# About Filter Criteria for Feature Selection in Regression

Alexandra Degeest[1,2(✉)], Michel Verleysen[2], and Benoît Frénay[3]

[1] Haute-Ecole Bruxelles Brabant - ISIB, 150 rue Royale, 1000 Brussels, Belgium
adegeest@he2b.be
[2] UCLouvain Machine Learning Group - ICTEAM,
Place du Levant 3, 1348 Louvain-La-Neuve, Belgium
michel.verleysen@uclouvain.be
[3] Faculty of Computer Science, NADI Institute - PReCISE Research Center,
Université de Namur, Rue Grandgagnage 21, 5000 Namur, Belgium
benoit.frenay@unamur.be

**Abstract.** Selecting the best group of features from high-dimensional datasets is an important challenge in machine learning. Indeed problems with hundreds of features have now become usual. In the context of filter methods, the selected relevance criterion used for filtering is the key factor of a feature selection method. To select an appropriate criterion among the numerous existing ones, this paper proposes a list of six necessary properties. This paper describes then three relevance criteria, the mutual information, the noise variance and the adjusted R-squared, and compares them in the view of the aforementioned properties. Any new, or popular, criterion could be analysed in the light of these properties.

**Keywords:** Feature selection · Relevance criteria · Regression

## 1 Introduction

High-dimensional datasets appear now frequently in various domains such as healthcare, marketing or social media, especially in regression. Selecting the most relevant subset of features in high-dimensional datasets has therefore become essential for many purposes: to increase the interpretability of features, to facilitate the learning process, to visualise data, to alleviate the curse of dimensionality, etc [14,16].

Many works, in a variety of domains, focus on methods to reduce the number of features in datasets [2,10,12,17,19,22,23,27]. These methods can be roughly categorised into filters, wrappers and embedded methods. This paper focuses on filter methods, which have the advantage to be fast because they do not require to train any model during the feature selection process, contrarily to wrappers [18,19] and embedded methods [7].

Filter methods rely on a relevance criterion to reduce the set of features to only the most relevant ones. This relevance criterion is therefore the key factor


I. Rojas et al. (Eds.): IWANN 2019, LNCS 11507, pp. 579–590, 2019.
https://doi.org/10.1007/978-3-030-20518-8_48

of a successful filter-based feature selection process. Several relevance criteria exist and are used in feature selection methods on various datasets.

This paper focuses on the necessary properties of a criterion used in filter methods for feature selection in regression. What is needed in a filter criterion in order to obtain the best subset of features with respect to the target or prediction goal (in classification or regression tasks)? Existing criteria are often designed to fulfill a unique purpose: for instance to measure a nonlinear relation or to estimate the noise variance of the distribution. An efficient criterion should probably combine various goals.

This paper does not intend to propose a new filter criterion but, instead, focuses on the diverse properties that make a good relevance criterion, in order to be able to select one among the numerous existing ones. These important properties are listed and discussed in Sect. 3, after an introduction to feature selection in regression with filters in Sect. 2.

It is essential to analyse relevance criteria in view of these properties and these goals in order to understand the strengths and the weaknesses of each of them, and to understand their behaviour according to the type of dataset at hand. Existing criteria are described in Sect. 4 and compared with respect to these properties in Sect. 5. Finally, conclusions are given in Sect. 6.

## 2 Feature Selection with Filter Methods

Feature selection is an important task in machine learning. It helps to reduce the dimension of the dataset by eliminating redundant and less useful features.

In the context of filter methods for feature selection in regression, a good relevance criterion is necessary to select the most relevant features among all the available ones. The relevance criterion aims at measuring the existing relationship between a feature, or a set of features, and the variable to predict. There exist several relevance criteria based on different measures such as entropy or noise variance.

Filter methods also need a search procedure to find the best feature subset among an exponential number (exponential to the dimensionality of the dataset) of all possible ones that could be extracted from the complete dataset [16]. During the search procedure, the filter criterion is again a strategic factor because it is used to evaluate the relevance of each subset with respect to the target. Implicitly, the search procedure is also used to measure the redundancy between different features or groups of features.

The properties of a filter criterion are therefore essential because they determine the success of a good feature selection process. Understanding why a filter criterion is better for a specific dataset or less good for another one is also important in order to choose the best criterion for every situation.

