INFORMATION-THEORETIC FEATURE SELECTION IN MICROARRAY DATA USING VARIABLE COMPLEMENTARITY

PATRICK E. MEYER, COLAS SCHRETTER AND GIANLUCA BONTEMPI

{pmeyer,cschrett,gbonte}@ulb.ac.be
http://www.ulb.ac.be/di/mlg/
Machine Learning Group
Computer Science Department
Université Libre de Bruxelles
CP 212 - 1050 Brussels - Belgium

Abstract. The paper presents an original filter approach for effective feature selection in microarray data characterized by a large number of input variables and a few samples. The approach is based on the use of a new information-theoretic selection, the Double Input Symmetrical Relevance (DISR), which relies on a measure of variable complementarity. This measure evaluates the additional information that a set of variables provides about the output with respect to the sum of each single variable contribution. We show that a variable selection approach based on DISR can be formulated as a quadratic optimization problem - the Dispersion Sum Problem. To solve this problem, we use a strategy based on Backward Elimination and Sequential Replacement (BESR). The combination of BESR and the DISR criterion is compared in theoretical and experimental terms to recently proposed information-theoretic criteria. Experimental results on a synthetic dataset as well as on a set of 11 microarray classification tasks show that the proposed technique is competitive with existing filter selection methods.

Information-theoretic feature selection, variable interaction, variable complementarity

1. Introduction

Statisticians and data-miners are used to building predictive models and infer dependencies between variables on the basis of observed data. In microarray analysis, they are facing datasets characterized by a very large number of features, a large amount of noise, nonlinear dependencies and, often, only a few hundreds of samples. A well-known example is cancer classification where the goal is, for example, to distinguish between tumor classes or predict the effects of medical treatments on the basis of gene expression

Date: Corrected version (2009) of the IEEE JSTSP article.

profiles. Here, the number of input variables, represented by the number of gene probes, is huge (around several thousands) while the number of samples, represented by the clinical trials, is very limited (a few tens). In this context, the detection of functional relationships as well as the design of effective classifiers appear to be a major challenge. Because of interpretability, numerical and statistical accuracy issues, it appears to be necessary to reduce the number of variables before starting a learning procedure. This is also relevant in clinical terms since the selection of an appropriate subset of genes (aka cancer signature) can return prognostic subgroups that are relevant to patient management and to the prediction of therapy response ([41][42]). In machine learning jargon, gene selection is an example of f*eature selection task* whose objective is to select, among a set of input variables, the ones that will lead to the best predictive model. Feature selection can also increase the intelligibility of a model, meanwhile, decreasing the measurements and storage requirements ([18]). A number of experimental studies ([21, 7, 34]) have shown that irrelevant and redundant features can dramatically reduce the predictive accuracy of models built from data.

Feature selection is a stochastic optimization problem where the search for the best subset among the features is selected on the basis of statistical estimates of the accuracy of a given subset. Two well-known approaches in feature selection combine a search strategy with a stochastic evaluation function: the *wrapper approach* and the *filter approach* (see [7, 21]). In the wrapper approach, the evaluation function is the validation outcome (e.g. by leave-one-out) of the learning algorithm itself. In the filter approach, examples of evaluation functions are probabilistic distance, interclass distance, information-theoretic or probabilistic dependence measures ([11]). These measures are often considered as intrinsic properties of the data, because they are calculated directly on the raw data instead of requiring a learning model that smoothes distributions or reduces the noise.

This paper will focus on filter techniques which use mutual information for feature selection from microarray data in supervised classification tasks. Mutual information is a well-known statistical measure which has been used in data analysis for defining concepts like variable *relevance* ([37, 4, 22]), *redundancy* ([26, 43, 45]) and *interaction* ([26, 20, 22]). Redundancy and relevance notions have been used in several filter algorithms like Ranking ([16]), Markov blanket filter ([23]), Fast Correlation Based Filter (FCBF) ([45]), Relevance Filter ([3, 4]), Minimum Redundancy Maximum Relevance (MRMR) ([32]) and the Conditional Mutual Information Maximization (CMIM) ([17]) filter.

We present an original filter approach for microarray data based on a new information theoretic selection criterion, called the *Double Input Symmetrical Relevance* (DISR) and its effective implementation, called MASSIVE (Matrix of Average Sub-Subset Information for Variable Elimination).

