technische universität
dortmund

Modelling with Feature Costs under a Total Cost Budget Constraint

Dissertation

presented to the Faculty of Statistics, TU Dortmund

in fulfillment of the requirements for the degree

"Doktor der Naturwissenschaften"

by

Rudolf Jagdhuber

First Referee:	**Prof. Dr. Jörg Rahnenführer**
Second Referee:	**Dr. Uwe Ligges**
Day of Submission:	**2020-08-17**
Day of Disputation:	**2020-10-28**

ACKNOWLEDGEMENTS

This thesis was started as a cooperation of the Department of Statistics at the TU Dortmund and the numares AG Regensburg. Multiple research projects at numares yielded interesting statistical problems that could be targeted by a doctoral theses. This led me to a discussion with my superior Dr. Sindy Neumann and later with the CEO of numares Dr. Volker Pfahlert, who I would both like to thank for their open mind about this concept and for laying the foundation, that allowed me to work on this thesis. I would furthermore like to thank Dr. Jana Fruth for highly contributing to the next steps by recommending Prof. Dr. Jörg Rahnenführer as my supervisor and initiating the first contact.

On the university side, my greatest thanks of course go to Prof. Dr. Jörg Rahnenführer for providing me the opportunity to pursue this work and for his high engagement in supervising this dissertation, which I highly appreciated. As my main reference person for statistical issues, he always provided technical guidance when necessary and assisted me in every step along the way. I would also like to thank Michel Lang. Whenever I visited TU Dortmund I could be sure that he would find time for me and provide great input on any given problem.

My final thanks go to my wife Kristin Jagdhuber, who put up with being repeatedly parted for longer time-spans during the course of this thesis. With her personal guidance especially in difficult times, she played the most essential non-technical role for the successful outcome of this thesis.

Abstract

In modern high-dimensional data sets, feature selection is an essential pre-processing step for many statistical modelling tasks. The field of *cost-sensitive feature selection* extends the concepts of feature selection by introducing so-called *feature costs*. These do not necessarily relate to financial costs, but can be seen as a general construct to numerically valuate any disfavored aspect of a feature, like for example the run-time of a measurement procedure, or the patient harm of a biomarker test. There are multiple ideas to define a cost-sensitive feature selection setup. The strategy applied in this thesis is to introduce an additive *cost-budget* as an upper bound of the total costs. This extends the standard feature selection problem by an additional constraint on the sum of costs for included features. Main areas of research in this field include adaptations of standard feature selection algorithms to account for this additional constraint. However, cost-aware selection criteria also play an important role for the overall performance of these methods and need to be discussed in detail as well.

This cumulative dissertation summarizes the work of three papers in this field. Two of these introduce new methods for cost-sensitive feature selection with a fixed budget constraint. The other discusses a common trade-off criterion of performance and cost. For this criterion, an analysis of the selection outcome in different setups revealed a reduction of the ability to distinguish between information and noise. This can for example be counteracted by introducing a hyperparameter in the criterion. The presented research on new cost-sensitive methods comprises adaptations of Greedy Forward Selection, Genetic Algorithms, filter approaches and a novel Random Forest based algorithm, which selects individual trees from a low-cost tree ensemble. Central concepts of each method are discussed and thorough simulation studies to evaluate individual strengths and weaknesses are provided. Every simulation study includes artificial, as well as real-world data examples to validate results in a broad context. Finally, all chapters present discussions with practical recommendations on the application of the proposed methods and conclude with an outlook on possible further research for the respective topics.

CONTENTS

CHAPTER

1

INTRODUCTION

"Nichts auf dieser Welt ist umsonst. Selbst der Tod kostet das Leben."
(Nothing in this world is for free. Even death costs your life.)

German Proverb

In times of digital data acquisition and the trend towards "Big Data", statistical modeling tasks are more than ever faced with problems arising from high dimensional feature spaces. These can reach from an increased prediction uncertainty created by uninformative features ("noise"), over multicollinearity problems from redundancies in the data, that result in convergence issues, to general model fitting problems for example in situations, where the number of features extends the number of observations. In these scenarios, selecting a suitable subset of all available features describes an essential pre-processing step. The corresponding field of *feature selection* (Guyon and Elisseeff, 2003) is widely researched and provides a large set of methods and tools designed specifically for this purpose.

The work presented in this thesis extends the idea of feature selection by introducing so-called *feature costs*. These costs can be seen as a general construct to numerically valuate a disfavored aspect of a feature. In financial scenarios, feature costs can refer

to actual costs and describe the price of obtaining a feature. This can be a relevant aspect, if for example the goal of the analysis is to build a diagnostic test in a competitive field, where one of the defining criteria is the final market price. However, costs can also be seen in a more abstract sense, like for example as success or failure rates in obtaining a feature. This has special relevance in modelling applications, where a missing value in a model component directly results in a missing value in the final prediction. An example for this is again the diagnostic test setup. As official regulations typically require a general prove of efficacy, adaptive ideas or imputation approaches are still very uncommon in this field. Therefore, avoiding missing values is a highly relevant secondary objective here as well. The medical setting also includes a third possible use-case for costs, which is patient harm. Standard feature selection algorithms do not distinguish between a feature requiring a painful biopsy and a simple non-invasive urine test, if both show similar predictive performance. In practice, however, this difference may oftentimes be even more important than slight advantages in accuracy. One last example of a feature cost definition is prediction time. In certain online applications, relevant features need to be computed in real-time along an input of the user. Modern search engines for instance provide intelligent guesses of the intended search term while the user is typing. The quality of these guesses of course is one important factor. Yet, the time to obtain these guesses is also essential. A model that requires more than a second for its suggestions would be worthless in practice, irregardless of its performance.

All of these applications illustrate the need for feature selection algorithms, which do not only focus on optimizing predictive performance, but also account for an aspect of feature costs. The corresponding field of research is commonly referred to as "cost-sensitive learning" (Tan, 1993). There are three main strategies to integrate feature costs into statistical model selection problems, which are described shortly in the following.

