Smoothness Bias in Relevance Estimators for Feature Selection in Regression

Alexandra Degeest, Michel Verleysen, Benoît Frenay

Smoothness Bias in Relevance Estimators 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]Université catholique de Louvain, Machine Learning Group - ICTEAM
Place du Levant 3 - 1348 Louvain-La-Neuve, Belgium,
michel.verleysen@uclouvain.be,
[3]Université de Namur - Faculty of Computer Science,
NADI Institute - PReCISE Research Center
Rue Grandgagnage 21 - 5000 Namur, Belgium,
benoit.frenay@unamur.be

Abstract. Selecting features from high-dimensional datasets is an important problem in machine learning. This paper shows that in the context of filter methods for feature selection, the estimator of the criterion used to select features plays an important role; in particular the estimators may suffer from a bias when comparing smooth and non-smooth features. This paper analyses the origin of such bias and investigates whether this bias influences the results of the feature selection process. Results show that non-smooth features tend to be penalised especially in small datasets.

Keywords: feature selection, smoothness, filter methods, mutual information, noise variance.

1 Introduction

High-dimensional datasets are now ubiquitous. Selecting a subset of the most relevant features is useful to ease the learning process, to alleviate the curse of dimensionality, to increase the interpretability of features, to visualise data, among others. Many works focus on methods to reduce the number of features in datasets [1–7]. These methods can be roughly categorised into filter methods, wrappers and embedded methods that all have their respective advantages and drawbacks [1]. 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 [6] and embedded methods [8].

Filters use a relevance criterion during the feature selection process. Three popular relevance criteria used to select features in regression tasks are the correlation, the mutual information and the noise variance. This paper focuses on mutual information and noise variance because of their property to be able to detect features that have

nonlinear relationships with the variable to predict. It shows that the statistical estimators of mutual information and of noise variance both suffer from a bias, mostly when small samples are considered, and that this bias may affect the selection of the features. The paper also shows that this bias disappears in large datasets, but faster when using noise variance than when using mutual information.

The remaining of the paper is organised as follows. Feature selection in regression with filters is detailed in Section 2. Section 3 analyses the behaviour of mutual information and Delta Test, and discusses the potential bias for small sample datasets. In order to confirm the bias and its consequences, simple experiments are described in Section 4 and their results are shown in Section 5. Finally, conclusions are given in Section 6.

2 Feature Selection with Filters

In the context of filter methods for feature selection, a 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. Correlation is the simplest one, but it is only able to detect linear relationships between random variables, and it is restricted to the univariate case (sets of features can only be evaluated individually, which prevents to take into account the possible relations between the features themselves). In this paper, we focus on nonlinear and multivariate relationships between a set of random input variables and one random output variable. For this type of relationships, mutual information (MI) and noise variance are both popular measures used as relevance criteria for filter methods. Both need to be estimated in practice on a finite set of data: traditional estimators are the Kraskov estimator for the former and the Delta Test for the latter. These criteria have been repeatedly used for feature selection in regression problems [9,10].

This section reviews the mutual information (MI) and noise variance criteria, and their Kraskov and Delta Test estimators. Both estimators are based on k-nearest neighbours. The next sections show that these estimators implicitly take into account a measure of smoothness (Section 3), which could lead to a bias in the choice of features during the feature selection process (Sections 4 and 5).

2.1 Feature Selection with Mutual Information

Mutual information (MI) is a popular criterion for filter methods [5,11–14]. Based on entropy, it is a symmetric measure of the dependence between random variables, introduced by Shannon in 1948 [15]. MI measures the information contained in a feature, or in a group of features, with respect to another one. It has been shown to be a reliable criterion to select relevant features in classification [16] and regression [9,10,17,18]. This paper focuses on regression problems.

Let X and Y be two random variables, 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

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

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 addition to the criterion, feature selection needs a search procedure to find the best feature subset among all possible ones. Given the exponential number of possible subsets, search procedures such as greedy search or genetic algorithms are used to find the best subset of features without having to compute the selection criterion between all subsets of variables and the output. Among these subsets, the one maximising the MI with the output is selected.

