Research Highlights

- Reading grid of six properties for regression filter criteria analysis.
- Analysis of a nonlinear adjusted R^2 criterion for feature selection in regression.
- Comparison of three relevance criteria for regression problems.

Pattern Recognition Letters
journal homepage: www.elsevier.com

Reading grid for feature selection relevance criteria in regression

Alexandra Degeest[a,b], Benoît Frénay[c], Michel Verleysen[b]

[a] *Haute-Ecole Bruxelles Brabant - ISIB, 150 rue Royale, 1000 Brussels, Belgium*
[b] *UCLouvain - Machine Learning Group - ICTEAM, Place du Levant 3, 1348 Louvain-La-Neuve, Belgium*
[c] *Université de Namur - Faculty of Computer Science - NADI Institute, Rue Grandgagnage 21, 5000 Namur, Belgium*

ABSTRACT

Feature selection is an important preprocessing step in machine learning. It helps to better understand the importance of some features and to reduce the dimensionality of a dataset, which improves machine learning and information extraction. Among the different existing methods for selecting features, filters are popular because they are independent from the model, which will be learnt afterwards, and computationally efficient. The efficiency of filter methods relies on a strategic choice: the choice of the relevance criterion. Many criteria exist; they exhibit various properties, which in turn result in selecting different features. The choice of the criterion is thus important and should ideally be linked to the properties of the data and to users' goals. This paper shows that six properties should be analysed when selecting a relevance criterion in the context of regression problems. It proposes a reading grid to analyse relevance criteria and to make a well-guided choice.

1. Introduction

High-dimensional data are ubiquitous in regression problems. In high-dimensional datasets, some features may be not relevant or redundant to the considered regression problem and should be ignored. In contrast, other features are strategic; focusing on the latter helps to model data and extract information. Feature selection is therefore an important preprocessing step in machine learning; it does not only intend to reduce the dimension of the data but also improves interpretability and reduces computational costs by selecting which features, or combinations of features, are really of interest.

Among the different categories of feature selection methods, filters are known to be faster than wrappers and embedded methods. But filter methods need a criterion to measure the relevance of features for the problem at hand. The relevance criterion is therefore the key ingredient of a good feature selection process.

Many questions may influence the choice of the most appropriate criterion for the problem at hand: does the criterion need to be multivariate? Nonlinear? How does it scale with large datasets? Is it robust enough with small datasets? What about the performance of its estimator? This paper focuses on these strategic questions to choose the best relevance criterion for a dataset in the context of regression problems, with real valued input variables. It aims at providing a reading grid detailing the key properties of relevance criteria, helping the user for choosing the most adequate one to be used with a specific dataset.

The following of this paper is structured as follows. Section 2 describes feature selection and the problem statement. Section 3 explores related works about filter feature selection and relevance criteria; while these works are important for introducing and comparing feature selection algorithms, they lack an in-depth study of their differences and of the properties of the relevance criteria. Section 4 presents six important properties to be carefully analysed before choosing a relevance criterion and its estimator. To illustrate and to show the importance of these six properties, Section 6 analyses three relevance criteria, previously introduced in Section 5. The summary of this analysis and a comparison of the relevance criteria are discussed in Section 7. Finally, conclusions are drawn in Section 8.

2. Feature Selection

The key idea of feature selection is that a good feature subset should contain the features that are the most relevant to the target. Contrarily to dimension reduction whose goal is to create new features, feature selection selects features among the

e-mail: adegeest@he2b.be (Alexandra Degeest)

original set, maintaining their interpretation. Many works focus on methods reducing the original set of features in datasets [13, 32]. Feature selection has many benefits: it fights the curse of dimensionality, decreases memory needs, improves prediction performances, reduces computational costs, and allows to better interpret the features and the model.

Feature selection methods are categorised into filters, wrappers and embedded methods. Unlike wrappers and embedded methods, filters are independent from the model to be learned. This is an advantage in terms of computational cost and speed because relevant features are selected without training numerous models [2]. This computational advantage allows to test more possible feature subsets than in a wrapper of embedded approach. This paper focuses on filters.

To select the most relevant subset of features, a filter method relies on two choices: the choice of the search procedure and the choice of the relevance criterion. As the number of possible subsets of features is exponential with the dimensionality (i.e., the number of original features), search procedures aim at reducing the number of subsets that are considered and evaluated. One of the most common search procedures is the forward search. During a forward search, the first step finds the most relevant feature with respect to the target. In the second step, it computes the relevance of every group of two features containing the one selected during the first step and a new one to be added. This procedure is extended to three, four, etc. features. The forward search procedure is used in this paper as an illustration for the feature selection process; however it can easily be replaced by a backward search or any other search procedure; the search procedure and the choice of the relevance criterion are indeed independent choices.

Besides the search procedure, the key factor of a successful feature selection is the relevance criterion, i.e., the measure of the relevance between a group of features and the target. The goal of this paper is to examine the properties of relevance criteria, and to provide the user a reading grid that can be used to adapt the choice of a criterion according to key properties of the data in regression problems.

3. Related Works

Several papers have recently benchmarked feature selection methods [4, 18, 29] where many relevance criteria are enumerated for classification, clustering or regression. Several benchmarks have been specifically realised to find the best filter criterion. For example, Bommert et al. analyse different filter methods with respect to runtime and accuracy on high-dimensional datasets, in a classification context [2]. Their work shows that there is no filter method that always outperforms all other ones, although some filter methods perform well on many of the datasets. There is no analysis on the reasons of the conclusions in the paper. Furthermore, they focus on classification. Another example is the work done by Shivan Darshan et al. for filters in classification [28]. It shows that, among many filter criteria, several ones behave better than others in terms of accuracy.

