



Implications on Feature Detection When Using the Benefit–Cost Ratio

Rudolf Jagdhuber¹ · Jörg Rahnenführer¹

Received: 20 July 2020 / Accepted: 15 May 2021 / Published online: 3 June 2021
© The Author(s) 2021

Abstract

In many practical machine learning applications, there are two objectives: one is to maximize predictive accuracy and the other is to minimize costs of the resulting model. These costs of individual features may be financial costs, but can also refer to other aspects, for example, evaluation time. Feature selection addresses both objectives, as it reduces the number of features and can improve the generalization ability of the model. If costs differ between features, the feature selection needs to trade-off the individual benefit and cost of each feature. A popular trade-off choice is the ratio of both, the benefit–cost ratio (BCR). In this paper, we analyze implications of using this measure with special focus to the ability to distinguish relevant features from noise. We perform simulation studies for different cost and data settings and obtain detection rates of relevant features and empirical distributions of the trade-off ratio. Our simulation studies exposed a clear impact of the cost setting on the detection rate. In situations with large cost differences and small effect sizes, the BCR missed relevant features and preferred cheap noise features. We conclude that a trade-off between predictive performance and costs without a controlling hyperparameter can easily overemphasize very cheap noise features. While the simple benefit–cost ratio offers an easy solution to incorporate costs, it is important to be aware of its risks. Avoiding costs close to 0, rescaling large cost differences, or using a hyperparameter trade-off are ways to counteract the adverse effects exposed in this paper.

Keywords Feature costs · Feature detection · Benefit–cost ratio · Feature selection · Cost-sensitive learning

Background

Feature selection is a common preprocessing step in many learning tasks, which aims to remove noise features and to identify a suitable subset of relevant information from an often high-dimensional data set. This way it can improve the generalization ability and reduces computational complexity of subsequent learning algorithms. The field of feature selection is widely studied. A thorough introduction of the main concepts can be found, e.g., in Guyon and Elisseeff [6]. A recent benchmark study comparing different filter algorithms was presented by Bommert et al. [2]. Cost-sensitive learning describes an extension of the general feature selection problem by introducing acquisition costs for selected features. Depending on the application field, these costs may not only refer to financial aspects, but could also represent

a time span to raise a feature or a patient harm during the sample taking process.

The general strategy to incorporate feature costs into a feature selection framework depends on the problem at hand. If a fixed total feature cost limit can be defined, the problem reduces to an additional optimization constraint for the feature selection problem. Many example applications of fixed budget costs can be found. Min et al. [13, 14] presents cost-sensitive feature selection heuristics and also provide a thorough problem definition in the context of rough sets. Jagdhuber et al. [7] and Liu et al. [11] further extend this idea and propose genetic algorithms with fixed feature cost budgets. For situations without a fixed cost limit, the goal may be to harmonize costs of features and costs of prediction errors by identifying an optimal trade-off. Research on these flexible solutions can be found, e.g., in Liu et al. [20], who incorporate individual feature costs in the generation of the base trees from random forests to produce lower cost solutions on average, or in Zhou et al. [1], who develop a framework to include individual costs in standard filter methods. A third situation is given when feature acquisition is undertaken sequentially. In such situations, tests can take

✉ Jörg Rahnenführer
rahnenuuehrer@statistik.tu-dortmund.de

¹ Department of Statistics, TU Dortmund, Vogelpothsweg 87, 44227 Dortmund, Germany

advantage of intermediate results and reduce total costs by only requesting further features if the benefit justifies the additional cost, see, e.g., the work of Kusner et al. [9] and Xu et al. [18, 19], who developed a method named “tree of classifiers”. In this method, test inputs traverse along individual paths of the tree, which include different features and thus allow to reduce the average prediction costs of the total population.

A common factor for all mentioned tasks is the need to somehow trade-off the benefit of a feature with its cost. As these measures are on different scales, very popular options to combine them are either to optimize the ratio of both [5, 10, 12–15], or to trade-off a weighted sum [1, 8, 9, 17–19].

In this paper, we take a close look at the first of these two mentioned alternatives. We specifically analyze the consequences of using a simple benefit–cost ratio (BCR) with respect to the discrimination between relevant features and features with no information. Especially for small effect sizes, penalizing an information criterion can obfuscate performance measures below noise level. We aim to assess important factors that influence this effect and raise awareness for the consequences for the feature detection rate when using this popular measure. To clearly illustrate the influence of the BCR we restrict the analysis to the basic scenario of a single feature selection step. We do not consider general feature selection strategies and do not assess the relevance of whole feature sets.

We start by defining the general cost-sensitive feature selection problem and discussing the theoretical implications of using the BCR. In the following section we perform simulation studies to analyze the influence of multiple data parameters and feature cost settings on the feature detection rate. Finally, we present the obtained results, discuss the general applicability of the basic BCR and provide recommendations for alternative trade-off measures.

Problem Definition

Given is a data set with n observations $D_i, i = 1, \dots, n$ and p features $x_{ij}, j = 1, \dots, p$ for observation i , and continuous response y_i for observation i . Assume that the true relation is given by $y_i = \beta_0 + \sum_{j=1}^{p_{\text{rel}}} x_{ij}\beta_j + \epsilon_i$, with $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. In this data p_{rel} features are assumed to have an influence on the response, while all other p_{noise} features are independent of y . Then the goal of feature selection is to identify the subset of relevant features.