The DISR criterion combines two well-known properties of feature selection: first, a combination of variables can return more information on the

output class than the sum of the information returned by each of the variables taken individually. This property is made explicit by the notion of *variable complementarity* ([26, 20]). Secondly, in the absence of any further knowledge on how subsets of d variables should combine, it is intuitive to assume a combination of the best performing subsets of $d-1$ variables as the most promising set. This property will be made formal by the computation of a lower-bound on the information of a set of variables expressed as the average of information of all its subsets. The variable selection problem using DISR can be reduced to a well-known optimization problem called the *DSP* (*Densest Subgraph Problem* or *Dispersion Sum Problem*, see [2, 33]). For that reason our implementation (called MASSIVE) adopts a *Backward Elimination combined with Sequential Replacement* (BESR) strategy, an heuristic known to be efficient for DSP problems ([6]). It will be shown that this implementation, though based on a backward strategy, has a cost comparable to that of state-of-the-art filter algorithms, typically based on forward selection search strategies.

The MASSIVE strategy has been benchmarked against several state-of-the-art filters in an experimental classification session. In particular, we considered: a Ranking algorithm ([16]), the Fast Correlation Based Filter ([45]) and three filters based on the same search strategy (forward selection): the Relevance filter ([3, 4]), the Minimum Redundancy Maximum Relevance filter ([32]) and the Conditional Mutual Information Maximization filter ([17]). The assessment is done by measuring the classification accuracy of two learning algorithms which use as inputs the variables returned by each of the filter methods. We use a synthetically generated dataset and 11 public-domain multi-class microarray gene expression datasets in order to compare the filters. The experimental results show that the proposed technique is competitive with existing filter selection methods.

The outline of the paper is the following. Section 2 will introduce the basic notations and information-theoretic notions used in the paper. Section 3 introduces three state-of-the-art definitions: the relevance, the redundancy and the complementarity. Section 4 recalls five information-theoretic feature selection techniques which have been proposed in literature to deal with the variable selection problem. Section 5 provides the theoretical justification of the DISR information-theoretic criterion and its strengthes for dealing with microarray data. This section also formulates the optimization problem as a known problem: the DSP. Section 6 illustrates the experimental results and the conclusions are summarized in Section 7.

2. Framework

This paper deals with cancer classification from microarray data, hence supervised and multi-class, where all the variables are either discrete or quantized. We will denote by Y the discrete output random variable representing the type of cancer and by X the multi-dimensional discrete input random

vector representing the expressions of the genes. The list of notations that we use follows:
$X = \{X_j : \ j \in A = \{1, ..., n\}\}$ is the set of input variables.
$X_S = \{X_j : \ j \in S \subset A\}$ is a subset of input variables.
$X_{i,j} = \{X_i, X_j\}$ is the subset made of X_i and X_j.
$X_{i,S} = \{X_i, X_S\}$ is a the subset composed of X_i and all the variables of X_S.
$X_{S-i} = \{X_j : \ j \in \{S \setminus \{i\}\}$ is the subset of input variables X_S with the variable X_i put aside.
$X_{-i} = \{X_j : \ j \in A \setminus \{i\}\}$ is the whole set of input variables with X_i removed.
$X_{-S} = \{X_j : \ j \in A \setminus S\}$ is the whole set of input variables with those in X_S removed.
$X_{-(i,S)}$ is the whole set of input variables with those in $X_{i,S}$ removed.
Now, we review some elementary notions of information theory (see [10]) which will be used in the following.

Definition 1. The entropy of a discrete random variable Y with domain (or alphabet) $\mathcal{Y}$ and probability distribution $p(y) = \Pr\{Y = y\}, \ y \in \mathcal{Y}$ is defined by:

$$H(Y) = -\sum_{y \in \mathcal{Y}} p(y) \log p(y) \tag{2.1}$$

Definition 2. The conditional entropy $H(Y|X)$ of Y given X with respective alphabet $\mathcal{Y}$ and $\mathcal{X}$ and joint probability distribution $p(y, x)$ is defined as:

$$H(Y|X) = -\sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} p(y, x) \log p(y|x) \tag{2.2}$$

Note that the following property holds:

Theorem 3. *Conditioning reduces entropy*

$$H(Y|X) \leq H(Y) \tag{2.3}$$

with equality if X and Y are independent.