The next section details some essential properties of a relevance criterion.

## 3 Properties of a Relevance Criterion for Feature Selection

This section introduces the important properties of a good relevance criterion for feature selection in regression. It also justifies why these properties are important. An analysis of these properties with respect to current filter criteria is realised in Sect. 5.

### 3.1 Property 1: Ability to Detect Nonlinear Relationships

A good relevance criterion should be able to detect nonlinear relationships between variables (features and target variables) [11,15]. This ability allows the criterion to detect the relevance between a group of features and the target, but also to detect the redundancy between features, even when the relationship is nonlinear, which is most generally the case with real datasets.

### 3.2 Property 2: Ability to Detect Multivariate Relationships

An efficient relevance criterion must be able to detect any relationship between two variables or, more importantly, between two groups of variables. Indeed, measuring the univariate relation between a single input feature and the target is not sufficient, as some features only contributing to the output when they are combined would not be detected (an obvious example of that phenomenon is a problem where the target is determined by the product of two features).

The necessity for a multivariate criterion is also a direct consequence of the use of greedy search procedures to find the most effective subset of features, such as forward and backward search, genetic algorithms, etc [6,16,24,27].

### 3.3 Property 3: Estimator Behaviour

Machine learning methods are always used on finite datasets. However, relevance criteria are generally defined in terms of integrals over the data space. In order to use them in practice, an estimator of the criteria, defined on a finite set of data, is needed. The computational complexity and the statistical properties of the estimators are important characteristics that should be taken into account when one needs to choose a criterion for selecting a reduced set of features [3].

### 3.4 Property 4: Estimator Parameters

In addition to the statistical properties of the estimators, the latter usually require to adjust a parameter whose influence on the quality of the estimation might be important. For example, nearest-neighbours based estimators require to choose the number of neighbours used in the estimation.

The choice of the parameters is sometimes underestimated in the literature, while in practice this choice may be crucial. Criteria whose estimators that do not rely on any parameter, or that rely on parameters having only low influence on the estimation, are therefore more appropriate.

### 3.5 Property 5: Estimator Behaviour in Small Sample Datasets

The ratio "number of instances/dimensionality" is a very important concept in all machine learning methods. A small sample dataset is a dataset with few instances with respect to the number of features. Many estimators do not work well with these datasets and need many instances to estimate correctly the relevance of features [3,5]. Unfortunately it is not always possible to collect more instances. Therefore, this property of behaving well in small sample scenarios is essential as well for the estimator. Section 5.5 analyses how the different relevance criteria behave in small sample situations.

### 3.6 Property 6: Invariant Estimator

Among the estimators of relevance criteria, some are not completely invariant to the gradient of the relation between the features and the target, especially in small sample scenarios. Depending on the scaling method or the normalisation method used during the process, the gradient of the relation may vary. However the importance of a feature, or a group of features, with respect to the target should not depend on this gradient. The consequence in feature selection could make a relevance criterion prefer a feature over another one only because of the gradient of their relation with the target, which should not happen.

Section 5.6 shows practically that some relevance criteria are influenced by this gradient of the function in small sample, and some are not.

## 4 Description of Three Popular Criteria

This section reviews three filter criteria, and their most frequently used estimators, in order to illustrate the strategic properties of a good relevance criterion as listed in Sect. 3.

### 4.1 Mutual Information

Mutual Information (MI) is a popular criterion for feature selection with filter methods [1,3,4,14,17,26]. It is a symmetric measure of the dependence between random variables (or sets of variables), based on entropy, introduced by Shannon in 1948 [25].

Let $X$ and $Y$ be two random variables, where $X$ represents the set of features and $Y$ the target. MI measures the reduction in the uncertainty on $Y$ when $X$ is known

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

where

$$H(Y) = -\int_Y p_Y(y) \log p_Y(y) dy \tag{2}$$

is the entropy of $Y$ and

$$H(Y|X) = \int_X p_X(x) H(Y|X = x) dx \tag{3}$$

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 $I(X;Y) = H(Y)$.

