Comparison Between 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] Machine Learning Group - ICTEAM, UCLouvain, 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. High-dimensional data are ubiquitous in regression. To obtain a better understanding of the data or to ease the learning process, reducing the data to a subset of the most relevant features is important. Among the different methods of feature selection, filter methods are popular because they are independent from the model, which makes them fast and computationally simpler than other feature selection methods. The key factor of a filter method is the filter criterion. This paper focuses on which properties make a good filter criterion, in order to be able to select one from the numerous existing ones. Six properties are discussed, and three filter criteria are compared with respect to the aforementioned properties.

Keywords: Feature selection · Filter criteria · Regression

1 Introduction

In regression problems, datasets are often high-dimensional. Therefore, selecting a reduced subset of the most relevant features is an important challenge for data analysis, to help fighting the curse of dimensionality, to ease the learning process, to increase interpretability of the original data, etc. Many works, in diverse domains such as healthcare, spam detection or marketing, focus on methods reducing the original subset of features in datasets [9,13,22,26,28]. Feature selection can roughly be classified into filters, wrappers and embedded methods. This paper focuses on filter methods, which have the advantage to be fast thanks to their independence from the model used. Indeed filters do not need to train any model during the selection process, contrarily to wrappers [16,17] or embedded methods [20]. They also lead to a selection of features that is independent from the models that will be subsequently used. To select the subset of the

I. V. Tetko et al. (Eds.): ICANN 2019, LNCS 11728, pp. 59–71, 2019.
https://doi.org/10.1007/978-3-030-30484-3_5

most relevant features for the problem at hand, filter methods need a sound relevance criterion. The choice of this criterion is therefore strategic for a successful filter-based feature selection process. The problem is that there exist many filter criteria with various properties. Among these criteria, this paper does not intend to present the best criterion to be used on every dataset but focuses instead on the essential properties of what makes a good filter criterion for feature selection in regression, and what are its strengths or weaknesses with respect to these properties. These properties are described and discussed in Sect. 3, after introducing feature selection with filters in Sect. 2. Several interesting filter criteria are described in Sect. 4 and compared in the light of these properties in Sect. 5. Finally, conclusions on these comparisons are given in Sect. 6.

2 Feature Selection with Filter Criteria

Feature selection helps to reduce the dimension of a dataset by removing redundant or less useful features. In the context of filter methods for feature selection, a good relevance criterion must first be selected. The relevance criterion measures the importance of the relationship between a feature, or a group of features, and the target. The target is the output variable depending on the problem at hand. This criterion is therefore the key factor of a successful dimension reduction process with filter methods.

Filter criteria are very diverse: for example, some are based on entropy, such as the mutual information (Sect. 4.1), while some are based on noise variance (Sect. 4.2) or regression quality (Sect. 4.3). The question raised by this paper is: "What is needed in a good filter criterion in order to obtain the subset of the most relevant features for the problem at hand?" These properties are listed in Sect. 3 and used for comparison of several relevance criteria in Sect. 5.

A complete feature selection process needs mainly two ingredients: a filter criterion, as already discussed, but also a search procedure, in order to find the best subset of features among all possible subsets (whose number grows exponentially with the number of initial features), without having to compute the filter criterion between all subsets of variables and the target. Several methods exist, such as forward search, greedy search, genetic algorithms, mRMR [23], etc. During this search procedure, the chosen relevance criterion is, again, strategic. For example, those search methods need a multivariate criterion to be able to measure the relevance between two features but also between a subset of features and the target. Indeed it is often believed that filters use univariate criteria only; this is not correct: several multivariate filter criteria exist, such as the mutual information (Sect. 4.1) and the noise variance (Sect. 4.2). This property is discussed in Sect. 3.2 and is used for comparison in Sect. 5.2.

3 Properties of a Relevance Criterion for Feature Selection in Regression

The choice of the relevance criterion used for feature selection, among the various existing ones, is strategic. This section describes the six essential properties of

a good relevance criterion used as a filter during a feature selection process in regression. It also discusses the reasons why these properties are important. In order to demonstrate how to use these six properties, a comparison of the filter criteria presented in Sect. 4, with respect to these properties, is realised in Sect. 5.

3.1 Property 1: Ability to Detect Nonlinear Relationships

In real datasets, the relationship between variables is most generally nonlinear. Therefore, a good filter criterion for feature selection in regression should be able to detect nonlinear relations between variables, or groups of variables [11,14].