Although the above papers usually do not focus on regression, they illustrate the large range of existing filter criteria and the lack of a single, ideal one that fits all possible needs. There exist indeed many different relevance criteria for filter methods: the correlation factor, the mutual information and its variants (MI, NMI, JMI, etc.) [3], the noise variance [8], the $mRMR$ [7], the coefficient of determination [23], etc.

However, these papers do not provide an in-depth discussion of the properties of the criteria. As no criterion outperforms all other ones from every aspect, the analysis of the properties provides a reading grid that may be used to choose the criterion that is best adapted to a regression problem and a dataset at hand.

4. Properties of Relevance Criteria

The choice of the relevance criterion in a feature selection process is strategic. As detailed in Section 3, there exist many such criteria. This shows the need to establish a reading grid to allow the user to make a choice for the problem at hand.

This section details six fundamental properties that can be used to analyse any relevance criterion and understand its behaviour for a specific dataset. It describes why a relevance criterion used in regression should be multivariate (Section 4.1) and nonlinear (Section 4.2), and the important properties of the estimator of the criterion itself, with one property focusing on the estimator complexity in Section 4.4 and three properties focusing on stability properties in Sections 4.3, 4.5 and 4.6.

4.1. Property 1: Multivariate Criterion

When performing a search during a feature selection process, e.g., with a forward search, one needs to evaluate the relevance of *groups of features* (subsets of the initial set) with respect to the target. Indeed, some features do not bring any information on their own but bring information when coupled to other features; an example is when the target depends on the product of two features but is independent from each of the latter taken individually. For this reason, using a relevance criterion able to detect multivariate relationships is essential [29]. This property is regarded as mandatory: a criterion that does not meet this requirement will not be considered in this paper. This property is used in Section 6.1 to analyse relevance criteria.

4.2. Property 2: Nonlinear Criterion

Most regression datasets present nonlinearities between variables. A relevance criterion used for feature selection in regression must therefore be able to detect nonlinear relationships. Similarly to Property 1, this property is regarded as mandatory: a criterion that does not meet this requirement will not be considered in this paper. This property is analysed in Section 6.2.

4.3. Property 3: Estimator Parameters

Most filter criteria are defined as a statistical property of the data averaged over the domain space. Because their evaluation on a finite dataset relies on numerical integration over the domain space, the criteria can usually not be evaluated exactly: an estimator is needed. The latter usually requires one or several parameters to be tuned. The parameters may greatly influence the quality of the estimator, and therefore the quality of the feature selection process itself. For example, nearest-neighbour-based estimators, such as the Kraskov estimator [20] detailed in

Section 5, need to choose the number of neighbours used in the estimation. This choice of parameter might be crucial [10]; its influence is often underestimated in the literature. In addition, filter feature selection is an unsupervised process: the regression model is not used in filters, which means that the ground truth would be the true value of the criterion, not of the regression; this ground truth is unknown, preventing the optimisation of estimator parameters. A relevance criterion should therefore be provided with an estimator that is as independent as possible from the choice of user-defined meta-parameters. This property is discussed in Section 6.3.

4.4. Property 4: Estimator Complexity

When choosing a criterion, the computational complexity of its estimator must be taken into account [24]. Indeed depending on the search procedure to find the best subset of features, the number of estimations of the criterion may be large (up to $2^d - 1$ for an exhaustive search, where d is the maximum number of selected features). An estimator with a high computational complexity would seriously hinder the benefits of filters for feature selection (with respect to wrappers). A relevance criterion should therefore have a low computational complexity; this property is analysed in Section 6.4.

4.5. Property 5: Estimator Sample Robustness

As estimators work with finite datasets, how they behave with small samples, and how they are sensitive to small variations in sets, are important questions.

First, while datasets with many instances are available in some fields, the size of datasets remains limited in other fields for many reasons, such as the occurrence of events or the cost of collection, even when the number of features is large [27]. The *number of instances/dimensionality* ratio is therefore important in machine learning. A dataset with a low *number of instances/dimensionality* ratio is called a small sample dataset. For such datasets, feature selection is a necessity, e.g., to facilitate the estimation of the model. An estimator of the relevance criterion which is robust to small samples is thus necessary.

Here, *robustness* first means that the estimator should be as unbiased as possible and have a small variance, when the number of available data is limited. In addition, it might be accepted that, in a feature selection process (e.g., forward search), the bias of the estimator is less important than the preservation of ranks, as a feature is selected when it maximises or minimises the criterion, regardless of its actual value. This second, weaker definition of *robustness* can be formulated as the variation in feature selection results due to small changes in the dataset [25]. It is also often called *stability* in some papers [4].

In order to guarantee these two properties, the estimator should have as small as possible bias and variance, when the size of the dataset is small. Section 6.5 discusses this property.

4.6. Property 6: Estimator Noise Robustness

Real datasets can be noisy, which can influence the result of the feature selection process. The extent of this influence is an important matter and depends on the level and the type of the noise. The value of an estimator should be as stable as possible with respect to a certain level of noise. The way how stability is measured is identical to Property 5. This property is discussed and analysed for regression problems in Section 6.6.

