Entropy **2010**, *12*, 2144-2170; doi:10.3390/e12102144

OPEN ACCESS

entropy

ISSN 1099-4300

www.mdpi.com/journal/entropy

Article

Increasing and Decreasing Returns and Losses in Mutual Information Feature Subset Selection

Gert Van Dijck * **and Marc M. Van Hulle**

Laboratorium voor Neuro- en Psychofysiologie, Computational Neuroscience Research Group, Katholieke Universiteit Leuven, O&N II Herestraat 49 - bus 1021, 3000 Leuven, Belgium; E-Mail: marc.vanhulle@med.kuleuven.be

* Author to whom correspondence should be addressed; E-Mail: gert.vandijck@med.kuleuven.be; Tel.: +32-16-330428; Fax: +32-16-345960.

Received: 28 August 2010 / Accepted: 19 September 2010 / Published: 11 October 2010

Abstract: Mutual information between a target variable and a feature subset is extensively used as a feature subset selection criterion. This work contributes to a more thorough understanding of the evolution of the mutual information as a function of the number of features selected. We describe decreasing returns and increasing returns behavior in sequential forward search and increasing losses and decreasing losses behavior in sequential backward search. We derive conditions under which the decreasing returns and the increasing losses behavior hold and prove the occurrence of this behavior in some Bayesian networks. The decreasing returns behavior implies that the mutual information is concave as a function of the number of features selected, whereas the increasing returns behavior implies this function is convex. The increasing returns and decreasing losses behavior are proven to occur in an XOR hypercube.

Keywords: Bayesian networks; bit parity; conditional entropy; conditional mutual information; decreasing losses; decreasing returns; feature subset selection; increasing losses; increasing returns

---

1. Introduction

Feature subset selection [1] is an important step in the design of pattern recognition systems. It has several advantages. First, after feature selection has been performed off-line, predictions of a target

variable can be made faster on-line, because predictions can be performed in a lower-dimensional space and only a subset of the features needs to be computed. Secondly, it can lead to a decrease in hardware costs, because smaller patterns need to be stored or sensors that do not contribute to the prediction of a target variable can be eliminated (see e.g. [2] for a recent application of sensor elimination by means of feature subset selection). Thirdly, it can increase our understanding of the underlying processes that generated the data. Lastly, due to the curse of dimensionality [1,3], more accurate predictions can often be obtained when using the reduced feature set. Depending on the context, 'feature selection' is sometimes called 'characteristic selection' [4] or 'variable selection' [5].

Mutual information between the target variable, denoted as 'C', and a set of features, denoted as $\{F_1,...F_n\}$, is a well-established criterion to guide the search for informative features. However, some pitfalls in mutual information based feature subset selection should be avoided. A well-known property of entropy, *i.e.*, conditioning reduces uncertainty, does not necessarily hold for the mutual information criterion in feature selection. It would imply that conditioning on more features reduces the information that a feature contains about the target variable. Conditioning reduces information has been assumed to hold sometimes in the approximation of the high-dimensional mutual information by means of lower-dimensional mutual information estimations [6,7]. However, we show in this paper, using some counterexamples, related to the bit parity problem, that conditioning can increase the mutual information of a feature or a feature set about a target variable. We show in Section 3 that this can hold for both discrete, either binary or non-binary, and continuous features. Most lower-dimensional mutual information estimators may perform weak when dealing with probability distributions in which conditioning can increase information, see Section 4.4.

It has been observed sometimes that increments in mutual information become smaller and smaller in a sequential forward search in [8]. In fact, in this paper, we prove by means of a counterexample that the opposite behavior can also occur: the increments in mutual information become larger and larger in the SFS. This increasing returns behavior could be proven to occur in a (2n+1)-(2n-1)-...5-3 XOR hypercube in Section 4.1. The decreasing returns behavior, *i.e.*, increments in mutual information become smaller and smaller, could be proven to occur in some Bayesian networks in Section 4.2. We show in Section 5.1 that the 'increasing returns' has a comparable 'decreasing losses' implication in the sequential backward search. In Section 5.2, we show that the 'decreasing returns' has a comparable 'increasing losses' implication in the sequential backward search. All our theoretical claims are supported and illustrated by experiments.

2. Background and Definitions

2.1. Historical Background

Lewis was among the first to apply mutual information as a feature selection criterion almost half a century ago [4]. Mutual information was used to select good characteristics in a letters and numerals prediction problem. The following years the criterion was applied by Kamentsky and Liu in [9] and Liu

in [10] to similar character recognition problems. Although at that time Lewis did not call his criterion 'mutual information', but a 'measure of goodness' G_i:

$$G_i = \sum_{f_i,c} p(f_i, c) log\left(\frac{p(f_i|c)}{p(f_i)}\right) \tag{1}$$

Multiplying both the numerator and the denominator within the logarithm with p(c), this leads to the mutual information criterion for feature selection, used extensively since [11]:

$$MI(F_i; C) = \sum_{f_i,c} p(f_i, c) log\left(\frac{p(f_i, c)}{p(f_i).p(c)}\right) \tag{2}$$

This expression for mutual information is also the one used in [12]. For a subset of features $\mathbf{F}_S$ = {F_{S1},F_{S2}, ... F_{Sn1}}, this is:

$$MI(\boldsymbol{F}_S; C) = \sum_{\boldsymbol{f}_s,c} p(\boldsymbol{f}_s, c) log\left(\frac{p(\boldsymbol{f}_s, c)}{p(\boldsymbol{f}_s).p(c)}\right) \tag{3}$$

The mutual information can be expanded in entropy terms H(C) and H(C|$\mathbf{F}_S$) as:

$$MI(\boldsymbol{F}_S; C) = H(C) - H(C|\boldsymbol{F}_S) \tag{4}$$

Throughout the article, we use bold style to denote vectors, capitals to denote variables and lowercase letters to denote values of variables.

2.2. *Conditional Mutual Information*

The conditional mutual information of feature F_i and C, given F_j, is defined as [12]:

$$MI(F_i; C|F_j) = \sum_{f_i,f_j,c} p(f_i, f_j, c) log\left(\frac{p(f_i, c|f_j)}{p(f_i|f_j).p(c|f_j)}\right) \tag{5}$$

Due to the chain rule for information [12], the conditional mutual information is equal to a difference in (unconditional) mutual information:

$$MI(F_i; C|F_j) = MI(F_i, F_j; C) - MI(F_j; C) \tag{6}$$

3. Conditioning Increases Information

One particular example where conditioning on additional variables can increase information was given in [12] (see page 35). The example was Z = X + Y, with X and Y binary independent distributed variables. Then, it was shown that MI(X;Y) < MI(X;Y|Z). We provide examples in this section that are more tailored to the feature subset selection problem, and also show that it holds for binary, non-binary discrete and continuous features. The n-bit parity problem and the checkerboard pattern were also mentioned in [5] to indicate that a variable which is useless by itself can be useful together with others. We derive the implications for these examples for conditioning in mutual information based feature selection. We provide results for 'n' features in general and derive a general 'conditioning increases information' result in inequality (11) under the conditions given in (7) and (9).

3.1. n-bit Parity Problem

The bit parity problem is frequently used as a benchmark problem, e.g. in neural network learning [13]. Consider 'n' independent features F_1,... F_n that are binary: $F_i \in \{0,1\}$ and $p(F_1,F_2, ...F_n) = \prod_{i=1}^{n} p(F_i)$.

The target variable in the case of the n-bit parity problem, which is an XOR problem for n = 2, is then defined as:

$$C = mod(\sum_{i=1}^{n} F_i, 2) \tag{7}$$

We denote the modulo 2 computation by mod(.,2). The target variable is equal to 1 in case the n-tuple (f_1,f_2, ...f_n) contains an odd number of 1's and is 0 otherwise. The mutual information based on the full feature set is equal to:

$$\begin{aligned} MI(F_1, F_2, ...F_n; C) &= H(C) - H(C|F_1, F_2, ...F_n) \\ &= H(C) \end{aligned} \tag{8}$$

where Equation (8) is due to the result that the uncertainty left about C after observing F_1, F_2, ...F_n is equal to 0: $H(C|F_1,F_2, ...F_n) = 0$. The probability of class 0, p(c=0), and class 1, p(c=1), is equal to 1/2 for $n \geq 2$, this implies that according to Equation (8) the mutual information $MI(F_1,F_2,...F_n;C) = 1$ bit. In the previous computation, we used the base 2 logarithm. In case the target variable takes 2 values, $C \in \{0,1\}$, the mutual information will be maximally 1 bit, in fact, in (3) one can choose the base of the logarithm. For any strict subset $\mathbf{F}_S \subset \{F_1,F_2,...F_n\}$ excluding $\emptyset$, it can be verified that $p(\mathbf{f}_S,c) = p(\mathbf{f}_S).p(c)$. From this, it follows that, using the definition of mutual information in Equation (3), $MI(\mathbf{F}_S;C) = 0$. This leads us to the following result. Suppose that:

$$\begin{aligned} \{F_1, F_2, ...F_n\} &= \boldsymbol{F}_{S1} \cup \boldsymbol{F}_{S2}, \\ \text{with } \boldsymbol{F}_{S1} \neq \emptyset, \boldsymbol{F}_{S2} \neq \emptyset \quad &\text{and} \quad \boldsymbol{F}_{S1} \cap \boldsymbol{F}_{S2} = \emptyset \end{aligned} \tag{9}$$

Because $\mathbf{F}_{S1}$ and $\mathbf{F}_{S2}$ are strict subsets of the full feature set, it holds that $MI(\mathbf{F}_{S1};C) = 0$ and $MI(\mathbf{F}_{S2};C) = 0$ in the n-dimensional XOR problem. For the conditional mutual information, it holds that:

$$\begin{aligned} MI(\boldsymbol{F}_{S2}; C|\boldsymbol{F}_{S1}) &= MI(\boldsymbol{F}_{S1}, \boldsymbol{F}_{S2}; C) - MI(\boldsymbol{F}_{S1}; C) \\ &= 1 - 0 \end{aligned} \tag{10}$$

From this, we can derive following general result:

$$MI(\boldsymbol{F}_{S2}; C|\boldsymbol{F}_{S1}) > MI(\boldsymbol{F}_{S2}; C) \tag{11}$$

Similarly, we can conclude: $MI(\mathbf{F}_{S1};C|\mathbf{F}_{S2}) > MI(\mathbf{F}_{S1};C)$.

A case derived from the n-bit parity problem is obtained when the variable F_j and C are interchanged in Equation (7):

$$F_j = mod(\sum_{i=1, i \neq j}^{n} F_i + C, 2) \tag{12}$$

with all the variables $F_{i,i\neq j}$ and C independent and binary. For this case it holds that MI(F_1,...F_{j-1},F_{j+1},...F_n;C) = 0, because $\{F_1,...F_{j-1}, F_{j+1},...F_n\}$ and C are independent by construction. However, after conditioning on F_j we obtain:

$$MI(F_1, ...F_{j-1}, F_{j+1}, ...F_n; C|F_j) = H(C|F_j) - H(C|F_1, F_2, ...F_n) \quad (13)$$

$$= H(C|F_j) = 1 \quad (14)$$

In Equation (13) H(C|F_1,F_2,...F_n) = 0, because after all other features are observed, C can be perfectly predicted. In Equation (14) H(C|F_j) = H(C), because F_j on its own contains no information about C. Hence, for Equation (12) we conclude: MI(F_1,...F_{j-1},F_{j+1},...F_n;C|F_j) > MI(F_1,...F_{j-1},F_{j+1},...F_n;C). In fact, for Equation (12), the more general result of inequality (11) also holds under the conditions of (9), regardless whether $F_j \in \mathbf{F}_{S1}$ or $F_j \in \mathbf{F}_{S2}$.

