# Quantitative Score for Assessing the Quality of Feature Rankings

Ivica Slavkov, Matej Petković, Dragi Kocev and Sašo Džeroski
Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, Slovenia
Jožef Stefan International Postgraduate School, Jamova cesta 39, 1000 Ljubljana, Slovenia
E-mail: saso.dzeroski@ijs.si

**Keywords:** feature ranking, feature selection, evaluation methodology, high-dimensional data

**Received:** January 18, 2018

*Feature ranking is a machine learning task that is related to estimating the relevance (importance) of individual features in a dataset. Relevance estimates can be used to induce an ordering of the features from a dataset, also called a feature ranking. In this paper, we consider the problem of the evaluation of different feature rankings. For that purpose, we propose an intuitive evaluation method, based on iterative construction of feature sets and their evaluation by learning predictive models. By plotting the obtained predictive performance of the models, we obtain error curves for each feature ranking. We then propose a scoring function to quantitatively assess the quality of the feature ranking. To evaluate the proposed method, we first define a synthetic setting in which we analyse the method and investigate its properties. By using the proposed method, we next perform an empirical comparison of several feature ranking methods on datasets from different domains. The results demonstrate that the proposed method is both appropriate and useful for comparing feature rankings of varying quality.*

*Povzetek: Rangiranje značilk je naloga strojnega učenja, povezana z ocenjevanjem pomembnosti značilk v podatkih. Značilke lahko nato uredimo glede na dobljene ocene in tako dobimo ureditev, ki ji prav tako pravimo rangiranje značilk. V tem delu obravnavamo problem evalvacije različnih metod za urejanje značilk. Predlagamo postopek, ki temelji na iterativni konstrukciji množic značilk ter njihovi evalvaciji s pomočjo napovednih modelov. Če dobljene ocene natančnosti modelov narišemo na graf, dobimo krivulje natančnosti za vsako rangiranje značilk. Te krivulje s predlaganim postopkom pretvorimo v kazalec, ki poda kakovost rangiranja številsko. Metodo najprej evalviramo na sintetičnih podatkih, nato pa jo preizkusimo še na realnih podatkih iz različnih domen. Rezultati pokažejo, da je predlagana metoda uporabna za razločevanje rangiranj značilk različnih kvalitet.*

## 1 Introduction

In many application domains, such as bioinformatics and computer vision, supervised learning methods are becoming more frequently applied to high-dimensional problems. In such problems, one typically expects only a relatively small proportion of all input features to be relevant for predicting the output. Also, all relevant features are not equally important. In many practical applications, the problem of discovering the relevant features and/or qualitatively assessing their relative importance can be the main objective of the application of machine learning techniques, even taking precedence over the need to obtain the best possible predictive model. In bioinformatics, for example, the main objective of the analysis of microarray datasets is to identify genes whose expression is, individually or jointly, indicative of some biological state of interest (e.g., a disease), with the goal of improving the understanding of this biological state.

There are two machine learning tasks related to the analysis of feature relevance, namely *feature selection* (FS) and *feature ranking* (FR) [9]. The purpose of feature selection is typically to solve the so-called minimal-optimal problem [15], i.e., to find a minimal subset of features that best explain the output [8]. Feature ranking, on the other hand, solves the so-called all-relevant problem [15], i.e., providing an ordered list of the features from the most to the least important according to a given notion of relevance. Feature ranking methods range from univariate techniques, that assess the relevance of each feature independently of the others (e.g., using mutual information or p-values derived from some statistical test), to multivariate techniques, that derive more complex feature importance scores taking into account potential interactions among the features (e.g., ReliefF [18] or Random forests [2]). These two problems of feature selection and feature ranking are linked: A solution to the feature selection problem can be found by setting a cut-off point on a feature ranking.

The present paper focuses on feature ranking, and more specifically addresses the challenging problem of the evaluation of the output of feature ranking algorithms. Feature selection as stated above is a well-defined optimization problem and as a consequence, the output of two different feature selection methods can be directly compared according to the predictive performance of a model trained from the selected features and/or according to to the size

of the selected subsets. The problem of feature ranking, on the other hand, can not be as easily formulated as an optimisation problem, mainly because there is no commonly accepted notion of feature relevance. Actually, feature ranking methods typically correspond to different definitions of relevance or assumptions of dependence (e.g., univariate versus multivariate, linear versus non-linear). As a consequence, when run on the same dataset, different methods will typically provide different rankings of the features. Determining the best ranking among several ones for a problem at hand is thus a practically very relevant question. For specific problems, this question can be addressed by using domain specific knowledge. In the general case, however, this is an unsolved problem that we would like to address in this paper.

The remainder of this paper is organized as follows. We start by discussing related work in Section 2. We then propose a new algorithmic procedure for evaluating feature rankings that does not require any prior knowledge and can thus be applied on real problems. Following previous works, this method is based on the evaluation of the predictive performance of models trained from nested feature subsets derived from the rankings (described in Section 3). More precisely, two error curves are constructed: the forward feature addition curve (FFA) and the reverse feature addition curve (RFA). They depict the performance of models built on nested feature subsets obtained by taking features from either the top or the bottom of the ranking. Next, we propose a score based on the differences between the FFA and RFA curves as a way to compare different feature ranking methods. We investigate the performance of the proposed method on a wide range of datasets. We start by experiments on the synthetic datasets (Section 4) and proceed with a description of its use on real-world benchmark datasets (Section 5). Section 6 concludes with a summary of our contributions and an outline of possible directions for further work.

## 2 Related work

The evaluation of feature ranking methods has been typically performed on artificial problems, where the relevant and irrelevant features are known by construction. In such a setting, feature ranking algorithms can be evaluated based on their capability to delineate relevant from irrelevant features. This capability can be measured, for example, through a ROC curve showing the trade-off achieved by the algorithm between assigning high ranks to relevant features and low ranks to irrelevant ones [11]. In the context of the ReliefF algorithm [18] the concepts of *separability* and *usability* are defined to evaluate feature rankings. Separability measures how well the algorithm separates the relevant from the irrelevant features by the difference between the lowest estimated relevance of the relevant features and the highest relevance of the irrelevant features. Usability, on the other hand is defined as the difference between the highest estimated relevance of the relevant features and the highest estimated relevance of the irrelevant features. When a ground truth ranking of the features is known (and not only which features are relevant/irrelevant), finer measures can be used to compare a learnt feature ranking to the ground truth, such as the Spearman's rank correlation.