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. The estimator introduced by Kraskov et al. [19] is based on a *k*-nearest neighbour method and results from the Kozachenko-Leonenko entropy estimator [20]

$$\hat{H}(X) = -\psi(k) + \psi(N) + \log c_d + \frac{d}{N} \sum_{i=1}^{N} \log \epsilon_k(i), \tag{4}$$

where k is the number of neighbours, N is the number of instances in the dataset, d is the dimensionality, $c_d = (2\pi^{d/2})/\Gamma(d/2)$ 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 ψ is the digamma function.

Kraskov estimator (4) of the mutual information is then

$$\hat{I}(X;Y) = \psi(N) + \psi(K) - \frac{1}{k} - \frac{1}{N} \sum_{i=1}^{N} (\psi(\tau_x(i)) + \psi(\tau_y(i))) \tag{5}$$

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.

2.2 Feature Selection with Noise Variance

Noise variance is another filter criterion used for feature selection. Its definition is even more intuitive than mutual information. With this filter criterion, the noise represents the error in estimating the output variable by a function of the input variables, under the hypothesis that this function could be built (by a machine learning regression model). It is a filter criterion because it does not require building a regression model, but it is close to the idea of a wrapper method because the goal is to evaluate how good a model could be.

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

$$y_i = f(\mathbf{x}_i) + \epsilon_i \quad i = 1, ..., N \tag{6}$$

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 which lead to the lowest prediction error, or lowest noise variance [17].

In practice the noise variance has to be estimated, e.g. with the Delta Test [18]. The Delta Test δ is defined as

$$\delta = \frac{1}{2N} \sum_{i=1}^{N} \left[y_{NN(i)} - y_i \right]^2 \tag{7}$$

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

Similarly to the use of mutual information for feature selection, when using the Delta Test the relationships between several subsets of features and Y are computed, again with a search procedure such as a greedy search. Among these subsets of features, the one minimising the value of δ with Y will be selected. The Delta Test has also been widely used for feature selection [21,22].

3 Behaviour of *k*NN-based Estimators of Relevance Criteria in Small Sample Scenarios

This section analyses the behaviour of the mutual information and noise variance estimators in small datasets.

3.1 Mutual Information Analysis

The Kraskov estimator (5) can be used to estimate MI in regression. However, as a *k*NN-based estimator of $I(X;Y) = H(Y) - H(Y|X)$, it is affected by the degree of smoothness of the relationship between the target and the considered features. Indeed, the Kraskov estimator assumes that the conditional distribution $p(Y|X)$ is stationary in the k-neighbourhood of x. However, if the neighbourhood of x is large, which is the

case when the sample is small, this hypothesis does not hold anymore and the interval of observed values for *Y* will widen. The Kraskov estimator will consequently overestimate the conditional entropy *H(Y|X)* and underestimate *I(X;Y)*. This underestimation will be more severe for non-smooth functions, as the interval of *Y* in the neighbourhood of *x* is larger in this case. Consequently, when two features will be compared, the one that has the smoother relation to *Y* will tend to be favoured in the feature selection.

3.2 Delta Test Analysis

To estimate the variance of the noise in regression problems, the Delta Test uses a 1-nearest neighbour method by looking for the nearest neighbour of each point of the dataset and by computing a variation in target values between the point and its nearest neighbour.

The Delta Test, already defined in (7), can be rewritten with (6) as

$$\delta = \frac{1}{2N} \sum_{i=1}^{N} \left[f\left(\mathbf{x}_{NN(i)}\right) + \epsilon_{NN(i)} - f\left(\mathbf{x}_i\right) - \epsilon_i \right]^2 \tag{8}$$

where noise ϵ_i is i.i.d. The average behaviour of the Delta test can be characterised using a first order approximation $f(\mathbf{x}) \approx f(\mathbf{x}_i)+\nabla f(\mathbf{x}_i)^T(\mathbf{x} - \mathbf{x}_i)$, based on the assumption that the nearest neighbour is close enough to make this approximation sufficiently accurate. The expected value of the Delta Test is then approximated as