5. Relevance Criteria Description

The goal of this paper is to provide a reading grid of properties that can guide the user when choosing a feature relevance criterion. Numerous such criteria may be used for filter feature selection. However, many of them do not satisfy Properties 1 and 2 (ability to evaluate the relevance of groups of features, and discovering nonlinear dependencies). This paper only considers criteria that meet these two basic properties. For example, most of the criteria benchmarked by Bommert et al. [2] are univariate, except those based on mutual information. This is also the case for the correlation [33], the Fisher score [14] or the Laplacian score [17]. To illustrate the reading grid this paper covers three (families of) criteria that meet the multivariate and nonlinearity properties; other criteria can easily be cast in that framework.

As mutual information is at the root of many feature relevance criteria (see Brown et al. [3] for details), the first criterion is mutual information (as estimated by the Kraskov estimator, see Section 5.1). The second criterion is the noise variance, which is especially suited for regression, as it is closely related to the mean squared error of the best possible model (as estimated by the delta test estimator, see Section 5.2). Eventually, the third criterion is the adjusted coefficient of determination and its extension to nonlinear dependencies.

5.1. Mutual Information

Mutual information originates from information theory and has been introduced by Shannon [26]. This entropy-based criterion is widely used in feature selection processes, in classification [21] and regression [11]. This paper focuses on regression; a similar analysis could be done for classification.

Let X be a random vector of features and Y a random variable (or vector) of targets, whose respective probability density functions are p_X and p_Y. The mutual information $I(X;Y)$ measures the reduction in the uncertainty on Y when X is known

$$I(X;Y) = H(Y) - H(Y|X) \tag{1}$$

where $H(Y)$ is the entropy of Y and $H(Y|X)$ is the conditional entropy of Y given X. The mutual information between X and Y is equal to zero if and only if they are independent. If Y can be perfectly predicted as a function of X, then $I(X;Y) = H(Y)$. Notice that in Equation (1), both X and Y can be multidimensional random vectors; if X gathers a subset of features and Y is the target variable, $I(X;Y)$ can be directly used to measure the relevance of this subset of features.

With real datasets, $I(X;Y)$ cannot be directly computed because it is defined in terms of probability density functions, which are unknown when only a finite sample of data is available. Therefore, $I(X;Y)$ has to be estimated [12].

The Kraskov estimator of the mutual information, introduced by [20], is based on the nearest-neighbour-based Kozachenko-Leonenko entropy estimator [19]. Its definition is

$$\hat{I}(X;Y) = \psi(N) + \psi(k) - \frac{1}{k} - \frac{1}{N}\sum_{i=1}^{N}\left(\psi\left(\tau_x(i)\right) + \psi\left(\tau_y(i)\right)\right) \quad (2)$$

where ψ is the digamma function, N is the number of instances in the dataset, k is the number of neighbours, $\tau_x(i)$ is the number of points located no further than the distance $\epsilon_X(i,k)/2$ from the i^{th} observation in the $\mathcal{X}$ space, $\tau_y(i)$ is the number of points located no further than $\epsilon_Y(i,k)/2$ from the i^{th} observation in the $\mathcal{Y}$ space and where $\epsilon_X(i,k)/2$ and $\epsilon_Y(i,k)/2$ are the projections into the $\mathcal{X}$ and $\mathcal{Y}$ subspaces of the distance between the i^{th} observation and its k^{th} neighbour. The intuition behind this estimator is to measure whether the number of instances that are neighbours is similar depending on whether X and Y are considered together or separately (see Fig. 1 in [20] and details therein).

During a search procedure, such as the forward search, in order to find the most relevant features for the problem at hand, the subset with the highest value of $\hat{I}(X;Y)$ (2) is selected.

5.2. Noise Variance

The noise variance is another frequently used filter criterion in regression [8]. This criterion evaluates the level of noise in a finite dataset. In the context of regression, the noise level in the target estimation represents the error that a regression model would make, based on the currently selected input features.

Considering a finite dataset of N instances, D features X_j, a target Y and N input-output pairs $(\mathbf{x}_i, y_i)$, the relationship between these input-output pairs is

$$y_i = f(\mathbf{x}_i) + \epsilon_i \qquad i = 1, ..., N \quad (3)$$

where f is the unknown function between $\mathbf{x}_i$ and y_i, and ϵ_i is the noise, or *prediction error*, when estimating f. For feature selection, one selects the subsets of features X_j which lead to the lowest prediction error, or the lowest noise variance [8].

With real finite datasets, the noise variance has to be estimated, e.g. with the delta test [15]

$$\delta = \frac{1}{2N}\sum_{i=1}^{N}[y_{NN(i)} - y_i]^2 \quad (4)$$

where N is the size of the dataset and $y_{NN(i)}$ is the output associated to $\mathbf{x}_{NN(i)}$, the nearest neighbour of the instance $\mathbf{x}_i$.

During the feature selection search procedure, the subset with the lowest value of δ (4) is selected at each step.

5.3. Coefficient of Determination

The coefficient of determination R^2 is the proportion of the variance in the output variable Y that can be explained from the input variables X_j; it ranges from 0% (unpredictable) to 100% (totally predictable). The definition of R^2 is

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \quad (5)$$

where $SS_{res} = \sum_i (y_i - f(\mathbf{x}_i))^2$ and $SS_{tot} = \sum_i (y_i - \bar{y})^2$ with $i = 1, ..., N$, with f being a linear regression and with $\bar{y}$ being the mean of the observed data. This coefficient statistically measures how well regression approximates the target. Because R^2 automatically increases when features are added to the model, its alternative, the adjusted R^2, or R^2_{adj}, is used in this work:

$$R^2_{adj} = 1 - \frac{SS_{res}/(N-d-1)}{SS_{tot}/(N-1)} \quad (6)$$

where d is the number of selected features in the model and N the sample size. A low R^2_{adj} indicates that observed data are not close to the regression and a high R^2_{adj} indicates the opposite.