In practice, MI cannot be directly computed because it is defined in terms of probability density functions. These probability density functions are unknown when only a finite sample of data is available. Therefore, MI has to be estimated from the dataset [13]. The estimator introduced by Kraskov et al. [21] is based on a $k$-nearest neighbour method and results from the Kozachenko-Leonenko entropy estimator [20] $\hat{H}(X) = -\psi(k) + \psi(N) + \log c_d + \frac{d}{N} \sum_{i=1}^{N} \log \epsilon_k(i)$, where $k$ is the number of neighbours, $N$ is the number of instances in the dataset, $d$ is the dimensionality, $c_d = (2\pi^{\frac{d}{2}})/\Gamma(\frac{d}{2})$ is the volume of the unitary ball of dimension $d$, $\epsilon_k(i)$ is twice the distance from the $i^{th}$ instance to its $k^{th}$ nearest neighbour and $\psi$ is the digamma function. Kraskov estimator of the mutual information is then

$$\hat{I}(X;Y) = \psi(N) + \psi(K) - \frac{1}{k} - \frac{1}{N} \sum_{i=1}^{N} (\psi(\tau_x(i)) + \psi(\tau_y(i))) \tag{4}$$

where $\tau_x(i)$ is the number of points located no further than the distance $\epsilon_X(i,k)/2$ from the $i^{th}$ observation in the X space, $\tau_y(i)$ is the number of points located no further than the distance $\epsilon_Y(i,k)/2$ from the $i^{th}$ observation in the Y space and where $\epsilon_X(i,k)/2$ and $\epsilon_Y(i,k)/2$ are the projections into the X and Y subspaces of the distance between the $i^{th}$ observation and its $k^{th}$ neighbour.

When using MI for feature selection, the relationships between several subsets of features and the target Y are computed with a search procedure. Among these subsets, the one maximising the value of $\hat{I}(X;Y)$ (4) is selected.

### 4.2 Noise Variance

Noise variance is another popular relevance criterion, whose aim is to estimate the level of noise in a finite dataset. In the context of regression, the noise may be considered as the error in estimating the target as a function of the input features, under the hypothesis that a model could be built.

Let us consider a dataset with $N$ 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

$$y_i = f(\mathbf{x}_i) + \epsilon_i \text{ where } i = 1, ..., N \tag{5}$$

where $f$ is the unknown function between $\mathbf{x}_i$ and $y_i$, and $\epsilon_i$ is the noise, or prediction error, when estimating $f$. The principle is to select the subsets of features which lead to the lowest prediction error, or lowest noise variance [15].

In practice the noise variance has also to be estimated. One widely used estimator is the Delta Test [8,9,28]. The definition of the Delta Test $\delta$ is

$$\delta = \frac{1}{2N} \sum_{i=1}^{N} [y_{NN(i)} - y_i]^2 \tag{6}$$

where $y_{NN(i)}$ is the output associated to $x_{NN(i)}$, $x_{NN(i)}$ being the nearest neighbour of the point $x_i$.

For selecting features with the noise variance, the same procedure as for the mutual information can be used. But instead of selecting the group of features with the highest mutual information estimation $\hat{I}(X;Y)$, the search procedure selects the group of features with the lowest value of the noise variance estimator $\delta$ (6).

### 4.3 $R^2$ and Adjusted $R^2$

$R^2$, also called the coefficient of determination, is the proportion of the variance in the output variable that can be explained from the input variables; it ranges between 0% (unpredictable) and 100% (totally predictable). The definition of $R^2$ is

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \tag{7}$$

where $SS_{res} = \sum_i (y_i - f(\mathbf{x}_i))^2$ and $SS_{tot} = \sum_i (y_i - \overline{y})^2$ with i = 1,...,n. This coefficient is a statistical measure of how well the regression approximates the target. The $R^2$ measure automatically increases when features are added to the model. This is the reason why we use its alternative, Adjusted $R^2$, or $R^2_{adj}$, for feature selection in regression, more suitable for small sample sizes. Its definition is

$$R^2_{adj} = 1 - \frac{SS_{res}/(n-d-1)}{SS_{tot}/(n-1)} \tag{8}$$

where $d$ is the number of selected features in the model and $n$ the sample size. A low $R^2_{adj}$ indicates that the data are not close to the fitted regression line. 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. In practice, for each feature of the dataset, for each point of the function $f$, a linear regression is computed with a number of neighbours $k$ from 4 to $(n-1)$. The $R^2_{adj}$ is computed for every regression. For each value of $k$, an average of the $R^2_{adj}$ on every point of $f$ is computed. The best mean $R^2_{adj}$ is then selected; it corresponds to a specific number of neighbours $k$. The feature with the highest value of mean $R^2_{adj}$ is then selected. This is the univariate feature selection strategy, the first step of a search method. The multidimensional feature selection strategy can be implemented similarly.