3.2. *Non-binary Discrete Features*

Consider the checkerboard pattern shown in Figure 1.

Figure 1. Checkerboard pattern.

								F_2
1	0	1	0	1	0	1	0	8
0	1	0	1	0	1	0	1	7
1	0	1	0	1	0	1	0	6
0	1	0	1	0	1	0	1	5
1	0	1	0	1	0	1	0	4
0	1	0	1	0	1	0	1	3
1	0	1	0	1	0	1	0	2
0	1	0	1	0	1	0	1	1
F_1 1	2	3	4	5	6	7	8	

The target variable C is noted in the squares and depends on whether the square is black (c = 0) or white (c = 1). There are two features F_1 and F_2, corresponding to the column and row indicators, respectively. Variables F_1 and F_2 are 8-ary and take values $\{1,2,3,...8\}$, while C is binary. How does this checkerboard pattern relate to the n-bit parity problem? The pattern in Figure 1 can be seen as a natural extension of the n-bit parity problem and can also be expressed by Equation (7) with the same requirement of independence between features. In the case of m-ary features, with 'm' even, it holds that any strict subset of features contains no information about the target variable. The requirement of 'm' even arises from the fact that, for any subset of features, we need an equal number of 0 and 1 cells in order for strict subsets not to be informative, this can only be achieved when 'm' is even (see also Section 4). The full feature set for 'm' even contains 1 bit of information. This shows that the general result of inequality (11) also holds for non-binary discrete features.

A concept related to mutual information is the n-way interaction information introduced by McGill [14]. This n-way interaction information has been used to characterize statistical systems such as spin glasses in [15], where the concept was introduced as 'higher-order mutual information'. The n-way

interaction information for class variable C and features F_1, F_2,...F_n, written as $I_{n+1}(C,F_1,F_2,...F_n)$, can be expressed in terms of the mutual information as follows:

$$I_{n+1}(C, F_1, F_2, ...F_n) = \sum_{k=1}^{n} (-1)^{(k+1)} \sum_{i_1<...<i_k} MI(F_{i_1}, F_{i_2}, ...F_{i_k}; C), \tag{15}$$

with $\{i_1, i_2,...i_k\} \in \{1,2,...n\}$. As opposed to the usual mutual information, the higher-order mutual information can be negative. The statistical system is termed 'frustrated' in that case [15]. The XOR problem presented in Figure 1 is an example of a frustrated system, because $I_{n+1}(C,F_1,F_2,...F_n)$ = -MI(F_1,F_2;C)=-H(C). However, it is observed that the XOR problem appears as 'frustrated' if the number of features is even. In case of an odd number of features, the higher-order mutual information is positive, e.g. for n = 3 features $I_{n+1}(C,F_1,F_2,...F_n)$ = MI(F_1,F_2,F_3;C) = H(C). The n-way interaction information can be useful in feature selection to verify whether any of the features is independent of all other features and the target variable in a single test. This can be seen from Equation (15): the n-way interaction information is symmetric (see [14]), one can exchange the class variable with any feature. Hence, if for any F_j MI(F_1,...F_{j-1},F_{j+1},...F_n,C;F_j) = 0, then $I_{n+1}(C,F_1,F_2,...F_n)$ = 0. Three-way interaction information has been used to derive causal relationships between neurons in [16] and more recently between features and the target variable in [17].

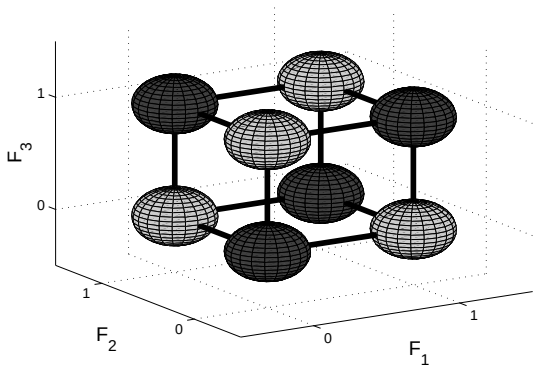
All examples above were given for discrete features, the question is whether the 'conditioning increases information' result can also hold for continuous features.

3.3. Continuous Features: Mixture Models

Consider a d-dimensional Gaussian mixture model (GMM) or t mixture model (tMM), with the number of Gaussians or t distributions equal to 2^d, with half of the Gaussians or t distributions assigned to class 0, and the other half to class 1, and where each distribution has a mixing proportion of $1/(2^d)$.

An example is provided in Figure 2, where the Gaussians are positioned on a cube with corners $(\mu_1, \mu_2, \mu_3) \in \{(0,0,0), (1,0,0),...(1,1,1)\}$, the covariance matrices are spherical and d = 3.

Figure 2. Gaussian distributions with spherical covariance matrices on cube corners.



Suppose that we assign the Gaussians for which mod(μ_1 + μ_2 + μ_3,2) = 1 to class 1 (light grey in Figure 2) and the Gaussians for which mod(μ_1 + μ_2 + μ_3,2) = 0 to class 0. (dark grey). This Gaussian mixture model can be written as:

$$\sum_{\substack{i=1:2^{(3-1)},\\ mod(\mu_{i1}+\mu_{i2}+\mu_{i3},2)=0}} \frac{1}{2^3}\mathcal{N}\Big(\begin{bmatrix}\mu_{i1}\\ \mu_{i2}\\ \mu_{i3}\end{bmatrix}, \Sigma\Big) + \sum_{\substack{j=2^{(3-1)}+1:2^{(3)},\\ mod(\mu_{j1}+\mu_{j2}+\mu_{j3},2)=1}} \frac{1}{2^3}\mathcal{N}\Big(\begin{bmatrix}\mu_{j1}\\ \mu_{j2}\\ \mu_{j3}\end{bmatrix}, \Sigma\Big) \tag{16}$$

The marginal distributions of a multivariate Gaussian distribution or multivariate t distribution are Gaussian and t distributed again [18], respectively. Hence, when we select 2 features, the marginal distributions are Gaussian again, and their centers will now be located on the corners of the square (0,0), (0,1), (1,0), (1,1). Moreover, on each corner, there will be 1 Gaussian from both classes, and in the case of equal covariance matrices, their distributions will be equal. Equal distributions for both classes implies that the mutual information with the target variable is equal to 0. This can be seen from the continuous version of mutual information:

$$MI(F_1, F_2, ...F_d; C) = \sum_c \iiint_{f_1,f_2,...f_d} p(f_1, f_2, ...f_d, c) log \frac{p(f_1, f_2, ...f_d, c)}{p(f_1, f_2, ...f_d).p(c)} df_1\, df_2 ...\, df_d \tag{17}$$

The numerator in Equation (17), after subset selection with $sn1$ features, with $sn1 <$ d and $sn1 > 0$, is equal to p(f_{s1},...f_{sn1},c), and the denominator is equal to:

$$\begin{aligned} p(f_{s1}, ...f_{sn1}).p(c) &= \big(p(f_{s1}, ...f_{sn1}|c=0).p(c=0) \\ &+ p(f_{s1}, ...f_{sn1}|c=1).p(c=1)\big).p(c) \\ &= p(f_{s1}, ...f_{sn1}, c) \end{aligned} \tag{18}$$

In Equation (18), we have that p(f_{s1},...f_{sn1}|c=0) = p(f_{s1},...f_{sn1} |c=1) = p(f_{s1},...f_{sn1}|c). Hence, the numerator and the denominator are equal in the MI definition and MI(F_1,...F_{sn1};C) = 0 holds. This can be extended to more than 3 dimensions. Again the general result of inequality (11) applies. It is important to note that we only require the covariance matrices for the Gaussian distributions to be equal, not necessarily spherical, and in the case of multivariate t distributions, one additionally requires their degrees of freedom to be equal to guarantee that MI(F_1,...F_{sn1};C) = 0.

4. Increasing and Decreasing Returns

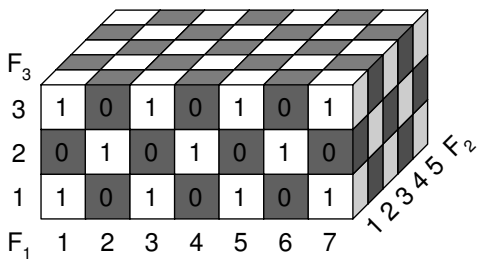
4.1. Increasing Returns

Previously, it was found in [8] that in the sequential forward search (SFS), when features are selected later in the forward search, they contribute less in the increase of information of the target variable compared to previously selected ones. Let us illustrate this with an example. Suppose that we dispose of 3 features F_1, F_2 and F_3 and that F_1 is the first feature for which MI(F_1;C) > MI(F_2;C) > MI(F_3;C).

In the first iteration of the SFS, the feature, for which the objective MI(F_i;C) is the highest, is selected. In this case, the selected feature will be F_1. Suppose that in the second iteration of the SFS, we have: MI(F_2;C|F_1) > MI(F_3;C|F_1). The second selected feature will be F_2. In the third iteration, the only feature left is F_3 and the incremental increase in information is: MI(F_3;C|F_1,F_2). The 'decreasing returns' (*i.e.* every additional investment in a feature results in a smaller return) is then observed as: MI(F_1;C) > MI(F_2;C|F_1) > MI(F_3;C|F_1,F_2).

However, this is not always true. We show with a counterexample that the opposite behavior can occur: although the order of selected features is F_1, F_2 and finally F_3, it can hold that MI(F_1;C) < MI(F_2;C|F_1) < MI(F_3;C|F_1,F_2), *i.e.* we observe 'increasing returns' (every additional investment in a feature results in an increased return) instead of decreasing returns. Consider a possible extension of the checkerboard to 3 dimensions in Figure 3.

Figure 3. 7-5-3 XOR Cube. Extension of checkerboard to 3 dimensions, the number of values that each feature can take is odd and different for each feature.



Here, the three features F_1, F_2 and F_3 take an odd number of values: 7, 5 and 3 respectively. We will refer to this example as '7-5-3 XOR'. As opposed to 'm' even, now each feature individually, as well as each subset of features, contains information about the target variable. We computed the conditional entropies for this example in Table 1.