The first of these is to harmonize costs of misclassifications and feature costs of a final model, while not defining any hard boundaries for either objective. Hence, any model that is not dominated in both aspects simultaneously by another model is a valid final candidate. The main motivation for this strategy is that hard cost limits are unusual in practical applications and a certain budget flexibility is often possible if the corresponding benefit is worthwhile. Research in this adaptive field can for example be found in Q. Zhou, H. Zhou, and Li (2016) or Bolón-Canedo et al. (2014). While the idea of flexible budgets is intuitive, the downside, however, is that the output of a feature selection algorithm generally still includes a large number of non-dominated models, and ultimately, a manual

decision from a Pareto-efficient frontier is necessary. In practice, this final choice may be to some extent arbitrary, as decision-makers often have no exact valuation of the trade-off of a performance measure and costs.

The second popular strategy for cost-sensitive learning focuses on average prediction costs of new observations. The main difference for this approach is that models do not use a fixed set of features, but decide which features to use for prediction for each observation individually. This means that for some easy-to-predict observations, a cheap set of a few features may be sufficient, while for other more difficult observations, a larger set of features can be used. Altogether, the goal is to minimize the average feature costs of all expected predictions. M. J. Kusner (2016) provides a good overview of this field, which is referred to as "resource-efficient learning". Further work can also be found in Z. Xu, M. Kusner, et al. (2013), Z. Xu, M. J. Kusner, et al. (2014), and M. Kusner et al. (2014). Of course this strategy is only possible, if features are obtained online at the time of prediction and costs arise from this data collection process. An example for a corresponding setup could be search engine suggestions, where feature costs refer to the time required to obtain a certain relevant metric. Complex user inputs could be predicted more accurately by obtaining additional metrics, while simple requests could be dealt with quickly. Nevertheless, for most of the practical problems and cost setups described earlier, this specific strategy is not applicable.

The third and final strategy is similar to the first and again aims to identify a fixed model, which harmonizes a trade-off between costs and predictive performance. However, to avoid the manual selection from a large set of models, this strategy introduces a fixed feature cost budget. With this limitation, a single optimal solution is always defined. The overall cost-sensitive approach extends the standard feature selection problem only by introducing an additional constraint on the total sum of costs. A formal definition of this problem for features X_j with individual costs c_j is given in the following. The feature subset $\hat{s}$ that is optimal with respect to a performance measure Q and holds a cost budget $c_{\max}$ is selected by

$$\hat{s} = \arg\max_{s} \{Q(s)\} \qquad \text{subject to} \qquad \sum_{j:X_j \in s} c_j \leq c_{\max}. \tag{1.1}$$

Because of the increasing complexity in higher-dimensional feature spaces, approximate heuristics are required to solve this problem in practice. Min, Hu, and Zhu (2014) present examples for such feature selection heuristics and also provides an alternative problem

definition in the context of rough sets. Other research on this topic can for example be found in Leskovec et al. (2007), or Min and J. Xu (2016). The main downside of this specific strategy is the missing flexibility that results from a fixed cost budget. If a model with only slightly higher costs would be able to dramatically boost performance, one would still not be able to identify it here. In practice, however, this problem can for instance be attenuated by performing multiple feature selection runs with different values of $c_{\max}$. The main advantages on the contrary are a clean definition of the feature selection problem and avoiding arbitrary manual decisions after the selection. Therefore, this strategy provides a general and easy to apply solution for cost-sensitive feature selection problems.

This thesis focuses on this third approach and uses a fixed feature cost budget limit. As most existing feature selection methods commonly do not naturally include options for a secondary constraint, introducing adequate cost-adaptations of these algorithms is one aspect of research in this field. Ensuring that methods hold a given feature cost budget is, however, only a first technical step. The (heuristic) search strategy of each algorithm essentially guides the individual selection steps. Cost-sensitive adaptations of this search strategy are therefore a promising approach to further improve the final result. A typical idea here is to modify the main selection criterion, which is often purely performance based, and introduce an alternative custom trade-off measure including feature costs. Discussing the consequences and implications of these criteria on the selection result is another aspect of research. Besides adapting existing methods and discussing selection criteria, a third option is to introduce completely new approaches. These can be specifically tailored to a cost-constrained setup and provide solutions for situations, where simple adaptations are not feasible.

The following chapters include examples for all of these mentioned ideas. Chapter 2 presents the work of Jagdhuber, Lang, Stenzl, et al. (2020), who adapt common feature selection algorithms like Greedy Forward Selection, Genetic Algorithms and filter methods to handle a fixed feature cost budget. Furthermore, a custom cost-sensitive selection measure is introduced and compared with a standard performance-based approach. A large scale simulation study on artificial and real-world data is conducted to evaluate all proposed methods. Finally, practical recommendations for the application of feature cost methods are provided.

Chapter 3 presents results of Jagdhuber and Rahnenführer (2020), who discuss a cost-sensitive trade-off measure that is commonly used in feature cost scenarios. Negative

implications of the uncontrolled version of this measure are illustrated on a practical example for multiple parameter setups. These problems can for example be avoided by introducing a hyperparameter.

The final Chapter 4 summarizes the work of Jagdhuber, Lang, and Rahnenführer (2020), who introduce a novel feature selection method tailored to cost-sensitive Random Forest problems. Additionally, adaptations of common alternative approaches for this setup are proposed. An artificial simulation study, as well as a thorough analysis on six real-world data sets from different fields of application are used to compare these methods and identify strengths and weaknesses in different setups.

Each chapter ends with an outlook for possible further research in its topic.

CHAPTER

2

COST-CONSTRAINED FEATURE SELECTION IN BINARY CLASSIFICATION: ADAPTATIONS FOR GREEDY FORWARD SELECTION AND GENETIC ALGORITHMS

2.1 Contributed Material

Rudolf Jagdhuber, Michel Lang, Arnulf Stenzl, Jochen Neuhaus, and Jörg Rahnenführer (2020). "Cost-Constrained feature selection in binary classification: adaptations for greedy forward selection and genetic algorithms". In: *BMC bioinformatics* 21.1, pp. 1–21. DOI: 10.1186/s12859-020-3361-9. URL: https://doi.org/10.1186/s12859-020-3361-9

Authors' contribution
Rudolf Jagdhuber developed and implemented the proposed methods, designed and executed the simulation studies, interpreted the results and wrote the manuscript. Michel Lang contributed to the design of the simulation settings, to the interpretation of the results, and corrected and approved the manuscript. Arnulf Stenzl and Jochen Neuhaus contributed to the acquisition of the plasmode data set samples and approved the manuscript. Jörg Rahnenführer supervised the project, initiated the feature-cost topic, contributed to the design of the simulation settings and to the interpretation of the results, and corrected and approved the manuscript.