Definition 4. The mutual information is a symmetric measure defined as:

$$I(Y; X) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log \left(\frac{p(x, y)}{p(x).p(y)} \right) \tag{2.4}$$

Note that the following relation holds

$$I(Y; X) = H(Y) - H(Y|X) = H(X) - H(X|Y) \tag{2.5}$$

A conditional mutual information $I(X; Y|Z)$ can also be defined as follows.

Definition 5. The conditional mutual information of a random variable X and Y given Z is defined as:

$$I(X; Y|Z) = H(Y|Z) - H(Y|X, Z) \tag{2.6}$$

$$= H(X|Z) - H(X|Z, Y) = I(Y; X|Z) \tag{2.7}$$

This quantity measures the reduction of uncertainty of Y (repectively X) due to the variable X (respectively Y), when Z is known. According to Theorem 3, this measure is never negative.

Theorem 6. *(Non-negativity of conditional mutual information)*

$$I(X;Y|Z) \geq 0 \tag{2.8}$$

with equality if $p(Y|Z) = p(Y|Z,X)$.

Note also that the mutual information between a set of random variables $X_1, X_2, ..., X_n$ and the random variable Y can be expressed as a sum of conditional mutual information terms (chain rule):

Theorem 7. *(Chain rule for mutual information) The mutual information between the set of random variables* $X_1, ...X_n$ *and* Y *can be decomposed as follows:*

$$\begin{array}{c} I(X_1, X_2, ..., X_n; Y) = I(X_{(1,...,n)}; Y) \\ = \sum_{i=1}^{n} I(X_i; Y|X_{i-1}, X_{i-2}, ..., X_1) \end{array} \tag{2.9}$$

3. On Relevance, Redundancy and Complementarity

Feature selection algorithms often rely on information theory concepts like *relevance*, *redundancy* and *complementarity* of variables. A concise review of these notions is the topic of this section. In particular, the difference between the definition of a relevance/redundancy measure and the definition of a variable as a relevant/redundant one will be pointed out.

3.1. Relevance.

The mutual information quantity introduced in Definition 4 can be seen as a measure of dependency between random variables. This measure applies to categorical as well as numeric variables and makes no assumption of linearity. The *mutual information* $\mathrm{I}(X_S; Y)$ has been used in ([37, 22]) for defining the relevance measure of a set X_S of random variables to another random variable Y.

As a result, the measure of relevance of an input variable is the conditional mutual information

$$I(X_i; Y|X_S) = I(X_{S,i}; Y) - I(X_S; Y) \tag{3.1}$$

where X_S is a set of variables previously selected. This quantity returns how much a variable X_i is relevant to the output Y given X_S and its normalized version (i.e. constrained to range between zero and one) has been introduced by ([4]) in order to select relevant features. Other normalized measures of mutual information were reviewed in ([44]).

The measure $I(X_i; Y|X_S)$ stresses one major aspect of feature selection: relevance is *conditionally dependent* on the context X_S. This makes difficult to decide whether a variable should be labeled as *relevant* or not. In fact, a variable can have a significant relevance given a context and a null relevance given another one, as shown by the following examples.

Example 8. Let $Y = X_i + X_j$ and $X_i = \frac{X_M}{2}$. Suppose also that $I(X_i; Y|X_j) > 0$.

In this example, X_i and X_M have both an unconditional positive relevance to Y once taken individually. Also, X_i has a positive relevance to Y given X_j but a null relevance to Y given X_M. Therefore, conditioning can reduce relevance.

Meanwhile, it is possible to increase the relevance of an input variable by conditioning, as shown by the following example from [10].

Example 9. Let Y and X_i be independent random variables, and $X_j = X_i + Y$.

Since $I(X_i; Y) = 0$, the conditional mutual information is higher than the mutual information, i.e. $I(X_i; Y|X_j) > I(X_i; Y)$. It follows that X_i has a positive relevance to Y given X_j but no relevance to Y once taken alone.

