cases.

A natural question directly motivated by this work concerns the optimal number of models to include in the evaluation study. In a post-hoc analysis of our simulation data, we found that after a rapid gain in expected model performance when increasing the number of models from a single one to a few, this gain vanishes again when including even further models. This is intuitive, as evaluating too many models will lead to a shift of the selection bias issue from the validation to the test data. Based on another recent work, we plan to model the expected utility, e.g. the final model performance, more explicitly in a Bayesian framework at the time point of model selection for evaluation (Westphal, 2019).

An important extension of the present work will be the simultaneous assessment of more than one performance measure for multiple models. At least for medical diagnosis applications, the most important example in this regard is the assessment of sensitivity and specificity of a binary classifier as co-primary endpoints. That is to say, the goal of the evaluation study is to show superiority in both these endpoints for at least one model under control of family-wise error rate. We are looking forward to adapt our recent multiple testing approach to this more complex setting in the future (Westphal et al., 2019).

Acknowledgements

The authors would like to thank the reviewers for their helpful comments. This project was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project number 281474342/GRK2224/1.

References

Boulesteix, A.-L. Over-optimism in bioinformatics research. *Bioinformatics*, 26(3):437–439, 2009.

Boulesteix, A.-L. and Strobl, C. Optimal classifier selection and negative bias in error rate estimation: an empirical study on high-dimensional prediction. *BMC medical research methodology*, 9(1):85, 2009.

Boulesteix, A.-L., Lauer, S., and Eugster, M. J. A plea for neutral comparison studies in computational sciences. *PloS one*, 8(4):e61562, 2013.

Breiman, L. Bagging predictors. *Machine learning*, 24(2): 123–140, 1996.

Cath, C., Wachter, S., Mittelstadt, B., Taddeo, M., and Floridi, L. Artificial intelligence and the 'good society': the us, eu, and uk approach. *Science and engineering ethics*, 24(2):505–528, 2018.

Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., Kalinin, A. A., Do, B. T., Way, G. P., Ferrero, E., Agapow, P.-M., Zietz, M., Hoffman, M. M., et al. Opportunities and obstacles for deep learning in biology and medicine. *Journal of The Royal Society Interface*, 15 (141):20170387, 2018.

Dickhaus, T. *Simultaneous statistical inference*. Springer, 2014.

Dietterich, T. G. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, 10(7):1895–1923, 1998.

Eugster, M. J., Hothorn, T., and Leisch, F. Exploratory and inferential analysis of benchmark experiments. 2008.

Friedman, J., Hastie, T., and Tibshirani, R. *The elements of statistical learning*, volume 2. Springer series in statistics New York, 2009.

Friedman, J. H. Multivariate adaptive regression splines. *The annals of statistics*, pp. 1–67, 1991.

García, S., Fernández, A., Luengo, J., and Herrera, F. Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power. *Information Sciences*, 180(10):2044–2064, 2010.

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., and Hothorn, T. *mvtnorm: Multivariate Normal and t Distributions*, 2018. URL https://CRAN.R-project.org/package=mvtnorm. R package version 1.0-8.

Géron, A. *Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems*. " O'Reilly Media, Inc.", 2017.

Gómez, E., Castillo, C., Charisi, V., Dahl, V., Deco, G., Delipetrev, B., Dewandre, N., González-Ballester, M. Á., Gouyon, F., Hernández-Orallo, J., et al. Assessing the impact of machine intelligence on human behaviour: an interdisciplinary endeavour. *arXiv preprint arXiv:1806.03192*, 2018.

Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y. *Deep learning*, volume 1. MIT press Cambridge, 2016.

Hothorn, T., Leisch, F., Zeileis, A., and Hornik, K. The design and analysis of benchmark experiments. *Journal of Computational and Graphical Statistics*, 14(3):675–699, 2005.

Hothorn, T., Bretz, F., and Westfall, P. Simultaneous inference in general parametric models. *Biometrical journal*, 50(3):346–363, 2008.

Japkowicz, N. and Shah, M. *Evaluating learning algorithms: a classification perspective*. Cambridge University Press, 2011.

Jelizarow, M., Guillemot, V., Tenenhaus, A., Strimmer, K., and Boulesteix, A.-L. Over-optimism in bioinformatics: an illustration. *Bioinformatics*, 26(16):1990–1998, 2010.

Jensen, D. D. and Cohen, P. R. Multiple comparisons in induction algorithms. *Machine Learning*, 38(3):309–338, 2000.

Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., Wang, Y., Dong, Q., Shen, H., and Wang, Y. Artificial intelligence in healthcare: past, present and future. *Stroke and vascular neurology*, 2(4):230–243, 2017.

Knottnerus, J. A. and Buntinx, F. *The evidence base of clinical diagnosis: theory and methods of diagnostic research*. BMJ Books, 2009.

Kuhn, M. and Johnson, K. *Applied predictive modeling*, volume 26. Springer, 2013.

Lang, M., Bischl, B., and Surmann, D. batchtools: Tools for r to work on batch systems. *The Journal of Open Source Software*, 2(10), feb 2017. doi: 10.21105/joss.00135. URL https://doi.org/10.21105/joss.00135.

LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *nature*, 521(7553):436, 2015.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A., Van Ginneken, B., and Sánchez, C. I. A survey on deep learning in medical image analysis. *Medical image analysis*, 42: 60–88, 2017.

Miotto, R., Wang, F., Wang, S., Jiang, X., and Dudley, J. T. Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*, 2017.

Nadeau, C. and Bengio, Y. Inference for the generalization error. In *Advances in neural information processing systems*, pp. 307–313, 2000.

Olhede, S. and Wolfe, P. The growing ubiquity of algorithms in society: implications, impacts and innovations. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2128): 20170364, 2018.

Pepe, M. S. *The statistical evaluation of medical tests for classification and prediction*. Medicine, 2003.

Pesapane, F., Volonté, C., Codari, M., and Sardanelli, F. Artificial intelligence as a medical device in radiology: ethical and regulatory issues in europe and the united states. *Insights into imaging*, pp. 1–9, 2018.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. URL http://www.R-project.org/.

Raschka, S. Model evaluation, model selection, and algorithm selection in machine learning. *arXiv preprint arXiv:1811.12808*, 2018.

Reed, C. How should we regulate artificial intelligence? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2128): 20170360, 2018.

Shalev-Shwartz, S. and Ben-David, S. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Westphal, M. Simultaneous Inference for Multiple Proportions: A Multivariate Beta-Binomial Model. *Manuscript submitted for publication*, 2019.

Westphal, M. and Brannath, W. Evaluation of multiple prediction models: a novel view on model selection and performance assessment. *Statistical Methods in Medical Research*, forthcoming 2019.

Westphal, M., Zapf, A., and Brannath, W. A multiple testing framework for diagnostic accuracy studies with coprimary endpoints. *Manuscript in preparation*, 2019.

Zheng, A. *Evaluating Machine Learning Models—A Beginner's Guide to Key Concepts and Pitfalls*. O'Reilly Media, 2015.