Evaluating feature ranking methods on artificial problems is very useful to assess a newly proposed ranking algorithm or to highlight overall differences between methods. In practice however, the best method is expected to be problem dependent. This stresses the need for methods to quantitatively assess feature ranking methods in real-world scenarios, where it is not known a-priory which features are relevant and which are irrelevant. In such settings, the performance of feature ranking algorithms has been mostly evaluated from the point of view of their predictive performance associated to feature subsets derived from the rankings.

A way to assess feature rankings is to estimate the predictive performance obtained by using subsets of feature derived from these rankings. For example, for a given number of features $k$, a ranking $A$ could be considered better than a ranking $B$ if a model trained from the top-$k$ features of ranking $A$ is more accurate than a model trained from the top-$k$ features of ranking $B$. Variations of this evaluation procedure are given in [9, 7, 16, 21] where the predictive models are compared for different numbers of top-$k$ features.

## 3 Evaluation method for feature rankings

In general, the purpose of feature ranking algorithms is to solve the all-relevant feature selection problem [15]. However, besides delineating relevant from irrelevant features, a feature ranking algorithm should also provide a proper ordering of features according to their relevance to some target concept. An ideal feature ranking algorithm should produce the ground truth ranking. In reality however, the ranking methods provide only an approximation of it.

A good feature ranking method would produce a ranking that is well ordered. This means that the more relevant features would have a higher rank, i.e., all of them are concentrated at the beginning of the feature ranking. In contrast, a bad feature ranking method is not necessarily the one that produces an inverse ground truth ranking. Namely, we consider as a worst-case scenario if the feature ranking produces a random ranking. In this case the relevant features are uniformly distributed in the ranked list. Estimating and comparing this distribution of relevant features across a ranking is the intuition on which we base our evaluation approach.

### 3.1 Evaluation method definition

Formally, we would like to evaluate a feature ranking algorithm $r(\cdot)$. The input to the algorithm is a dataset $\mathcal{D}$, consisting of a set of $n$ input features $\mathcal{F}$ and a target $F_t$, while the output is a feature ranking $\mathbf{R} = r(\mathcal{D})$, i.e., a list whose $i$-th component gives us the rank of $i$-th feature.

For an arbitrary feature subset $\boldsymbol{S} \subseteq \mathcal{F}$, we can assess if it contains relevant features by constructing and evaluating predictive models $\mathcal{M}(\boldsymbol{S}, F_t)$. We evaluate them, obtain the value of error measure $\mathrm{err}(\mathcal{M}(\boldsymbol{S}, F_t))$, and decide whether the set $\boldsymbol{S}$ contains important features or not.

The error estimates should provide insight into the correctness of the feature ranking and constitute an evaluation thereof, thus we construct the feature subsets of two types. The sets of the first type, denoted by $\boldsymbol{S}^i$, contain the top $i$ ranked features, $1 \leq i \leq n$. The sets of the second type, denoted by $\boldsymbol{S}_i$, contain the bottom $i$ features. Note that in the special case where $i = n$, i.e., the number of features, we have $\boldsymbol{S}^n = \boldsymbol{S}_n$

For each constructed feature subset $\boldsymbol{S}$, $\boldsymbol{S} = \boldsymbol{S}^i$ or $\boldsymbol{S} = \boldsymbol{S}_i$, we build predictive models $\mathcal{M}(\boldsymbol{S}, F_t)$, and evaluate their prediction errors. In that way, we obtain two curves: the *forward feature addition* (FFA) curve consists of points $(i, \mathrm{FFA}(i)) = (i, \mathcal{M}(\boldsymbol{S}^i, F_t))$ (see Fig. 1a), while the *reverse feature addition* (RFA) curve consists of points $(i, \mathrm{RFA}(i)) = (i, \mathcal{M}(\boldsymbol{S}_i, F_t))$ (see Fig. 1b).

In practical scenarios, if the number of features $n$ is high, running the algorithm might be costly. One simple way for a speed up would be to avoid forming all the feature subsets, and instead add $\Delta i > 1$ features to the set $\boldsymbol{S}^i$ to obtain $\boldsymbol{S}^{i+\Delta i}$. The rationale behind this is that in real-world scenarios involving high-dimensional data, only a small portion of the features are relevant. Therefore, the values of $\mathrm{FFA}(i)$ would not change much when adding more features to a relatively large number of features $i$. Also, the number of features added can be dependent on $i$. In the same manner, we can form the set $\boldsymbol{S}_{i+\Delta i}$ from the set $\boldsymbol{S}_i$.

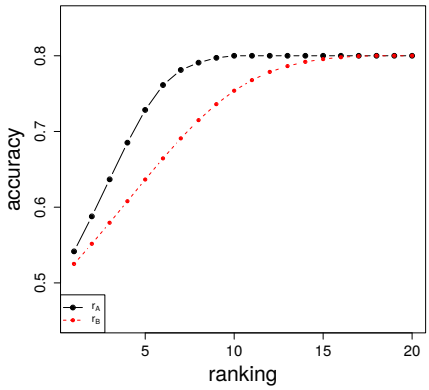
### 3.2 Quantitative comparison of two rankings

A visual inspection of the curves can only provide a qualitative intuition about which ranking method is better. For quantification purposes, it would be necessary to have a function which provides a cumulative assessment of the differences between the error estimates at different points of the curves. In the most general case, this would be an aggregation function $\mathrm{agg} : \mathbb{R}^n \to \mathbb{R}$, which would take a sequence of the weighted point-wise differences between two curves for its argument. For the FFA curve, we would have