3.2 Property 2: Ability to Detect Multivariate Relationships

When looking for the best subset of features, a search procedure, such as a forward search, needs to evaluate the relations between groups of two, or more, features and the target. Indeed, measuring the relationship between one input feature and the target is not enough, as some features only contribute to the target when combined. An example of that would be when the target is determined by the product of two features. To achieve this multidimensional measurement, the filter criterion needs to be able to measure multivariate relations between groups of variables. This property is essential for feature selection methods with a filter criterion [6,24].

3.3 Property 3: Estimator Behaviour

Most filter criteria are defined as a statistical property of the data integrated over the domain space. The criteria are repeatedly evaluated at each step of the search procedure. However, the criteria themselves can usually not be evaluated exactly because of the integration over the domain space. An estimator is thus needed, whose computational complexity and statistical properties are therefore important and must be taken into account when choosing a criterion [3].

3.4 Property 4: Estimator Parameters

Filter criteria estimator often require an optimization parameter to be tuned. The influence of this parameter on the quality of the estimation, and therefore on the quality of the feature selection itself, might be important. For example, nearest-neighbours based estimators, such as the Kraskov estimator (see Sect. 4.1), need to choose the number of neighbours used in the estimation. This choice of parameter might be crucial, while its influence is sometimes underestimated in the literature. When comparing different estimators, the stability of the estimator with respect to an optimization parameter is an important property to take into account. This property is discussed in Sect. 5.4.

3.5 Property 5: Estimator Behaviour in Small Datasets

Data are ubiquitous. In many fields the number of data that can be collected might be huge, leading to the so-called "big data". In other fields however the number of data remains limited (for example because of collection costs), while the number of dimensions (features) might be large for every data [15]. The ratio "number of instances/dimensionality" is therefore an important concept in all machine learning methods. A small sample dataset is one with a low ratio "number of instances/dimensionality". When working with a small sample dataset, the estimator of the filter criterion needs to behave well to select the best subset of features. This property is discussed, and filter criteria are compared with respect to it, in Sect. 5.5, for small sample situations.

3.6 Property 6: Invariant Estimator

When comparing features or groups of features having a linear, or quasi-linear, relationship, some estimators evaluate differently linear relationships with different gradients of the relationship between the features and the target. However, the importance of a feature or a group of features should not depend on this gradient but only on the predictability, i.e. the quality of the information in the features used to predict the output. In feature selection, this would have the consequence to prefer some features with a greater gradient than features with a lower gradient. The independence of the estimator towards this gradient is discussed in Sect. 5.6.

4 Description of Three Popular Criteria

This section reviews three filter criteria, the mutual information, the noise variance and the adjusted coefficient of determination, with their most frequently used estimators, in order to illustrate the important properties of a good filter criterion presented in Sect. 3.

4.1 Mutual Information

Mutual information (MI) is a frequently used criterion for filter methods, in regression or classification [1,2,13,21,27]. Based on entropy, it is a symmetric measurement of the dependence between random variables, introduced by Shannon in 1948 [25]. It evaluates the importance of the relationship between a feature, or a group of features, and another one. It has been shown to be an efficient criterion to select relevant features in classification [11,23] and in regression [10,14]. This paper focuses on regression problems.

Let X and Y be two random variables, whose respective probability density functions are p_X and p_Y, and where X represents the 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 $H(Y) = -\int_Y p_Y(y) \log p_Y(y) dy$ is the entropy of Y and $H(Y|X) = \int_X p_X(x) H(Y|X = x) dx$ 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 $I(X;Y) = H(Y)$.

With real datasets, MI 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, MI has to be estimated from the dataset [12]. The estimator introduced by Kraskov et al. [19] is based on a k-nearest neighbour method and results from the Kozachenko-Leonenko entropy estimator [18] $\hat{H}(X) = -\psi(k) + \psi(N) + \log c_d + \frac{d}{N} \sum_{i=1}^{N} \log \epsilon_k(i)$, 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, Γ is the gamma function, $\epsilon_k(i)$ is twice the distance from the i^{th} instance to its k^{th} nearest neighbour and ψ is the digamma function. The Kraskov estimator of the mutual information is

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

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 the estimated mutual information $\hat{I}(X;Y)$ (2) is selected.

4.2 Noise Variance

Noise variance is another popular filter criterion used for feature selection. Its aim is to evaluate the level of noise in a finite dataset. In regression, the noise represents the error in estimating the output variable as function of the input variables, under the hypothesis that a model could be built (by a machine learning regression model).