The mutual information and the conditional mutual information can be derived from the conditional entropies and are shown on the right side of the table. Clearly, the first feature that will be selected is F_1, as this feature contains individually the most information about the target variable. The next feature selected is F_2, because conditioned on F_1, F_2 contains the most information. Finally, F_3 will be selected with a large increase in information: MI(F_3;C|F_1,F_2) ≈ 0,9183 bits. This increasing returns behavior can be shown to hold more generally for a $(2n+1) - ... - 7 - 5 - 3$ XOR hypercube, with 'n' the number of features. The total number of cells (feature value combinations) in such a hypercube is equal to $(2n+1).(2n-1).(2n-3)...(3)$. This can be written as a double factorial:

$$(2n+1)!! = (2n+1).(2n-1).(2n-3)...3 \tag{19}$$

This is an odd number of cells. $((2n+1)!! - 1)/2$ of the cells have been assigned a 0 or a 1 value. The entropy H(C) can therefore be written as:

$$\begin{aligned} H(C) = & -\frac{(2n+1)!!-1}{2.(2n+1)!!} log\left(\frac{(2n+1)!!-1}{2.(2n+1)!!}\right) \\ & -\frac{(2n+1)!!+1}{2.(2n+1)!!} log\left(\frac{(2n+1)!!+1}{2.(2n+1)!!}\right) \end{aligned} \tag{20}$$

Table 1. 7-5-3 XOR Cube. Entropies and Mutual Information for the SFS, NA = not available.

Entropy	value(bit)	Mutual Inf.	value(bit)
H(C)	$-\frac{53}{105}\log_2\frac{53}{105}$ $-\frac{52}{105}\log_2\frac{52}{105}$ $\approx 0{,}9999$	NA	NA
$H(C\|F_1)$	$-\frac{8}{15}\log_2\frac{8}{15}$ $-\frac{7}{15}\log_2\frac{7}{15}$ $\approx 0{,}9968$	$MI(F_1;C)$	$\approx 3{,}143.10^{-3}$
$H(C\|F_2)$	$-\frac{10}{21}\log_2\frac{10}{21}$ $-\frac{11}{21}\log_2\frac{11}{21}$ $\approx 0{,}9984$	$MI(F_2;C)$	$\approx 1{,}571.10^{-3}$
$H(C\|F_3)$	$-\frac{18}{35}\log_2\frac{18}{35}$ $-\frac{17}{35}\log_2\frac{17}{35}$ $\approx 0{,}9994$	$MI(F_3;C)$	$\approx 5{,}235.10^{-4}$
$H(C\|F_1,F_2)$	$-\frac{2}{3}\log_2\frac{2}{3}$ $-\frac{1}{3}\log_2\frac{1}{3}$ $\approx 0{,}9183$	$MI(F_2;C\|F_1)$	$\approx 7{,}850.10^{-2}$
$H(C\|F_1,F_3)$	$-\frac{3}{5}\log_2\frac{3}{5}$ $-\frac{2}{5}\log_2\frac{2}{5}$ $\approx 0{,}9710$	$MI(F_3;C\|F_1)$	$\approx 2{,}584.10^{-2}$
$H(C\|F_2,F_3)$	$-\frac{4}{7}\log_2\frac{4}{7}$ $-\frac{3}{7}\log_2\frac{3}{7}$ $\approx 0{,}9852$	$MI(F_3;C\|F_2)$	$\approx 1{,}314.10^{-2}$
$H(C\|F_1,F_2,F_3)$	0	$MI(F_3;C\|F_1,F_2)$	$\approx 0{,}9183$

In every step of the sequential forward search, the feature that takes the largest number of feature values will be selected first, because this will decrease the conditional entropy (and hence increase the mutual information) the most. This can be observed from Table 1: first F_1 (which can take 7 values) is selected, subsequently conditioned on F_1, F_2 (which can take 5 values) is selected. Finally, conditioned on F_1 and F_2, F_3 (which can take 3 values) is selected. The conditional entropy conditioned on k variables needs to be computed over hypercubes with dimension (n-k), each containing $(2n+1-2k)!!$ cells. Again

$((2n+1-2k)!!-1)/2$ cells have been assigned a 0 or a 1 value. Therefore the conditional entropy after k steps, $k \leq n-1$, of the SFS can be computed as:

$$H(C|F_1, F_2, ...F_k) = -\frac{(2(n-k)+1)!!-1}{2.(2(n-k)+1)!!} log\left(\frac{(2(n-k)+1)!!-1}{2.(2(n-k)+1)!!}\right) - \frac{(2(n-k)+1)!!+1}{2.(2(n-k)+1)!!} log\left(\frac{(2(n-k)+1)!!+1}{2.(2(n-k)+1)!!}\right) \quad (21)$$

4.2. Decreasing Returns

Next, we ask ourselves under what condition the decreasing returns holds. Suppose that the selected subset found so far is **S**, and that the feature selected in the current iteration is F_x. In order for the decreasing returns to hold, one requires for the next selected feature F_y: MI(F_x;C|**S**) > MI(F_y;C| **S**,F_x). First, we expand MI(F_x,F_y;C| **S**) in two ways by means of the chain rule of information:

$$\begin{aligned} MI(F_x, F_y; C|\boldsymbol{S}) &= MI(F_x; C|\boldsymbol{S}) + MI(F_y; C|\boldsymbol{S}, F_x) \\ &= MI(F_y; C|\boldsymbol{S}) + MI(F_x; C|\boldsymbol{S}, F_y) \end{aligned} \quad (22)$$

In the sequential forward search, F_x was selected before F_y, thus, it must be that: MI(F_x;C| **S**) > MI(F_y;C|**S**). In the case of ties, it may be possible that MI(F_x;C|**S**) $\geq$ MI(F_y;C|**S**), we focus here on the case where we have a strict ordering >. Then, in Equation (22) we have that:

$$MI(F_y; C|\boldsymbol{S}, F_x) < MI(F_x; C|\boldsymbol{S}, F_y) \quad (23)$$

Hence, a sufficient condition in order for the decreasing returns to hold is that: MI(F_x;C|**S**,F_y) $\leq$ MI(F_x;C|**S**). This means that additional conditioning on F_y decreases (or equals) information of F_x about C.

A first dependency structure between variables for which the decreasing returns can be proven to hold in the SFS is when all features are child nodes of the class variable C. This means that all features are conditionally independent given the class variable. This dependency structure is shown in Figure 4.

Lemma 4.1. *Suppose that the order in which features are selected by the SFS is: firstly F_1 subsequently F_2 next F_3 until F_n. If all features are conditionally independent given the class variable,* i.e. *$p(F_1,F_2, ...F_n|C) = \prod_{i=1}^{n} p(F_i|C)$, then the decreasing returns behavior holds: $MI(F_1;C) > MI(F_2;C|F_1) > MI(F_3;C|F_1,F_2) > ... > MI(F_n;C|F_1,F_2...F_{n-1})$.*

Proof. First, we show that MI(F_1;C) > MI(F_2;C|F_1).

$$MI(F_2; F_1, C) = MI(F_2; C) + MI(F_2; F_1|C) \quad (24)$$

$$= MI(F_2; F_1) + MI(F_2; C|F_1) \quad (25)$$

In Equation (24) it holds that, due to the conditional independence of the features given the class variable, MI(F_2;F_1|C) = 0. In Equation (25), we have that MI(F_2;F_1) $\geq$ 0. Comparing Equations (24) and (25), we obtain that: MI(F_2;C) $\geq$ MI(F_2;C|F_1). Because, F_1 was selected before F_2, we have that MI(F_1;C) > MI(F_2;C) (again we assume a strict ordering among variables, in case of ties we have MI(F_1;C) $\geq$ MI(F_2;C)). Hence, we obtain MI(F_1;C) > MI(F_2;C) $\geq$ MI(F_2;C|F_1).

Using a similar reasoning as above, we can show that it holds in general: MI(F_{k-1};C|F_1,F_2,...F_{k-2}) > MI(F_k;C|F_1,F_2,...F_{k-1}). We start with the generalization of Equation (24):

$$MI(F_k; F_1, F_2, ...F_{k-1}, C) = MI(F_k; F_1, F_2, ...F_{k-2}, C) + MI(F_k; F_{k-1}|C, F_1, ...F_{k-2}) \quad (26)$$

In appendix (B), we prove that the conditional independence of the variables given C implies that MI(F_k;F_{k-1}|C,F_1,...F_{k-2}) = 0. Further expansion of the left and the right hand sides in Equation (26) results in:

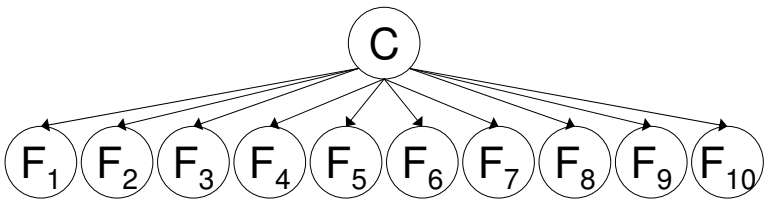
$$MI(F_k; F_1, F_2, ...F_{k-1}, C) = MI(F_k; F_1, F_2, ...F_{k-2}, C) \quad (27)$$

$$MI(F_k; F_1, ...F_{k-1}) + MI(F_k; C|F_1, ...F_{k-1}) = MI(F_k; F_1, ...F_{k-2}) + MI(F_k; C|F_1, ...F_{k-2}) \quad (28)$$

Because MI(F_k;F_1,...F_{k-1}) ≥ MI(F_k;F_1,...F_{k-2}) in Equation (28), we see that MI(F_k;C|F_1,...F_{k-1}) ≤ MI(F_k;C|F_1,...F_{k-2}). But, F_{k-1} was selected before F_k, *i.e.* MI(F_k;C|F_1,...F_{k-2}) < MI(F_{k-1};C|F_1,...F_{k-2}), from which we obtain what needed to be proven: MI(F_k;C|F_1,...F_{k-1}) < MI(F_{k-1};C|F_1,...F_{k-2}). □

In Figure 4 we show a Bayesian network [19,20] where the class variable C has 10 child nodes. This network has 21 degrees of freedom: we can randomly choose p(c=0) ∈ [0,1] and for the features we can choose p(f_i=0|c=0) ∈ [0,1] and p(f_i=0|c=1) ∈ [0,1]. We generated a Bayesian network where the probability p(c=0) and the conditional probabilities p(f_i=0|c=0) and p(f_i=0|c=1) are generated randomly following a uniform distribution within [0,1]. According to Lemma 4.1, we should find the decreasing returns behavior if we apply the SFS to this network.

Figure 4. Example of class conditional independence of the features given the class variable C. The joint probability distribution can be factorized as: p(F_1,F_2,...F_{10},C) = $\left(\prod_{i=1}^{10}\right.$ p(F_i |C)$\left.\right)$.p(C).