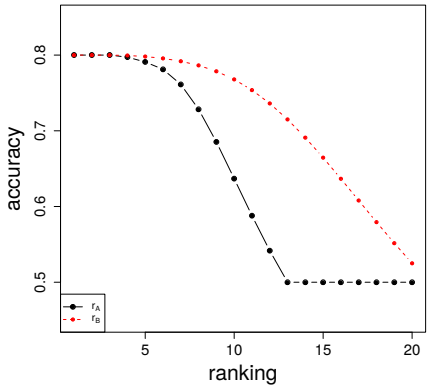
$$\mathrm{FFA}_\delta(r_A, r_B) = \underset{i}{\mathrm{agg}}\ w_i(\mathrm{FFA}_{r_A}(i) - \mathrm{FFA}_{r_B}(i)), \quad (1)$$

while for the RFA curve we would have

$$\mathrm{RFA}_\delta(r_A, r_B) = \underset{i}{\mathrm{agg}}\, w_i(\mathrm{RFA}_{r_A}(i) - \mathrm{RFA}_{r_B}(i)). \quad (2)$$




(a) Comparison of FFA curves




(b) Comparison of RFA curves

Figure 1: Comparison of different ranking methods $r_A$ and $r_B$

There are several sensible choices for instantiations of the aggregation function agg. The choice depends on the specific task at hand. Considering that we are comparing feature rankings, two aspects are important. The first is the position of most of the relevant features in the ranking. The second relates to the position of the "most" relevant features. In a comparative sense, the first aspect relates to the position of the FFA/RFA curves differences, while the second relates to the magnitude of these differences.

Differences between the FFA/RFA curves of two ranking methods at the beginning of the curves are more important than differences at the end of the curves. Namely, if two FFA curves are different at the beginning, this means that one of the ranking methods is not putting the most relevant features at the top of the ranking. Correspondingly, for the RFA curves, differences at the beginning of the curve (at the bottom of the ranking), mean that one of the feature ranking methods is giving low ranks to features which are relevant. The second aspect is related more to the magnitude of the differences between the FFA/RFA curves than to their position. The intuition is that if "more" relevant features are misranked, then this is worse than "fewer" relevant features being misranked.

From a technical perspective, in order to emphasise the importance of position, the weighting function from

Eqs. 1 and 2, should be a function of the position, $i$, namely $w_i = f(i)$. In the same manner, in order to emphasise the importance of magnitude, the weighting function should depend on the size of the difference, namely $w_i = f(\delta_i)$ with $\delta_i$ the difference between the two compared curves at $i$. In addition, it is also possible to construct a weighting function that takes into account both position and magnitude, $w_i = f(i, \delta_i)$. To this end, we define four instantiations of Eq. 1 and Eq. 2, which we use to calculate the difference between the FFA/RFA curves from Fig. 1. We consider the following weighting functions:

- $w_i = 1$, equal weight for all differences;
- $w_i = f(i) = 1/|\boldsymbol{S}_i|$, weight inverse to feature subset size;
- $w_i = f(\delta_i) = |\delta_i|$, weight proportional to the difference magnitude;
- $w_i = f(i, \delta_i) = |\delta_i|/|\boldsymbol{S}_i|$, weight which includes both position and magnitude.

The aggregation function used for summarising the differences (in all of the four instantiations) is the weighted average:

$$\underset{i}{\text{agg}}\, w_i\delta_i = \frac{\sum_{i=1}^{n} w_i\delta_i}{\sum_{i=1}^{n} w_i}. \tag{3}$$

The obtained values are given in Table 1. They are calculated for the FFA/RFA examples in Fig. 1a and Fig. 1b. The difference is calculated for $r_A$ with respect to $r_B$. As seen in Table 1, the values for the FFA curves are positive, which can be interpreted as "$r_A$ is better than $r_B$". While the values for the RFA curves are negative, the interpretation is the same: "ranking method $r_A$ is better than ranking method $r_B$".

In order to obtain a single number that quantifies the difference between two feature ranking algorithms, we can combine both values into a single value by calculating the so-called error curve average (ECA)

$$\text{ECA}_\delta(r_A, r_B) = \frac{\text{FFA}_\delta(r_A, r_B) - \text{RFA}_\delta(r_A, r_B)}{2}. \tag{4}$$

Note that the minus sign in the equation is due to the inverse interpretation of negative values for the RFA curve. Namely, if $r_A$ is better than $r_B$, then the differences of the RFA curves should be negative. This places the overall interpretation of the $ECA_\delta$ value on the positive scale. Namely, if $r_A$ is better than $r_B$, then the overall score should be positive.

| $w_i$ | 1 | $1/\lvert\boldsymbol{S}_i\rvert$ | $\lvert\delta_i\rvert$ | $\lvert\delta_i\rvert/\lvert\boldsymbol{S}_i\rvert$ |
|---|---|---|---|---|
| $\text{FFA}_\delta$ | 0.018 | 0.019 | 0.032 | 0.03 |
| $\text{RFA}_\delta$ | -0.042 | -0.054 | -0.08 | -0.077 |

Table 1: Different quantitative comparisons of error curves

### 3.3 Quantitative score for a single ranking

In real-world scenarios, the ground truth ranking is not known. Therefore, when evaluating just a single ranking algorithm, the FFA/RFA curve of the algorithm can not be compared to the one of the ground truth ranking. However, the opposite to the ground truth ranking is the uniformly random ranking, for it is the least informative. The motivation for introducing the random ranking FFA/RFA curves is the following: If we can not say how good a single ranking $\mathbf{R}$ is, maybe we can say how close to random it is.

At the point $i$, the expected value of the FFA/RFA curve, which belong to the uniformly random ranking $\mathbf{R}_{\text{RND}}$, produced by the algorithm $r_{\mathcal{R}}$, is dependent solely on the $i$ and properties of the dataset under consideration. Moreover, it is the same for both the FFA and the RFA curve. For simplicity reasons, we refer to these curves as *expected curves*.