## 5 Analysis and Comparison

This section analyses the six strategic properties of a relevance criterion for feature selection, given in Sect. 3. In the view of these properties, the mutual information, the noise variance and the Adjusted $R^2$ are compared.

### 5.1 Comparison with Property 1: Non-linearity

As explained in Sect. 3, a good relevance criterion must be able to detect non-linear relationships between variables. Mutual information and noise variance are both able to measure nonlinear relationships between variables. Intrinsically, $R^2_{adj}$ only estimates the quality of a linear regression. However, the method used with $R^2_{adj}$, described in Sect. 4.3, uses local approximations of the regression and is thus suitable for nonlinear relations between the features and the target.

This property is therefore non-discriminant for the three relevance criteria compared in this section. They can all be used with nonlinear relations between variables (features and target).

### 5.2 Comparison with Property 2: Multivariate Criterion

As shown in their respective equation, mutual information (1), noise variance (5) and $R^2_{adj}$ (8) can all be used to measure the relation between groups of features.

This property is therefore also non-discriminant for the three relevance criteria compared in this section.

### 5.3 Comparison with Property 3: Estimator Behaviour

The estimators of the three relevance criteria compared in this paper are all based on a $k$-nearest neighbour method. Therefore, this property is non-discriminant for them, because the time-complexity is approximately the same.

On the other hand, these estimators behave differently in small sample. This is discussed in Sect. 5.5.

### 5.4 Comparison with Property 4: Estimator Parameters

As explained in Sect. 4, the estimators of the three relevance criteria are all based on a $k$-nearest neighbour method. Nonetheless, this method is applied differently for each estimator.

The Kraskov estimator has only this $k$ parameter to adjust. Usually it is set to a number between 6 and 8 for good results [5,21]. The Delta Test sets by definition its $k$ to 1 [8,28]. Therefore, this estimator does not have any parameter to adjust. With Adjusted $R^2$, the range of $k$ is much larger, depending on the size of the dataset and the variables.

In view of this property, Adjusted $R^2$ has the most complex $k$ parameter to adjust and the Delta Test is the easiest to adjust.

### 5.5 Comparison with Property 5: Estimator in Small Sample

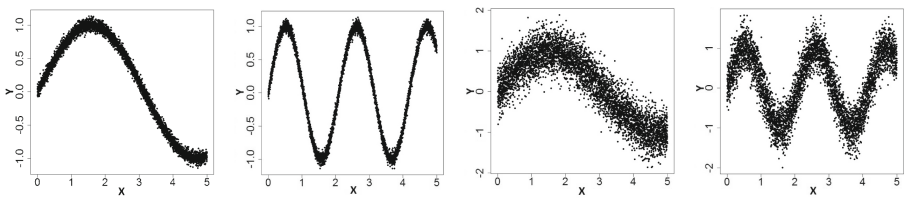
As discussed in [5], Kraskov estimator and the Delta Test suffer from a bias when comparing smooth and non-smooth features, especially in small sample. An overestimation of the noise variance and an underestimation of the mutual information can occur in small datasets when the function to estimate is not smooth. The biases in the estimations are much more severe when using mutual information than when using the noise variance [5,8].

Adjusted $R^2$ also underestimates non-smooth functions in small datasets and behaves approximately as the Delta Test, in the sense that the minimal size needed to estimate correctly the same nonlinear relation is approximately the same for Delta Test than for Adjusted $R^2$.