$$\begin{aligned} \mathrm{E}\left[\delta\right] &= \mathrm{E}\left[\frac{1}{2N} \sum_{i=1}^{N} \left[f\left(\mathbf{x}_{NN(i)}\right) + \epsilon_{NN(i)} - f\left(\mathbf{x}_i\right) - \epsilon_i \right]^2 \right] \\ &\approx \mathrm{E}\left[\frac{1}{2N} \sum_{i=1}^{N} \left[\nabla f(\mathbf{x}_i)^T \left(\mathbf{x}_{NN(i)} - \mathbf{x}_i\right) + \epsilon_{NN(i)} - \epsilon_i \right]^2 \right] \\ &= \mathrm{E}\left[\epsilon^2\right] + \frac{1}{2}\mathrm{E}\left[\left[\nabla f(\mathbf{x})^T \left(\mathbf{x}_{NN} - \mathbf{x}\right) \right]^2 \right]. \end{aligned} \tag{9}$$

The first term of (9) is the noise variance, but the second term is related to the smoothness of *f* and is independent from the noise variance: it measures how much *f* changes on average from an instance $\mathbf{x}$ to its closest neighbour $\mathbf{x}_{NN}$. This second term is affected by two factors. First, if the gradient is small (i.e. the function is smooth), the second term remains small. Second, if instances and their closest neighbours are close (i.e. the dataset is quite large), the second term also remains small. Hence, for small datasets, the second term penalises non-smooth functions.

3.3 Discussion

In small datasets, smooth relations between features and output will have, on average, a smaller Delta Test or a higher MI result. On the opposite, a nonsmooth relation will

have, on average, a larger Delta Test or smaller MI result, even with the same level of target noise. As discussed above, estimators based on *k*-nearest neighbours methods seem to be biased by the smoothness of functions. The two estimators make the assumption that the function does not vary too much in the proximity of the neighbours. However, in small sample and with non-smooth functions, this assumption is violated, which introduces a bias in the estimators. It is thus anticipated that smooth features will tend to be selected first when comparing two features that have the same level of information content to predict output *Y*. However, this short analysis does not answer the question whether this estimation bias has a real influence during the feature selection process, nor if the problem is more severe with MI or with noise variance. The next section evaluates these questions by experiments.

4 Experimental Settings

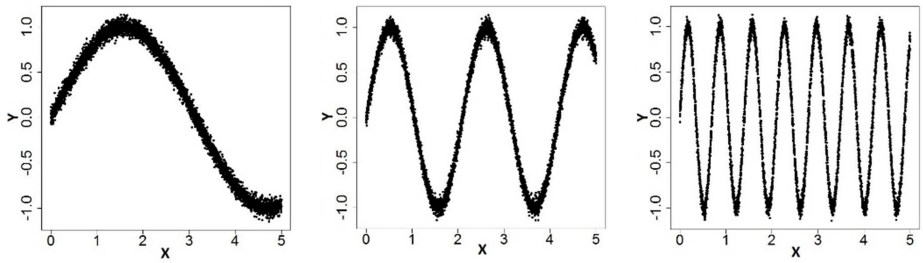
In order to study how much the smoothness can be a bias for selection criteria such as the mutual information or Delta Test in regression, experiments performed in this paper consider several functions with various smoothnesses and several sizes of datasets. These experiments are conducted to give some insights to the questions raised in the previous section, i.e. does the estimation bias has an influence while comparing features, and is the problem more severe with MI or with noise variance.

Six different periodic functions have been generated with different frequencies and different levels of noise:

$$\begin{aligned}
y_1 &= f_1(\mathbf{x}) = sin(\mathbf{x}) + \epsilon \quad && \text{where } \epsilon \sim N(0,0.05) \\
y_2 &= f_2(\mathbf{x}) = sin(3\mathbf{x}) + \epsilon \quad && \text{where } \epsilon \sim N(0,0.05) \\
y_3 &= f_3(\mathbf{x}) = sin(9\mathbf{x}) + \epsilon \quad && \text{where } \epsilon \sim N(0,0.05) \\
y_4 &= f_4(\mathbf{x}) = sin(\mathbf{x}) + \epsilon \quad && \text{where } \epsilon \sim N(0,0.3) \\
y_5 &= f_5(\mathbf{x}) = sin(3\mathbf{x}) + \epsilon \quad && \text{where } \epsilon \sim N(0,0.3) \\
y_6 &= f_6(\mathbf{x}) = sin(9\mathbf{x}) + \epsilon \quad && \text{where } \epsilon \sim N(0,0.3)
\end{aligned} \tag{10}$$

Figures 1(a), 1(b), 1(c), 1(d), 1(e) and 1(f) represent the six functions f_1, f_2, f_3, f_4, f_5 and f_6, respectively. In theory, features associated to f_1, f_2 and f_3 (resp. f_4, f_5, f_6) should be selected equally in a feature selection process, as prediction errors (or levels of noise) are identical.