One obvious approach to ensure finding this optimal subset is an exhaustive search, i.e. to consider all possible subsets. However, this approach is usually not feasible for high-dimensional feature spaces. Thus, heuristic selection algorithms like greedy sequential forward selection (SFS)

are used. SFS iteratively adds the single most promising feature to the current result set. A typical way to estimate the importance of a feature x_j when added to a given set s is to calculate the performance gain of a statistical model including the feature $M(s \cup \{x_j\}|D)$ compared to a baseline model without it $M(s|D)$. The feature with the highest gain in performance is then selected. Assuming a performance criterion Q , for which the optimal value is the minimal value, we can formulate one feature selection step of SFS by

$$\hat{m} = \arg \max_j \{Q(M(s|D)) - Q(M(s \cup \{x_j\}|D))\} =: \arg \max_j \Delta Q_j. \quad (1)$$

In many real-world scenarios, obtaining a feature x_j may cause individual feature costs c_j . Cost-sensitive feature selection aims to incorporate these costs into the selection process to find cheap and well performing models. A popular method is to adapt the problem of Eq. (1) to

$$\hat{m} = \arg \max_j \frac{\Delta Q_j}{c_j}. \quad (2)$$

This ratio of benefit and cost leads to a simple trade-off optimization, which relates the importance of a feature to its cost. In the following we describe negative implications of this simple and popular method when discriminating between relevant and noise features.

The true performance gain of a noise feature is a value smaller or equal to zero, as it has no relation to the response but may create additional uncertainty. The true performance gain of a relevant feature is typically a value greater than 0. Nevertheless, the actual performance gain estimated on a sample data set does not always result in these true values. It can rather be seen as a random variable following a certain unknown distribution around the true value:

$$\Delta Q_j \sim \mathcal{V}_j(\cdot). \quad (3)$$

For a real world data situation, the theoretical distributions of ΔQ_j for different j can be assumed to overlap to some extent. That means, for one given sample data set, the actual estimated performance gain of a noise feature may be higher than the one of the relevant feature and thus an irrelevant feature may be selected.

When incorporating cost according to Eq. (2), the performance gain distribution of feature x_j is scaled by a positive factor c_j , which increases and broadens \mathcal{V}_j , if $c_j < 1$, and decreases and narrows it, if $c_j > 1$. Increasing and broadening the distribution of a noise feature, while not altering the one of a relevant feature increases the overlap of both distributions. Therefore, the probability of falsely selecting the noise feature increases. In some situations this problem may be negligible. In others, the cost-sensitive feature selection procedure can completely obfuscate any relevant feature.

The actual magnitude of the cost influence depends on many factors including the sample size n , the true effect size of relevant features β , the residual variance σ^2 , the statistical model, and the performance measure Q . The goal of this paper is to analyze this problem and describe multiple parameter settings and their influence on the feature detection rate. We focus on linear regression models and use the root mean squared error (RMSE) on independent data to assess the quality of models. The RMSE is defined as

$$RMSE = \sqrt{\sum_{i=1}^n \left(y_i - \hat{\beta}_0 - \sum_{j \in S} x_{ij} \hat{\beta}_j \right)^2}, \tag{4}$$

with $\hat{\beta}_0$ and $\hat{\beta}_j$ estimated on training data and x_{ij} and y_i denoting observations of an independent test data set. By using such an independent test data set, the RMSE also allows a result of no improvement after adding a feature.

In the following, for ease of presentation, we describe a single feature selection step of SFS from a pool of p_{rel} relevant and p_{noise} noise features. We also define this single step to be the first selection step, i.e. we define our baseline model to be the intercept model and compare the quality of all one-feature models. The final selection result of this one step can either be ‘noise selected’, ‘relevant feature selected’, or ‘no feature selected’. Similarly to Definition (1), in the following we denote the gain in RMSE for feature j by $\Delta RMSE_j$. The corresponding distribution $\mathcal{V}_j(\cdot)$ has no analytical form. In the artificial simulation study, we overcome this problem by numerically approximating this distribution and computing selection probabilities on the empirical distribution.

Simulation Studies

Two simulation studies are performed. The first uses artificially generated data and evaluates a broad spectrum of simulation parameters, while the second is based on a real data set to also analyze a setup observed in the real world. In the following subsections, both studies are introduced in detail.

Artificial Data Simulation

The main goal of both simulation studies is to assess the detection rate of a cost-sensitive feature selection step in multiple parameter settings. Additionally for our artificial setup we aim to analyze the empirical distribution of our performance measure to further illustrate effects of cost scaling. We consider a linear regression scenario. Our response variable, as well as all p features are assumed to be normally distributed. We define p_{rel} features to be relevant and the remaining $p_{noise} = p - p_{rel}$

features to be noise. The individual costs of features can be seen as a relative scaling between the respective $\Delta RMSE_j$ values of the features. To simplify our analyses, we do not consider individual costs for all features, but define only one single scaling factor θ for the relevant features. Hence, we implicitly define equal costs for the group of noise features and equal costs for the group of relevant features. We only differentiate between costs for information and costs for noise. To thoroughly assess the influence on the detection rate, we vary the feature cost scaling factor θ between 1, 10, 100 and 1000, the number of relevant features p_{rel} between 1, 2, 5 and 10, the number of noise features p_{noise} between 1, 10 and 50, and the effect size of the relevant feature β between 0, 0.01, ..., 0.5. For multiple relevant features, we do not vary the effect size and define $\beta_j := \beta$.

For each parameter combination, $B = 1000$ training ($n_{train} = 100$) and test data sets ($n_{test} = 1000$) are generated as follows. In a first step, features are drawn from a p -dimensional normal distribution

$$x_1, \dots, x_p \sim \mathcal{N}_p(\mathbf{0}, \mathbf{I}_p), \tag{5}$$

where \mathbf{I}_p is the p -dimensional identity matrix. Next, the response is drawn from the normal distribution

$$y_i \sim \mathcal{N}\left(\beta_0 + \sum_{j=1}^{p_{rel}} x_{ij} \beta, \sigma^2\right). \tag{6}$$

We set the intercept to $\beta_0 = 1$ and the residual variance to $\sigma^2 = 1$ for all settings.