To illustrate this behaviour, experiments have been performed on simple synthetic datasets. Four different periodic functions have been generated with two different frequencies and two levels of noise :

$$\begin{aligned} y_1 = f_1(\mathbf{x}) = sin(\mathbf{x}) + \epsilon \quad & \text{where } \epsilon \sim \mathrm{N}(0,0.05) \\ y_2 = f_2(\mathbf{x}) = sin(3\mathbf{x}) + \epsilon \quad & \text{where } \epsilon \sim \mathrm{N}(0,0.05) \\ y_3 = f_3(\mathbf{x}) = sin(\mathbf{x}) + \epsilon \quad & \text{where } \epsilon \sim \mathrm{N}(0,0.3) \\ y_4 = f_4(\mathbf{x}) = sin(3\mathbf{x}) + \epsilon \quad & \text{where } \epsilon \sim \mathrm{N}(0,0.3) \end{aligned} \tag{9}$$

Figures 1(a), (b), (c), (d) represent the four functions $f_1$, $f_2$, $f_3$, $f_4$, respectively.



(a) $f_1(\mathbf{x}) = sin(\mathbf{x}) + \epsilon \sim N(0, 0.05)$. (b) $f_2(\mathbf{x}) = sin(3\mathbf{x}) + \epsilon \sim N(0, 0.05)$. (c) $f_3(\mathbf{x}) = sin(\mathbf{x}) + \epsilon \sim N(0, 0.3)$. (d) $f_4(\mathbf{x}) = sin(3\mathbf{x}) + \epsilon \sim N(0, 0.3)$

**Fig. 1.** Experimental data generated with two different frequencies and two levels of noise.

Results are presented in Fig. 2. Figure 2(a) shows that the mutual information underestimates the non-smooth function $f_2$ (lower level of noise) over the smooth function $f_3$ (higher level of noise). Figures 2(b) and (c) show, respectively, that the Delta Test and the Adjusted $R^2$ overestimate the non-smooth function $f_2$ over the smooth function $f_3$. Figure 2 also shows that the Delta Test and Adjusted $R^2$ converge quickly than the mutual information.

### 5.6 Comparison with Property 6: Estimator Stability

In order to study the estimator stability with respect to the gradient of the relation between the features and the target, illustrative experiments performed

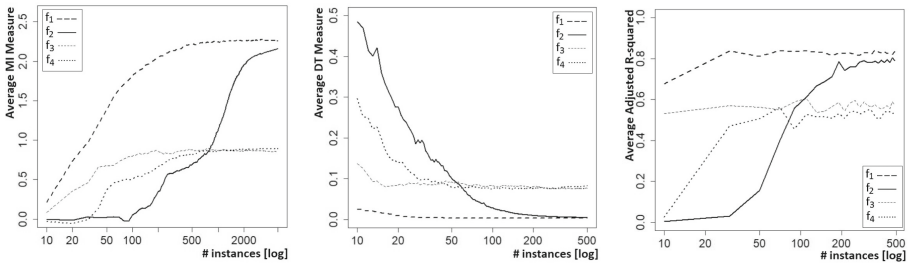



**Fig. 2.** Average values of (a) MI measures, (b) Delta Test, (c) Adjusted $R^2$, for two functions with a low level of noise and for two functions with a higher level of noise.

in this paper consider three linear functions with various slopes (Fig. 3). These illustrative experiments are conducted to show the importance of the estimator stability with respect to this gradient and to compare the three criteria described in Sect. 4.

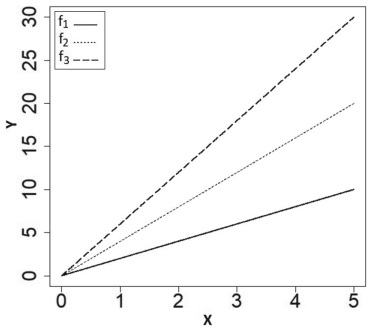
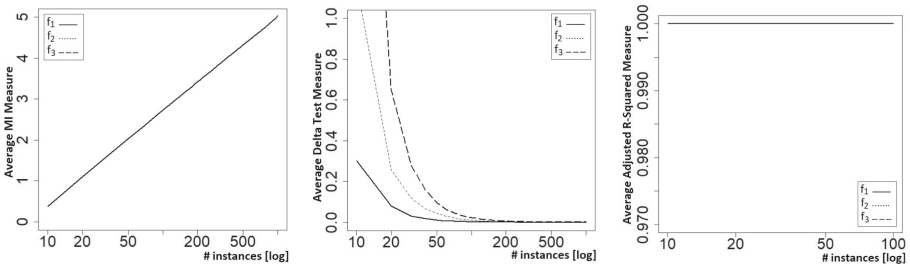



**Fig. 3.** Experimental data generated with three different slopes.

They have been performed with various sizes of samples, from extremely small to large ones. For each size of the sample, an estimator of the three decision criteria has been used. Results are shown in Fig. 4. The mutual information shows the same results for the three different slopes (Fig. 4(a)), the three measures are superposed, which means that there is no influence of the function slope on its result. The Delta Test shows an influence of the slope of the results in small datasets (Fig. 4(b)). This influence tends to disappear when the size of the datasets sufficiently increases. Adjusted $R^2$ shows (Fig. 4(c)) no influence of the function slope, the three functions are also superposed, even in small sample scenarios.