The expected value of the error measure err, is the average of the error estimations of all possible subsets $\boldsymbol{S} \subseteq \mathcal{F}$, whose cardinality equals $i$, i.e.,

$$E[\text{err}(\mathcal{M}(\boldsymbol{S}, F_t))] = \frac{1}{\binom{n}{i}} \sum_{\substack{\boldsymbol{S}\subseteq\mathcal{F} \\ |\boldsymbol{S}|=i}} \text{err}(\mathcal{M}(\boldsymbol{S}, F_t)) \tag{5}$$

Calculating the expected curves by following Eq. 5 to the letter is intractable, especially for high-dimensional spaces, as we have to consider an exponentially high number of feature subsets. However, for practical purposes, this problem can be circumvented by sampling the space of possible feature subsets for each $i$.

Suppose we have somehow calculated or approximated the expected FFA/RFA curve. If we have a ranking algorithm $r$ that produces a good (mostly correct) ranking, its FFA curve would be above the expected FFA curve. For the RFA curve, the opposite would apply and the algorithm's curve would be below the expected RFA curve. The score $\text{ECA}_\delta(r, r_{\mathcal{R}})$ between the FFA/RFA curves of this ranking versus the expected curves can thus be used as an absolute quantitative measure of the quality of this ranking. It should be noted that when calculating $\text{ECA}_\delta(r, r_{\mathcal{R}})$ by using $w_i = 1$, it is not necessary to compute the expected curve in order to calculate this $ECA_\delta$ score. Indeed, $\text{ECA}_\delta$ can be simply computed as the sum over all positions of the difference between the FFA and RFA curves we want to evaluate:

$$\begin{aligned} \text{ECA}_\delta(r, r_{\mathcal{R}}) &= \frac{(\text{FFA}_\delta(r, r_{\mathcal{R}}) - \text{RFA}_\delta(r, r_{\mathcal{R}}))}{2} \\ &= \frac{1}{2}\left(\sum_{i=1}^{n} \frac{\text{FFA}_r(i) - \text{RFA}_r(i)}{n}\right), \end{aligned}$$

since $\text{FFA}_{r_{\mathcal{R}}}(i) = \text{RFA}_{r_{\mathcal{R}}}(i)$.

## 4 Evaluation on synthetic data

The goal of the experiments presented in this section is to demonstrate the usefulness of our feature ranking evaluation method. As previously mentioned, feature ranking

methods provide an approximation of the ground truth ranking that can be viewed as a noisy ground truth. A noisier ranking is more distant from the ground truth ranking and therefore of worse quality.

An evaluation method should be sensitive to the amount of noise and should provide a corresponding quality estimate of the feature ranking. For that purpose, we design experiments to demonstrate that our evaluation method is sensitive to the addition of noise to the ground truth ranking. We first generate noisy feature rankings and then construct FFA/RFA curves from them.

## 4.1 Generating synthetic data

We first perform an empirical evaluation of the proposed notion of FFA/RFA curves in a controlled setting by using synthetic datasets. The main advantage of using synthetic data is the possibility of defining a good baseline ranking that allows us to assess our proposed feature ranking evaluation method.

The complete statistics of the generated datasets and their feature interaction sets are summarized in Table 2. All of the datasets consisted of 1000 instances and 100 features in total. Among the 100 features, the "single" dataset has 9 relevant features, the "pair" dataset contains 18 relevant features and the "combined" dataset contains 27 relevant features. In all three datasets, every set $\mathcal{F}_{\text{int}}$ of relevant features has two additional redundant copies. Irrelevant features are realized independently of each other. More details on the generation of the datasets are available in [20].

For each dataset, we would like to define a good baseline ranking against which to compare feature ranking methods. We define this ranking from feature relevance scores $\text{rel}(F_i, F_t)$ for each feature $F_i$, calculated directly from the specified feature interaction structure, by using the following equation:

$$\text{rel}(\text{F}_\text{i}, \text{F}_\text{t}) = \frac{\text{I}(\mathcal{F}_{\text{int}}; F_t)}{|\mathcal{F}_{\text{int}}|},$$

where $\mathcal{F}_{\text{int}}$ is the (unique) interaction set that contains $F_i$ and $I(\mathcal{F}_{\text{int}}; F_t)$ is the mutual information between features in $\mathcal{F}_{\text{int}}$ and the target $F_t$. By dividing the mutual information by the number of features, we distribute the information equally between all features in an interaction set. As a consequence, features that brings information about the target $F_t$ individually are considered more relevant than features that bring the same amount of information about the target only in conjunction with other features.

Note that this baseline ranking is not guaranteed to be optimal in terms of the FFA and RFA curves for a given learning algorithm, but is nevertheless expected to be close to optimal. In our experiments, we will consider this ranking as a ground truth ranking, denoted $R_{\text{GT}}$, against which we will compare other rankings.

| n | $\lvert\mathcal{F}_{\text{int}}\rvert$ | $f(\mathcal{F}_{\text{int}})$ | $P$ |
|---|---|---|---|
| 3 | 1 | $F_i$ | 0.8 |
| 3 | 1 | $F_i$ | 0.7 |
| 3 | 1 | $F_i$ | 0.6 |
| 91 | 1 | $F_i$ | 0.5 |

(a) "single" dataset

| n | $\lvert\mathcal{F}_{\text{int}}\rvert$ | $f(\mathcal{F}_{\text{int}})$ | $P$ |
|---|---|---|---|
| 3 | 2 | $\text{XOR}(F_i, F_j)$ | 0.8 |
| 3 | 2 | $\text{XOR}(F_i, F_j)$ | 0.7 |
| 3 | 2 | $\text{XOR}(F_i, F_j)$ | 0.6 |
| 82 | 1 | $F_i$ | 0.5 |