For every data set obtained in this way, we fit the baseline intercept model and all one-feature models separately and obtain

$$\begin{aligned} M_0 &: y = \hat{\beta}_0 + \epsilon, \\ M_j &: y = \hat{\beta}_0 + x_j \hat{\beta}_j + \epsilon, \quad j = 1, \dots, p. \end{aligned} \tag{7}$$

We then compute the increase in RMSE for all features by $\Delta RMSE_j = RMSE(M_0) - RMSE(M_j), \quad j = 1, \dots, p.$ (8)

As we are only interested in the question if a noise feature or a relevant feature is selected, we define the RMSE gain of noise and relevant features as our target variables. The best $\Delta RMSE$ value indicates the candidate that is selected from the noise and the relevant features, respectively.

$$\begin{aligned} \Delta RMSE_{rel} &= \max(\{\Delta RMSE_j : j = 1, \dots, p_{rel}\}) \\ \Delta RMSE_{noise} &= \max(\{\Delta RMSE_k : k = p_{rel} + 1, \dots, p\}). \end{aligned} \tag{9}$$

As described earlier, we define our cost setting by a single factor θ , which scales relevant features. Hence, the assessed measure of RMSE gain for relevant features actually results in $\frac{\Delta RMSE_{rel}}{\theta}$.

The final feature selection on a single data set can lead to three different outcomes \hat{m} . We only consider increases in ΔRMSE . Therefore, if neither relevant, nor noise features result in a positive RMSE gain, then no feature is selected.

$$\hat{m} = \arg \max \left(\frac{\Delta\text{RMSE}_{\text{rel}}}{\theta}, \Delta\text{RMSE}_{\text{noise}}, 0 \right). \quad (10)$$

As every setting is repeated 1000 times with newly simulated data sets, we can estimate the probability for each selection result m by looking at the relative frequency among those 1000 runs. We can further obtain empirical distributions of ΔRMSE for relevant and for noise features in different settings. The results for both of these analyses are presented in the following section.

Plasmode Simulation on Real-World Data

To assess the detection rate of a feature selection algorithm, it is essential to know, which of the available features are actually relevant and which are noise. For a real-world data set, this information is typically unknown. To solve this problem we perform a so-called plasmode simulation study [16]. A plasmode study uses a data set generated from natural processes but adds a simulated aspect to the data [4]. For this paper, we use the well-known Spambase dataset from the UCI machine learning repository [3] as basis of our plasmode simulation. It contains data of 4601 E-mails with 57 numeric features including word and character frequencies as well as further general numeric measures on the text composition. To create a controlled scenario, which allows an objective assessment of the detection rate, the real relationship between features and response variable needs to be known. Hence, the response variable is generated from a fixed set of six features, corresponding to approximately 10% of all features, that are defined to be relevant. All features are standardized, and the response $y_i, i = 1, \dots, 4601$, is drawn from the distribution

$$y_i \sim \mathcal{N} \left(1 + \sum_{j=1}^6 x_{ij} \beta_j; 1 \right), \quad (11)$$

which introduces a linear relation with similar intercept and residual variance as in the artificial simulation. We define $\beta_j = 0.25$ and analyze cost scalings between 1 and 5 to create a challenging setup.

One thousand simulation runs are performed as follows. The data are split randomly into approximately $\frac{2}{3}$ of all observations (3067) used for training and approximately $\frac{1}{3}$ of all observations (1534) used for testing. For every training data set, we fit the baseline intercept model and all one-feature models separately and obtain the resulting ΔRMSE on test data analogously to the artificial simulation. The maximum

value of this difference reveals if a relevant feature, a noise feature, or no feature at all is detected. We analyze the detection rates of relevant features for different cost-scaling values and perform a one-sided two-proportions z -test, which tests the null hypothesis that the detection rate at a given value of θ is not smaller than the detection rate at $\theta = 1$ (corresponding to no cost-scaling).

Results

Artificial Data Simulation

This section comprises the analysis of the selection probabilities with main results presented in Figs. 1 and 2, and the analysis of the empirical distribution of the selection criterion presented in Fig. 3. To provide comprehensive illustrations, both analyses focus mainly on the setting with one relevant feature, and only a small analysis to describe the effects of different numbers of relevant features is added. Corresponding illustrations of all settings can be found in the Supplementary material.