For this property, Adjusted $R^2$ and the mutual information behave better than the Delta Test, in the sense that they offer the same value for the three different functions $f_1$, $f_2$ and $f_3$.




(a) (b) (c)

**Fig. 4.** Average values of MI measures 4(a), Delta Test measures 4(b) and Adjusted $R^2$ measures 4(c) for three functions $f_1$, $f_2$ and $f_3$, with three different slopes (see Fig. 3).

### 5.7 Discussion

Table 1 shows a summary of the comparison realised for the three relevance criteria with respect to the six properties presented in this paper.

**Table 1.** Comparison of the mutual information with Kraskov, the Delta Test and the adjusted $R^2$. A '+' indicates a good behaviour of the criterion towards this property. A '-' indicates a weakness of the criterion towards this property. The signs '++' or '- -' are only there to show a difference between two criteria with a good (or bad) behaviour towards the property, when one of them is better (or worse) than the other one.

| Properties | MI with Kraskov | Noise variance with DT | Adjusted $R^2$ |
|---|---|---|---|
| P1: Non-linearity | + | + | + |
| P2: Multivariate | + | + | + |
| P3: Estimator Behaviour | + | + | + |
| P4: Estimator Parameters | + | + + | - |
| P5: Estimator in Small Sample | - - | - | - |
| P6: Estimator Stability | + | - | + |

The three filter criteria proposed in Sect. 4 all respect the two first properties, which make them good candidates for feature selection. The four last properties help to decide between the three criteria, depending on the dataset and the problem at hand. Indeed when comparing the three criteria with the fourth property (P4), the Delta Test does not have any parameter to adjust, which makes it easier to use with respect to the mutual information and the adjusted $R^2$. In a small sample scenario (P5), the adjusted $R^2$ seems to behave as the Delta Test for non-smooth functions, which is better than the mutual information. Finally when comparing the criteria with the sixth property (P6), the adjusted $R^2$ and the mutual information are more stable with respect to the gradient of the function between the features and the target than the Delta Test.

## 6 Conclusions

This paper proposes six strategic properties of a good relevance criterion for feature selection in regression: two properties for the relevance criterion itself and four properties for the estimator of the relevance criterion. To illustrate the importance of these properties, this paper describes three interesting relevance criteria and compares them with the aforementioned properties. Any relevance criterion used for filters in feature selection could be analysed in the light of these properties.