The R^2_{adj} criterion used with a linear regression model cannot capture the nonlinear relationships between the features and the target. In order to use the R^2_{adj} in a nonlinear context, local linear approximations are considered [5], computed as follows. For each instance a linear regression is performed on a defined number of neighbours k around the instance. The R^2_{adj} is computed for every regression and the average on all observed data is taken. This process is repeated for increasing values of k (starting from 4 in our experiments). The best mean R^2_{adj} is then selected; it corresponds to a specific number of neighbours k. Indeed, depending on the dataset, locally linear relationships will be detected and measured at different scales, i.e., for a different number of neighbours k. This best mean R^2_{adj} used for nonlinear functions is called the nonlinear adjusted R^2, or NLR^2, in the remaining sections of this paper.

If a forward search is used, at each step of the selection process, the group of features that corresponds to the highest NLR^2 value is selected. In that process, the local linear regressions are successively fit in spaces of increasing dimensionality [6].

6. Analysis of Relevance Criteria

In this section, the three relevance criteria previously introduced in Section 5 are analysed: the mutual information (MI) with the Kraskov estimator (2), the noise variance (NV) with the delta test estimator (4) and the adjusted coefficient of determination with the NLR^2 estimator (6). They are analysed with respect to the six properties previously described in Section 4, on three informative features from three real-world datasets: Anthrokids, Poland and Santa Fé (Fig. 1). The simulations are illustrative only and aim at providing a basis for discussion of the properties. It extends preliminary work presented in [5, 6]. The code for all experiments is available on `https://github.com/alexdegeest/FeatureSelection_ReadingGrid`.

The Anthrokids dataset represents the results of a three-year study on 3900 infants and children representative of the U.S. population of year 1977, ranging in age from newborn to 12 years of age. The dataset comprises 121 variables and the target variable to predict is children's weight. As this dataset presents many missing values, a prior sample and variable discrimination has been performed; the final set without missing values contains 1019 instances, 53 input variables and one output (weight) [16]. The Poland electricity load dataset consists of 1370 samples with 30 continuous features. The original time series is transformed into a regression problem, where the 30

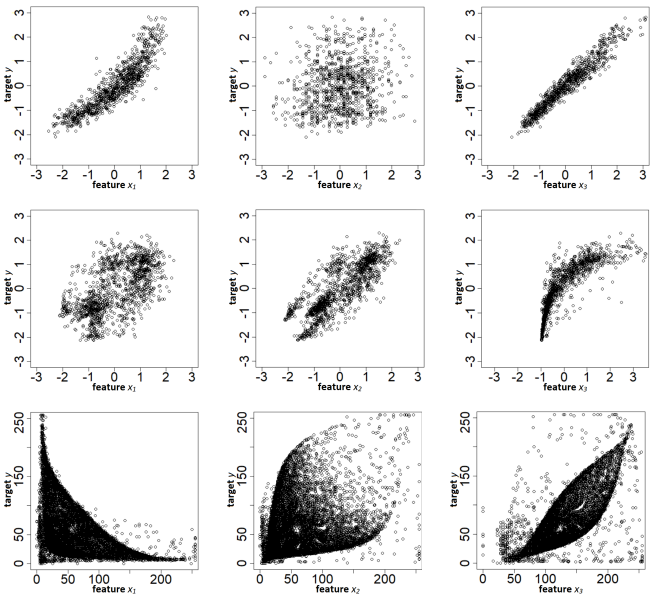

Fig. 1. Target variable (y axis) with respect to each of the three selected features (x axis) for the Anthrokids (row 1), Poland (row 2) and Santa Fé (row 3) datasets. See text for details.

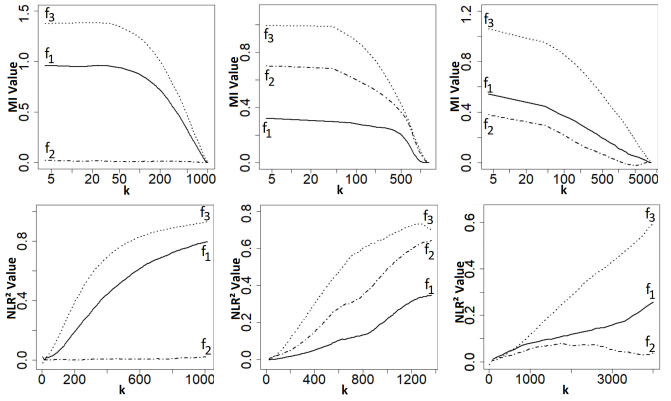

Fig. 2. Relevance criterion scores for datasets Anthrokids (column 1), Poland (column 2) and Santa Fé (column 3), for increasing values of k from 4 to the total number of instances in the dataset. MI is represented in the first row and NLR^2 in the second row. The first row has a logarithmic scale and the second row a linear one, in order to focus on the small values for the first and the large ones for the latter; see text for details.

past values are used to predict the electricity load of the next day [30]. The Santa Fé laser dataset consists of 10081 samples with 12 continuous features [31].