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

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

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. The principle is to select the subsets of features X_j which lead to the lowest prediction error, or lowest noise variance [14].

With real finite datasets, the noise variance has to be estimated, e.g. with the Delta Test, which is a widely used estimator [7,8]. The Delta Test δ is defined as

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

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

For feature selection using the noise variance, and therefore the Delta Test estimator, the relationships between several subsets of features and Y are computed, again with a search procedure such as a greedy search. The search procedure selects the subset of features with the lowest value of the estimator δ.

4.3 Coefficient of Determination

The coefficient of determination R^2 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

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \tag{5}$$

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, and with $\overline{y}$ being the mean of the observed data. This coefficient statistically measures how well the regression approximates the target. Because R^2 automatically increases when features are added to the model, we use its alternative, the adjusted R^2, or R^2_{adj}, for feature selection in regression, which is more suitable for small sample sizes. Its definition is

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

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 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]. In practice, for each point of the function f, a linear regression is computed with a number of neighbours k starting from 4. 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 first step of the search procedure, the univariate step, selects the feature with the highest mean R^2_{adj} [5]. The multidimensional feature selection strategy, where the group of features with the highest mean R^2_{adj} is selected, is implemented similarly, but instead of evaluating the regression in two dimensions, R^2_{adj} evaluates the fitness of the local, multivariate regressions in three, four, five dimensions, depending on the step of the search process.

5 Analysis and Comparison

This section analyses and compares, in the context of feature selection in regression, the mutual information, the noise variance and the coefficient of determination (described in Sect. 4), with respect to the six properties listed and discussed in Sect. 3.

5.1 Comparison with Property 1: Non-linearity

The three filter criteria described in this paper can evaluate a nonlinear relationship between a group of features and the target. For the mutual information and the noise variance, it is intrinsically true. For R^2_{adj}, the implementation described in Sect. 4.3 uses local approximations of the regression and is therefore suitable for nonlinear relations between the features and the target as well. This important property is therefore non-discriminant for the three filter criteria compared in this section. This paper does not consider dependencies between features X_j and focuses on evaluating their interest for predicting a given target. Several approaches have been proposed to deal with redundancy in feature subsets, like e.g. mRMR [23].

5.2 Comparison with Property 2: Multivariate Criterion

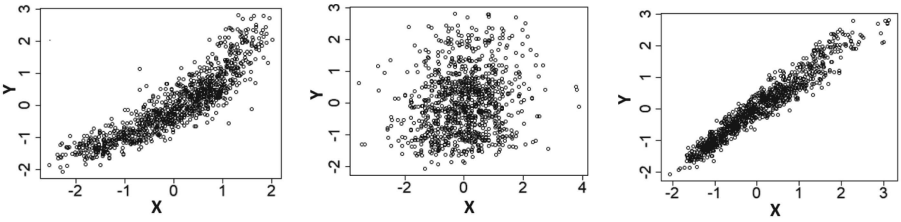
The three filter criteria compared in this section are all able to evaluate the relationship between a group of features and the target for relevance, and between groups of features for redundancy, as shown in their respective equations (see Sect. 4). This second property is therefore also non-discriminant. This shows that these three criteria are all worth considering when doing feature selection.

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, their time complexity and their statistical properties are similar and less discriminant. Their respective behaviour during a forward search method for selecting features with small sample datasets is discussed in Sect. 5.5.

5.4 Comparison with Property 4: Estimator Parameters

Being all based on a k-nearest-neighbour method, the three estimators of the compared criteria all have one optimization parameter k. For the Kraskov estimator of the mutual information, this parameter is usually set between 6 and 8 [19]. For the Delta Test, the k parameter is, by definition, set to 1 and therefore does not need to be modified. For the mean R^2_{adj}, k is usually set higher than for the Kraskov estimator, in order to obtain stable results during the feature selection process.

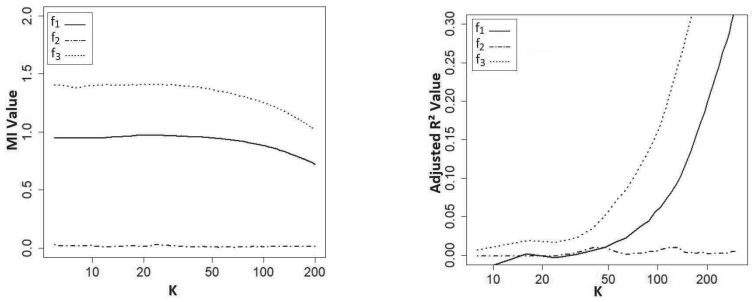


(a) Feature f_1. (b) Feature f_2. (c) Feature f_3.