The above examples show that, as pointed out by ([21, 45]), a variable can be labeled as "relevant" only with respect to a given set of variables. In order to instantiate the notion of relevant variable in an information-theoretic framework, we reformulate the well-known definitions of ([21, 45]) as follows

Definition 10. Given a set of input variables X:
A variable X_i is irrelevant to Y iff:

$$\forall X_S \subseteq X_{-i} \ : \ I(X_i; Y|X_S) = 0 \tag{3.2}$$

A variable X_i is strongly relevant to Y iff:

$$I(X_i; Y|X_{-i}) > 0 \tag{3.3}$$

A variable X_i is weakly relevant to Y iff :

$$\begin{array}{c} I(X_i; Y|X_{-i}) = 0 \quad \text{AND} \\ \exists X_S \subset X_{-i} \ : \ I(X_i; Y|X_S) > 0 \end{array} \tag{3.4}$$

In simple words, strongly relevant variables are necessary since they contain information on Y that no other variables of X owns, while weakly relevant variables may be replaced by other variables without diminishing the relevance of X to the output.

For instance, in the set $\{X_M, X_i, X_j\}$ of example 8 the first two variables are weakly relevant while X_j is strongly relevant.

Note that the reformulation of the ([21]) definitions in information-theoretic terms makes easier the detection of relevant variables since the new definition relies on the estimation of a scalar value rather than on a whole probability distribution. At the same time, it establishes a link between the Kohavi's work and the relevance measure introduced by ([37, 4, 22]).

3.2. **Redundancy.** The following redundancy measures based on information theory can be found in ([26, 43]).

Definition 11. The redundancy among n random variables $X_1, ..., X_n$ is defined by

$$R(X_i; ...; X_n) = \sum_{i=1}^{n} H(X_i) - H(X_{1,..,n}) \tag{3.5}$$

This measure is nonlinear, symmetric, non-negative and non-decreasing with the number of variables. The monotonicity is justified by the fact that, unlike relevance, the amount of redundancy of a variable can never decrease when more variables are added.

For $n = 2$ the redundancy measure takes the following form:

Definition 12. The redundancy between two random variables X_i and X_j, is defined by

$$R(X_i; X_j) = H(X_i) + H(X_j) - H(X_{i,j}) = I(X_i; X_j) \tag{3.6}$$

This definition was used in ([32, 38]) as a founding criterion of a filter selection technique.

Note that while the relevance measure concerns the relation between inputs and output, the redundancy measure applies exclusively to input variables. This redundancy measure is also called *total correlation* ([20]) or *multiinformation* ([36]).

The simplest case of redundant variable is encountered when the measure of redundancy (3.6) takes its maximum value. In the discrete case, a variable X_i is redundant with respect to a set X_S if $R(X_i; X_S) = I(X_i; X_S) = H(X_i)$ or equivalently if $H(X_i|X_S) = 0$. This is due to the fact that $H(X_i|X_S) = 0$ is equivalent to $I(X_i; Y|X_S) = 0$ as shown by the following derivation

$$I(X_i; Y|X_S) = H(X_i|X_S) - H(X_i|Y, X_S) \tag{3.7}$$

$$= -H(X_i|Y, X_S) \geq 0 \Leftrightarrow \tag{3.8}$$

$$H(X_i|Y, X_S) = 0 \Leftrightarrow I(X_i; Y|X_S) = 0 \tag{3.9}$$

which holds because of the non negativity of the entropy (Theorem 6). In other terms, if $H(X_i|X_S) = 0$ we may drop the variable X_i since it is not able to bring any further conditional information to Y.

However, this does not hold anymore when redundancy is positive yet lower than the maximum value $H(X_i)$. In this case, $I(X_i; Y|X_S)$ is no more necessarily null.

A more complex redundancy condition relies on the notion of Markov blanket. The notion of Markov blanket ([23, 45]) can be rephrased into information-theoretic terms as follows (see proof in Appendix):

Definition 13. Let X be a set of features containing the variable X_i, X_M some subset of X_{-i} and $X_{-(i,M)}$ the set of remaining features. X_M is said a Markov Blanket for X_i, iff $I(X_i; (Y, X_{-(i,M)})|X_M) = 0$.

In other words, a Markov blanket X_M subsumes not only the information that X_i has about Y but also about all the remaining variables $X_{-(i,M)}$ in the set X. Although the existence of a Markov blanket allows to drop a variable without any risk of information loss, it is not sufficient to distinguish between irrelevant and redundant variables. The additional condition to label a variable as redundant is introduced in ([45]):

Definition 14. Given a set of variables X, a variable X_i is redundant, iff it is weakly relevant and has a Markov blanket in the set X.