Only three features of each dataset have been chosen to illustrate at best the differences between the analysed filter criteria: features 1, 11 and 19 for Anthrokids; features 1, 24 and 34 for Poland; features 1, 4 and 5 for Santa Fé; these numbers correspond to the feature numbers in references [16, 30, 31]. In order to increase the nonlinearity of features in this analysis, again for illustrative purposes, one of the features of Poland and Santa Fé has been modified: f_{24} of Poland is replaced by $e^{f_{24}}$ and f_5 of Santa Fé is replaced by $16 \times \sqrt{f_5}$. Features are normalised to avoid issues in distance computations when using k-nearest-neighbour-based methods.

The following of this section presents experimental results of the six properties detailed in Section 4 on these features. Section 7 will discuss the results.

6.1. Property 1: Multivariate Criterion

The filter criteria considered here for analysis are all multivariate, in the sense that they can compute a measure of the relevance of a group of features with respect to the target. The first step of a forward search is univariate, but all criteria are multivariate from the second step.

6.2. Property 2: Nonlinear Criterion

The filter criteria presented in Section 5 are all able to evaluate the relevance of nonlinear relationships, such as the relationship between the third feature x_3 and the target y of Poland (right feature in the second row of Fig. 1). Mutual information and the noise variance are intrinsically nonlinear. For the adjusted coefficient of determination, the implementation, called NLR^2 and described in Section 5.3, uses local approximations of the regression, which makes it also able to analyse nonlinear relationships between the feature subsets and the target.

6.3. Property 3: Estimator Parameters

As detailed in Section 4.3, filter criteria need an estimator for finite datasets. The behaviour of the estimator is therefore essential during feature selection. The estimators of the three compared criteria are all based on a k-nearest-neighbour search, but they show very different properties with respect to the choice of k. The mutual information requires to set k, and as the ground truth (the real value of MI) is not known, there is no possibility to supervise the choice. In the delta test (4), k is set to 1, which means that there is no parameter. About the NLR^2, a procedure has been detailed in Section 5.3 to supervise the choice of the best value of k. Hence the method does not depend on k anymore, but it remains interesting to check whether a reasonable value of k is selected as a large k can reduce the sensitivity towards noise but can miss local nonlinearities.

To compare the Kraskov estimator and NLR^2, experiments have been performed with the three features of the three datasets, for increasing values of k from 4 to the total number of instances in the dataset. The score of the two compared relevance criteria (MI and NLR^2) are shown in Figure 2. The MI displayed on the first row of the figure is stable when k remains small with respect to the number of instances in the dataset. With high values of k, the MI value drops abruptly. This result illustrates that MI is adequate for feature selection with a fixed small k, e.g., around 6 such as advised in [20], in the sense that its value remains stable in that range.

On the other hand, NLR^2, displayed on the second row of the figure, can only be used above a certain k (around 200) for a successful feature selection process; below this threshold, NLR^2 is unable to reflect the relevance of features. Above this threshold, the ranking of the features is stable, even if the NLR^2 value is not stable itself. In addition, the maximum value of NLR^2, selected as detailed in Section 5.3, is most often close to the value obtained with the maximum number of instances. In this case, NLR^2 gets closer to a linear criterion, which contradicts Property 2.

6.4. Property 4: Estimator Complexity

The estimators (2) (4) (6) of the three relevance criteria used in this work are all based on a k-nearest-neighbour-based search; this search is the most computationally demanding operation for the three estimators, as detailed below.

For the Kraskov estimator, the k-nearest-neighbour-based search is performed once for each of the N instances, in order to set a distance ϵ. Then, a search of the number of points within ϵ is done twice, once in the $\mathcal{X}$ subspace and once in the $\mathcal{Y}$ subspace. The difference $\psi(k) - \psi(\tau_x(i)) - \psi\left(\tau_y(i)\right)$ is then computed for each instance and averaged. In total, this gives a complexity of $O(N \log N)$ with an efficient search algorithm, such as a k-d tree [1] or a ball tree [22]. For the delta test estimator, a nearest-neighbour-based search (i.e., $k = 1$) is performed N times: once for each of the N instances in the dataset. The difference $y_{NN(i)} - y_i$ between the target value for the instance and its neighbour is then computed for each instance and averaged. Again, it results in a complexity of $O(N \log N)$. The above analysis focuses on the computational cost of the k-nearest-neighbour-based search for a fixed value of k, but the comparison of the computational complexities also depends on the value of k. For the delta test, k is set to 1 by definition, which makes its computational cost lower than the one of the Kraskov estimator, although the difference will not be too large as the latter works well with a low value of k, as mentioned in Section 6.3.

For the nonlinear adjusted R^2 (NLR^2), the k-nearest-neighbour-based search is performed $m \times N$ times, where m is the number of tested k values, resulting in a complexity of $O(m \times N \log N)$. As explained in Section 6.3, the NLR^2 needs to use several values for k and it works better with a higher value of k, what means that this relevance estimator has the worst computational complexity between the three estimators.

6.5. Property 5: Estimator Sample Robustness

When applying feature selection on a small sample dataset, one should select the most appropriate filter criterion. But what happens to those criteria when the number of samples is low with respect to the number of features? And are they robust to small variations in the datasets?