Indeed, this decreasing returns behavior can be observed in Figure 5 using the generated Bayesian network: Lemma 4.1 predicts that the conditional mutual information decreases with an increasing number of features being selected. This implies that the mutual information is a concave function in function of the number of features selected. This can be seen from the fact that the mutual information can be written as a sum of conditional mutual information terms:

$$MI(F_1, F_2, ...F_n; C) = MI(F_1; C) + M(F_2; C|F_1) + ... + MI(F_n; C|F_1, F_2, ...F_{n-1}) \quad (29)$$

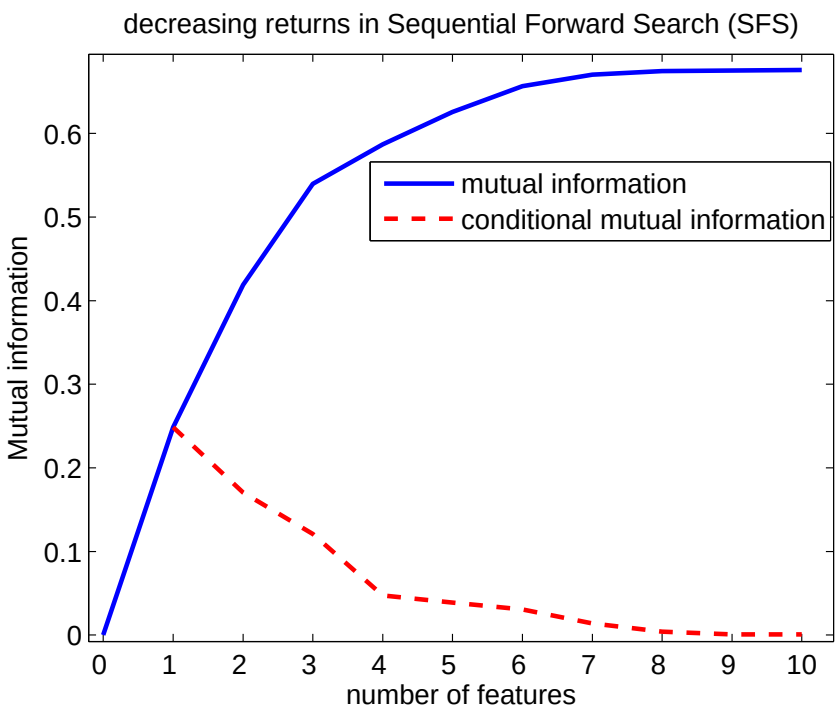
with every next term smaller than the previous one. A particular case of Figure 4 is obtained if besides class conditional independence among features also independence is assumed. In that case, it can

be shown [21] that the high-dimensional mutual information, can be written as a sum of marginal information contributions:

$$MI(F_1, F_2, ...F_n; C) = MI(F_1; C) + M(F_2; C) + ... + MI(F_n; C) \tag{30}$$

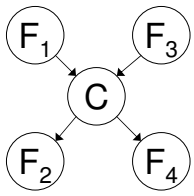
The SFS, which is in general not optimal, can then be shown to be optimal in mutual information sense. Indeed, at the k'th step of the SFS, *i.e.* after 'k' features have been selected, there is no subset of 'k' or less than 'k' features out of the set of 'n' features that leads to a higher mutual information than the set that has been found with the SFS at step 'k', if MI(F_k;C) > 0. Independence and class conditional independence will often not be satisfied for data sets. Nevertheless, for gene expression data sets that typically contain up to 10,000 features, overfitting in a wrapper search can be alleviated if the features with the lowest mutual information are removed before applying the wrapper search [22].

Figure 5. Evolution of the mutual information in function of the number of features selected with the SFS. A Bayesian network according to Figure 4 was created with probability p(c=0), conditional probabilities p(f_i=0|c=0) and p(f_i=0|c=1) drawn randomly following a uniform distribution within [0,1]. The conditional mutual information at 1 feature is MI(F_1;C) at 2 features MI(F_2;C|F_1),... and finally at 10 features MI(F_{10};C|F_1,F_2,...F_9). Lemma 4.1 predicts that the conditional mutual information decreases with an increasing number of features selected. This implies that the mutual information is concave in function of the number of features selected.



It can be shown that also in more complex settings when there are both child and parent nodes the decreasing returns behavior can still hold. In Figure 6 an example of dependencies between 4 features is provided for which the decreasing returns holds, if the parent and child nodes are selected alternately. This leads to the following lemma.

Figure 6. Example of dependencies between features for decreasing returns. The joint probability distribution can be factorized as: $p(F_1,F_2,F_3,F_4,C) = p(F_2|C).p(F_4|C).p(C|F_1,F_3).p(F_1).p(F_3)$. This factorization implies that: $MI(F_1,F_3;F_2,F_4|C) = 0$.



Lemma 4.2. *Suppose that the order in which features are selected by the SFS is: firstly F_1 subsequently F_2 next F_3 until F_n. Assume that the odd selected features,* i.e. *F_1, F_3, F_5 ..., are parents of C and the even selected features,* i.e. *F_2, F_4, F_6 ..., are children of C, then the decreasing returns behavior holds: $MI(F_1;C) > MI(F_2;C|F_1) > MI(F_3;C|F_1,F_2) > ... > MI(F_n;C|F_1,F_2,...F_{n-1})$.*

Let us first prove the result for the case with 4 features, as shown in Figure 6. The order of selected features in the SFS is F_1, F_2, F_3 and F_4, respectively. We show that $MI(F_3;C|F_1,F_2) < MI(F_2;C|F_1)$.

$$MI(F_3;C,F_1,F_2) = MI(F_3;F_1,F_2) + MI(F_3;C|F_1,F_2) \quad (31)$$

$$= MI(F_3;F_1) + MI(F_3;C|F_1) + MI(F_3;F_2|C,F_1) \quad (32)$$

In Equation (32) $MI(F_3;F_2|C,F_1) = 0$, this follows from the fact that $MI(F_1,F_3;F_2,F_4|C) = 0$, see appendix (A). We have in Equation (31) and Equation (32) that $MI(F_3;F_1,F_2) \geq MI(F_3;F_1)$. Hence, combining previous 2 results yields: $MI(F_3;C|F_1,F_2) \leq MI(F_3;C|F_1)$. Because feature F_2 is selected before F_3 we have: $MI(F_2;C|F_1) > MI(F_3;C|F_1)$. Finally, we obtain that $MI(F_3;C|F_1,F_2) < MI(F_2;C|F_1)$. Similar expansions for $MI(F_2;C,F_1)$ and $MI(F_4;C,F_1,F_2,F_3)$ as in Equations (31) and (32), enable us to prove that $MI(F_2;C|F_1) < MI(F_1;C)$ and $MI(F_4;C|F_1,F_2,F_3) < MI(F_3;C|F_1,F_2)$ respectively. Hence, we can conclude that the decreasing returns holds.

Now let us prove the result for any 'k' in general and regardless whether F_k is a parent node or a child node. Apply a similar expansion as in Equation (31).

$$MI(F_k;C,F_1,F_2,...F_{k-1}) = MI(F_k;F_1,F_2,...F_{k-1}) + MI(F_k;C|F_1,F_2,...F_{k-1}) \quad (33)$$

$$= MI(F_k;F_1,F_2,...F_{k-2}) + MI(F_k;C|F_1,F_2,...F_{k-2}) + MI(F_k;F_{k-1}|C,F_1,...F_{k-2}) \quad (34)$$

Comparing Equations (33) and (34), we have that: $MI(F_k;F_1,F_2,...F_{k-1}) \geq MI(F_k;F_1,F_2,...F_{k-2})$. Moreover, in Equation (34): $MI(F_k;F_{k-1}|C,F_1,...F_{k-2}) = 0$, due to the fact that parent and child nodes are independent when conditioned on C. Hence, we conclude that $MI(F_k;C|F_1,F_2,...F_{k-1}) \leq MI(F_k;C|F_1,F_2,...F_{k-2})$. Because F_{k-1} was selected before F_k, we have that $MI(F_k;C|F_1,F_2,...F_{k-2}) < MI(F_{k-1};C|F_1,F_2,...F_{k-2})$. Hence, finally this yields what is to be proven: $MI(F_k;C|F_1,F_2,...F_{k-1}) < MI(F_{k-1};C|F_1,F_2,...F_{k-2})$. Note that we did not need to specify whether F_k is a parent or a child node, we only needed that one node F_k or F_{k-1} was a parent node and the other a child node.

Because, the proof is independent regardless F_k is a child or a parent node, we obtain following corollary of Lemma 4.2.

Corollary 4.3. *Suppose that the order in which features are selected by the SFS is: firstly F_1 subsequently F_2 next F_3 until F_n. Assume that the odd selected features,* i.e. *F_1, F_3, F_5 ..., are children of C and the even selected features,* i.e. *F_2, F_4, F_6 ..., are parents of C, then the decreasing returns behavior holds: $MI(F_1;C) > MI(F_2;C|F_1) > MI(F_3;C|F_1,F_2) > ... > MI(F_n;C|F_1,F_2,...F_{n-1})$.*

We performed an experiment to verify whether it is plausible that parent and child nodes may become selected alternately in the SFS, as Lemma 4.2 and Corollary 4.3 require. We generated 10,000 Bayesian networks with 2 parent and 2 child nodes as shown in Figure 6. This network contains 10 degrees of freedom. The following probabilities can be chosen freely: the prior probabilities $p(f_1=0)$ and $p(f_3=0)$, the conditional probability $p(c=0|f_1,f_3)$ for all 4 combinations of F_1 and F_3, $p(f_2=0|c)$ for the 2 values of C and $p(f_4=0|c)$ for the 2 values of C. In each of the 10,000 networks, probabilities were drawn following a uniform distribution within [0,1]. It can be shown that randomly assigning conditional distributions in this way, this will result almost always in joint distributions that are faithful to the directed acyclic graph (DAG). This means that no conditional independencies are present in the joint distribution that are not entailed by the DAG based on the Markov condition, see e.g., [20] on page 99. Next, the SFS was applied to each of the 10,000 networks. In 943 out of 10,000 cases, a parent node was selected first and parent and child nodes were selected alternately. In 1,125 out of 10,000 cases, a child node was selected first and parent and child nodes were selected alternately. In Table 2 we show the probabilities of a Bayesian network in which first a parent node was selected and the parent and child nodes were selected alternately. The evolution of the mutual information, when the SFS is applied to the Bayesian network with probabilities shown in Table 2, is shown in Figure 7.

4.3. Selection Transitions

We show that Lemmas 4.1 and 4.2 and Corollary 4.3 can be put in a more global theory of allowed selection transitions between features to achieve a decrease in return. When the target variable has both parent and child features, four elementary selection transitions can occur as shown in Figure 8.

Lemma 4.4. *Suppose that feature F_k is just selected after feature F_{k-1} in the SFS. Assume a network with child and parent variables of the target variable C as shown in Figure 8 then a decrease in return must hold, i.e. $MI(F_k;C|F_1,F_2,...F_{k-1}) < MI(F_{k-1};C|F_1,F_2,...F_{k-2})$, in: case 1) F_{k-1} is a child node and F_k is a child node, in case 2) F_{k-1} is a child node and F_k is a parent node and case 3) F_{k-1} is a parent node and F_k is a child node.*

Proof. Case (1) F_{k-1} is a child node and F_k is a child node. The proof proceeds in a similar way as in Lemma 4.2 by starting from the same expansions as in Equations (33) and (34). If F_k and F_{k-1} are both child nodes then it can be proven that $MI(F_k;F_{k-1}|C,F_1,...F_{k-2}) = 0$, holds, even in the case when there are parent nodes. This is shown in Appendix C. For the rest, the proof proceeds the same as in Lemma 4.2.
Case (2) F_{k-1} is a child node and F_k is a parent node. This result was already obtained at the end of the proof of Lemma 4.2 starting from Equations (33) and (34).
Case (3) F_{k-1} is a parent node and F_k is a child node. This result was already obtained at the end of the proof of Lemma 4.2 starting from Equations (33) and (34). □

Table 2. Probabilities for the network shown in Figure 6. These probabilities were obtained from one of the 10,000 Bayesian networks that were generated randomly. Applying the SFS to the network with these probabilities leads to the selection of parent and child nodes alternately.