3.3. **Interaction and Complementarity.** The definition of variable interaction in information-theoretic terms can be found in the seminal paper of [26] and, more recently, in [20].

Definition 15. The interaction among n sets of random variables, $X_1, X_2, ..., X_n$ is defined as:

$$C(X_1; X_2; ...; X_n) = \sum_{k=1}^{n} \sum_{S \subseteq \{1,...,n\}:|S|=k} (-1)^{k+1} H(X_S) \tag{3.10}$$

Given a set of random variables X and a random variable Y the link between $I(X;Y)$ and multivariate interaction is made explicit by the following definition from ([26]) and ([22]) .

Definition 16. The Möbius representation of mutual information between Y and any subset X, is given by:

$$\begin{array}{c} I(X;Y) = \sum_{i \in A} C(X_i;Y) - \sum_{i,j \in A} C(X_i;X_j;Y) \\ + \sum_{i,j,k \in A} ... + (-1)^{n+1} C(X_1;X_2;...;X_n;Y) \end{array} \tag{3.11}$$

with $C(X_i;Y) = I(X_i;Y)$.

In plain words, mutual information can be seen as a series where higher order terms are corrective terms that represent the effect of the multivariate interaction. In most cases the measure of interaction is positive and it indicates that the n sets of variables share a common information (redundancy). However, interaction can be negative. In the latter case, the variables are said to be complementary ([28]).

Definition 17. $X_1, X_2, ..., X_n$ are said to be complementary if

$$(-1)^n C(X_1; X_2; ...; X_n) > 0$$

The complementarity effect was mentioned in several papers ([21, 18]) and has been explicitely used in feature selection algorithms only recently ([22, 28, 46])

In particular, the complementarity between two random features X_i and X_j and the output Y ($n = 3$)

$$\begin{array}{c} -C(X_i;X_j;.Y) = -H(X_i) - H(X_j) - H(Y) \\ +H(X_i,X_j) + H(X_i,Y) + H(X_j,Y) - H(X_i,X_j,Y) \end{array} \tag{3.12}$$

$$= I(X_{i,j}; Y) - (I(X_i; Y) + I(X_j; Y)) \tag{3.13}$$

has been exploited in ([28]). This quantity measures, in bits, the gain resulting from using the joint mutual information of two variables X_i and X_j instead of the sum of the univariate informations. It is well-known, indeed, that the joint information of two random variables, i.e., $I(X_{i,j}; Y)$ can be higher than the sum of their individual information $I(X_i; Y)$ and $I(X_j; Y)$. Example 9 is an illustration of complementarity because in that case: $I(X_{i,j}; Y) > \underbrace{I(X_i; Y)}_{0} + I(X_j; Y)$.

Another well-known illustration of this phenomenon is the XOR problem as pointed out by ([21]):

Example 18. Xor problem:

X_1	X_2	$Y = X_1 \oplus X_2$
1	1	0
1	0	1
0	1	1
0	0	0

One can see that X_1 and X_2 have a null relevance individually, i.e. $I(X_1; Y) = 0$, $I(X_2; Y) = 0$, whereas together $X_{1,2}$ has maximal relevance, i.e. $I(X_{1,2}; Y) = H(Y) > 0$. Complementarity explains why a combination of apparently irrelevant variables can perform efficiently in a learning task.

4. State-of-the-art filter algorithms using mutual information

Feature selection aims at finding the subset of variables that allows the most accurate prediction of the output Y. For n inputs, however, the number of possible subsets is 2^n and the task of feature selection is known to be hard ([12, 8]). In front of several thousands of variables as in microarray data, feature selection algorithms have often recourse to sub-optimal search strategy having a polynomial complexity. The most used strategies are the ranking of features ([16]) and the greedy forward selection ([45, 32, 17]).

The second particularity of microarray data lies in the low number of noisy samples (from several tens to several hundreds). As a result of this low number of samples, estimating large multivariate densities is completely inaccurate and highly computationaly costly. This led to the proposition of several new information-theoretic approaches presented in this section. The main idea consists in estimating the information of subpart of a subset rather than the information of the subset itself. However, it worth noting that by using this approach, complementary subsets of variables can be missed.