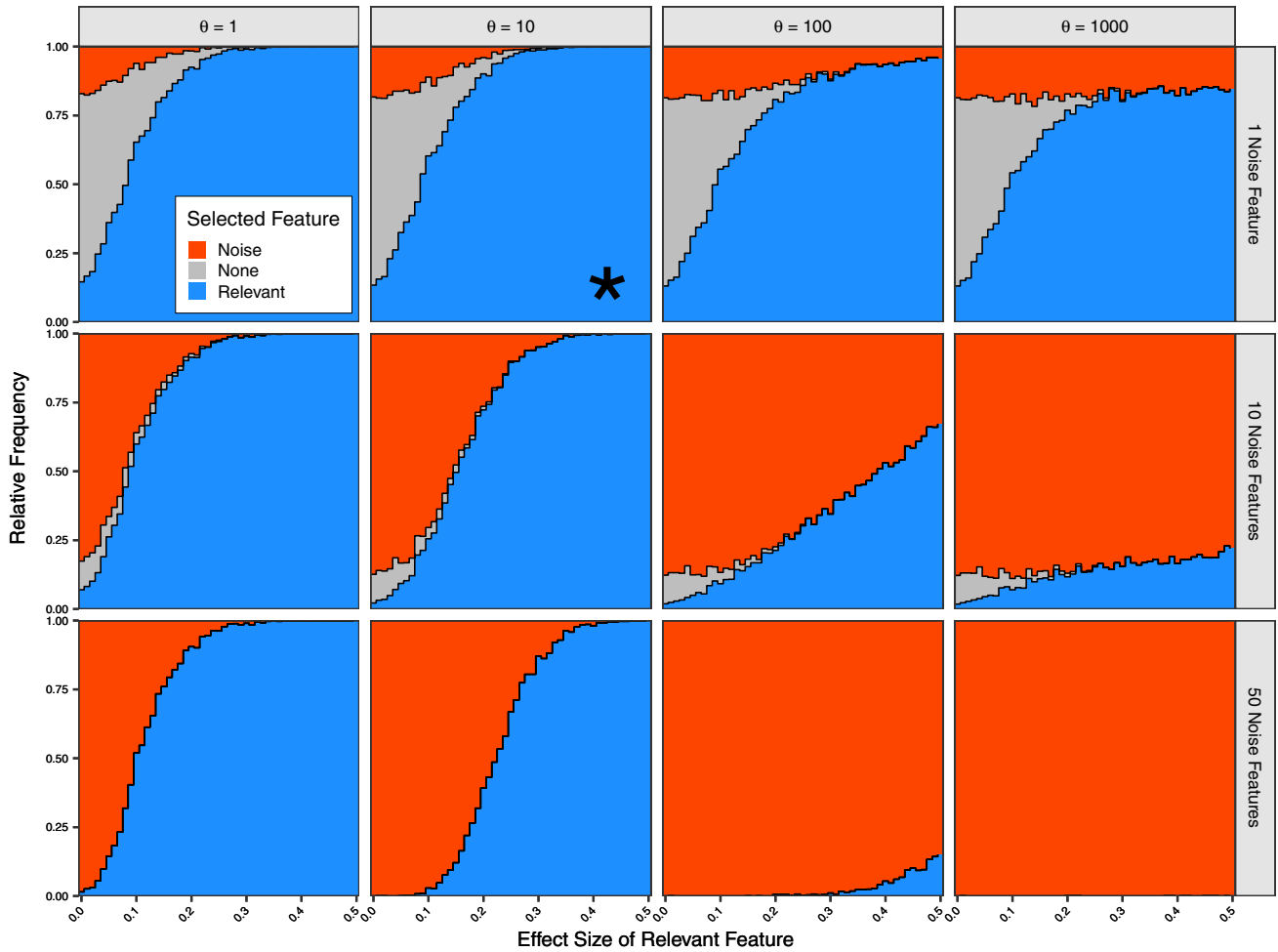
The individual plots of Fig. 1 illustrate the estimated probabilities for the three selection outcomes ‘relevant feature selected’, ‘noise feature selected’ and ‘no feature selected’ along multiple effect sizes of the true effect β . Rows of the main plot matrix relate to different numbers of noise features, while columns represent the extent of cost-scaling applied to the relevant feature.

The top-left plot describes a setting with one relevant and one noise feature. No cost scaling is applied, which could refer to a setting without or with equal costs, respectively. At an effect size of $\beta = 0$, where both features can be considered noise, their selection probability is approximately equal. In almost 70% of the cases, neither of them is selected. When increasing the effect size β , the selection probability for the relevant feature rises, while the probabilities for both other outcomes decrease. From around $\beta = 0.3$ onward, the relevant feature is identified approximately 100% of times.

Increasing the number of noise features (rows 2 and 3) changes this result in multiple ways. The main difference can be seen in the number of times that no feature is selected. This value is reduced greatly for ten noise features and disappears completely for 50 noise features. The other difference is that the selection curve of the relevant feature starts at a lower value and reaches 100% selection slightly later. These differences are however more subtle.

The main focus of our paper lies on the effect of incorporating costs and thus scaling the performance distribution of the relevant features. This scaling factor corresponds to the columns of the plot matrix. When increasing the factor, the decrease in selection probability of noise for higher effect sizes becomes smaller, eventually resulting

Scaling Analysis (1 Relevant Feature)



Relevant Feature Analysis ($\theta = 10$)

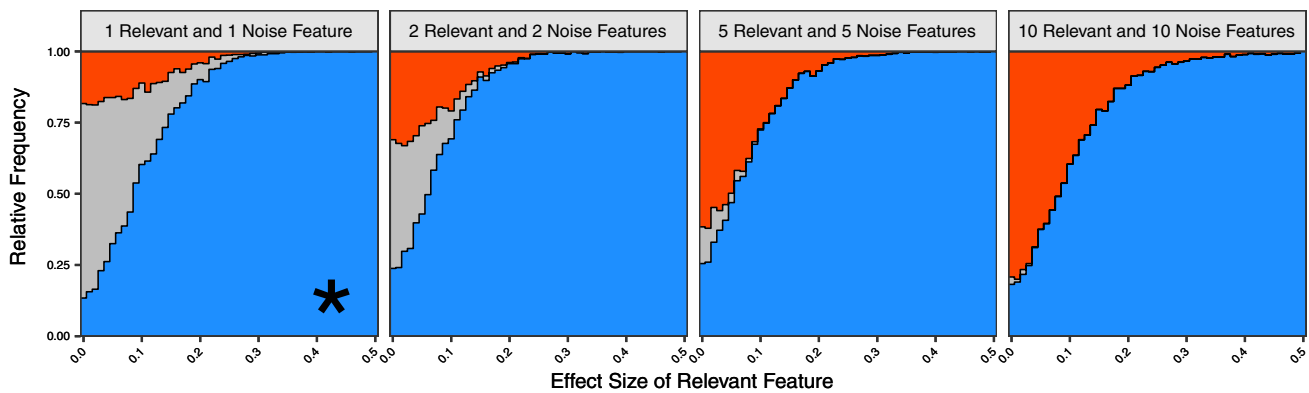


Fig. 1 Selection probabilities of relevant, noise, or no feature, along multiple values of the cost-scaling factor θ (columns), the number of noise features p_{noise} (rows) and the effect size of relevant features β (x-axis per plot). The main 3×4 plot matrix analyzes the setting

of $p_{rel} = 1$. An additional bottom row illustrates corresponding plots for different numbers of relevant features p_{rel} at a fixed scaling level $\theta = 10$. The plots annotated with a star are identical