To analyse the behaviour of the criteria with respect to the number of samples, experiments have been performed on the Anthrokids and Poland datasets. Each relevance criterion has been evaluated ten times, with the number of samples increasing from few instances (20) to the complete dataset. Small samples have been obtained by random subsampling of the whole dataset. Results (average values and standard deviations over ten repetitions) are represented in Figure 3. Regarding the small sample robustness (average values), the three criteria behave quite well above 200 instances. Below 200, the resulting feature ranking could differ for NLR^2. Regarding the robustness to variations in the datasets, the standard deviation is the largest for the noise variance (NV), especially when the number of samples is low. The standard deviation of mutual information (MI) is smaller than for other criteria, even in small samples. Below 100 instances, for the Poland dataset, the ranking of features may be disturbed, especially with the noise variance.

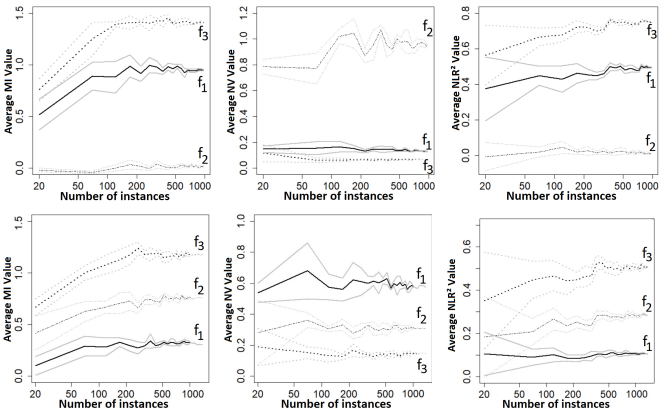

Fig. 3. Average value (in black), and standard deviation (in grey) over ten repetitions, of the MI (left column), the NV (centre column) and the NLR^2 (right column) for the three features of the datasets Anthrokids (row 1) and Poland (row 2), for sample sizes from 20 to the total number of instances.

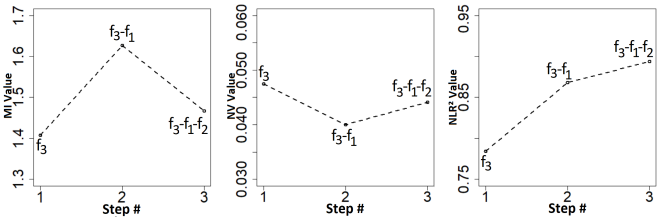

Fig. 4. Feature selection performed on the three illustrative features of Anthrokids. Relevance score for groups of features: MI (left), NV (centre) and NLR^2 (right), for the three steps (1, 2 and 3) of the forward search.

A related interesting behaviour of a relevance criterion estimator is its ability to correctly handle the successive steps of a feature selection procedure. Indeed, in a forward search for example, the successive estimations of the criterion are performed in spaces of increasing dimensionality, with increasing risk of estimation error due to the curse of dimensionality. Figure 4 shows an illustrative 3-feature selection process performed on Anthrokids. It shows a decrease in the mutual information (MI) score between the second step and the third step of the forward search. If the true value of MI was illustrated, such decrease would not be observed: even if an additional feature does not add any information, the MI score should remain constant and not decrease. The decrease is thus clearly linked to the quality (bias or variance) of the estimator when comparing groups of features in different dimensions. Note that the lack of theoretical maximum means that another stopping criterion must be used in a forward search (see, for example, [9]). The noise variance (NV) score shows the same problem. Interestingly, NLR^2 does not show this phenomenon. But even with these estimation errors, the resulting ranking of features of MI, NV and NLR^2 are identical: f_3 first, then f_1 and finally f_2.

6.6. Property 6: Estimator Noise Robustness

To compare the robustness of the three relevance criteria with respect to noise, uniform noise with three different amplitudes has been added, in a first experiment to the target Y (Fig. 5) and in a second experiment to the features X (Fig. 6). The number of affected data ranges from 0 to 40% of the whole dataset.

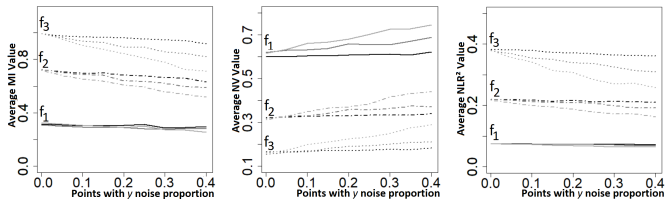

Fig. 5. Average value for *MI* (left), *NV* (centre) and NLR^2 (right) when the proportion of points with *y*-noise changes from 0% to 40%. Three amplitudes of noise are represented (small amplitude in black, moderate amplitude in grey, high amplitude in light grey).

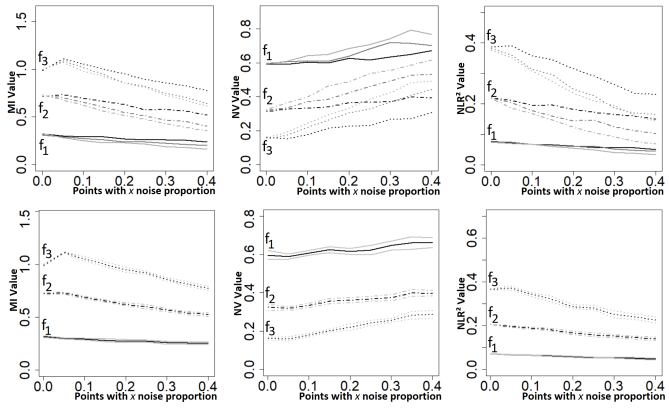

Fig. 6. Relevance value for *MI* (left), *NV* (centre) and NLR^2 (right) when the proportion of points with *x*-noise changes from 0% to 40%. First row: average values with three amplitudes of noise (small amplitude in black, moderate amplitude in grey, high amplitude in light grey). Second row: Average values over 10 times (black) and standard deviation values (grey) with a high amplitude of noise.