F_1	F_2	F_3	F_4	C	p(.)
0					$p(F_1) = 0.6596$
1					$p(F_1) = 0.3404$
		0			$p(F_3) = 0.5186$
		1			$p(F_3) = 0.4814$
0		0		0	$p(C\|F_1,F_3) = 0.9730$
0		0		1	$p(C\|F_1,F_3) = 0.0270$
1		0		0	$p(C\|F_1,F_3) = 0.6490$
1		0		1	$p(C\|F_1,F_3) = 0.3510$
0		1		0	$p(C\|F_1,F_3) = 0.8003$
0		1		1	$p(C\|F_1,F_3) = 0.1997$
1		1		0	$p(C\|F_1,F_3) = 0.4538$
1		1		1	$p(C\|F_1,F_3) = 0.5462$
	0			0	$p(F_2\|C) = 0.4324$
	1			0	$p(F_2\|C) = 0.5676$
	0			1	$p(F_2\|C) = 0.8253$
	1			1	$p(F_2\|C) = 0.1747$
			0	0	$p(F_4\|C) = 0.0835$
			1	0	$p(F_4\|C) = 0.9165$
			0	1	$p(F_4\|C) = 0.1332$
			1	1	$p(F_4\|C) = 0.8668$

Let us remark that case (4) does not necessarily exclude a decreasing return. This occurs e.g., when the probability distribution is not faithful to the directed acyclic graph (DAG). In that case it occurs that $MI(F_k;F_{k-1}|C,F_1,...F_{k-2}) = 0$ and hence the decreasing returns holds. This independence is not entailed by the DAG based on the Markov condition [20]. Now let us reinterpret Lemmas 4.1 and 4.2 and Corollary 4.3 in light of Lemma 4.4. Lemma 4.1 only consists of the selection transitions of case 1 and hence the decreasing returns is guaranteed. Lemma 4.2 starts with a parent, next a child is selected

Figure 7. Evolution of the mutual information in function of the number of features selected with the SFS. A Bayesian network according to Figure 6 was created with the probabilities set to values listed in Table 2. The conditional mutual information at 1 feature is MI(F_1;C) at 2 features MI(F_2;C|F_1),... and finally at 4 features MI(F_4;C|F_1,F_2,F_3). Lemma 4.2 predicts that the conditional mutual information decreases with an increasing number of features selected. This implies that the mutual information is concave in function of the number of features selected.

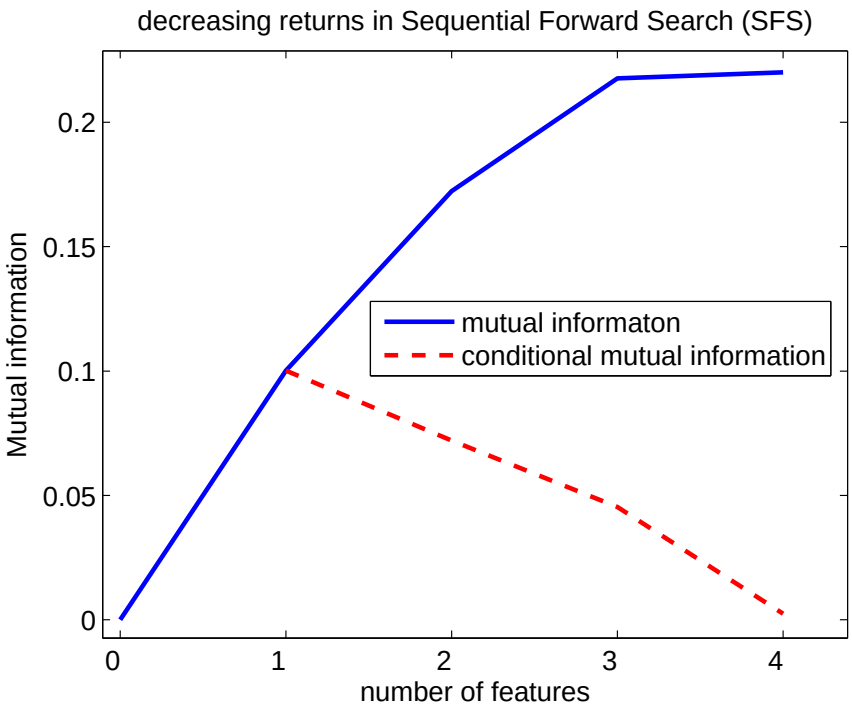
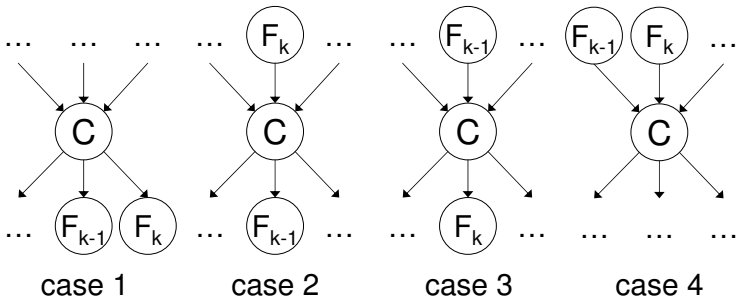


Figure 8. Four elementary selection transitions in the SFS. F_{k-1} is the feature selected at step k-1, F_k is the feature selected at step k. Case 1: F_{k-1} is a child and F_k is a child. Case 2: F_{k-1} is a child and F_k is a parent. Case 3: F_{k-1} is a parent and F_k is a child. Case 4: F_{k-1} is a parent and F_k is a parent.



(*i.e.*, a case 3 transition), next a parent is selected (*i.e.*, a case 2 transition) and so on. Hence, case 3 and case 2 transitions are alternated. Corollary 4.3, starts with a child, next a parent is selected (*i.e.*, a case 2 transition), next a child is selected (*i.e.*, a case 3 transition) and so on. Hence case 2 and case 3 transitions are alternated. Let us remark that also other combinations of cases are possible to guarantee the decreasing returns behavior. A case 3 (parent $\rightarrow$ child) transition is also allowed to be followed by a

case 1 transition (child → child), and a case 1 (child → child) transition is allowed to be followed by a case 2 (child → parent) transition. Finally, we remark that the increasing returns behavior illustrated by the XOR hypercube in Section 4.1 is an example of case 4 transitions.

4.4. Relevance-redundancy Criteria

To avoid the estimation of mutual information in high-dimensional spaces, Battiti [11] proposed a SFS criterion that selects in each iteration the feature with the largest marginal relevance penalized with a redundancy term. Suppose that the set of features selected thus far is **S** and that F_i is a candidate feature to be selected, then the feature F_i is selected for which following criterion is maximal.

$$Crit = MI(F_i;C) - \beta \sum_{F_s \in \boldsymbol{S}} \alpha(F_i, F_s, C) MI(F_i;F_s) \tag{35}$$

In Battiti's work β is a user defined parameter and α (F_i,F_s,C) = 1. Similar criteria were proposed in [23] (for which α (F_i,F_s,C) = $\frac{MI(F_s;C)}{H(F_s)}$), in [24] (for which α (F_i,F_s,C) = 1 and β is adaptively chosen as 1/|S|) and in [25] (for which α (F_i,F_s,C) = 1/min{H(F_i),H(F_s)} and β is adaptively chosen as 1/|S|). All these criteria will not be informative for the examples shown in Sections 3.1, 3.2 and 3.3. These criteria will return for each feature in Equation (35) Crit = 0, because MI(F_i;C) = 0 and MI(F_i;F_s) = 0. Therefore, these criteria may be tempted to include no features at all, despite the fact that all features are strongly relevant. For the 7-5-3 XOR cube all criteria will select the features in the same order: first F_1, then F_2 and then F_3. This is due to the fact that F_1 individually contains more information than F_2 about the target variable, see Table 1. Also F_2 contains more information than F_3 about the target variable, see Table 1. Moreover for the 7-5-3 XOR cube all variables are independent, hence MI(F_i;F_s) = 0. However, from the criterion values Crit = MI(F_1;C), then Crit = MI(F_2;C) and finally Crit = MI(F_3;C) the increasing returns cannot be observed. Another criterion that uses lower-dimensional conditional mutual information to select features was proposed in [26]. This selection algorithm proceeds in 2 stages:

$$F_{s1} = \arg\max_{1 \le i \le n} MI(F_i;C) \tag{36}$$

$$F_{sk} = \arg\max_{j} \left\{ \min_{1 \le i \le k-1} MI(F_j;C|F_{si}) \right\} \tag{37}$$

In the first step in Equation (36) the feature which bears individually most information about the target variable is selected, *i.e.,* F_{s1}. Next, in the k'th step of the second stage, *i.e.,* Equation (37), the feature is selected which contributes most, conditioned on the set of already selected features F_{s1}, F_{s2}, ...F_{sk-1}. The contribution for feature F_j is estimated conservatively as $\min_{1 \le i \le k-1} MI(F_j;C|F_{si})$. This algorithm will be able to detect the increasing returns in the XOR problem in case there are only 2 features. However, it would fail to detect the strongly relevant features in case there are at least 3 features in the XOR problem. To overcome the limitations of the lower-dimensional mutual information estimators, higher-dimensional mutual information estimators for classification purposes were proposed [21,22,27,28]. In [27] the authors proposed a density-based method: the probability density is estimated by means of Parzen windows and the mutual information is estimated from this probability density estimate. In [21,22] the mutual information was estimated based on pair-wise distances between data points. A similar estimator can be used for regression purposes [29]. In [28] the mutual information is estimated also

based on distances between data points, but this time from a minimal spanning tree that is constructed from the data points.