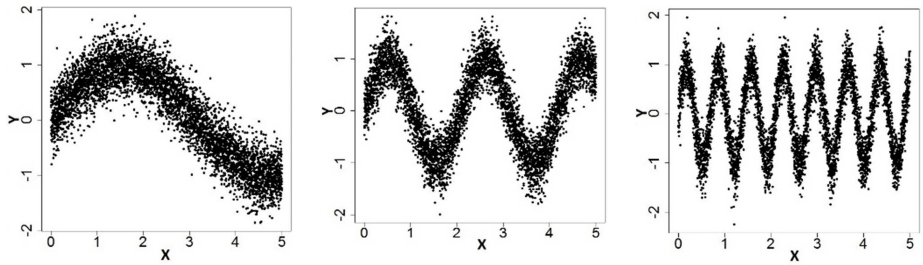
The experiments have been performed with various sizes of samples, from extremely small to large ones. For each size of the sample, an estimator of the two decision criteria, the mutual information and the noise variance, has been used to drive the selection process, in order to show the influence of the bias introduced by the smoothness of the function on both criteria. For the noise variance, the estimator used is the Delta Test, based on a *k*-NN method with 1-nearest neighbour, and described in Section 2.2. For the mutual information, the estimator used is the one introduced by Kraskov et al. and described in Section 2.1, also based on a *k*-NN method (with k=6 as suggested in [19]). All experiments have been repeated 10 times; averages are reported.



(a) $f_1(\mathbf{x}) = sin(\mathbf{x}) + \epsilon$ where $\epsilon \sim N(0,0.05)$

(b) $f_2(\mathbf{x}) = sin(3\mathbf{x}) + \epsilon$ where $\epsilon \sim N(0,0.05)$

(c) $f_3(\mathbf{x}) = sin(9\mathbf{x}) + \epsilon$ where $\epsilon \sim N(0,0.05)$



(d) $f_4(\mathbf{x}) = sin(\mathbf{x}) + \epsilon$ where $\epsilon \sim N(0,0.3)$

(e) $f_5(\mathbf{x}) = sin(3\mathbf{x}) + \epsilon$ where $\epsilon \sim N(0,0.3)$

(f) $f_6(\mathbf{x}) = sin(9\mathbf{x}) + \epsilon$ where $\epsilon \sim N(0,0.3)$

Fig.1. Experimental data generated with various frequencies and different levels of noise variance.