2.2 Overview of Methods and Simulations

Jagdhuber, Lang, Stenzl, et al. (2020) consider the cost-constrained feature selection problem in binary classification and propose multiple extensions of common heuristic feature selection methods in this context. The first of these algorithms is Greedy Forward Selection. A cost-constrained version of this method can be obtained by subsetting the feature candidates in every iteration to only include those that do not exceed the budget if added. While this adaptation (named `FS`) does produce admissible results, the decision on which feature to include at each iteration is purely based on performance. A more sophisticated idea is to also include the relative costs of a feature for this decision. Therefore, a hyperparameter-controlled trade-off criterion – the benefit-cost ratio (BCR) – is introduced. Using Akaike's Information Criterion (AIC) (Akaike, 1974) as measure for the performance of a model $M(\cdot)$, the BCR of adding a feature X_j with cost c_j to a candidate set s is given by

$$\text{BCR}_\xi = \frac{\text{AIC}(M(s)) - \text{AIC}(M(s \cup X_j))}{c_j + \xi}, \tag{2.1}$$

where ξ is a hyperparameter to guide the trade-off. The corresponding adaptation of the Greedy Forward Selection algorithm, which chooses features according to this criterion, is labeled `cFS`.

The second class of methods adapted in this paper are Genetic Algorithms (GA) (Holland, 1973). GAs propose candidate feature combinations by applying a set of so-called "genetic operators", which translate the evolutionary ideas of survival of the fittest, genetic crossover and random mutation. The proposed candidate population is then evaluated by a fitness function, which assigns a real number assessing the suitability of a candidate set. There are two general ideas to adapt this method for a feature cost setup. The first idea (`fGA`) alters the fitness function of the GA. For a feature set, whose total costs are within the given budget, the fitness value corresponds to the predictive performance. However, for feature sets violating the cost constraint, a negative fitness value specifying the extent of the cost violation is used. This way, the GA is able to evolve from higher constraint violations to lower ones, eventually finding valid candidate sets. The second idea (`cGA`) alters the genetic operators to only propose feature combinations within the budget in the first place. For this, Jagdhuber, Lang, Stenzl, et al. (2020) propose a cost-constrained population initialization algorithm, a cost-constrained crossover operator and

a cost-constrained mutation operator. In combination with the existing *lrSelection* operator, this set of genetic operators defines an adapted setup, that can be used with any fitness function.

To assess the quality of the proposed methods `FS`, `cFS`, `fGA` and `cGA`, four filter methods are furthermore implemented as baseline approaches. Filters compute a measure of importance for every feature that in a second step can be used to select a suitable feature subset. The selected methods were chosen according to the recommendations of a benchmark study by Bommert et al. (2020) and included the methods `Filter.tTest`, `Filter.Symuncert`, `Filter.PraznikJMIM` and `Filter.RangerImpurity`. To select a final set from the feature rankings of the filter methods, a top-down approach is used. Features are added to the model in order of their rank according to the filter, but only, if the cost of the resulting model does not exceed the budget. The process is stopped, if the additional cost of any remaining feature would exceed the budget.

All proposed methods are compared in eleven artificial simulation settings and two simulation studies based on real-world data sets. The aim is to analyze a wide spectrum of data situations for a thorough performance overview. The artificial simulations include six main settings with different combinations of the number of relevant and noise features, effect sizes and feature cost budgets. The remaining five artificial settings focus on special design traits, like for example a correlation of feature costs with effect sizes, or features originating from mixture distributions. The real-world simulations comprise two biological data sets. For the first of these, a so-called plasmode approach (Vaughan et al., 2009) is used. Plasmodes take a data set generated from natural processes and add a simulated aspect to the data (Franklin et al., 2014). To obtain a controlled scenario, this simulated aspect is the binary response variable, which is computed from features that are defined to be relevant. As this provides only a partial real-world application, a second simulation on unmodified real-world data is performed as well. Analyses focus on predictive performance, run-time, and detection of relevant features where applicable.

2.3 Main Results and Conclusions

Figure 2.1 illustrates performance and detection rate results for an exemplary setting of the artificial simulation study. This setting includes a total of $p = 300$ features, of which $p^{(\text{rel})} = 30$ are considered relevant with moderate effects. The budget limit is set to $\gamma = \frac{1}{3}$