Fig. 1. Three features of the dataset *Anthrokids*, where f_1 is the 1^{st} feature, f_2 is the 11^{th} feature and f_3 is the 19^{th} feature of the dataset.

In order to evaluate the relation between the value of k and the value of the filter criterion, and in order to understand how to select k when using these estimators, the first step of a feature selection method has been performed on 3 features of the *Anthrokids*[1] dataset (Fig. 1), with a value of k neighbours ranging from 5 to approximately 200, both for the mutual information and the noise variance. Let us remark that this experiment has not been performed for the Delta Test, the value of k being set to 1 by definition.

Results are shown in Fig. 2: The MI value for the 3 features f_1 to f_3 in Fig. 2(a) and the mean R^2_{adj} for the same features in Fig. 2(b). Figure 2(a) shows a good stability of the mutual information estimator towards the value of k for these features. Figure 2(b) shows that the value of the highest mean R^2_{adj} is less stable than the values of the mutual information. But for a value of k higher than 50, the feature selection results are stable for R^2_{adj} as well, as the ranking of the features does not change for a k between 50 and 200, the feature f_1 being the first to be selected and the feature f_3 being the last to be selected by both criteria.



(a) (b)

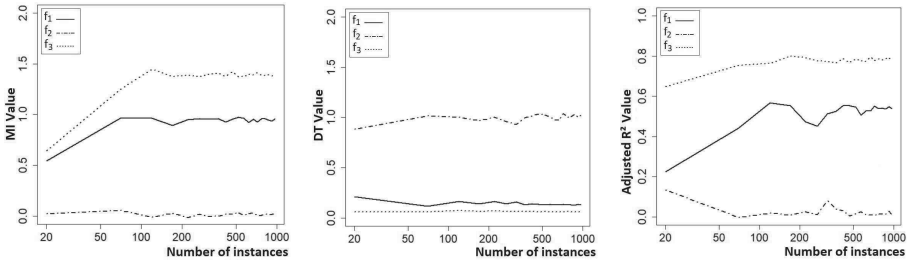
Fig. 2. Evolution of (2(a)) MI values and (2(b)) R^2_{adj} values for three features f_1, f_2 and f_3 (corresponding respectively to the features 1, 11 and 19 of the dataset *Anthrokids* presented in Fig. 1) when k varies.

[1] https://research.cs.aalto.fi/aml/datasets.shtml.

5.5 Comparison with Property 5: Estimator Behaviour in Small Sample Datasets

Methods such as k-nearest-neighbours methods can suffer from a bias when comparing smooth and non-smooth features, especially in small sample [4,5]. However, the biases in the estimations are much more severe with the mutual information than with the noise variance or the adjusted R^2 [4,7].

Experiments have been performed on the real-world dataset *Anthrokids*. A forward search has been realised with three features of this dataset (Fig. 1). Results are presented in Fig. 3. The best feature selected in Fig. 3(a) is the one maximizing the mutual information value. In Fig. 3(b), it is the one minimizing the noise variance. And in Fig. 3(c), it is the one maximizing the adjusted R^2. For the three criteria, the first feature selected is the feature f_3. Results also show that the value of the mutual information and of the adjusted R^2 stabilises itself around a hundred instances (Figs. 3(a) and (c)). The values of the Delta Test measure are more stable, even in small sample. When comparing Figs. 3(a), (b) and (c), the adjusted R^2 measure and the MI measure presents the advantage to distinguish better the values between the features f_1 and f_3. For the Delta Test, these values are almost the same, contrarily to the adjusted R^2.