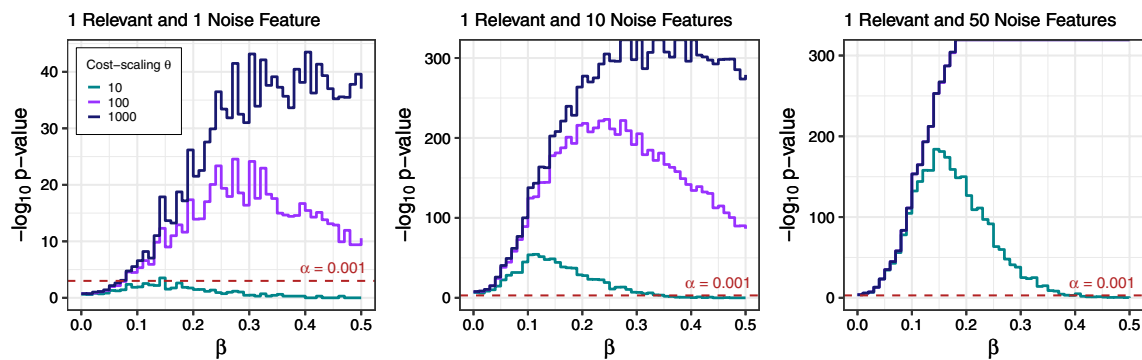


Fig. 2 Three plots with different setups of relevant and noise features corresponding to the three rows of Fig. 1. Each y-axis shows negative base-10 logarithms of p values from a one-sided two-proportions z -test, which tests the null hypothesis that the detection rate at a given

in an approximately constant noise selection probability at $\theta = 1000$. As the initial selection probability for noise increases with a larger number of noise features, the combined effect results in always selecting noise at the bottom-right plot.

The effects of increasing the number of relevant features is illustrated for a fixed scaling factor $\theta = 10$ and an equal number of noise and relevant features in the additional bottom row of Fig. 1. The main observation is that the extent of selecting no feature reduces with increasing p_{rel} and instead a noise feature is selected. The probability of selecting a relevant feature does not seem to be strongly influenced, it is only slightly pushed back by noise and reaches the area of 100% selection for slightly larger effect sizes. Full illustrations including multiple values of θ and non-identical p_{rel} and p_{noise} are given in Additional file 1.

To test the significance of the observed effects, Fig. 2 illustrates negative base-10 logarithms of p values from a one-sided two-proportions z test, which tests the null hypothesis that the detection rate of a cost-scaled relevant feature is not smaller than the detection rate without cost-scaling.

For one noise feature and one relevant feature with very small effect size ($\beta < 0.075$), the left plot indicates no significant ($\alpha = 0.001$) differences for any analyzed cost-scaling. Apart from that region, scalings of $\theta = 100$ and $\theta = 1000$ always significantly lower the detection rate of relevant features. For $\theta = 10$, the number of noise features is an important factor for the resulting p values. With only one noise feature (left plot in Fig. 2), almost no significant decreases in the detection rate can be observed. With 10 and 50 noise features, however, significant differences can be observed for all medium-sized values of β and the differences only disappear for large β values, when the detection rate approaches 1. Altogether the notable decrease in the detection rate observed in Fig. 1 can also be considered

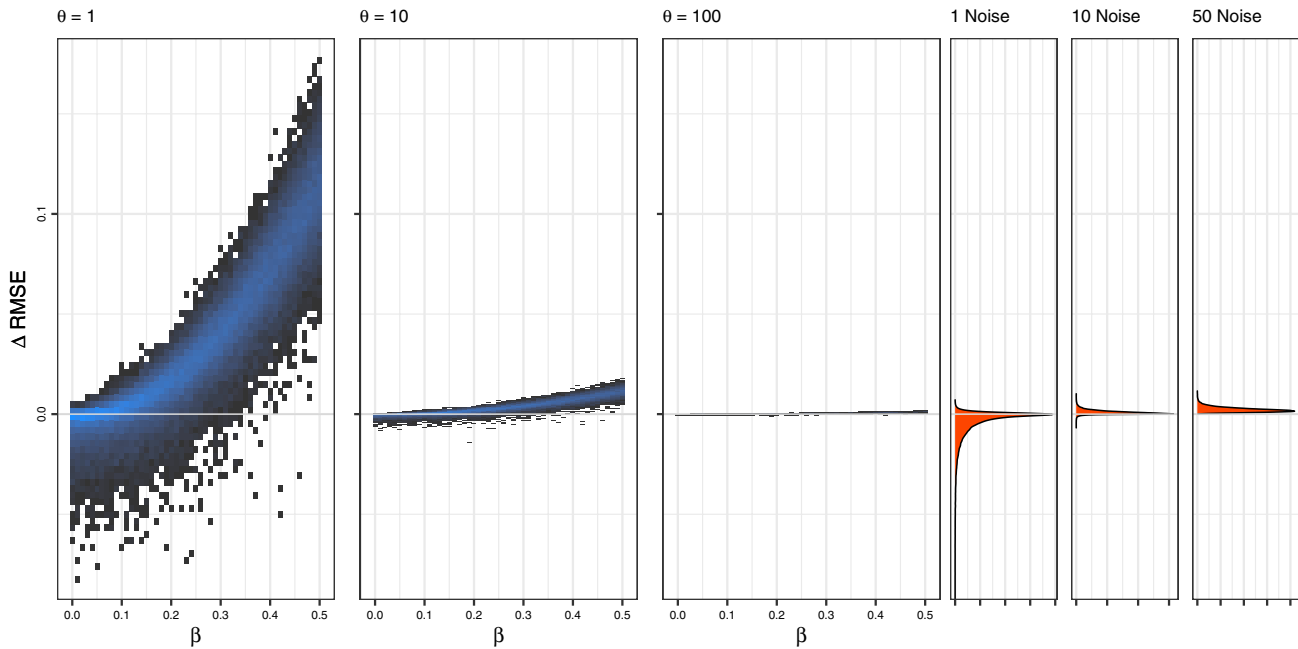
value of θ and β is not smaller than the detection rate at the same β with $\theta = 1$ (no cost-scaling). On the x -axis different values of β are indicated. The line color corresponds to different values of θ . A horizontal dashed red line indicates the significance level $\alpha = 0.001$