4.5. Importance of Increasing and Decreasing Returns

The importance of the decreasing and increasing returns lies in that we can compute an upper bound and a lower bound on the probability of error, without having to compute the mutual information for higher dimensions. Suppose that the mutual information has been computed up to a subset of $n1$ features $\mathbf{F}_{n1}$, with mutual information MI($\mathbf{F}_{n1}$;C). Suppose that the last increment in going from a subset of $n1-1$ features $\mathbf{F}_{n1-1}$ to $\mathbf{F}_{n1}$ equals ΔMI = MI($\mathbf{F}_{n1}$;C)-MI($\mathbf{F}_{n1-1}$;C). For the mutual information of a subset of n2 features $\mathbf{F}_{n2}$, with $\mathbf{F}_{n2} \supset \mathbf{F}_{n1}$, it holds, under the decreasing returns, that:

$$MI(\boldsymbol{F}_{n1};C) \leq$$
$$MI(\boldsymbol{F}_{n2};C) \leq MI(\boldsymbol{F}_{n1};C) + (|n2| - |n1|)\Delta MI \tag{38}$$

and under the increasing returns that:

$$MI(\boldsymbol{F}_{n2};C) \geq MI(\boldsymbol{F}_{n1};C) + (|n2| - |n1|)\Delta MI \tag{39}$$

For the example shown in Figure 5, it can be seen that ΔMI (the conditional mutual information) at 4 features is representative for the conditional mutual information at 5, 6 and 7 features. The conditional mutual information at 8 features is representative for the ones at 9 and 10 features. From the inequalities in (38) and (39) one can constrain the probability of error that can be achieved by observing the ($|n2|$-$|n1|$) additional features. This can be obtained by exploiting upper and lower bounds that were established for the equivocation H(C|F). These upper and lower bounds can be restated in terms of the mutual information. In Figure 9 the upper bounds are restated in terms of the mutual information as follows for the Hellman-Raviv upper bound [30]:

$$\frac{(H(C) - MI(F;C))}{2} \geq P_e \tag{40}$$

and for the Kovalevsky upper bound [31]:

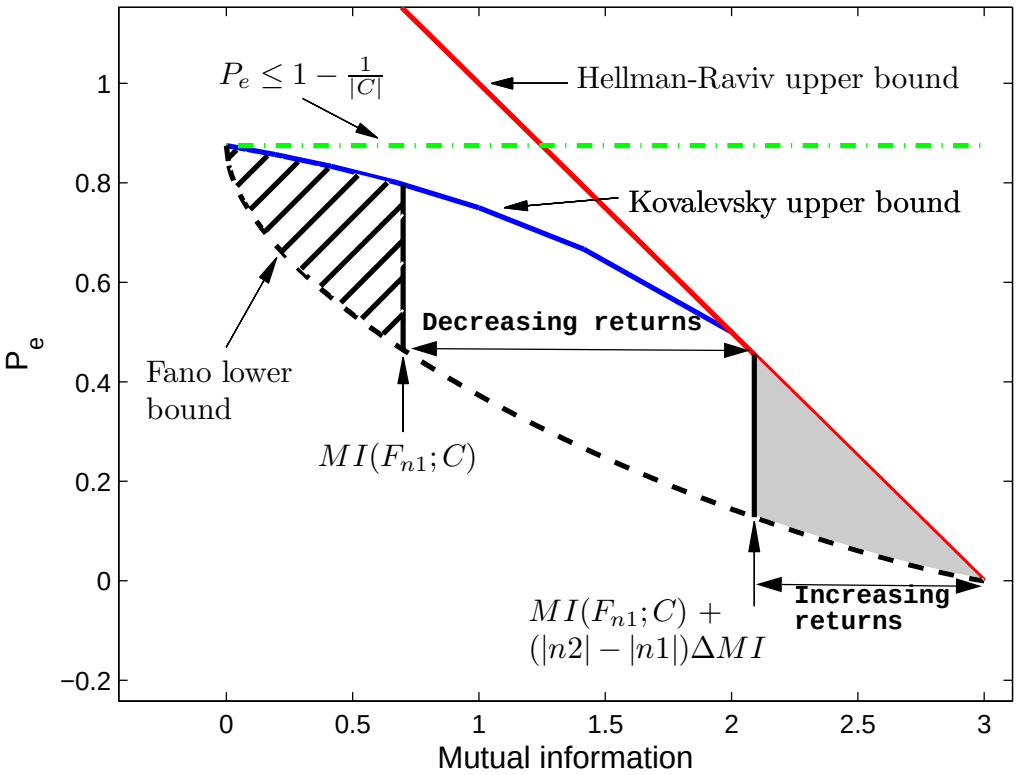
$$H(C) - MI(F;C) \geq \tag{41}$$
$$log_2(i) + i(i+1)(log_2\tfrac{i+1}{i})(P_e - \tfrac{i-1}{i})$$

with 'i' an integer such that $(i-1)/i \leq P_e \leq i/(i+1)$ and 'i' smaller than the number of classes |C|. Let us remark that some of the bounds on the probability of error have been established independently by different researchers. The Hellman-Raviv upper bound has also been found in [32]. The Kovalevsky upper bound on the probability of error has been proposed at least 3 times: first in [31,32] and later in [33]; see also the discussion in [34]. The lower bound in Figure 9 is solved using the Fano lower bound [12,35]:

$$H(C) - MI(F;C) \leq H(P_e) + P_e log_2(\mid C \mid -1) \tag{42}$$

Due to (38) it must be that the probability of error corresponding with $\mathbf{F}_{n2}$ falls within the white area under the decreasing returns, and, due to (39), within the dark grey area under the increasing returns.

Figure 9. Bounds on the probability of error. For subset $\mathbf{F}_{n1}$ the mutual information equals MI($\mathbf{F}_{n1}$;C), for which the probability of error falls between the Fano lower bound and the Kovalevsky upper bound. The white area represents the possible combinations of probability of error and mutual information for the decreasing returns in the selection of (|n2|-|n1|) additional features, because MI($\mathbf{F}_{n2}$;C) $\leq$ MI($\mathbf{F}_{n1}$;C) + (|n2|-|n1|)ΔMI. The grey area is the possible area for the increasing returns. The hatched area is not possible, because adding features can only increase the information. This figure illustrates the case when the number of classes |C| is equal to 8 and when all prior probabilities of the classes are equal.



5. Decreasing Losses and Increasing Losses

We show that the increasing returns and the decreasing returns for the SFS has a comparable implication in the sequential backward search (SBS): the comparable behavior for the increasing returns is the decreasing losses in the SBS, the comparable behavior for the decreasing returns is the increasing losses in the SBS.

5.1. Decreasing Losses

In the SBS, the feature, for which the information loss is minimal, is removed in every iteration. Starting from the example in Figure 3, one computes the 3 information losses: MI(F_1,F_2,F_3;C) $-$ MI(F_2,F_3;C) = MI(F_1;C|F_2,F_3), MI(F_2;C|F_1,F_3) and MI(F_3;C|F_1,F_2). One then removes the feature F_i for which MI(F_i;C|F_j,F_k) is minimal. We show the computations of the mutual information for the 7-5-3 XOR cube in Table 3. In the SBS for this example, we first remove F_3, because MI(F_3;C|F_1,F_2) is

Table 3. 7-5-3 XOR Cube. Mutual Information for the SBS.

Mutual Information	value(bit)
MI(F_1,F_2,F_3;C)	$\approx$ 0,9999
MI(F_1;C|F_2,F_3)	$\approx$ 0,9852
MI(F_2;C|F_1,F_3)	$\approx$ 0,9710
MI(F_3;C|F_1,F_2)	$\approx$ 0,9183
MI(F_1;C|F_2)	$\approx 8{,}007.10^{-2}$
MI(F_2;C|F_1)	$\approx 7{,}850.10^{-2}$
MI(F_1;C)	$\approx 3{,}143.10^{-3}$

the smallest information loss for the set of 3 features. Next F_2 is removed, because MI(F_2;C|F_1) is the smallest for sets of 2 features and, finally, feature F_1 remains. Instead of the increasing returns in the SFS, we observe now 'decreasing losses' in the SBS: MI(F_3;C|F_1,F_2) $\approx$ 0,9183 $>$ MI(F_2;C|F_1) $\approx 7{,}850.10^{-2} >$ MI(F_1;C) $\approx 3{,}143.10^{-3}$. The 7-5-3 XOR cube also illustrates that, for this type of problems, the SBS can outperform the SFS. The initial small increments in the SFS are close to 0 (in the order of 10^{-3}): for small values, it may be tempting to stop the SFS too early. In XOR-type problems, when the number of values the features can take are even, e.g., in the 8-8-8 XOR cube, the situation is even worse. All increments in the SFS are equal to 0, except only in the last iteration of the SFS, a large increment in the mutual information is observed. In the SBS, this problem is not encountered. In the first iteration of the SBS, a large information loss is observed immediately ($\approx$ 0,9183). One concludes immediately correctly that one should not remove any features at all.

5.2. *Increasing Losses*

Similar as in the case of decreasing returns in the SFS, it can be questioned under which conditions 'increasing losses' can be observed. Suppose that, in our SBS example, F_y is removed before F_x. We can use the 2 expansions of (22). When F_y is removed before F_x, it must be that: MI(F_y;C|**S**,F_x) $<$ MI(F_x;C|**S**,F_y). Combining this inequality with (22), it is clear that: MI(F_x;C|**S**) $>$ MI(F_y;C|**S**). Hence, under the condition that MI(F_y;C|**S**) $\geq$ MI(F_y;C|**S**,F_x) one obtains an 'increasing losses' behavior: MI(F_x;C|**S**) $>$ MI(F_y;C|**S**,F_x).

Lemmas comparable to Lemma 4.1 and Lemma 4.2 are obtained for the SBS. For Lemma 4.1 this leads in the SBS to the following lemma.

Lemma 5.1. *Suppose that the order in which features are removed by the SBS is: firstly F_n subsequently F_{n-1} next F_{n-2} until F_1. If all features are conditionally independent given the class variable, i.e., $p(F_1,F_2, ...F_n|C) = \prod_{i=1}^{n} p(F_i|C)$, then the increasing losses behavior holds: $MI(F_n;C|F_1,F_2,...F_{n-1}) < MI(F_{n-1};C|F_1,F_2,...F_{n-2}) < MI(F_{n-2};C|F_1,F_2,...F_{n-3}) < ... < MI(F_2;C|F_1) < MI(F_1;C)$.*

The proof proceeds in a similar way as in Lemma 4.1, but starts from a different mutual information than Equation (26).

$$MI(F_{k-1}; F_1, F_2, ...F_{k-2}, F_k, C) = MI(F_{k-1}; F_1, F_2, ...F_{k-2}, C) + MI(F_{k-1}; F_k|C, F_1, F_2, ...F_{k-2}) \quad (43)$$

Similar as in Lemma 4.1, it holds that: MI(F_{k-1};F_k|C,F_1,F_2,...F_{k-2}) = 0. Further expanding of the left and the right hand sides of Equation (43) results in:

$$MI(F_{k-1}; F_1, F_2, ...F_{k-2}, F_k, C) = MI(F_{k-1}; F_1, F_2, ...F_{k-2}, C) \quad (44)$$

$$\begin{aligned} MI(F_{k-1}; F_1, F_2, ...F_{k-2}, F_k) + MI(F_{k-1}; C|F_1, F_2, ...F_{k-2}, F_k) &= MI(F_{k-1}; F_1, F_2, ...F_{k-2}) \\ &+ MI(F_{k-1}; C|F_1, F_2, ...F_{k-2}) \end{aligned} \quad (45)$$

Because MI(F_{k-1};F_1,F_2,...F_{k-2},F_k) $\geq$ MI(F_{k-1};F_1,F_2,...F_{k-2}) in Equation (45), we see that MI(F_{k-1};C|F_1,F_2,...F_{k-2},F_k) $\leq$ MI(F_{k-1};C|F_1,F_2,...F_{k-2}). But, F_k was removed before F_{k-1}, *i.e.*, MI(F_k;C|F_1,F_2,...F_{k-1}) < MI(F_{k-1};C|F_1,F_2,...F_{k-2},F_k), from which we obtain what needed to be proven: MI(F_k;C|F_1,F_2,...F_{k-1}) < MI(F_{k-1};C|F_1,F_2,...F_{k-2}).

We generated a Bayesian network as in Figure 4 with the free parameters randomly drawn following a uniform distribution within [0,1]. According to Lemma 5.1, we should find the increasing losses behavior if we apply the SBS to this network. The result of the SBS is shown in Figure 10.