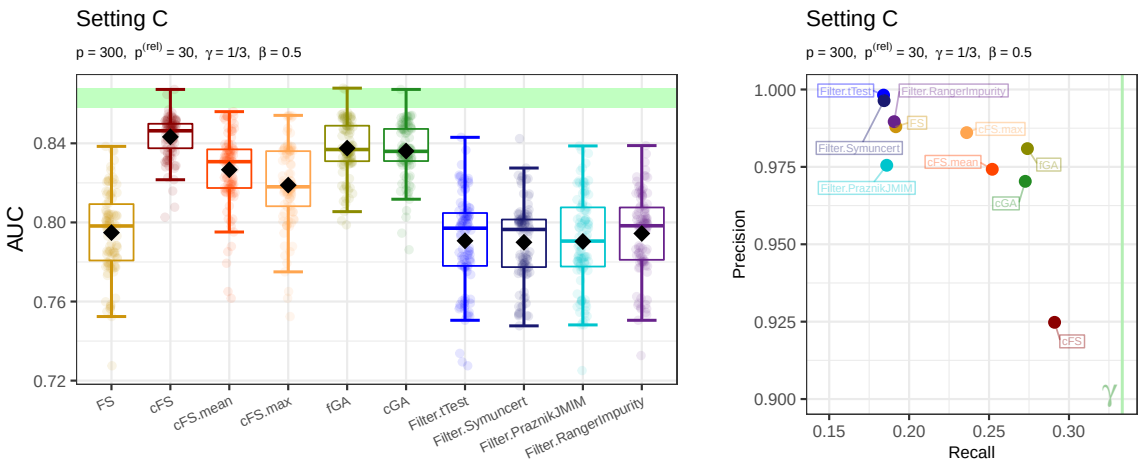

Figure 2.1: Left: Boxplots for every feature selection method illustrate the distribution of the values for the area under the receiver operating characteristic (AUC) obtained in the 100 simulation data sets (transparent dots). The black diamonds depict the mean AUC values. A horizontal green bar highlights the area between the 0.05 and 0.95 quantile of AUC values when always selecting the cheapest subset of relevant features that fit in the budget. Right: Precision-recall plot comparing the analyzed feature selection methods. Precision corresponds to the ratio of detected relevant features divided by the total number of features in the model. Recall shows the ratio of detected relevant features divided by the total number of relevant features. The cost budget defines an upper limit for the recall in the simulations. It is highlighted by a green line.

of the cost of all relevant features combined, and hence defines a notable constraint.

Compared to the baseline filter methods and FS, the proposed adaptations of Forward Selection and Genetic Algorithms typically result in better performing feature sets. The only exception to this occurs at a specific simulation setting, where the budget allows to include twice the total cost of all relevant features. In these situations without a *true* budget constraint, the BCR-based cFS method notably falls behind. However, a proper choice of the BCR hyperparameter can overcome this problem and achieves good results in all settings. The GA adaptations generally rank within the best methods and provide an overall versatile method, yet also require an at least five-fold higher run-time compared to all other methods. The analysis of the feature detection rate highlighted that in this case with respect to performance, it is more important to include relevant features than to reduce the number of noise features. The precision-recall analysis of Figure 2.1 is a good example for this. In this setting, cFS has the lowest precision, but highest recall, and turns out to be the best performing method, while the filter approaches have the highest precision but lowest recall and show the lowest performances. The results obtained on real-world data reinforced the presented findings of the simulations with artificial data.

Altogether, it is recommended to use cGA for its robustness in a wide variety of data settings and its generalized implementation allowing to define a completely unconstrained fitness function.

2.4 Outlook

Beyond the scope of this work, many extensions of the proposed methods are possible. In the presented analyses, the selection of a feature subset is performed according to the AIC. Other performance measures, which for instance evaluate a cross-validated setup, could further improve the general predictive performance. Moreover, this would also broaden the field of applicable modelling methods, which is currently only limited by the applicability of the AIC. While the presented paper specifically analyses a binary classification task, none of the proposed methods is technically limited to this setup. In principle, each approach is also valid in many further supervised learning tasks, and research on the effects for example in regression setups could be of interest as well. Finally, besides Forward Selection and Genetic Algorithms, adaptations of other feature selection approaches are a general option for further research to extend the spectrum of available methods in this field.

CHAPTER

3

IMPLICATIONS ON FEATURE DETECTION WHEN USING THE BENEFIT-COST RATIO

3.1 Contributed Material

Rudolf Jagdhuber and Jörg Rahnenführer (2020). *Implications on Feature Detection when using the Benefit-Cost Ratio*. arXiv: 2008.05163v2 [stat.ML]

Authors' contribution
Rudolf Jagdhuber initiated the topic, formulated and discussed the problem, designed and executed the simulation studies, interpreted the results, and wrote the manuscript. Jörg Rahnenführer supervised the project, contributed to the problem definition, the design of the simulation study and to the interpretation of the results, and corrected and approved the manuscript.

3.2 Problem Definition

A common cost-sensitive strategy to valuate the importance of a feature X_j is to use the ratio of a metric for performance gain ΔQ_j and the additional costs c_j that this gain requires.

$$\mathrm{BCR}(X_j) = \frac{\Delta Q_j}{c_j} \tag{3.1}$$

This popular statistic, which is referred to as the benefit-cost ratio (BCR), can for example be found in Min, He, et al. (2011), Min, Hu, and Zhu (2014), Min and J. Xu (2016), Leskovec et al. (2007), and Grubb and Bagnell (2012). It applies a scaling to the estimated gain in performance relative to the induced costs. For uninformative noise features, this gain is theoretically zero and the BCR should not be affected from high or low costs. With finite data, however, the estimation uncertainty can also randomly lead to positive values of ΔQ_j, which may then be amplified in specific cost settings. Jagdhuber and Rahnenführer (2020) discuss this problem and analyze the practical consequences of an individual cost-scaling with respect to distinguishing relevant information from noise.

A simulation study with a single feature selection step for a linear regression setup is conducted in multiple parameter constellations to analyze effects on the detection rate of relevant features. By defining equal costs for relevant and noise features, respectively, the BCR can be formulated as a scaling of the performance gain, which differs between both classes only by a single parameter θ. The cost-sensitive feature selection step thus identifies the maximal value from $\left(\frac{\Delta Q_{\text{relevant}}}{\theta}, \Delta Q_{\text{noise}}, 0\right)$ and chooses either a relevant feature, a noise feature, or no feature at all. Influences of the scaling parameter θ, the number of relevant and noise features, and the effect size of relevant parameters on this selection result are analyzed.

3.3 Key Results and Conclusions

The first observation of the obtained simulation results is that cost-scaling does not influence the case of deciding for no feature. As features are only selected if their performance measure is greater than zero and θ does not change the sign of the measure, the absolute size of the BCR is irrelevant for this aspect. Thus, the only critical scenario occurs, if both relevant and noise features produce positive values of ΔQ. In these situations, it can be observed that increasing values of θ result in a decreasing detection rate. This