This experiment has been performed on the three same datasets but, for a matter of space, only the results for Poland are represented in Figure 5 and the first row of Figure 6, respectively; the behaviour of the three criteria is similar with the two other datasets and is therefore not exhibited. The figures show average values of four repetitions. *MI* and NLR^2 have a similar behaviour: their score for the "best" feature is more affected by the amplitude of noise than the "worst" feature. The *NV* score, on the other hand, is affected in the same way by the amplitude of noise for every feature, informative or not.

The second row of Figure 6 shows the same experiment as in the first row, with only the highest level of noise (light grey lines in the top row), but now with the standard deviation (estimated on ten repetitions). All three criteria exhibit a low standard deviation (even though the standard deviation for feature f_1 in the *NV* experiment is slightly larger). This shows therefore a stability towards *x*-noise for all three criteria.

7. Discussion

To illustrate the proposed reading grid of the six properties described in Section 4, an analysis of three interesting filter criteria has been done in Section 6. Table 1 shows a summary of the results of this analysis: the ability for a filter criterion to be multivariate (P1) and nonlinear (P2), the sensitivity to parameters (P3) and complexity (P4), and their robustness towards samples (P5) and noise (P6). With respect to each property, it indicates whether the behaviour of the criterion is excellent, good or fair.

The two first properties P1 and P2 are not discriminant for the three considered criteria; they remain essential if other criteria are considered. P2 has however been rated 'Fair' for the NLR^2 criterion, because P2 and P3 cannot really be guaranteed jointly (see comment about the value of *k* in Section 6.3).

The next four properties are discriminant in this study. For the comparison of the estimator parameters (P3), the delta test is the easier estimator to set up for an experiment as it has no parameter. Kraskov is simple as well: it has a single parameter, but setting a reasonably low value is sufficient: no optimisation is required. NLR^2 is more complex, in terms of parameters because the best value for parameter *k* is hard to find, problem-dependent and needs optimisation. Concerning the complexity of the estimator (P4), the delta test has the lowest computational complexity, the NLR^2 has the highest one and the Kraskov estimator offers a trade-off.

Regarding the small-sample robustness (P5), there is no real difference in the number of samples the three methods need in order to reach acceptable estimations. However NLR^2 easily fails in ranking correctly the features if the number of samples is too low. In addition, what concerns the robustness to small variations in the datasets, the standard deviation of the estimators shows that the mutual information is more robust and the noise variance is the less robust method.

The sixth property (P6) shows that the amplitude of noise on the data has less influence on the values of *MI* and NLR^2 than on the values of the noise variance, especially for *x*-noise.

The properties are qualitatively summarised in Table 1. Table 1 shows that, generally speaking, in the context of the experiments developed in this paper, the mutual information has less drawbacks than the two other criteria: it shows a good, or an excellent, behaviour with regards to all properties, while the noise variance is less robust to small samples and noise, and the NLR^2 is both sensitive to its parameter and too close to a linear criterion if the parameter is set automatically. It is important to remember that the purpose of this paper is to provide a methodology and a reading grid; depending on the application, some of the properties will be more (or less) preponderant. Hence, the results in the reading grid will need to be balanced for each criterion depending on the specific needs. In practice, for a specific dataset, it is suggested to (1) rely by default on multivariate and nonlinear criteria (Properties P1 and P2); (2) check the complexity of the estimator for this specific dataset as this can influence both the choice of a criterion itself, and of the subset search procedure as the latter is always a compromise between the coverage of the subset space and the computational resources at disposal; (3) check the robustness of the considered estimators along the guidelines provided in this paper (and prefer *MI*-based methods if the set of criteria is restricted to those covered in this paper).

8. Conclusions

Filter methods for feature selection need a relevance criterion adapted to the problem at hand, in regression and in classification. This paper focuses on regression problems and proposes

Table 1. Reading grid with three relevance criteria (see text for details).

Properties	*MI*	*NV*	NLR^2
P1: Multivariate	Excell.	Excell.	Excell.
P2: Nonlinearity	Excell.	Excell.	Fair
P3: Estimator parameters	Excell.	Excell.	Fair
P4: Estimator complexity	Good	Excell.	Good
P5: Sample robustness	Excell.	Fair	Good
P6: Noise robustness	Good	Fair	Good

a reading grid of six fundamental properties to analyse any relevance criterion, in order to find the criterion which is the most adapted to the current problem. To illustrate the proposed reading grid, three relevance criteria have been analysed in feature selection experiments (see Table 1 for a summary of the results).

This grid offers the advantage that it can be extended to other relevance criteria in order to choose the best criterion for the problem at hand. Furthermore, the grid only focuses on the choice of the relevance criterion and is, therefore, completely independent from the search strategy. On the other hand, the reading grid has only been applied to regression datasets; it could be extended to classification problems.

References

[1] Bentley, J.L., 1975. Multidimensional binary search trees used for associative searching. Commun. ACM 18, 509–517.

[2] Bommert, A., Sun, X., Bischl, B., Rahnenführer, J., Lang, M., 2020. Benchmark for filter methods for feature selection in high-dimensional classification data. Comput. Stat. and Data Analysis 143, 106839.

[3] Brown, G., Pocock, A., Zhao, M., Lujan, M., 2012. Conditional likelihood maximisation: A unifying framework for mutual information feature selection. J. of Machine Learning Research 13, 27–66.