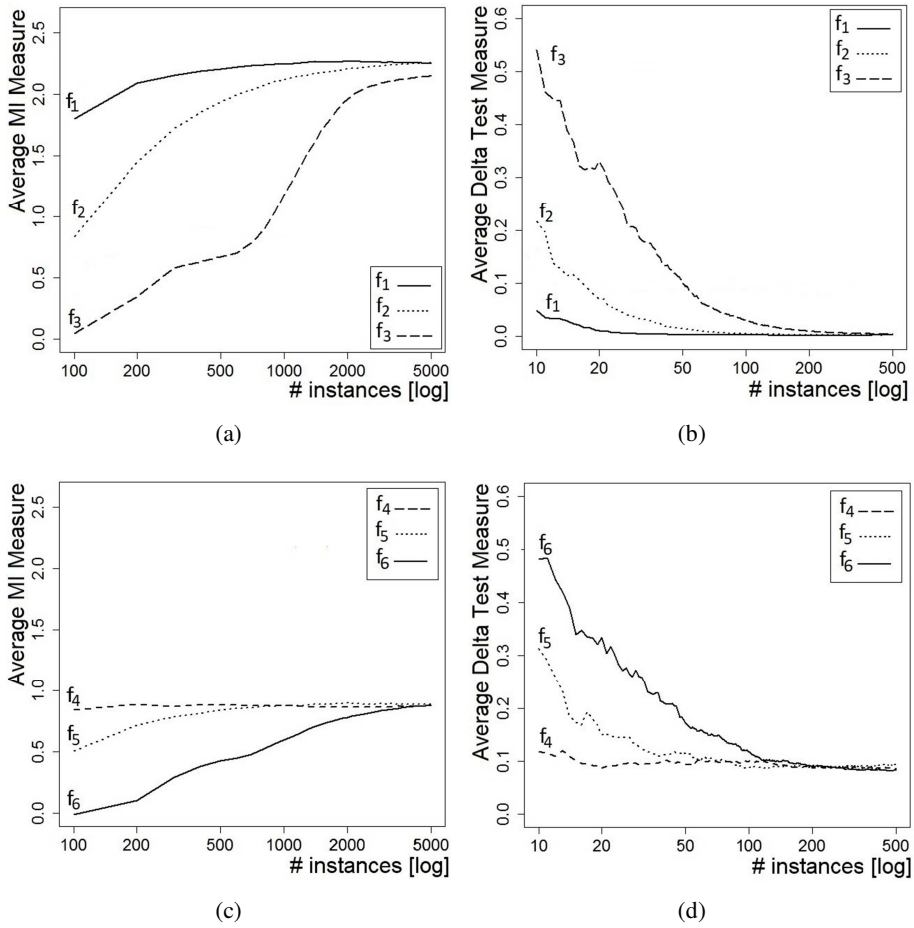


Fig.2. Average values of MI measures for 3 functions with a low level of noise (a) and for 3 functions with a higher level of noise (c), and of Delta Test for 3 functions with a low level of noise (b) and 3 functions with a higher level of noise (d).

5 Experimental Results

Figure 2 represents the average value (on 10 repetitions) of the mutual information estimator (2(a) and 2(c)) and the Delta Test estimator (2(b) and 2(d)), for increasing sizes of the dataset.

All figures show a clear effect in overestimating the noise variance and underestimating the mutual information in small datasets. The over- and underestimations are much more severe for non-smooth functions (f_3 and f_6). It is also clear that when the size of the dataset increases, the biases tend to disappear. What is more interesting to see is that the asymptotic values of the Delta Test are reached in this experiment when the dataset includes a few hundreds of instances, while for the

MI a few thousands of instances are necessary, in the same experiment (the horizontal logarithmic scales with the number of instances are different in the left and right figures). This is an argument in favour of using the noise variance rather than the mutual estimation.

When comparing the upper and lower parts of Figure 2 (both left -MI- and right -noise variance-), it is also interesting to see that for small samples, the order of selection between features can be inverted. For example, let us consider the Delta Test values in Figures 2(b) and 2(d) for three cases. First, for 40 instances, the 6 functions will be ranked in the following order: f_1, f_2, f_4, f_5, f_3 and f_6, given that the features with the lower Delta values are selected first. Without the bias effect shown in this paper, it would have been expected that f_1, f_2 and f_3 would be selected first, as their capacity to predict Y is higher (or their noise is lower) than for f_4, f_5 and f_6. Second, for 70 instances, the 6 functions will be ranked in the following order f_1, f_2, f_3, f_4, f_5 and f_6. In this case, the order of selection between features is not inverted anymore but the bias still remains. Finally, for approximately 300 instances, the bias disappears, the 3 functions f_1, f_2, f_3 obtain the same Delta value and the 3 functions f_4, f_5, f_6 obtain another unique Delta value, higher than the one for f_1, f_2, f_3. These cases show that, for small samples, the bias has an influence on the order of selection between features and that it disappears with a larger dataset. A similar behaviour can be observed for MI in figures 2(a) and 2(c).

6 Conclusion

To the best of our knowledge, no work in the literature focuses on the bias explicitly associated to the smoothness in a feature selection context. Wookey and Konidaris [23] use smoothness as a prior knowledge during feature selection, but only for data regularization.

This paper shows that 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. Experiments have been conducted with both criteria on functions with various smoothnesses and levels of noise, for different sizes of datasets. They confirm the theoretical discussion and show that the biases in the estimations are much more severe when using mutual information than when using the noise variance; this is an argument in favour of using the latter rather than the former.