Similar to Lemma 4.2 when there are both child and parent nodes this leads in the SBS to the following lemma.

Lemma 5.2. *Suppose that the order in which features are removed by the SBS is: firstly F_n subsequently F_{n-1} next F_{n-2} until F_1. Assume that odd removed features, F_{n-1}, F_{n-3},...F_3, F_1, are parents of C and even removed features, F_n, F_{n-2},...F_4, F_2, are children of C, then the increasing losses behavior holds: MI(F_n;C|F_1,F_2,...F_{n-1}) < MI(F_{n-1};C|F_1,F_2,...F_{n-2}) < MI(F_{n-2};C|F_1,F_2,...F_{n-3}) < ... < MI(F_2;C|F_1) < MI(F_1;C).*

Proof. The proof proceeds similar as in Lemma 4.2, but starts from a slightly different mutual information then Equation (33).

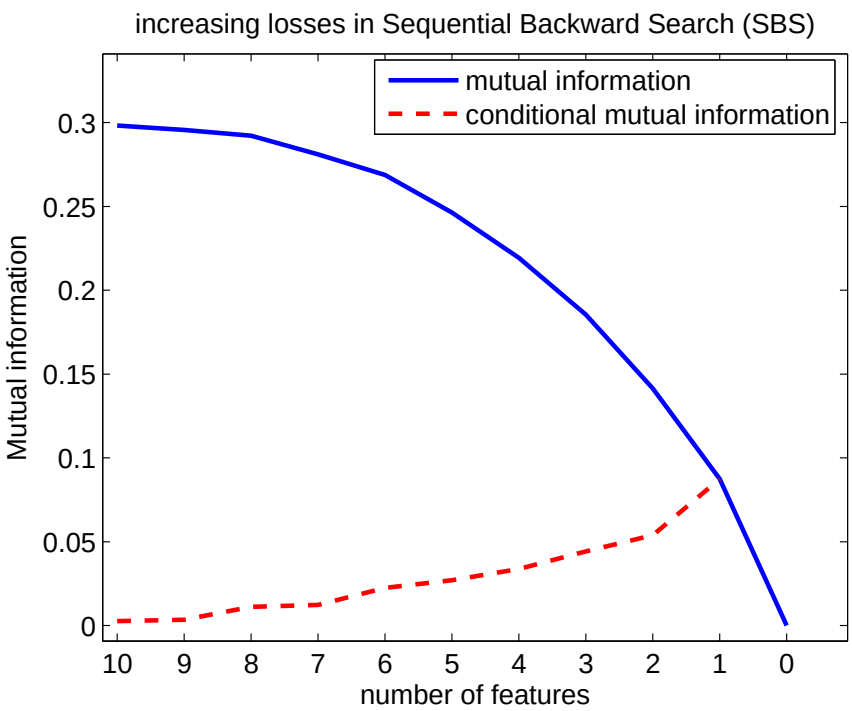
$$MI(F_{k-1}; C, F_1, F_2, ...F_{k-2}, F_k) = MI(F_{k-1}; F_1, F_2, ...F_{k-2}, F_k) + MI(F_{k-1}; C|F_1, F_2, ...F_{k-2}, F_k) \quad (46)$$

$$\begin{aligned} &= MI(F_{k-1}; F_1, F_2, ...F_{k-2}) + MI(F_{k-1}; C|F_1, F_2, ...F_{k-2}) \\ &\quad + MI(F_{k-1}; F_k|C, F_1, ...F_{k-2}) \end{aligned} \quad (47)$$

We do not need to specify that F_k is a parent or a child node, we only need that one node F_k or F_{k-1} is a parent node and the other a child node. Comparing Equations (46) and (47), we have that: MI(F_{k-1};F_1,F_2,...F_{k-2},F_k) $\geq$ MI(F_{k-1};F_1,F_2,...F_{k-2}). Moreover in Equation (47) MI(F_{k-1};F_k|C,F_1,...F_{k-2}) = 0 due to the fact that parent and child nodes are

independent when conditioned on C. Hence, we conclude that $MI(F_{k-1};C|F_1,F_2,...F_{k-2},F_k) \leq MI(F_{k-1};C|F_1,F_2,...F_{k-2})$. Because F_k was removed before F_{k-1}, we have that $MI(F_k;C|F_1,F_2,...F_{k-1}) < MI(F_{k-1};C|F_1,F_2,...F_{k-2},F_k)$. Hence, finally this yields what is to be proven: $MI(F_k;C|F_1,F_2,...F_{k-1}) < MI(F_{k-1};C|F_1,F_2,...F_{k-2})$. □

Figure 10. Evolution of the mutual information in function of the number of features selected with the SBS. A Bayesian network according to Figure 4 was created with probability p(c=0), conditional probabilities $p(f_i=0|c=0)$ and $p(f_i=0|c=1)$ drawn randomly from a uniform distribution within [0,1]. The conditional mutual information for 10 features is $MI(F_{10};C|F_1,F_2,...F_9)$, for 9 features $MI(F_9;C|F_1,F_2,...F_8)$,... and, finally for 1 feature $MI(F_1;C)$. Lemma 5.1 predicts that the conditional mutual information increases with increasing number of features removed. This implies that the mutual information is concave in function of the number of features selected.



Similar to Corollary 4.3, we obtain following corollary in the SBS.

Corollary 5.3. *Suppose that the order in which features are removed by the SBS is: firstly F_n subsequently F_{n-1} next F_{n-2} until F_1. Assume that odd removed features, F_{n-1}, F_{n-3},...F_3, F_1, are children of C and even removed features, F_n, F_{n-2},...F_4, F_2, are parents of C, then the increasing losses behavior holds: $MI(F_n;C|F_1,F_2,...F_{n-1}) < MI(F_{n-1};C|F_1,F_2,...F_{n-2}) < MI(F_{n-2};C|F_1,F_2,...F_{n-3}) < ... < MI(F_2;C|F_1) < MI(F_1;C)$.*

Finally, we can relate Lemmas 5.1 and 5.2 and Corollary 5.3 to the allowed selection transitions of Section 4.3.

Lemma 5.4. *Suppose that feature F_{k-1} is just removed after feature F_k has been removed in the SBS. Assume a network with child and parent variables of the target variable C as shown in Figure 8 then an*

increase in loss must hold, i.e. *MI(F_k;C|F_1,F_2,...F_{k-1}) < MI(F_{k-1};C|F_1,F_2,...F_{k-2}), in: case 1) F_{k-1} is a child node and F_k is a child node, in case 2) F_{k-1} is a child node and F_k is a parent node and case 3) F_{k-1} is a parent node and F_k is a child node.*

Proof. The proof is obtained starting from the expansions in Equations (46) and (47). It holds for all 3 cases that MI(F_{k-1};F_k|C,F_1,...F_{k-2}) = 0, then the proof proceeds similar as in Lemma 5.2. □

Interpreting Figure 8 now as 'F_{k-1} is removed just after F_k has been removed', following combinations can be made to guarantee an increasing losses behavior. Case 1 (child → child) can be followed by case 1 (child → child) and by case 3 (child → parent). Case 2 (parent → child) can be followed by case 1 (child → child) or by case 3 (child → parent). Case 3 (child → parent) can be followed by case 2 (parent → child).

6. Conclusions

This work contributes to a more thorough understanding of the evolution of the mutual information in function of the number of features that are selected in the sequential forward search (SFS) and in the sequential backward search (SBS) strategies. Conditioning on additional features can increase the mutual information about the target variable for discrete features (binary as well as non-binary) and continuous features. Increments in mutual information can become larger and larger in the sequential forward search, a behavior we described as 'increasing returns'. An example of increasing returns was constructed using a $(2n+1)-(2n-1)-...-5-3$ XOR hypercube. It was shown that, when conditioning on additional variables reduces information about the target variable, then this is a sufficient condition for the decreasing returns to hold in the sequential forward search. We provided examples of dependencies between features and the target variable from which the decreasing returns behavior could be proven to occur. If features are conditionally independent given the target variable, the decreasing returns behavior is proven to be guaranteed. Even in the case of more complex dependencies, when there are both child and parent variables, the decreasing returns was proven to occur when parent and child variables would be selected alternately by the SFS. The analogous behaviors in the mutual information based SBS are: 'decreasing losses' and 'increasing losses'. Similar to the SFS, if conditioning on additional variables reduces information about the target variable, then this is a sufficient condition for the increasing losses to hold in the SBS. If the features are conditionally independent given the target variable, the increasing losses behavior is proven to occur. If parent and child variables would be removed alternately by the SBS, the increasing losses behavior is also proven to occur. Lemmas were supported by experimental results.

Acknowledgments

GVD is supported by the CREA Financing (CREA/07/027) program of the K.U.Leuven. MMVH is supported by research grants received from the Excellence Financing program (EF 2005), the Belgian Fund for Scientific Research Flanders (G.0588.09), the Interuniversity Attraction Poles Programme Belgian Science Policy (IUAP P6/054), the Flemish Regional Ministry of Education (Belgium) (GOA 10/019), and the European Commission (IST-2007-217077).

References

1. Liu, H.; Motoda, H. *Computational Methods of Feature Selection*; Chapman & Hall/CRC: Boca Raton, FL, USA, 2007.
2. Van Dijck, G.; Van Vaerenbergh, J.; Van Hulle, M.M. Posterior probability profiles for the automated assessment of the recovery of patients with stroke from activity of daily living tasks. *Artif. Intell. Med.* **2009**, *46*, 233–249.
3. Duda, R.O.; Hart, P.E.; Stork, D.G. *Pattern Classification*, second ed.; John Wiley & Sons: New York, NY, USA, 2001.
4. Lewis II, P.M. The characteristic selection problem in recognition systems. *IEEE Trans. Inf. Theory* **1962**, *8*, 171–178.
5. Guyon, I.; Elisseeff, A. An introduction to variable and feature selection. *J. Mach. Learn. Res.* **2003**, *3*, 1157–1182.
6. Wang, G.; Lochovsky, F.H.; Yang, Q. Feature selection with conditional mutual information maximin in text categorization. Proceedings of the 13th ACM International Conference on Information and Knowledge Management (CIKM'04); Evans, D.A.; Gravano, L.; Herzog, O.; Zhai, C.; Ronthaler, M., Eds.; ACM Press: New York, NY, USA, 2004; pp. 342–349.
7. Guo, B.; Nixon, M.S. Gait feature subset selection by mutual information. *IEEE Trans. Syst. Man Cybern. Part A-Syst. Hum.* **2009**, *29*, 36–46.
8. Huang, D.; Chow, T.W.S.; Wa, E.W.M.; Li, J. Efficient selection of discriminative genes from microarray gene expression data for cancer diagnosis. *IEEE Trans. Circuits Syst. I-Regul. Pap.* **2005**, *52*, 1909–1918.
9. Kamentsky, L.A.; Liu, C.N. Computer-automated design of multifont print recognition logic. *IBM J. Res. Dev.* **1963**, *7*, 2–13.
10. Liu, C.N. A programmed algorithm for designing multifont character recognition logics. *IEEE Trans. Electron.* **1964**, *EC-13*, 586–593.
11. Battiti, R. Using mutual information for selecting features in supervised neural net learning. *IEEE Trans. Neural Netw.* **1994**, *5*, 537–550.
12. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*, second ed.; John Wiley & Sons: Hoboken, NJ, USA, 2006.
13. Liu, D.; Chang, T.; Zhang, Y. A constructive algorithm for feedforward neural networks with incremental training. *IEEE Trans. Circuits Syst. I-Regul. Pap.* **2002**, *49*, 1876–1879.
14. McGill, W.J. Multivariate information transmission. *IEEE Trans. Inf. Theory* **1954**, *4*, 93–111.
15. Matsuda, H. Physical nature of higher-order mutual information: Intrinsic correlations and frustration. *Phys. Rev. E* **2000**, *62*, 3096–3102.
16. Shiono, S.; Yamada, S.; Nakashima, M.; Matsumoto, K., Information theoretic analysis of connection structure from spike trains. In *Advances in Neural Information Processing Systems 5*; Hanson, S.J.; Cowan, J.D.; Giles, C.L., Eds.; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 1993; pp. 515–522.