statistically significant for most non-extreme cases according to Fig. 2.

The second aspect analyzed in the artificial simulation is the empirical distribution of RMSE gain for the relevant features. This distribution depends on the true effect β , the cost scaling parameter θ , and the number of relevant features in the model. For noise, it only depends on the numbers of noise and relevant features, as the true effect is 0 and no scaling of noise is performed. A comprehensive illustration of all analyzed distributions for $p_{\text{rel}} = 1$ is given in the top plot of Fig. 3. A heatmap describes the distributions of RMSE gain for relevant features along different effect sizes. Lighter colors correspond to higher densities. RMSE gains for noise features are depicted by three density curves for settings with 1, 10 and 50 noise features, respectively. A gray area highlights the decision boundary for not selecting any feature.

The given plots provide deeper insight into the selection decisions illustrated previously in Fig. 1. Analyzing the noise features, the distribution of RMSE gains of one single noise feature has the great majority of its probability mass within the gray area and would not be selected, regardless of the RMSE gain of the relevant feature. However, when increasing the number of noise features p_{noise} , the noise distribution steadily moves out of this area. For the relevant parameter, the unscaled distribution (top-left plot) increases superlinearly along β and completely passes any noise distribution at around $\beta = 0.4$. A cost-scaling however lowers the slope of this increase and decreases the variance of the relevant feature distribution. As a consequence of both, surpassing the noise distributions happens notably slower. For $\theta = 100$, the size of the relevant feature distribution compared to noise is shrunken down to a level making it almost invisible in the plot. The largest noise distribution is not surpassed at all in our range of β values. However, an important observation is that the total density of $\Delta\text{RMSE}_{\text{rel}}$ below

Scaling Analysis (1 Relevant Feature)



Relevant Feature Analysis ($\theta = 10$)

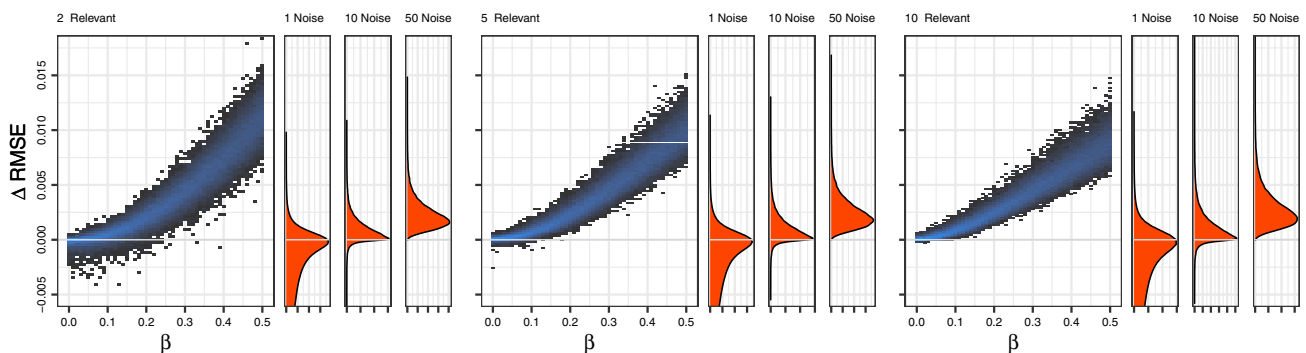


Fig. 3 Empirical distributions of $\Delta RMSE$ for noise features (red) and relevant features (blue). The latter are illustrated as heatmap along different values of the true effect size β . Lighter colors indicate higher densities. The first row describes a setting with $p_{rel} = 1$. Three plots

of relevant features for different values of the cost-scaling θ and three plots of noise features for different values of p_{noise} are given. The bottom row shows corresponding plots for different values of p_{rel} at a fixed cost-scaling of $\theta = 10$

or equal to zero is constant for any scaling. We omitted an illustration for $\theta = 1000$ as it is invisible on this scale. Rescaled versions for all distributions can be found in Additional file 2.

The bottom part of Fig. 3 depicts the effects of increasing the number of relevant features in the true model, for $\theta = 10$. Mainly, the general density mass below zero decreases when the number of relevant features increases. However, the maximum $\Delta RMSE$ value for $\beta = 0.5$ also decreases. The RMSE gain of noise features, on the other hand, results in almost identical density curves.

Plasmode Simulation

The plasmode simulation study analyzes a specific setup, where 6 out of 57 features from a real-world data set are defined to be relevant ($\beta = 0.25$) for the prediction of a simulated response variable. Our goal is to analyze the decrease in the detection rate of relevant features for different cost-scalings. Furthermore, we assess the significance of this decrease with a one-sided two-proportions z -test. The results of these analyses are illustrated in Fig. 4.