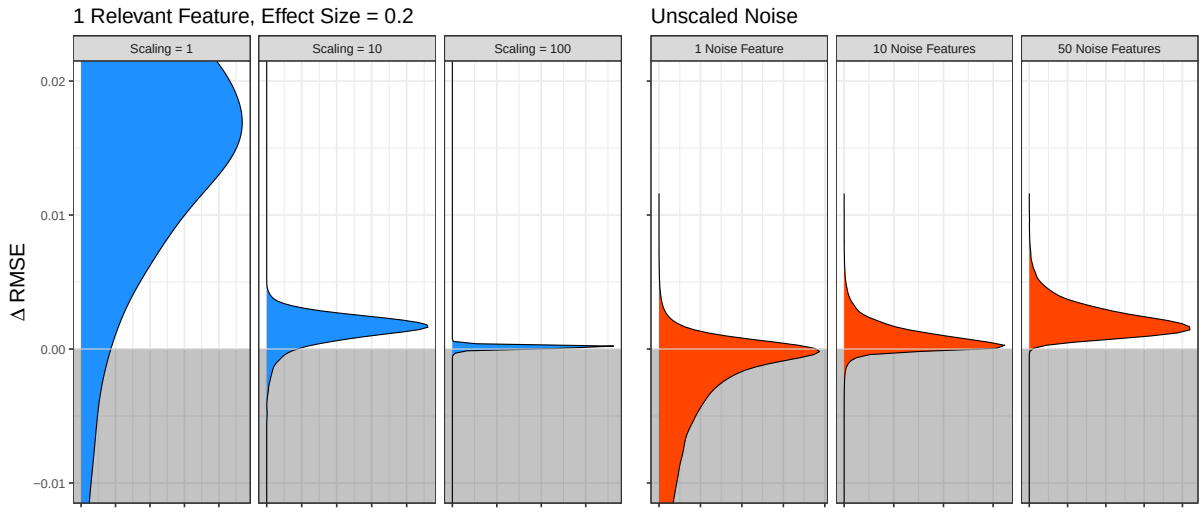

Figure 3.1: Empirical distributions of the gain in root mean squared error ΔRMSE for the simulation study in Jagdhuber and Rahnenführer (2020). Three different levels of cost-scaling for relevant features with effect size 0.2 are illustrated in blue. Three density plots corresponding to different numbers of noise features are shown in red. All plots share a common y-axis.

effect is only counteracted by the effect size of the relevant features. For very large θ, the detection rate is reduced to the probability of observing $\Delta Q \leq 0$ for all noise features and $\Delta Q > 0$ for at least one relevant feature. The effects of θ are therefore also associated with the number of noise features, which defines the probability for a positive ΔQ_{noise}. The described results are illustrated by the (scaled) densities of relevant and noise features shown in Figure 3.1. It can be seen that the cost-scaling shrinks the density of the performance gain for relevant features, while keeping the total positive and negative probability masses constant. A higher number of noise features increases the respective positive probability mass and hence the overall likelihood of selecting noise.

In conclusion, the simulation study showed a notable influence on the detection rate when using the BCR in settings with high relative cost differences. Such scenarios may not be unrealistic and can unintentionally occur in practice. Min, He, et al. (2011) for example suggested to solve mathematical problems arising from cost-free features by assigning small pseudo-costs to them. However, in an uncontrolled cost setup, this approach could easily create immense scalings. To address the adverse effects highlighted in this paper, one option is to manually avoid extreme cost ratios in the data. This can for instance be realized with an adequate transformation of the original costs. However, as a more flexible and overall superior option, it is recommendable to avoid the plain BCR criterion in general, and instead use a hyperparameterized alternative like for example $\frac{\Delta Q_j}{(c_j)^\xi}$ (Min, Hu, and Zhu, 2014), or $\frac{\Delta Q_j}{c_j + \xi}$ (Jagdhuber, Lang, Stenzl, et al., 2020). While the optimization of

a hyperparameter increases the computational complexity, the additional run-time is well spent. A tuned BCR criterion can notably improve the overall selection result (Min, Hu, and Zhu, 2014), and also prevent uncontrolled cost-scalings that impair detection rates by concealing relevant features.

3.4 Outlook

The benefit-cost ratio can be considered a suitable method to trade off costs and performance, if it is applied in a (hyperparameter) controlled setup. In situations, however, where the additional computational effort of tuning a hyperparameter is not feasible, research on a valid and comprehensible way to valuate both feature selection goals (for instance using expert knowledge) would be of high interest. This consequently also requires a suitable strategy for handling cost-free features. Even after deciding to use a hyperparameterized BCR criterion, researchers still are faced with multiple alternatives, like the criteria mentioned in the previous section, or also non-ratio based ideas such as $\Delta Q_j + \xi c_j$. A thorough comparison of the strengths and weaknesses of these methods could be a relevant guide for cost-sensitive analyses in practice. Finally, all implications described in this paper are demonstrated on a specific linear model setup only. Further research may also consider different model types, performance measures, feature distributions, and additional aspects.

CHAPTER

4

FEATURE SELECTION METHODS FOR COST-CONSTRAINED CLASSIFICATION IN RANDOM FORESTS

4.1 Contributed Material

Rudolf Jagdhuber, Michel Lang, and Jörg Rahnenführer (2020). *Feature Selection Methods for Cost-Constrained Classification in Random Forests*. arXiv: 2008.06298v2 [stat.ML]

Authors' contribution
Rudolf Jagdhuber developed all methods, designed and executed the simulation studies, interpreted the results, and wrote the manuscript. Michel Lang proposed the first idea of Tree Selection, contributed ideas for the evaluation, and corrected and approved the manuscript. Jörg Rahnenführer initiated the topic, supervised the project, contributed to the problem definition, the design of the simulation study and to the interpretation of the results, and corrected and approved the manuscript.

4.2 Overview of Methods and Simulations

The contributed paper by Jagdhuber, Lang, and Rahnenführer (2020) analyzes cost-sensitive feature selection problems for Random Forest applications. Due to the complex feature structure and relatively high computational complexity of this ensemble method, most popular feature selection approaches utilize filter methods. To provide a multivariate alternative for situations with feature costs and a limited budget, Shallow Tree Selection (STS) is introduced. This novel algorithm selects individual trees from Random Forests with limited tree depth to create a new tree ensemble, which on the one hand controls costs and on the other hand optimizes predictive performance. A greedy forward selection approach is used to iteratively add the most suited candidate tree to the result ensemble. The decision of which tree is most suited is based on the hyperparameterized BCR criterion

$$\mathrm{BCR}_\xi = \frac{\Delta Q_j}{(c_j)^\xi}. \tag{4.1}$$

Here, ΔQ_j describes the reduction in out-of-bag (OOB) error and c_j refers to the additional costs that the j-th tree generates. A full schematic overview of the STS algorithm including all relevant sub-steps is given in Figure 4.1. Detailed discussions and motivations for each design element can be found in Jagdhuber, Lang, and Rahnenführer (2020).