(a) (b) (c)

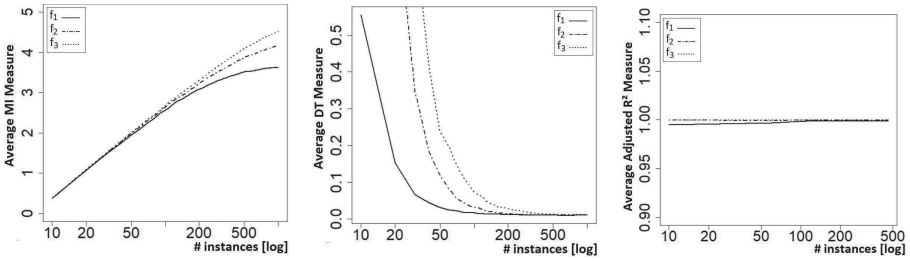
Fig. 3. Average values of (3(a)) MI measures, (3(b)) Delta Test measures and (3(c)) adjusted R^2 measures for three features of the dataset *Anthrokids*.

5.6 Comparison with Property 6: Estimator Invariance

In order to analyse the invariance of the estimator with respect to the gradient of the relation between a feature, or a group of features, and the target, an illustrative experiment has been realised on three linear functions with various slopes. Three different linear functions have been generated:

$$\begin{aligned} y_1 &= f_1(\mathbf{x}) = 2x_1 + x_2 + \epsilon \quad \text{where } \epsilon \sim \mathrm{N}(0,\, 0.1) \\ y_2 &= f_2(\mathbf{x}) = 4x_1 + 2x_2 + \epsilon \quad \text{where } \epsilon \sim \mathrm{N}(0,\, 0.1) \\ y_3 &= f_3(\mathbf{x}) = 6x_1 + 4x_2 + \epsilon \quad \text{where } \epsilon \sim \mathrm{N}(0,\, 0.1) \end{aligned} \tag{7}$$

Experiments have been performed with various sizes of samples, from small ones to larger ones. For each size, an average value of the estimators of the three criteria have been computed. The results of this experiment are presented in Fig. 4. The Delta Test (Fig. 4(b)) and the mutual information (Fig. 4(a)) show an influence of the gradient on the results. This influence disappears with a larger size of sample for the Delta Test. R^2_{adj} (Fig. 4(c)) shows no influence of the function gradient on the results as the three functions are almost superposed.



(a) (b) (c)

Fig. 4. Average value of (4(a)) the mutual information, (4(b)) the Delta Test and (4(c)) the adjusted R^2 for the three linear functions with different slopes described in (7).

For this property, the adjusted R^2 behaves better than the Delta Test and the mutual information in the sense that they are independent from the slope of the functions and they obtain the same value for the three functions f_1, f_2 and f_3. The Delta Test behaves better than the mutual information since the bias tends to disappear in larger samples.

5.7 Comparison Summary

Table 1 shows a summary of the comparison of the three filter criteria with respect to the six properties presented in Sect. 3, in a context of feature selection in regression. When analysing the ease of parameter optimization (P4), the Delta Test performs better than the mutual information and the adjusted R^2 because its only parameter k is set to 1 by definition. In the small sample scenario presented in this paper (P5), the Delta Test behaves better than the two other criteria. For the estimator invariance (P6), the adjusted R^2 behaves better than the Delta Test and the mutual information in the sense that this criterion is not influenced by the slope of the function between the features and the target.

Table 1. Comparison of mutual information with Kraskov, Delta Test and 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 Invariance	−	−	+

6 Conclusions

This paper focuses on which properties make a good filter criterion for feature selection in regression. To illustrate the importance of these properties, it compares three filter criteria, with artificial and real datasets, to show their respective strengths and weaknesses with respect to each other. These properties could be used to analyse any other filter criterion used for feature selection or they could be used in situations where there is redundancy like mRMR or missing values in feature subsets.

References

1. Battiti, R.: Using mutual information for selecting features in supervised neural net learning. IEEE Trans. Neural Netw. **5**, 537–550 (1994). https://doi.org/10.1109/72.298224
2. 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)
3. Chandrashekar, G., Sahin, F.: A survey on feature selection methods. Comput. Electr. Eng. **40**, 16–28 (2014). https://doi.org/10.1016/j.compeleceng.2013.11.024
4. Degeest, A., Verleysen, M., Frénay, B.: Smoothness bias in relevance estimators for feature selection in regression. In: Proceedings of AIAI, pp. 285–294 (2018). https://doi.org/10.1007/978-3-319-92007-825
5. Degeest, A., Verleysen, M., Frénay, B.: About filter criteria for feature selection in regression. In: Rojas, I., Joya, G., Catala, A. (eds.) IWANN 2019. LNCS, vol. 11507, pp. 579–590. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-20518-8_48
6. Doquire, G., Verleysen, M.: A comparison of multivariate mutual information estimators for feature selection. In: Proceedings of ICPRAM (2012). https://doi.org/10.5220/0003726101760185
7. Eirola, E., Lendasse, A., Corona, F., Verleysen, M.: The delta test: The 1-NN estimator as a feature selection criterion. In: Proceedings of IJCNN, pp. 4214–4222 (2014). https://doi.org/10.1109/IJCNN.2014.6889560