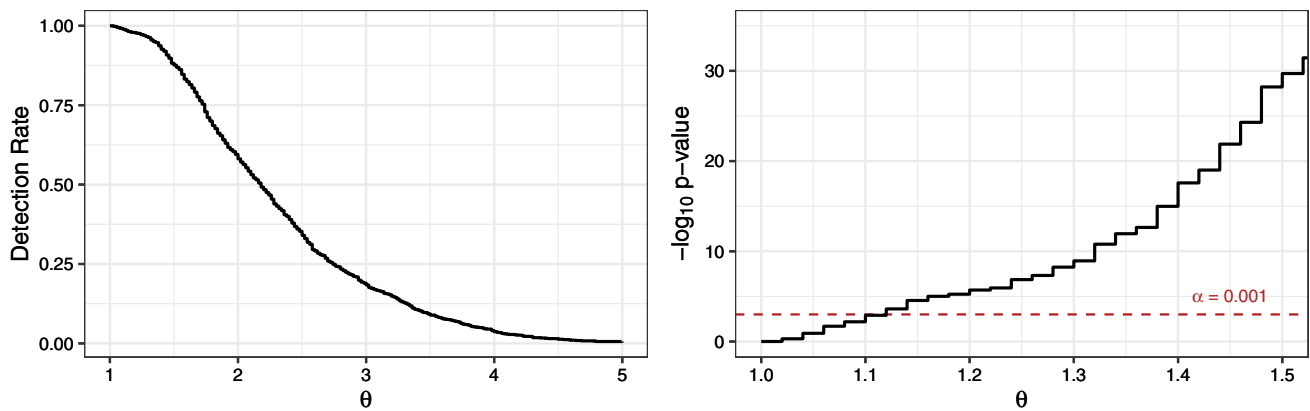


Fig. 4 Results of the plasmode simulation study. Left: detection rates of relevant features for different values of the cost-scaling factor θ . Right: negative base-10 logarithm of p values from a one-sided two-proportions z -test, which tests the null hypothesis that the detection

Without cost-scaling, the relevant features are detected in all 1000 simulation runs. However, at higher values of θ , this rate decreases to $\approx 50\%$ at $\theta = 2.16$ and 0.6% at $\theta = 5$. This means that even small cost-scalings can already highly impair the detection rates on an absolute level. The right plot of Fig. 4 puts these results into a statistical testing context. Assuming a significance level of $\alpha = 0.001$, the observed decrease can be considered significant for $\theta = 1.1$ and all larger values.

Discussion and Conclusion

The simulation studies revealed multiple consequences of cost-sensitive feature selection when using the popular benefit–cost ratio without a hyperparameter. In Fig. 1, we see that cost-scaling ΔAIC makes the selection probability of noise features more robust, especially for large true effects. With $\theta \rightarrow \infty$, this probability becomes independent of β . However, the frequency of selecting noise does not necessarily approach 1, but converges to a certain limit. For $\theta \rightarrow \infty$, this limit is given by $P(\Delta RMSE_{\text{noise}} > 0)$. Values with negative RMSE difference will never be selected, regardless of the scaling. With an increasing number of noise features, the probability that all estimated performance gains are negative decreases. Hence, the described limit for selecting noise rises. The third row of Fig. 1 illustrates the consequences of both effects, which eventually results in a noise selection probability of approximately 1 for all β values. Statistical test results given in Fig. 2 showed that this observed decrease of the detection of relevant features can be considered significant for almost all analyzed cost-scaling setups. The empirical distributions shown in Fig. 3 further describe this relation. With higher cost penalization, the slope and

rate at a given value of θ is not smaller than the detection rate at $\theta = 1$ (no cost-scaling). A dashed red line indicates the significance level $\alpha = 0.001$

variance of the RMSE gain distribution along β decreases. The probability regions favoring noise over the relevant features constantly become larger as θ increases, yet the probability masses above and below 0 stay constant, further illustrating the probability limit of noise selection. The effects of increasing the number of relevant features in the true model are more subtle. The selection probability plots mainly show the effects already observed when increasing the number of noise features. The differences in the empirical densities of RMSE gains of relevant features in Fig. 3 are the result of two effects. On the one hand, the maximum RMSE results in a higher value for a higher number of features. On the other hand, the relative share on the total information of a single feature decreases with higher p_{rel} . For small β , the distribution of $\Delta RMSE_{\text{rel}}$ is very skewed and the first effect dominates. For larger β , the distribution becomes less skewed and the latter effect has a higher impact. In total, this results in the observed trends with increasing p_{rel} . Extending the simulation setup to a real-world data set further highlighted the effects observed in the artificial setups. The study showed that even minor cost-scalings in the range of 10% can already significantly impair the ability to distinguish relevant information from noise.

Altogether, our paper addressed implications of using the benefit–cost ratio without an additional hyperparameter for cost-sensitive feature selection. As using this ratio is a typical approach to incorporate feature costs, it is important to understand possible problems resulting from it. We provided a thorough problem description, analyzed multiple parameter settings in an artificial simulation study and also evaluated detection rates on a real-world data set. Results from these studies illustrated that a strong cost-scaling, which may result from high relative cost differences between features, can notably influence the detection

limit of relevant features. This effect interacts with the number of noise features in the data.

To avoid this problem, we recommend using an adapted benefit–cost ratio, such as the ones proposed in Jagdhuber et al. [7] or Min et al. [13]. The main alternative solution to incorporate costs is a weighted linear combination as mentioned in the introduction of this paper. All these approaches share the idea of introducing a hyperparameter to control the trade-off between benefit and cost. This can reduce the problem, but it comes at the price of an additional estimation step. If the analysis requires the benefit–cost ratio without hyperparameter, we strongly recommend to thoroughly analyze the cost distribution of the given data set. If relative cost differences are high, transforming costs prior to applying the benefit–cost ratio may be beneficial. In practice, such extreme ratios may likely occur with some costs very close to 0, or from setting cost-free features to a cost of ϵ close to 0, as, e.g. recommended in Min et al. [13].