(b) "pair" dataset

| n | $\lvert\mathcal{F}_{\text{int}}\rvert$ | $f(\mathcal{F}_{\text{int}})$ | $P$ |
|---|---|---|---|
| 3 | 2 | $\text{XOR}(F_i, F_j)$ | 0.8 |
| 3 | 2 | $\text{XOR}(F_i, F_j)$ | 0.7 |
| 3 | 2 | $\text{XOR}(F_i, F_j)$ | 0.6 |
| 3 | 1 | $F_i$ | 0.8 |
| 3 | 1 | $F_i$ | 0.7 |
| 3 | 1 | $F_i$ | 0.6 |
| 73 | 1 | $F_i$ | 0.5 |

(c) "combined" dataset

Table 2: Synthetic datasets statistics: The feature interaction sets ($\mathcal{F}_{\text{int}}$) contained in each dataset; The interaction function for the feature sets ($f(\mathcal{F}_{\text{int}})$); The values $P(f(\mathcal{F}_{\text{int}}) = F_t)$ are denoted by $P$. The value of $n$ in the last row of each table corresponds to the number of irrelevant features in a dataset. In the other rows, $n$ denotes the number of copies of each interaction set, which are identically defined but independently realized (and differ in the random component).

## 4.2 Adding noise to the ground truth ranking

The noise is introduced into the ranking by selecting a proportion, $\theta$, of the features, which are randomly selected. For these features, random relevance values are assigned while the remaining features preserve their ground truth relevance. By considering these partially changed relevance values, a new noisy feature ranking, $\mathbf{R}_\theta$, is defined.

As the random relevance values can be distributed differently throughout the ranking, different FFA/RFA curves can be constructed for the same amount of noise.

We estimate the expected error values by sampling the space of possible FFA/RFA curves for a given $\theta$. First, we generate $N$ different noisy feature rankings and then construct $N$ FFA/RFA curves based on them. The expected values of FFA/RFA curve are estimated by averaging the

$N$ individual curves, namely

$$E[\text{FFA}]_\theta = \frac{1}{N}\sum_{i=1}^{N}\text{FFA}_{\theta,i}$$

$$E[\text{RFA}]_\theta = \frac{1}{N}\sum_{i=1}^{N}\text{RFA}_{\theta,i}$$

for a specified $N$ and $\theta$.

For estimating the error values of the FFA/RFA curves, SVMs with a polynomial (quadratic) kernel were used and a 10-fold cross validation was performed on the dataset under consideration. The epsilon parameter of the SVMs was set to 1.0E-12, while the complexity parameter was set to 0.1.

For our experiments, we consider several different amounts of noise $\theta$, namely: 5%, 10%, 15%, 20%, 30% and 50%, as well as the completely random ranking (100% of noise). Each noisy error curve was produced by summarizing the errors of 100 noisy rankings produced for a given $\theta$. We additionally constructed error curves based on the ground truth ranking.

The experiments were performed on the three synthetic datasets described in Section 4.1, each with its corresponding ground truth ranking, $\mathbf{R}_{GT}$.

## 4.3 Results on synthetic data

The results of our experiments are first plotted as graphs containing error curves. In Fig. 2, we only show the curves obtained on the "combined" dataset. These curves are representative of the curves obtained on the other datasets.

The FFA/RFA curves plotted on each graph are for rankings with different noise levels $\theta$, as well as for the ground truth $\mathbf{R}_{GT}$ and random rankings. In both Figs. 2a and 2b, the FFA and the RFA curves seem to be sensitive to the addition of noise. To begin with, the FFA/RFA curves of all the noisy rankings are located between the ground truth ranking FFA/RFA curve and the random ranking FFA/RFA curve. As noise is added to the ground truth ranking, the FFA/RFA estimates are slowly moving away from the ground truth FFA/RFA curve and towards the random ranking FFA/RFA curve.

Next, for performing quantitative analysis of the feature rankings, we begin by summarising the differences of the noisy rankings error curves w.r.t. the ground truth error curve. Additionally, some kind of baseline is required for comparison. As the ground truth ranking is known, the distance between the ground truth ranking and the noisy rankings can serve as a baseline.

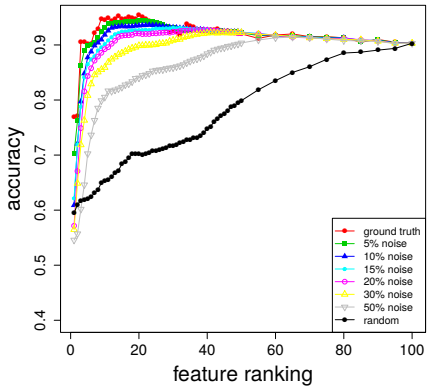
For summarising the differences between the noisy rankings FFA/RFA curves we use the ECA difference, calculated by using Eq. 4 from Section 3.2. For comparative purposes,when calculating the ECA differences, we use the different weighting functions as discussed in Section 3.2.

For calculating the baseline values, i.e., the distance between the ground truth ranking $\mathbf{R}_{GT}$ and the noisy rankings $\mathbf{R}_{\theta,i}$, we use the average Spearman rank correlation coefficient $\rho$ between the vectors $\mathbf{R}_{GT}$ and $\mathbf{R}_{\theta,i}$. The $i$-th component of such a vector gives the rank of the $i$-th feature in dataset. The distance between rankings is then computed as

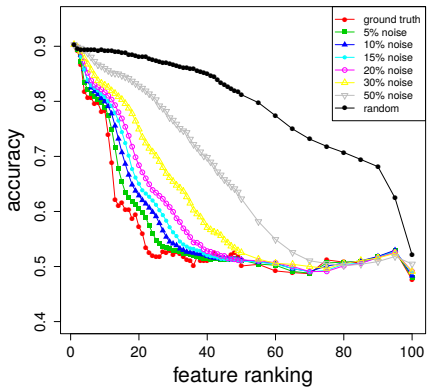
$$\text{dist}(\mathbf{R}_{GT},\mathbf{R}_\theta) = 1 - \bar{\rho}_{GT,\theta} = 1 - \frac{1}{N}\sum_{i=1}^{N}\rho(\mathbf{R}_{GT},\mathbf{R}_{\theta,i})$$

where $N$ is the number of different noisy rankings considered for a given $\theta$.