[4] Chandrashekar, G., Sahin, F., 2014. A survey on feature selection methods. Comput. Electr. Eng. 40, 16–28.

[5] Degeest, A., Verleysen, M., Frénay, B., 2019a. About filter criteria for feature selection in regression, in: Proc. of IWANN, Gran Canaria, Spain. pp. 579–590.

[6] Degeest, A., Verleysen, M., Frénay, B., 2019b. Comparison between filter criteria for feature selection in regression, in: Proc. of ICANN, Munich, Germany. pp. 59–71.

[7] Ding, C., Peng, H., 2003. Minimum redundancy feature selection from microarray gene expression data, in: Proc. of IEEE Bioinformatics Conference, Stanford, CA, USA. pp. 523–528.

[8] Eirola, E., Lendasse, A., Corona, F., Verleysen, M., 2014. The delta test: The 1-nn estimator as a feature selection criterion, in: Proc. of IJCNN, Beijing, China. pp. 4214–4222.

[9] François, D., Rossi, F., Wertz, V., Verleysen, M., 2007a. Resampling methods for parameter-free and robust feature selection with mutual information. Neurocomputing 70, 1276–1288.

[10] François, D., Wertz, V., Verleysen, M., 2007b. The concentration of fractional distances. IEEE Trans. on Knowl. and Data Eng. 19, 873–886.

[11] Frénay, B., Doquire, G., Verleysen, M., 2013. Is mutual information adequate for feature selection in regression ? Neural Networks 48, 1–7.

[12] Gao, W., Kannan, S., Oh, S., Viswanath, P., 2017. Estimating mutual information for discrete-continuous mixtures, in: Proc. of NIPS, Long Beach, CA, USA. pp. 5986–5997.

[13] Gao, W., Hu, L., Ping, Z., He, J., 2018. Feature selection considering the composition of feature relevancy. Pattern Recognition Letters 112, 70–74.

[14] Gu, Q., Li, Z., Han, J., 2011. Generalized fisher score for feature selection, in: Proc. of UAI, Arlington, Virginia, USA. pp. 266–273.

[15] Guillén, A., Arenas, M., van Heeswijk, M., Sovilj, D., Lendasse, A., Herrera, L., Pomares, H., Rojas, I., 2014. Fast feature selection in a gpu cluster using the delta test. Entropy 16, 854–869.

[16] Guillén, A., Sovilj, D., Lendasse, A., Mateo, F., Rojas, I., 2008a. Minimising the delta test for variable selection in regression problems. Int. J. of High Performance Systems Architecture 1, 269–281.

[17] He, X., Cai, D., Niyogi, P., 2005. Laplacian score for feature selection, in: Proc. of NIPS, Vancouver, Canada. pp. 507–514.

[18] Jović, A., Brkić, K., Bogunović, N., 2015. A review of feature selection methods with applications, in: Proc. of MIPRO, Croatia. pp. 1200–1205.

[19] Kozachenko, L.F., Leonenko, N., 1987. Sample estimate of the entropy of a random vector. Problems Inform. Transmission 23, 95–101.

[20] Kraskov, A., Stögbauer, H., Grassberger, P., 2004. Estimating mutual information. Phys. Rev. E 69, 066138.

[21] Nguyen, X.V., Chan, J., Romano, S., Bailey, J., 2014. Effective global approaches for mutual information based feature selection, in: Proc. of ACM SIGKDD, NY, USA. pp. 512–521.

[22] Omohundro, S.M., 1989. Five Balltree Construction Algorithms. Technical Report TR-89-063. International Computer Science Institute.

[23] Renaud, O., Victoria-Feser, M.P., 2010. A robust coefficient of determination for regression. J. of Stat. Plan. and Inference 140, 1852 – 1862.

[24] Ross, B.C., 2014. Mutual information between discrete and continuous data sets. PloS one 9, 1–5.

[25] Saeys, Y., Abeel, T., Van de Peer, Y., 2008. Robust feature selection using ensemble feature selection techniques, in: Proc. of ECML PKDD, Berlin, Heidelberg. pp. 313–325.

[26] Shannon, C.E., 1948. A mathematical theory of communication. Bell Systems Technical J. 27, 379–423.

[27] Shi, J., Zhou, S., Liu, X., Zhang, Q., Lu, M., Wang, T., 2016. Stacked deep polynomial network based representation learning for tumor classification with small ultrasound image dataset. Neurocomputing 194, 87–94.

[28] Shiva Darshan, S., Jaidhar, C., 2018. Performance evaluation of filter-based feature selection techniques in classifying portable executable files. Procedia Computer Science 125, 346 – 356.

[29] Vergara, J.R., Estévez, P.A., 2014. A review of feature selection methods based on mutual information. Neural Computing and Appl. 24, 175–186.

[30] Wei, C.C., Chen, T.T., Lee, S.J., 2013. k-nn based neuro-fuzzy system for time series prediction, pp. 569–574.

[31] Weigend, A., 1993. Time series prediction: Forecasting the future and understanding the past.

[32] Xue, B., Zhang, M., Browne, W., Yao, X., 2015. A survey on evolutionary computation approaches to feature selection. IEEE Trans. on Evolutionary Computation 20, 606–626.

[33] Yu, L., Liu, H., 2003. Feature selection for high-dimensional data: A fast correlation-based filter solution, in: Proc. of ICML, Washington, DC, USA. pp. 856–863.