The popularity of the benefit–cost ratio shows the need for simple methods to incorporate costs without an additional parameter tuning step. Beyond the scope of this work, solving this problem with a comprehensible way to specify the trade-off between costs and performance with expert knowledge, instead of tuning a black-box hyperparameter, would be of great interest. This would allow the user to specify the intended relation of costs and performance, which may differ greatly between fields of application. Our work covers a specific task in predictive modelling and tries to raise awareness of the problem. Nevertheless, many other modelling approaches or machine learning methods may be considered in future work. Further research may also deal with classification problems, or with different performance measures. Comparisons of the influences or possible biases from the choices regarding these aspects may also be relevant extensions of this work.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s42979-021-00705-6>.

Author Contributions RJ initiated the topic, formulated and discussed the problem, designed and executed the simulation studies, interpreted the results, and wrote the manuscript. JR 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.

Funding Open Access funding enabled and organized by Projekt DEAL. This work was supported by Deutsche Forschungsgemeinschaft (DFG), Project RA 870/7-1, and Collaborative Research Center SFB 876, A3. The authors acknowledge financial support by Deutsche Forschungsgemeinschaft and Technische Universität Dortmund within the funding programme Open Access Publishing.

Data Availability The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Code Availability The full code used during the current study is available from the corresponding author on reasonable request.

Declarations

Conflict of Interest The authors declare that they have no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Bolón-Canedo V, Porto-Díaz I, Sánchez-Marroño N, Alonso-Betanzos A. A framework for cost-based feature selection. *Pattern Recognit.* 2014;47(7):2481–9.
2. Bommert A, Sun X, Bischl B, Rahnenführer J, Lang M. Benchmark for filter methods for feature selection in high-dimensional classification data. *Comput Stat Data Anal.* 2020;143:106839.
3. Dua D, Graff C. UCI machine learning repository. 2017. <http://archive.ics.uci.edu/ml>. Accessed 4 Apr 2020.
4. Franklin JM, Schneeweiss S, Polinski JM, Rassen JA. Plasmode simulation for the evaluation of pharmacoepidemiologic methods in complex healthcare databases. *Comput Stat Data Anal.* 2014;72:219–26.
5. Grubb A, Bagnell D. Speedboost: anytime prediction with uniform near-optimality. *Artif Intell Stat.* 2012;22:458–66.
6. Guyon I, Elisseeff A. An introduction to variable and feature selection. *J Mach Learn Res.* 2003;3(Mar):1157–82.
7. Jagdhuber R, Lang M, Stenzl A, Neuhaus J, Rahnenführer J. Cost-constrained feature selection in binary classification: adaptations for greedy forward selection and genetic algorithms. *BMC Bioinform.* 2020;21(1):1–21.
8. Kong G, Jiang L, Li C. Beyond accuracy: learning selective Bayesian classifiers with minimal test cost. *Pattern Recognit Lett.* 2016;80:165–71.
9. Kusner M, Chen W, Zhou Q, Xu ZE, Weinberger K, Chen Y. Feature-cost sensitive learning with submodular trees of classifiers. In: Twenty-eighth AAAI conference on artificial intelligence; 2014.
10. Leskovec J, Krause A, Guestrin C, Faloutsos C, Faloutsos C, VanBriesen J, Glance N. Cost-effective outbreak detection in networks. In: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM; 2007. p. 420–429.
11. Liu J, Min F, Liao S, Zhu W. A genetic algorithm to attribute reduction with test cost constraint. In: 2011 6th International

- conference on computer sciences and convergence information technology (ICCCIT), IEEE; 2011. p. 751–754.
12. Min F, Juan X. Semi-greedy heuristics for feature selection with test cost constraints. *Granul Comput.* 2016;1(3):199–211.
 13. Min F, He H, Qian Y, Zhu W. Test-cost-sensitive attribute reduction. *Inf Sci.* 2011;181(22):4928–42.
 14. Min F, Qinghua H, Zhu W. Feature selection with test cost constraint. *Int J Approx Reason.* 2014;55(1):167–79.
 15. Paclík P, Duijn RPW, van Kempen GMP, Kohlus R. On feature selection with measurement cost and grouped features. In: Joint IAPR international workshops on statistical techniques in pattern recognition (SPR) and structural and syntactic pattern recognition (SSPR), Springer; 2002. p. 461–469.
 16. Vaughan LK, Divers J, Padilla MA, Redden DT, Tiwari HK, Pomp D, Allison DB. The use of plasmodes as a supplement to simulations: a simple example evaluating individual admixture estimation methodologies. *Comput Stat Data Anal.* 2009;53(5):1755–66.
 17. Xu Z, Weinberger K, Chapelle O. The greedy miser: learning under test-time budgets. arXiv preprint. 2012. arXiv:1206.6451.
 18. Xu Z, Kusner M, Weinberger K, Chen M. Cost-sensitive tree of classifiers. *Int Conf Mach Learn.* 2013;28:133–41.
 19. Zhixiang X, Kusner MJ, Weinberger KQ, Chen M, Chapelle O. Classifier cascades and trees for minimizing feature evaluation cost. *J Mach Learn Res.* 2014;15(1):2113–44.
 20. Zhou Q, Zhou H, Li T. Cost-sensitive feature selection using random forest: selecting low-cost subsets of informative features. *Knowl Based Syst.* 2016;95:1–11.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.