We obtain the results for all of the three synthetic datasets. Since there are no major differences among them, we show summarised results only for the "combined" dataset. Table 3 contains values calculated with respect to the ground truth ranking. The first row of the table refers to the distance $\text{dist}(\mathbf{R}_{GT},\mathbf{R}_\theta)$. The other rows are the ECA differences between the FFA/RFA curves of the GT ranking and the FFA/RFA curves of the noisy rankings. Each row containing the ECA differences refers to different weighting functions. All columns, except the last one, refer to different levels of noise, $\theta$. The final column gives the correlation between $\text{dist}(\mathbf{R}_{GT},\mathbf{R}_\theta)$ (row one) and the FFA/RFA curve distances (rows 2 to 5), across different




(a) FFA curves for the "combined" dataset




(b) RFA curves for the "combined" dataset

Figure 2: Plots comparing the FFA (on the left) and RFA (on the right) curves for the "combined" dataset. Each figure contains error curves for the ground truth ranking, rankings with different noise levels $\theta$ and the random ranking.

| | $\theta = 0.05$ | $\theta = 0.1$ | $\theta = 0.15$ | $\theta = 0.2$ | $\theta = 0.3$ | $\theta = 0.5$ | $\theta = 1$ | |
|---|---|---|---|---|---|---|---|---|
| dist | 0.1 | 0.171 | 0.252 | 0.32 | 0.432 | 0.652 | 1.048 | **corr.** |
| $w = 1$ | 0.009 | 0.02 | 0.027 | 0.037 | 0.061 | 0.117 | 0.223 | 0.992 |
| $w = 1/r$ | 0.018 | 0.042 | 0.047 | 0.064 | 0.084 | 0.132 | 0.178 | 0.991 |
| $w = \|\delta\|$ | 0.029 | 0.061 | 0.070 | 0.09 | 0.115 | 0.174 | 0.263 | 0.998 |
| $w = \|\delta\|$ /r | 0.044 | 0.091 | 0.095 | 0.121 | 0.142 | 0.199 | 0.254 | 0.982 |

Table 3: Comparison of different ECA values obtained by different weighting functions $w$. The ECA values are compared with the distance between the noisy rankings $R_\theta$ and the GT ranking $R_{GT}$. The final column of each table "corr." is the value of the correlation coefficient calculated between the ranking distance (first row) and each of the ECA difference rows.

noise levels $\theta$.

The final column gives an indication of how well the ECA differences relate to the distance between the ground truth ranking and the noisy rankings. As it can be seen, the curve distances correlate very well to the rank distances, regardless of which weighting function is used.

From this quantitative analysis, it can be concluded that the ECA difference derived from the error curves has the same sensitivity to noise as the actual distance between the ground truth and the noisy rankings. This implies that our method can be used in practical scenarios not just to qualitatively distinguish between different rankings, but also to quantify the difference between them. As for the specific weights used for calculating the ECA differences, it can be concluded that any of the considered weighting schemes can be used to properly compare the error curves.

# 5 Evaluation on real data

Thus far, our analysis only involved artificially generated problems. In this section, we want to illustrate the use of our feature ranking evaluation method on datasets originating from various real-life domains. The purpose of the experiments is to examine the quality of the feature rankings produced by several feature ranking methods on data with different characteristics.

The analysis is primarily a comparative one, performed solely by calculating the numeric scores derived from the FFA and RFA error curves. The datasets we consider are quite diverse, with unknown interaction structure and therefore unknown ground truth ranking. However, for each dataset, it is possible to generate the expected error curves of random rankings. These expected curves are used as a baseline for comparing the different feature ranking methods.

## 5.1 Datasets description

For our experiments, 28 diverse classification datasets with a single target class were selected. Most of them originate from the UCI data repository [14] and are from various domains. Of the remaining 3 datasets, one is from a medical study of acute abdominal pain in children (aapc) [4], while the remaining two ("water" and "diversity") are from an ecological study of river water quality [5].

Besides covering different domains (including biology, medicine, ecology etc.) these datasets have a wide range of different properties, including number/type of features and number of instances.

The main characteristics of each dataset are summarised in Table 4.

| Dataset | #Inst. | #Feat. | #Cl. |
|---|---|---|---|
| aapc | 335 | 84 | 3 |
| amlPrognosis | 54 | 12625 | 2 |
| arrhythmia | 452 | 279 | 16 |
| australian | 690 | 14 | 2 |
| bladderCancer | 40 | 5724 | 3 |
| breast-cancer | 286 | 9 | 2 |
| breast-w | 699 | 9 | 2 |
| breastCancer | 24 | 12625 | 2 |
| car | 1728 | 6 | 4 |
| chess | 3196 | 36 | 2 |
| childhoodAll | 110 | 8280 | 2 |
| cmlTreatment | 28 | 12625 | 2 |
| colon | 62 | 2000 | 2 |
| diversity | 292 | 86 | 5 |
| dlbcl | 77 | 7070 | 2 |
| german | 1000 | 20 | 2 |
| heart | 270 | 13 | 2 |
| heart-c | 303 | 13 | 2 |
| heart-h | 294 | 13 | 2 |
| ionosphere | 351 | 34 | 2 |
| leukemia | 72 | 5147 | 2 |
| mll | 72 | 12533 | 3 |
| prostate | 102 | 12533 | 2 |
| sonar | 208 | 60 | 2 |
| srbct | 83 | 2308 | 4 |
| tic-tac-toe | 958 | 9 | 2 |
| water | 292 | 80 | 5 |
| waveform | 5000 | 21 | 3 |

Table 4: Statistics for the benchmark datasets

## 5.2 Experimental setup

Four feature ranking methods were applied to each dataset:

- **Information gain**, calculating the information gain of each feature $F_i$ as $IG(F_t, F_i) = H(F_t) - H(F_t|F_i)$. This does not require any specific parameter setting.
- **SVM-RFE** is the recursive feature elimination (RFE) procedure that employs an SVM to evaluate the feature weights at each iteration. A linear SVM was employed [9] with the epsilon parameter set to 1.0E-12, while the complexity parameter was set to 0.1.
- **ReliefF** algorithm as proposed in [18]. The number of neighbours was set to 10 and all of the instances were used for estimating the relevance values.