The experiments also confirm that in a feature selection process, where a decision to select a feature is taken by comparing values of the criteria between different possible features or groups of features, the order of selection may be affected (a smooth feature with a low dependency to the output could be selected before a non-smooth with a high dependency). This is a serious shortcoming that should be taken into account when designing a feature selection algorithm. For example the noise variance could be explicitly estimated and used to remove the bias, or the selection process could be improved to favour non-smooth features.

References

1. I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3:1157–1182, March 2003.
2. D. François, F. Rossi, V. Wertz, and M. Verleysen. Resampling methods for parameter-free and robust feature selection with mutual information. *Neurocomputing*, 70(7-9):1276–1288, 2007.
3. M. Verleysen, F. Rossi, and D. François. Advances in feature selection with mutual information. In *Similarity-Based Clustering*, pages 52–69. 2009.
4. B. Frénay, M. van Heeswijk, Y. Miche, M. Verleysen, and A. Lendasse. Feature selection for nonlinear models with extreme learning machines. *Neurocomputing*, 102:111–124, 2013.
5. V. Gomez-Verdejo, M. Verleysen, and J. Fleury. Information-theoretic feature selection for functional data classification. *Neurocomputing*, 72(16-18):3580–3589, 2009.
6. R. Kohavi and G. H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97:273–324, 1997.
7. J. Paul, R. D'Ambrosio, and P. Dupont. Kernel methods for heterogeneous feature selection. *Neurocomputing*, 169:187–195, 2015.
8. B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32:407–499, 2004.
9. B. Frénay, G. Doquire, and M. Verleysen. Is mutual information adequate for feature selection in regression? *Neural Networks*, 48:1–7, 2013.
10. G. Doquire, B. Frénay, and M. Verleysen. Risk estimation and feature selection. In *Proceedings of the 21th International Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2013)*, 2013.
11. A. Degeest, M. Verleysen, and B. Frénay. Feature ranking in changing environments where new features are introduced. In *2015 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, July 2015.
12. G. Brown, A. Pocock, M. Zhao, and M. Lujan. Conditional likelihood maximisation: A unifying framework for mutual information feature selection. *Journal of Machine Learning Research*, 13:27–66, January 2012.
13. R. Battiti. Using mutual information for selecting features in supervised neuralnet learning. *IEEE Transactions on Neural Networks*, 5:537–550, 1994.
14. J. R. Vergara and P. A. Estévez. A review of feature selection methods based on mutual information. *Neural Computing and Applications*, 24:175–186, 2014.
15. C. E. Shannon. A mathematical theory of communication. *Bell Systems Technical Journal*, 27:379–423, 623–656, 1948.
16. B. Frénay, G. Doquire, and M. Verleysen. Theoretical and empirical study on the potential inadequacy of mutual information for feature selection in classification. *Neurocomputing*, 112:64–78, 2013.
17. A. Guillén, D. Sovilj, F. Mateo, I. Rojas, and A. Lendasse. New methodologies based on delta test for variable selection in regression problems. In *Workshop on Parallel Architectures and Bioinspired Algorithms, Toronto, Canada*, 2008.
18. Q. Yu, E. Séverin, and A. Lendasse. Variable selection for financial modeling. In *Proceedings of the CEF 2007, 13th International Conference on Computing in Economics and Finance, Montréal, Quebec, Canada*, pages 237–241, 2007.
19. A. Kraskov, H. Stögbauer, and P. Grassberger. Estimating mutual information.

Phys. Rev. E, 69:066138, 2004.
20. L. F. Kozachenko and N. Leonenko. Sample estimate of the entropy of a random vector. *Problems Inform. Transmission*, 23:95–101, 1987.
21. E. Eirola, E. Liitiäinen, A. Lendasse, F. Corona, and M. Verleysen. Using the delta test for variable selection. In *Proceedings of ESANN'08*, 2008.
22. E. Eirola, A. Lendasse, F. Corona, and M. Verleysen. The delta test: The 1-nnestimator as a feature selection criterion. In *Proceedings of the 2014 International Joint Conference on Neural Networks (IJCNN)*, pages 4214–4222, July 2014.
23. D. S. Wookey and G. D. Konidaris. Regularized feature selection in reinforcement learning. *Machine Learning*, 100(2):655–676, Sep 2015.