In addition to STS, three further methods are proposed, which extend common feature selection strategies in Random Forest applications. The first of these is a univariate filter approach based on the area under the receiver operating characteristic (AUC) (Hanley and McNeil, 1982). This approach assigns an individual rating to each feature using the BCR of Definition 4.1, with ΔQ_j referring to the normalized AUC of each feature. With this measure, a top-down approach similar to the filter strategies of Section 2.2 is used for the selection. The second adapted method utilizes the Permutation Feature Importance (pFI) (Breiman, 2001), which is a common metric to assess feature relevance in a Random Forest. While this measure can on the one hand be considered multivariate from the evaluation viewpoint, on the other hand it assigns only a single value to each feature. For the cost-sensitive adaptation, this value is computed using Definition 4.1, with ΔQ_j referring to the standard pFI. The result ensemble is then created similar to the AUC approach by a top-down strategy. The final proposed method is a Random Forest Forward Selection (FS). It starts with an empty feature set and iteratively evaluates all one-feature extensions of the current set by computing full Random Forests each. The

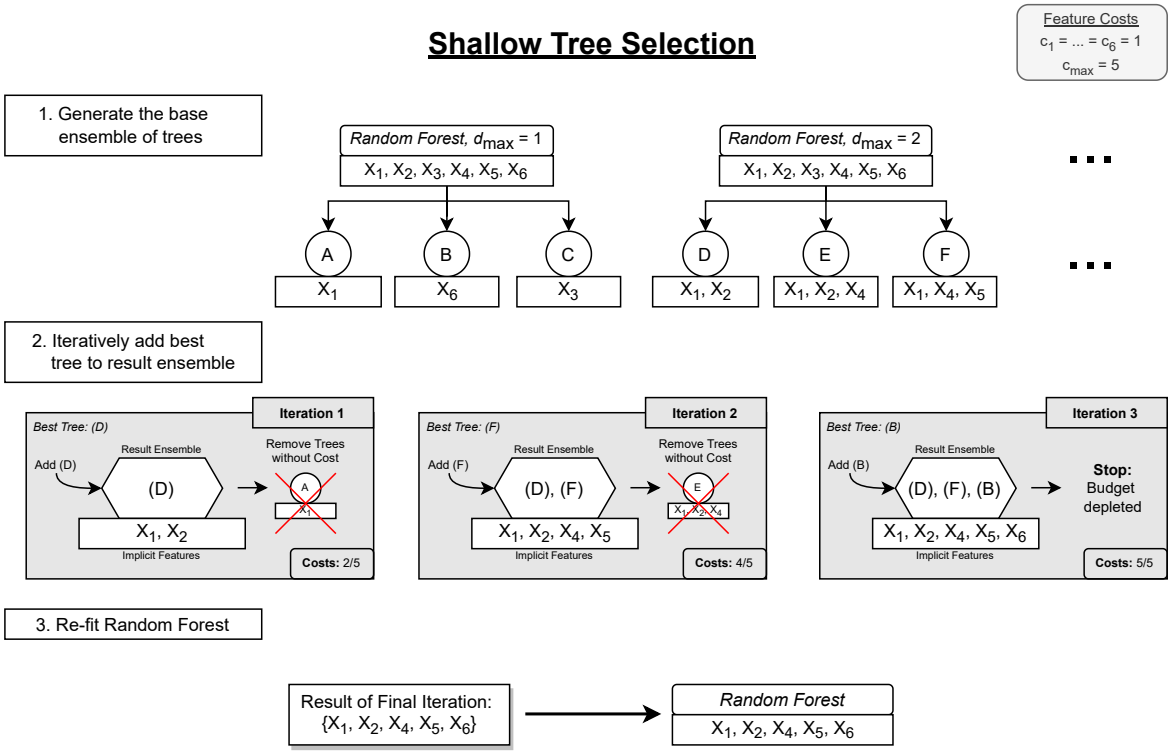

Figure 4.1: Schematic example of the Shallow Tree Selection method with six candidate features. In step 1, Random Forests with a maximum tree depth of one and two are fitted. Each forest generates three trees for a total of six candidate trees labeled (A) to (F). Step 2 describes the greedy forward selection, which starts with an empty result ensemble. In the first iteration, the current best candidate tree (D) is added to this result ensemble, which therefore now implicitly holds features X_1 and X_2. After that, candidate tree (A) only contains features that are already present in the result ensemble. It is thus removed from the list of candidate trees. In the second iteration, the most suited tree is (F). (F) is added to the result ensemble, which analogously to the first iteration leads to the removal of the now cost-free tree (E). The final iteration adds tree (B) and fills up the budget. This concludes the greedy forward selection. Step 3 uses the implicit feature set of the last iteration of the greedy forward selection to fit a new Random Forest with it. This step is important to overcome weaknesses, which result for instance from the limited tree depths.

BCR criterion of Definition 4.1, with ΔQ_j referring to the reduction of OOB error, is again used to choose the best candidate. Because of the immense computational complexity of the FS method, no tuning of ξ is considered here, and only the edge cases "cost-agnostic" ($\xi = 0$) and "simple BCR" ($\xi = 1$) are analyzed.