– **Random forests**, which can be used for estimating feature relevance as described in [2]. A forest of 100 trees was used, constructed by randomly choosing a $\log_2$ of the number of features.

To generate the error curves, SVMs with polynomial (quadratic) kernel, were employed as classifiers. The epsilon parameter was set to 1.0E-12, while the complexity parameter was set to 0.1. This classifier was showed to be appropriate in our previous experiments [20].

As a baseline of the comparison, *expected* FFA and RFA *curves* were used. They were produced by generating 100 random rankings for each dataset under consideration. This was done in a similar manner as described in Section 3.3.

## 5.3 Results on real data

The results summarizing the error curves average (ECA) differences are given in Table 5. The ECA differences are calculated by using Eq. 4 and the weighting function $w_i = 1$, i.e., as a standard mean value. Each row of Table 5 refers to a single dataset, while each column corresponds to a single feature ranking method. The ECA values in the table are calculated w.r.t. the baseline error curve, namely, the expected error curve. This gives an indication of how much each feature ranking method is better than a random ranking generator, but also allows for comparison between the quality of the feature rankings of the different methods.

A positive value of an ECA difference indicates that a feature ranking method performs better than the random ranking generator. The negative values, however, do not necessarily indicate that it performs worse than random, but that it provides a non-random ranking that is inverse to the correct one. A value close to zero means the feature ranking method provides rankings that are more (or less) random.

An intial inspection of the results in Table 5 reveals that random forests often have negative ECA values. The FFA and RFA curves of random forests, for these particular datasets, are below/over the expected FFA/RFA curves of random rankings. Upon closer inspection of their feature rankings (results not shown here due to space limitations) we find that they are inverse to those of the other feature ranking methods.

In order to summarise the results from Table 5 and to draw meaningful conclusions about the performance of the different ranking methods across the different datasets, we use statistical tests. We adopt the recommendations of Demšar [3] and use the Friedman [6] test for statistical significance with the correction by Iman [10]. If the null hypothesis $H_0$ that all ranking methods perform equally well, can be rejected, we use the Nemenyi post-hoc test [13] and additionally check between which feature ranking methods the statistically significant differences appear. The level of significance $p = 0.05$ was used.

When comparing the four feature ranking methods, statistically significant differences occur. We present the results with a critical distance diagram [3] in Fig. 3. In the diagram, the feature ranking methods are ordered according to which one is better on average (across all datasets). A method is better if it is positioned closer to the value one on the axis. It can be observed that ReliefF, Info Gain and SVM-RFE significantly outperform Random Forests, while not differing significantly among each other.

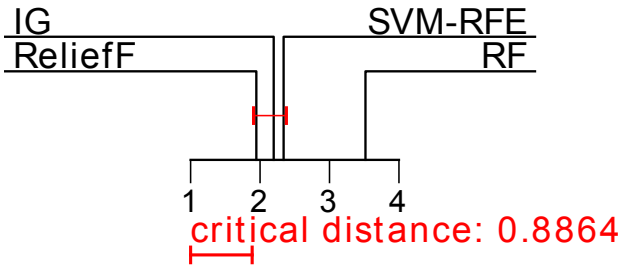



Figure 3: Critical distance diagrams representing the statistical comparison of the ECA differences of three ranking methods on the 28 datasets. The critical distance is calculated for a $p$ value of 0.05 and is represented by a horizontal line. If the feature ranking methods are connected by a line, then their performance is not statistically significantly different.

| dataset | IG | RF | ReliefF | SVM-RFE |
|---|---|---|---|---|
| aapc | 0.269 | 0.299 | **0.316** | 0.297 |
| amlPrognosis | **0.056** | 0.007 | 0.027 | 0.043 |
| arrhythmia | 0.041 | 0.041 | **0.057** | 0.053 |
| australian | **0.277** | 0.260 | 0.266 | 0.209 |
| bladderCancer | 0.125 | 0.059 | **0.167** | 0.161 |
| breast-cancer | **0.025** | 0.013 | 0.012 | −0.003 |
| breast-w | **0.246** | 0.206 | 0.190 | 0.194 |
| breastCancer | 0.050 | 0.037 | **0.128** | 0.110 |
| car | **0.085** | −0.081 | 0.079 | 0.066 |
| chess | 0.279 | −0.056 | **0.283** | 0.248 |
| childhoodAll | 0.083 | 0.040 | 0.033 | **0.154** |
| cmlTreatment | **0.028** | −0.009 | −0.026 | 0.004 |
| colon | 0.099 | 0.049 | **0.163** | 0.116 |
| diversity | 0.167 | 0.192 | **0.215** | 0.149 |
| dlbcl | 0.032 | 0.008 | 0.067 | **0.086** |
| german | **0.023** | −0.002 | 0.013 | 0.022 |
| heart | **0.159** | 0.039 | 0.150 | 0.130 |
| heart-c | **0.178** | 0.057 | 0.163 | 0.163 |
| heart-h | 0.146 | 0.058 | 0.110 | **0.147** |
| ionosphere | 0.116 | 0.088 | 0.041 | **0.136** |
| leukemia | 0.140 | 0.056 | **0.175** | 0.164 |
| mll | 0.118 | 0.045 | **0.355** | 0.281 |
| prostate | 0.212 | 0.067 | **0.236** | 0.232 |
| sonar | 0.066 | 0.060 | **0.096** | 0.070 |
| srbct | 0.142 | 0.084 | **0.292** | 0.261 |
| tic-tac-toe | 0.072 | −0.052 | **0.082** | 0.069 |
| water | 0.193 | 0.181 | **0.217** | 0.144 |
| waveform | 0.180 | −0.190 | 0.188 | **0.210** |

Table 5: ECA differences between the FFA/RFA curves of four feature ranking methods w.r.t. the curves of a random ranking. The missing values are due to SVM-RFE's inability to handle multi-valued discrete/nominal attributes. Boldfaced values are the largest ECA differences in each row.