8. Eirola, E., Liitiäinen, E., Lendasse, A., Corona, F., Verleysen, M.: Using the delta test for variable selection. In: Proceedings of ESANN (2008)
9. 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). https://doi.org/10.1016/j.neucom.2006.11.019
10. Frénay, B., Doquire, G., Verleysen, M.: Is mutual information adequate for feature selection in regression? Neural Netw. **48**, 1–7 (2013). https://doi.org/10.1016/j.neunet.2013.07.003
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). https://doi.org/10.1016/j.neucom.2012.12.051
12. Gao, W., Kannan, S., Oh, S., Viswanath, P.: Estimating mutual information for discrete-continuous mixtures. In: Advances in Neural Information Processing Systems, vol. 30, pp. 5986–5997. Curran Associates, Inc. (2017). http://arxiv.org/abs/1709.06212
13. Gómez-Verdejo, V., Verleysen, M., Fleury, J.: Information-theoretic feature selection for functional data classification. Neurocomputing **72**, 3580–3589 (2009). https://doi.org/10.1016/j.neucom.2008.12.035
14. 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, Canada (2008)
15. Helleputte, T., Dupont, P.: Feature selection by transfer learning with linear regularized models. In: Buntine, W., Grobelnik, M., Mladenić, D., Shawe-Taylor, J. (eds.) ECML PKDD 2009. LNCS (LNAI), vol. 5781, pp. 533–547. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-04180-8_52
16. 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). https://doi.org/10.5120/169-295
17. Kohavi, R., John, G.H.: Wrappers for feature subset selection. Artif. Intell. **97**, 273–324 (1997). https://doi.org/10.1016/S0004-3702(97)00043-X
18. Kozachenko, L.F., Leonenko, N.: Sample estimate of the entropy of a random vector. Prob. Inform. Trans. **23**, 95–101 (1987)
19. Kraskov, A., Stögbauer, H., Grassberger, P.: Estimating mutual information. Phys. Rev. E **69**, 066138 (2004). https://doi.org/10.1103/PhysRevE.69.066138
20. Lal, T.N., Chapelle, O., Weston, J., Elisseeff, A.: Embedded methods. In: Guyon, I., Nikravesh, M., Gunn, S., Zadeh, L.A. (eds.) Feature Extraction. Studies in Fuzziness and Soft Computing, vol. 207, pp. 137–165. Springer, Heidelberg (2006). https://doi.org/10.1007/978-3-540-35488-86
21. Liu, A., Jun, G., Ghosh, J.: A self-training approach to cost sensitive uncertainty sampling. In: Buntine, W., Grobelnik, M., Mladenić, D., Shawe-Taylor, J. (eds.) ECML PKDD 2009. LNCS (LNAI), vol. 5781, pp. 10–10. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-04180-8_10
22. Paul, J., D'Ambrosio, R., Dupont, P.: Kernel methods for heterogeneous feature selection. Neurocomputing **169**, 187–195 (2015). https://doi.org/10.1016/j.neucom.2014.12.098
23. Peng, H., Long, F., Ding, C.: Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans. Pattern Anal. Mach. Intell. **27**, 1226–1238 (2005). https://doi.org/10.1109/TPAMI.2005.159

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 (1948). https://doi.org/10.1002/j.1538-7305.1948.tb01338.x
26. Song, Q., Ni, J., Wang, G.: A fast clustering-based feature subset selection algorithm for high-dimensional data. IEEE Trans. Knowl. Data Eng. **25**(1), 1–14 (2013). https://doi.org/10.1109/TKDE.2011.181
27. Vergara, J.R., Estévez, P.A.: A review of feature selection methods based on mutual information. Neural Comput. Appl. **24**, 175–186 (2014). https://doi.org/10.1007/s00521-013-1368-0
28. Xue, B., Zhang, M., Browne, W., Yao, X.: A survey on evolutionary computation approaches to feature selection. IEEE Trans. Evol. Comput. **20** (2015). https://doi.org/10.1109/TEVC.2015.2504420