In an extensive simulation study, these four proposed methods are compared in multiple artificial data settings with varying effect sizes, feature costs, budget limits, univariate and multivariate effects on the response variable, and possible dependencies of costs and effect sizes. Additionally to the artificial settings, a comparative analysis on six real-world data sets from the OpenML repository (Vanschoren et al., 2013) is conducted. These data sets originate from various fields of application and differ in their feature count, number of observations, and covariance structure. They are each evaluated with different feature costs and in individual budget limit setups. The overall simulation study

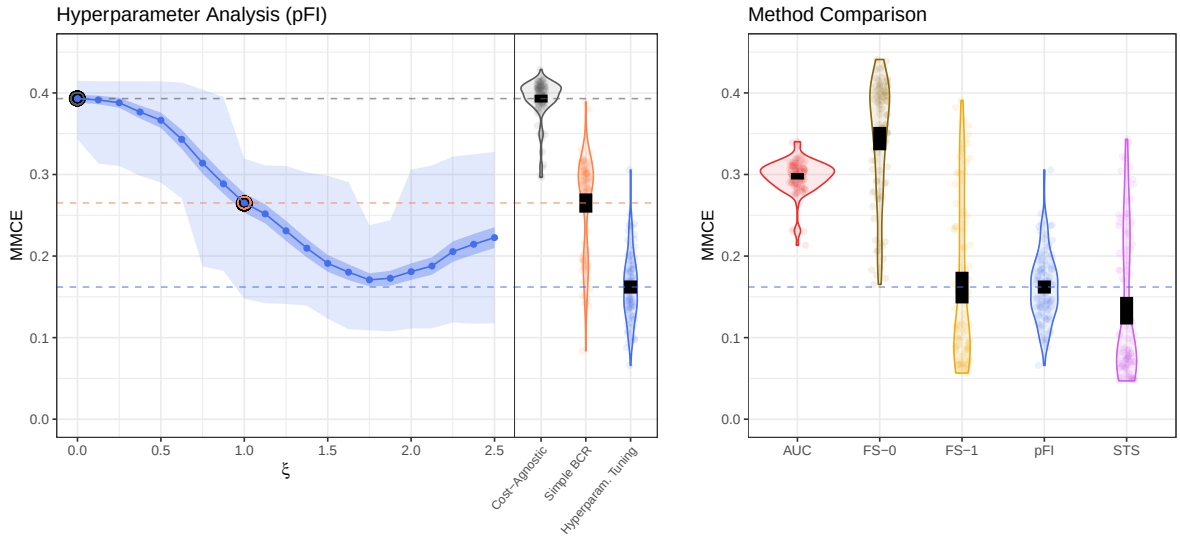

Figure 4.2: Artificial simulation setting D with multivariate effects on the response and correlations of effect sizes and feature costs. $c_{\text{max}} = 10$. Left: A blue line of connected points represents the mean MMCE over 100 simulation runs at a grid of ξ values. The dark shaded ribbon illustrates a 95%-CI around this mean value. The light shaded area shows the region between the 5% and 95% quantiles of the empirical distribution of the MMCE. The mean MMCE values at $\xi = 0$ and $\xi = 1$ are highlighted with a gray and an orange point, respectively. Violin plots on the right show the empirical MMCE distribution for the analyzed hyperparameter strategies. The 95%-CI region of the mean is given by a black box over the violins. The mean MMCE values of the three strategies are annotated with dashed lines over both subplots. Right: Violin plots with structure similar to the left plot for every feature selection method. Except for FS, the presented results for each method use the hyperparameter tuning approach.

has two main aims. The first is to understand the influence of the hyperparameter ξ in the BCR criterion. For this, a cost-agnostic approach, a simple BCR approach, and a hyperparameter tuning approach using Grid Search are compared. The second aim is to identify strengths and weaknesses of each of the proposed methods. These are analyzed with special focus on the known meta information of the given data scenarios to draw conclusions of the applicability of each approach.

4.3 Main Results and Conclusions

Results for both main objectives for one exemplary artificial simulation setting are illustrated in Figure 4.2. With respect to the hyperparameter choice, the simulations show that neither the cost-agnostic, nor the simple BCR approach is generally superior to the other. In the left plot of Figure 4.2, misclassification errors for the simple BCR choice are lower. This is, however, not consistent for different data setups and methods. Furthermore, the results also show that apart from these two analyzed choices, there is no other universally best fixed value of ξ. Therefore, tuning the hyperparameter is a relevant step to improve the general performance and should always be considered when applying the

BCR. Consequently, all following method comparisons are based on tuned results only.

The method comparisons show that in artificial settings with multivariate effects on the response variable, STS performs best (see Figure 4.2). In univariate setups, however, simple filter methods like AUC and pFI are superior. Correlations of costs and effect sizes do not influence these results notably. Larger budgets primarily result in smaller performance differences between the methods, but do not change the overall rankings. When applying the analyzed methods on real-world data sets, results vary strongly between different setups. While all methods rank higher than the purely performance-based baseline methods of this analysis, there is no one-fits-all solution among them. Every method is the best choice in at least one data setting and provides reasonable performance. These varying results could not be traced back to differences in budget limit, feature correlation, or field of application. They are therefore attributed to unknown underlying data generating mechanisms. In a global analysis of the average method rankings over all data sets, STS and pFI provided the best results.

In conclusion, all of the proposed methods can be successful strategies in specific data situations. As every approach follows fundamentally different basic ideas to generate a cost-sensitive feature subset, which may individually suit specific setups, there is no universal 'one-fits-all' approach. The novel STS method provides a fast and multivariate solution, which produces solid results in many analyzed situations, and - together with pFI - ranks best on a global average in the real-world simulations. Nevertheless, a thorough feature selection analysis should always base a final decision on a comparative evaluation of multiple methods to obtain the best results.

4.4 Outlook

The introduction of STS provides multiple ideas for further research. In the current method implementation, the main idea to control the number of features per tree is to limit the tree depth. This is a practical solution, which comes along with multiple downsides. First, a limited depth generally reduces predictive performance. Second, this approach does not guarantee that exactly the intended number of different features per tree is selected. There are multiple alternative ideas to this. One approach could be to limit the number of features a tree may use in a different way and allow the tree to fully grow with its selected set. Another idea would be to limit the cost of the tree directly

instead of limiting the feature count. As a third alternative, trees could also be grown in a full cost-sensitive manner that includes a trade-off of costs and performance for each split. The main hurdle for all of these ideas is that standard and fast implementations of Random Forests, such as for example the R-package *ranger* (Wright and Ziegler, 2015), could no longer be used out-of-the-box, as the internal tree generation algorithms would need to be altered. Nevertheless, a more general tree growing approach could for example help to reduce redundant trees or allow to emphasize custom multivariate structures.