# 6 Conclusions

In this paper, we focus on the problem of evaluating the output of feature ranking algorithms. We define and formalize an intuitive evaluation method for quantitative comparison of feature rankings. The method is based on iterative construction and evaluation of predictive models, resulting in so-called error curves: forward feature addition curve (FFA), starting from the top of a feature ranking, and the reverse feature addition curve (RFA), starting from the bottom of a ranking. From these two curves, we calculate the error curves average (ECA) difference that we propose as a numerical indicator for comparing different feature rankings.

We first test our method in a controlled environment on synthetic data. We compare feature rankings with different amount of added noise, starting from the known ground truth ranking and ending with completely random rankings. By comparing the different ECA values obtained for the different noise levels, we show that our method is sensitive to changes in the quality of the feature ranking.

In order to demonstrate the practical application of our evaluation method, we consider a collection of classification datasets from various domains with different properties. We compare the performance of four feature ranking methods across these different datasets and evaluate their outputs by using our proposed method. The analysis of the comparative evaluation shows that the best algorithm is often domain dependent and often simple approaches such as info gain can be used to produce a proper feature ranking.

Several directions of work can be taken to further develop the proposed evaluation methodology. The first is to directly use the feature relevance values produced by the ranking algorithm when inducing predictive models. This can be easily done in feature-weighted classifiers, such as weighted kNN. The second concerns feature ranking stability, another important aspect of the feature ranking process. Although we have not considered it explicitly in this work, we would like to include it in the feature ranking evaluation process, in a manner similar to that of [19]. Also, as structured data [1] are becoming increasingly common, we would like to adapt and investigate our method for different types of structured targets. To this end, we need to use a feature ranking method for structured targets and couple it with a predictive model for structured outputs [12, 17].

## Acknowledgements

IS would like to gratefully acknowledge the financial support of The Ad Futura Slovene Human Resources Development and Scholarship Fund. SD, DK, and MP have been supported by the Slovenian Research Agency through the program P2-0103, the project L2-7509, and a young researcher grant, respectively. The work has also been supported by the European Commission through the H2020 grant number 720270 (HBP SGA1).

# References

[1] Gökhan H. Bakır, Thomas Hofmann, Bernhard Schölkopf, Alexander J. Smola, Ben Taskar, and S. V. N. Vishwanathan, editors. *Predicting Structured Data*. The MIT Press, Cambridge, Massachusetts, 2007.

[2] Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

[3] Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

[4] Saso Džeroski, George Potamias, Vassilis Moustakis, and Giorgos Charissis. Automated revision of expert rules for treating acute abdominal pain in children. In *Artificial intelligence in medecine - AIME, LNCS 1211*, pages 98–109, 1997.

[5] Sašo Džeroski, Damjan Demšar, and Jasna Grbović. Predicting chemical parameters of river water quality from bioindicator data. *Applied Intelligence*, 13:7–17, 2000.

[6] Milton Friedman. A comparison of alternative tests of significance for the problem of m rankings. *Annals of Mathematical Statistics*, 11:86–92, 1940.

[7] Cesare Furlanello, Maria Serafini, Stefano Merler, and Giuseppe Jurman. Entropy-based gene ranking without selection bias for the predictive classification of microarray data. *BMC Bioinformatics*, 4:54, 2003.

[8] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, March 2003.

[9] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46:389–422, March 2002.

[10] Ronald Iman and James Davenport. Approximations of the critical region of the Friedman statistic. *Communications in Statistics - Theory and Methods*, 9:571–595, 1980.

[11] Kees Jong, Jérémie Mary, Antoine Cornuéjols, Elena Marchiori, and Michèle Sebag. Ensemble feature ranking. In *PKDD - LNCS 2302*, pages 267–278, 2004.

[12] Dragi Kocev, Ivica Slavkov, and Sašo Džeroski. More is better: ranking with multiple targets for biomarker discovery. In *Proc. Second International Workshop on Machine Learning in Systems Biology*, page 133, University of Liege, Belgium, 2008.

[13] Peter Bjorn Nemenyi. *Distribution-free Multiple Comparisons*. PhD thesis, Princeton University, Princeton, NY, USA, 1963.

[14] C.L. Blake D.J. Newman and C.J. Merz. UCI repository of machine learning databases, 1998. https://archive.ics.uci.edu/ml/datasets.html. Accessed on: 2015-12-13.

[15] Roland Nilsson, José M. Peña, Johan Björkegren, and Jesper Tegnér. Consistent feature selection for pattern recognition in polynomial time. *Journal of Machine Learning Research*, 8:589–612, December 2007.

[16] Silvano Paoli, Giuseppe Jurman, Davide Albanese, Stefano Merler, and Cesare Furlanello. Semisupervised profiling of gene expressions and clinical data. In *Proc. Sixth International Conference on Fuzzy Logic and Applications*, pages 284–289, 2005.

[17] Matej Petković, Sašo Džeroski, and Dragi Kocev. Feature ranking for multi-target regression with tree ensemble methods. In *Discovery Science*, pages 171–185, 2017.

[18] Marko Robnik-Šikonja and Igor Kononenko. Theoretical and empirical analysis of ReliefF and RReliefF. *Machine Learning*, 53:23–69, 2003.

[19] Yvan Saeys, Thomas Abeel, and Yves Van de Peer. Robust feature selection using ensemble feature selection techniques. In *ECML/PKDD, LNCS 5212*, pages 313–325, 2008.

[20] Ivica Slavkov. *An Evaluation Method for Feature Rankings*. PhD thesis, IPS Jožef Stefan, Ljubljana, Slovenia, 2012.

[21] Antanas Verikas, Adas Gelzinis, and Marija Bacauskiene. Mining data with random forests: A survey and results of new tests. *Pattern Recognition*, 44:330–349, 2011.