## References

1. Battiti, R.: Using mutual information for selecting features in supervised neural net learning. IEEE Trans. Neural Netw. **5**, 537–550 (1994)
2. Bing, X., Mengjie, Z., Will, N., B., Xin, Y.: A survey on evolutionary computation approaches to feature selection. IEEE Trans. Evol. Comput. 20(4), 606–626 (2016)
3. Brown, G., Pocock, A., Zhao, M., Lujan, M.: Conditional likelihood maximisation: a unifying framework for mutual information feature selection. J. Mach. Learn. Res. **13**, 27–66 (2012)
4. Degeest, A., Verleysen, M., Frénay, B.: Feature ranking in changing environments where new features are introduced. In: 2015 International Joint Conference on Neural Networks (IJCNN), pp. 1–8, July 2015
5. Degeest, A., Verleysen, M., Frénay, B.: Smoothness bias in relevance estimators for feature selection in regression. In: 2018 International Conference on Artificial Intelligence Applications and Innovations (IJCNN), pp. 285–294 (2018)
6. Doquire, G., Verleysen, M.: A comparison of multivariate mutual information estimators for feature selection. In: Proceeding of ICPRAM 2012 (2012)
7. Efron, B., Hastie, T., Johnstone, I., Tibshirani, R.: Least angle regression. Ann. Stat. **32**, 407–499 (2004)
8. Eirola, E., Lendasse, A., Corona, F., Verleysen, M.: The delta test: the 1-nn estimator as a feature selection criterion. In: Proceedings of the 2014 International Joint Conference on Neural Networks (IJCNN), pp. 4214–4222, July 2014
9. Eirola, E., Liitiäinen, E., Lendasse, A., Corona, F., Verleysen, M.: Using the delta test for variable selection. In: Proceedings of ESANN 2008 (2008)
10. François, D., Rossi, F., Wertz, V., Verleysen, M.: Resampling methods for parameter-free and robust feature selection with mutual information. Neurocomputing **70**(7–9), 1276–1288 (2007)
11. Frénay, B., Doquire, G., Verleysen, M.: Theoretical and empirical study on the potential inadequacy of mutual information for feature selection in classification. Neurocomputing **112**, 64–78 (2013)
12. Frénay, B., van Heeswijk, M., Miche, Y., Verleysen, M., Lendasse, A.: Feature selection for nonlinear models with extreme learning machines. Neurocomputing **102**, 111–124 (2013)
13. Gao, W., Kannan, S., Oh, S., Viswanath, P.: Estimating mutual information for discrete-continuous mixtures. In: Guyon, I., et al. (eds.) Advances in Neural Information Processing Systems, vol. 30, pp. 5986–5997. Curran Associates Inc, Red Hook (2017)

14. Gómez-Verdejo, V., Verleysen, M., Fleury, J.: Information-theoretic feature selection for functional data classification. Neurocomputing **72**(16–18), 3580–3589 (2009)
15. Guillén, A., Sovilj, D., Mateo, F., Rojas, I., Lendasse, A.: New methodologies based on delta test for variable selection in regression problems. In: Workshop on Parallel Architectures and Bioinspired Algorithms, Toronto, Canada (2008)
16. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. J. Mach. Learn. Res. **3**, 1157–1182 (2003)
17. Hancer, E., Xue, B., Zhang, M.: Differential evolution for filter feature selection based on information theory and feature ranking. Knowl.-Based Syst. **140**, 103–119 (2018)
18. Karegowda, A.G., Jayaram, M.A., Manjunath, A.S.: Feature subset selection problem using wrapper approach in supervised learning. Int. J. Comput. Appl. **1**(7), 13–17 (2010)
19. Kohavi, R., John, G.H.: Wrappers for feature subset selection. Artif. Intell. **97**, 273–324 (1997)
20. Kozachenko, L.F., Leonenko, N.: Sample estimate of the entropy of a random vector. Probl. Inform. Transm. **23**, 95–101 (1987)
21. Kraskov, A., Stögbauer, H., Grassberger, P.: Estimating mutual information. Phys. Rev. E **69**, 066138 (2004)
22. Li, J., et al.: Feature selection: a data perspective. ACM Comput. Surv. **50**(6), 94:1–94:45 (2017). https://doi.org/10.1145/3136625
23. Paul, J., D'Ambrosio, R., Dupont, P.: Kernel methods for heterogeneous feature selection. Neurocomputing **169**, 187–195 (2015)
24. Schaffernicht, E., Kaltenhaeuser, R., Verma, S.S., Gross, H.-M.: On estimating mutual information for feature selection. In: Diamantaras, K., Duch, W., Iliadis, L.S. (eds.) ICANN 2010. LNCS, vol. 6352, pp. 362–367. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15819-3_48
25. Shannon, C.E.: A mathematical theory of communication. Bell Syst. Tech. J. **27**(379–423), 623–656 (1948)
26. Vergara, J.R., Estévez, P.A.: A review of feature selection methods based on mutual information. Neural Comput. Appl. **24**, 175–186 (2014)
27. Verleysen, M., Rossi, F., François, D.: Advances in feature selection with mutual information. In: Biehl, M., Hammer, B., Verleysen, M., Villmann, T. (eds.) Similarity-Based Clustering. LNCS (LNAI), vol. 5400, pp. 52–69. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-01805-3_4
28. Yu, Q., Séverin, E., Lendasse, A.: Variable selection for financial modeling. In: Proceedings of the CEF 2007, 13th International Conference on Computing in Economics and Finance, Montréal, Quebec, Canada, pp. 237–241 (2007)