Apart from ideas relating particularly to the STS algorithm, there are also more general extensions to this work. Currently, all analyzed methods use Grid Search for hyperparameter tuning. Alternative strategies might be able to increase performance and reduce run-time. Especially for FS, with its immense computational complexity, such developments may be a crucial aspect for deciding if the method is feasible at all.

Finally, all current simulations specialize on binary classification. Yet, none of the proposed methods are technically limited to this setup. Analyzing the effects on different response and model types therefore also provides a good basis for further research in this field.

REFERENCES

Akaike, Hirotugu (1974). "A new look at the statistical model identification". In: *IEEE transactions on automatic control* 19.6, pp. 716–723.

Bolón-Canedo, Verónica, Iago Porto-Díaz, Noelia Sánchez-Maroño, and Amparo Alonso-Betanzos (2014). "A framework for cost-based feature selection". In: *Pattern Recognition* 47.7, pp. 2481–2489.

Bommert, Andrea, Xudong Sun, Bernd Bischl, Jörg Rahnenführer, and Michel Lang (2020). "Benchmark for filter methods for feature selection in high-dimensional classification data". In: *Computational Statistics & Data Analysis* 143, p. 106839.

Breiman, Leo (2001). "Random forests". In: *Machine learning* 45.1, pp. 5–32.

Franklin, Jessica M, Sebastian Schneeweiss, Jennifer M Polinski, and Jeremy A Rassen (2014). "Plasmode simulation for the evaluation of pharmacoepidemiologic methods in complex healthcare databases". In: *Computational statistics & data analysis* 72, pp. 219–226.

Grubb, Alex and Drew Bagnell (2012). "Speedboost: Anytime prediction with uniform near-optimality". In: *Artificial Intelligence and Statistics*, pp. 458–466.

Guyon, Isabelle and André Elisseeff (2003). "An introduction to variable and feature selection". In: *Journal of machine learning research* 3.Mar, pp. 1157–1182.

Hanley, James A and Barbara J McNeil (1982). "The meaning and use of the area under a receiver operating characteristic (ROC) curve." In: *Radiology* 143.1, pp. 29–36.

Holland, John H (1973). "Genetic algorithms and the optimal allocation of trials". In: *SIAM Journal on Computing* 2.2, pp. 88–105.

Jagdhuber, Rudolf, Michel Lang, and Jörg Rahnenführer (2020). *Feature Selection Methods for Cost-Constrained Classification in Random Forests*. arXiv: 2008.06298v2 [stat.ML].

Jagdhuber, Rudolf, Michel Lang, Arnulf Stenzl, Jochen Neuhaus, and Jörg Rahnenführer (2020). "Cost-Constrained feature selection in binary classification: adaptations for greedy forward selection and genetic algorithms". In: *BMC bioinformatics* 21.1, pp. 1–21. DOI: 10.1186/s12859-020-3361-9. URL: https://doi.org/10.1186/s12859-020-3361-9.

Jagdhuber, Rudolf and Jörg Rahnenführer (2020). *Implications on Feature Detection when using the Benefit-Cost Ratio*. arXiv: 2008.05163v2 [stat.ML].

Kusner, Matt J (2016). "Learning in the Real World: Constraints on Cost, Space, and Privacy". In: *Engineering and Applied Science Theses & Dissertations* 305, pp. 14–37.

Kusner, Matt, Wenlin Chen, Quan Zhou, Zhixiang Eddie Xu, Kilian Weinberger, and Yixin Chen (2014). "Feature-cost sensitive learning with submodular trees of classifiers". In: *Twenty-Eighth AAAI Conference on Artificial Intelligence*.

Leskovec, Jure, Andreas Krause, Carlos Guestrin, Christos Faloutsos, Christos Faloutsos, Jeanne VanBriesen, and Natalie Glance (2007). "Cost-effective outbreak detection in networks". In: *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 420–429.

Min, Fan, Huaping He, Yuhua Qian, and William Zhu (2011). "Test-cost-sensitive attribute reduction". In: *Information Sciences* 181.22, pp. 4928–4942.

Min, Fan, Qinghua Hu, and William Zhu (2014). "Feature selection with test cost constraint". In: *International Journal of Approximate Reasoning* 55.1, pp. 167–179.

Min, Fan and Juan Xu (2016). "Semi-greedy heuristics for feature selection with test cost constraints". In: *Granular Computing* 1.3, pp. 199–211.

Tan, Ming (1993). "Cost-sensitive learning of classification knowledge and its applications in robotics". In: *Machine Learning* 13.1, pp. 7–33.

Vanschoren, Joaquin, Jan N. van Rijn, Bernd Bischl, and Luis Torgo (2013). "OpenML: Networked Science in Machine Learning". In: *SIGKDD Explorations* 15.2, pp. 49–60. DOI: 10.1145/2641190.2641198. URL: http://doi.acm.org/10.1145/2641190.2641198.

Vaughan, Laura K, Jasmin Divers, Miguel A Padilla, David T Redden, Hemant K Tiwari, Daniel Pomp, and David B Allison (2009). "The use of plasmodes as a supplement to simulations: a simple example evaluating individual admixture estimation methodologies". In: *Computational statistics & data analysis* 53.5, pp. 1755–1766.

Wright, Marvin N and Andreas Ziegler (2015). "ranger: A fast implementation of random forests for high dimensional data in C++ and R". In: *arXiv preprint arXiv:1508.04409*.

Xu, Zhixiang, Matt J Kusner, Kilian Q Weinberger, Minmin Chen, and Olivier Chapelle (2014). "Classifier cascades and trees for minimizing feature evaluation cost". In: *The Journal of Machine Learning Research* 15.1, pp. 2113–2144.

Xu, Zhixiang, Matt Kusner, Kilian Weinberger, and Minmin Chen (2013). "Cost-sensitive tree of classifiers". In: *International Conference on Machine Learning*, pp. 133–141.

Zhou, Qifeng, Hao Zhou, and Tao Li (2016). "Cost-sensitive feature selection using random forest: Selecting low-cost subsets of informative features". In: *Knowledge-Based Systems* 95, pp. 1–11.