17. Bontempi, G.; Meyer, P.E., Causal filter selection in microarray data. In *Proceedings of the 27th International Conference on Machine Learning*; Fürnkranz, J.; Joachims, T., Eds.; Omnipress, 2010; pp. 95–102.
18. Kotz, S.; Nadarajah, S. *Multivariate t Distributions and Their Applications*; Cambridge University Press: Cambridge, UK, 2004; pp. 15–16.
19. Pearl, J. *Probabilistic Reasoning in Intelligent Systems*; Morgan Kaufmann: San Francisco, California, USA, 1988.
20. Neapolitan, R.E. *Learning Bayesian Networks*; Prentice-Hall, Inc.: Upper Saddle River, NJ, USA, 2003.
21. Van Dijck, G.; Van Hulle, M.M. Speeding up feature subset selection through mutual information relevance filtering. In *Knowledge Discovery in Databases: PKDD 2007*; Kok, J.; Koronacki, J.; Lopez de Mantaras, R.; Matwin, S.; Mladenic, D.; Skowron, A., Eds.; Springer Berlin / Heidelberg, 2007; Vol. 4702, *Lecture Notes in Computer Science*, pp. 277–287.
22. Van Dijck, G. Information Theoretic Approach to Feature Selection and Redundancy Assessment. PhD dissertation, Katholieke Universiteit Leuven, 2008.
23. Kwak, N.; Choi, C.H. Input feature selection for classification problems. *IEEE Trans. Neural Netw.* **2002**, *13*, 143–159.
24. 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.* **2005**, *27*, 1226–1238.
25. Estévez, P.A.; Tesmer, M.; Perez, C.A.; Zurada, J.M. Normalized mutual information feature selection. *IEEE Trans. Neural Netw.* **2009**, *20*, 189–201.
26. Fleuret, F. Fast binary feature selection with conditional mutual information. *J. Mach. Learn. Res.* **2004**, *5*, 1531–1555.
27. Kwak, N.; Choi, C.H. Input feature selection by mutual information based on Parzen window. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 1667–1671.
28. Bonev, B.; Escolano, F.; Cazorla, M. Feature selection, mutual information, and the classification of high-dimensional patterns: applications to image classification and microarray data analysis. *Pattern Anal. Appl.* **2008**, *11*, 309–319.
29. François, D.; Rossi, F.; Wertz, V.; Verleysen, M. Resampling methods for parameter-free and robust feature selection with mutual information. *Neurocomputing* **2007**, *70*, 1276–1288.
30. Hellman, M.E.; Raviv, J. Probability of error, equivocation, and the Chernoff bound. *IEEE Trans. Inf. Theory* **1970**, *IT-16*, 368–372.
31. Kovalevsky, V.A. The problem of character recognition from the point of view of mathematical statistics. In *Character Readers and Pattern Recognition*; Kovalevsky, V.A., Ed.; Spartan: New York, NY, USA, 1968.
32. Tebbe, D.L.; Dwyer III, S.J. Uncertainty and the probability of error. *IEEE Trans. Inf. Theory* **1968**, *IT-14*, 516–518.
33. Feder, M.; Merhav, N. Relations between entropy and error probability. *IEEE Trans. Inf. Theory* **1994**, *40*, 259–266.

34. Golić, J.D. Comment on "Relations between entropy and error probability". *IEEE Trans. Inf. Theory* **1999**, *45*, 372–372.
35. Fano, R.M. *Transmission of Information: A Statistical Theory of Communication*; John Wiley & Sons: New York, NY, USA, 1961.

Appendix A

Decompose MI(F_1,F_3;F_2,F_4 |C) = 0 in 2 steps as follows:

$$\begin{aligned} MI(F_1, F_3; F_2, F_4|C) &= MI(F_1; F_2, F_4|C) + MI(F_3; F_2, F_4|F_1, C) \\ &= MI(F_1; F_2, F_4|C) + MI(F_3; F_2|F_1, C) + MI(F_3; F_4|F_1, F_2, C) = 0 \end{aligned}$$

Given that the MI ≥ 0, it must be that every term in the last expression is equal to 0, hence MI(F_3;F_2|F_1,C) = 0.

Appendix B

Given that all features are class conditional independent p(F_1,F_2, ...F_n|C) = $\prod_{i=1}^{n}$ p(F_i|C), we can show that: MI(F_k;F_{k-1}|C,F_1,...F_{k-2}) = 0.

$$\begin{aligned} & MI(F_k; F_{k-1}|C, F_1, ...F_{k-2}) \\ &= \sum_{f_1, f_2, ...f_k, c} p(f_1, f_2, ...f_k, c) log\left(\frac{p(f_k, f_{k-1}|c, f_1, ...f_{k-2})}{p(f_k|c, f_1, ...f_{k-2}).p(f_{k-1}|c, f_1, ...f_{k-2})}\right) \end{aligned} \tag{48}$$

It can be shown that the fraction within the logarithm is always equal to 1.

$$\frac{p(f_k, f_{k-1}|c, f_1, ...f_{k-2})}{p(f_k|c, f_1, ...f_{k-2}).p(f_{k-1}|c, f_1, ...f_{k-2})}$$

$$= \frac{p(f_1, f_2, ...f_k, c).p(f_1, f_2, ...f_{k-2}, c)}{p(f_1, f_2, ...f_{k-2}, f_k, c).p(f_1, f_2, ...f_{k-1}, c)} \tag{49}$$

$$= \frac{\Big(p(f_1|c).p(f_2|c)...p(f_k|c).p(c)\Big).\Big(p(f_1|c).p(f_2|c)...p(f_{k-2}|c).p(c)\Big)}{\Big(p(f_1|c).p(f_2|c)...p(f_{k-2}|c).p(f_k|c).p(c)\Big).\Big(p(f_1|c).p(f_2|c)...p(f_{k-1}|c).p(c)\Big)} \tag{50}$$

$$= \frac{p(f_{k-1}|c)}{1}.\frac{1}{p(f_{k-1}|c)} \tag{51}$$

$$= 1 \tag{52}$$

This implies that the conditional mutual information must be 0.

Appendix C

Denote the set of the first k selected features by $\mathbf{F}_{1:k}$ = {F_1, F_2, ...F_k}. Denote the 'i'th selected parent of C within $\mathbf{F}_{1:k}$ by pa_i(c) and the 'j'th selected child of C within $\mathbf{F}_{1:k}$ by ch_j(c). Denote the set of all parents of C within $\mathbf{F}_{1:k}$ by $\mathbf{F}_{pa}$(c) = $\bigcup_{i=1}^{\#parents} pa_i(c)$ and the set of all children of C within $\mathbf{F}_{1:k}$ by $\mathbf{F}_{ch}$(c) = $\bigcup_{j=1}^{\#children} ch_j(c)$. We want to show that if F_k and F_{k-1} are children of C then

MI(F_k;F_{k-1}|C,F_1,...F_{k-2}) = 0. The definition of MI(F_k;F_{k-1}|C,F_1,...F_{k-2}) is equal to Equation (48). Starting from the term within the logarithm, Equation (49):

$$\frac{p(f_1, f_2, ...f_k, c).p(f_1, f_2, ...f_{k-2}, c)}{p(f_1, f_2, ...f_{k-2}, f_k, c).p(f_1, f_2, ...f_{k-1}, c)} \tag{53}$$

The four factors in Equation (53) can be factorized as:

$$p(f_1, f_2, ...f_k, c) = \Big(\prod_{\substack{ch_j(c) \\ \in \boldsymbol{F}_{ch}(c) \setminus \{F_{k-1}, F_k\}}} p(ch_j(c)|c) \Big).p(f_{k-1}|c).p(f_k|c).p(c|\boldsymbol{F}_{pa}(c)). \prod_{pa_i(c) \in \boldsymbol{F}_{pa}(c)} pa_i(c) \tag{54}$$

$$p(f_1, f_2, ...f_{k-2}, c) = \Big(\prod_{ch_j(c) \in \boldsymbol{F}_{ch}(c) \setminus \{F_{k-1}, F_k\}} p(ch_j(c)|c) \Big).p(c|\boldsymbol{F}_{pa}(c)). \prod_{pa_i(c) \in \boldsymbol{F}_{pa}(c)} pa_i(c) \tag{55}$$

$$p(f_1, f_2, ...f_{k-2}, f_k, c) = \Big(\prod_{ch_j(c) \in \boldsymbol{F}_{ch}(c) \setminus \{F_{k-1}, F_k\}} p(ch_j(c)|c) \Big).p(f_k|c).p(c|\boldsymbol{F}_{pa}(c)). \prod_{pa_i(c) \in \boldsymbol{F}_{pa}(c)} pa_i(c) \tag{56}$$

$$p(f_1, f_2, ...f_{k-1}, c) = \Big(\prod_{ch_j(c) \in \boldsymbol{F}_{ch}(c) \setminus \{F_{k-1}, F_k\}} p(ch_j(c)|c) \Big).p(f_{k-1}|c).p(c|\boldsymbol{F}_{pa}(c)). \prod_{pa_i(c) \in \boldsymbol{F}_{pa}(c)} pa_i(c) \tag{57}$$

The set of all children of C with F_{k-1} and F_k excluded was written in previous equations as the set difference between $\mathbf{F}_{ch}$ and the set { F_{k-1},F_k }: $\mathbf{F}_{ch}$(c) \ {F_{k-1},F_k}. We notice that all probabilities in Equations (54) to (57) have following factor in common:

$$\Big(\prod_{ch_j(c) \in \boldsymbol{F}_{ch}(c) \setminus \{F_{k-1}, F_k\}} p(ch_j(c)|c) \Big).p(c|\boldsymbol{F}_{pa}(c)). \prod_{pa_i(c) \in \boldsymbol{F}_{pa}(c)} pa_i(c) \tag{58}$$

After filling out Equations (54) to (57) in Equation (53) and removing the common factor of Equation (58), we obtain:

$$\frac{\big(p(f_{k-1}|c).p(f_k|c)\big)}{\big(p(f_k|c)\big).\big(p(f_{k-1}|c)\big)} = 1 \tag{59}$$

This implies that the conditional mutual information MI(F_k;F_{k-1}|C,F_1,...F_{k-2}) must be 0.

© 2010 by the authors; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution license http://creativecommons.org/licenses/by/3.0/.