We present the algorithms by stressing when and where the notion of relevance, redundancy and complementarity are used.

Before proceeding with the presentation of the filters, we state the objective of a feature selection method based on information theory:

Given a training dataset D_m of m samples and n input variables and an integer $d < n$, find the subset $S \subset \{1, ..., n\}$ of size d that maximizes the mutual information $I(X_S; Y)$.

In other words, the objective of *filters feature selection* consists in finding the subset X_S of $|S| = d$ variables such as:

$$X_S^{OPT} = \arg \max_{S \subset A : |S| = d} I(X_S; Y) \tag{4.1}$$

This equation is also known as the max-dependency criterion ([32]).

In this paper, we do not address the problem of the choice of d. We assume that, the number of variables d has to be determined by some a priori knowledge or has to be chosen in order to allow an internal dimensionality reduction of the learning procedure (or an additional wrapper approach) to reach the optimal subset in a reasonable amount of time. Note that filters often rank the variables according to their dependency measure. As a result, variables can be added one by one in a predictive model until a decrease appears in the performances. This procedure allows to reach an adequate number of variables for a given predictive model.

It should be noted that finding a set of genes that leads to the best classification accuracy does not necessarily mean finding the genes that regulate the biological process at hand. For instance, in network inference, where the subset of regulator genes is searched, variables such as coregulated genes are meant to be eliminated ([29]) whereas in a best classification accuracy perspective, dependent variables of the regulator genes can improve the robustness of the classification.

As discussed in the introduction, any filter algorithm requires a selection criterion and a search strategy. Three of the filters presented in the following adopts a forward selection search approach. *Forward selection* starts with an empty set of variables and progressively updates the solution by adding the variable that brings the best improvement (according to a given criterion). The hill-climbing search selects a subset of $d < n$ variables in d steps by exploring only $\sum_{i=0}^{d}(n - i)$ configurations. As a result, filters using forward selection can return a set of d variables in $O(n \times d)$ and a ranking of the variables in $O(n^2)$ (by selecting a subset composed of all the n variables).

4.1. **Forward Selection and Relevance Criterion (REL).** The relevance criterion is combined with a forward selection search strategy ([3, 4]). The approach consists of updating a set of selected variables X_S with the variable X_i featuring the maximum relevance $I(X_i; Y|X_S)$. This strategy prevents from selecting a variable which, though relevant to Y, is redundant with respect to a previously selected one.

In analytical terms, the variable X_i^{REL} returned by the relevance criterion at each update step is,

$$X_i^{REL} = \arg \max_{X_i \in X_{-S}} \{I(X_i; Y|X_S)\} \tag{4.2}$$

where $X_{-S} = X \setminus X_S$ is the set difference between the original set of inputs X and the set of variables X_S selected so far. Note that in ([4]), a normalized version of relevance is used, i.e. $rel(X_i, Y|X_S) = \frac{I(X_i,Y|X_S)}{H(Y|X_S)}$.

Although this method is appealing, it presents some major drawbacks. The estimation of the relevance requires the estimation of large multivariate densities, a problem known to be ill-posed. For instance, at the dth step of the forward search, the search algorithm asks for $n-d$ evaluations where each evaluation requires the computation of a $(d+1)$-variate density. It is known that for a large d, the estimations are poorly accurate and computationally expensive. For these reasons we recently assisted in filter literature to the adoption of selection criteria based on bi- and tri-variate densities at most ([14, 17, 45]).

4.2. **Variable Ranking - Relevance Selection (RANK).** The ranking method returns a ranking of variables on the basis of their individual mutual information with the output. This means that, given n input variables, the method first computes n times the quantity $I(X_i, Y)$, $i = 1, \ldots, n$, then ranks the variables according to this quantity and eventually discards the least relevant ones ([16]).

The main advantage of this method is the low computational complexity. Indeed, only n computations of mutual information are required. The main drawback derives from the fact that possible redundancies between variables are not taken into account. Indeed, two redundant variables, yet highly relevant taken individually, will be both well-ranked. On the contrary, two variables can be complementary to the output (i.e., highly relevant together) while each of them appears to be poorly relevant once taken individually (see Example 9 or Example 18). As a consequence, these variables could be badly ranked, or worse eliminated, by a ranking filter.

4.3. **Fast Correlation Based Filter (FCBF).** Fast Correlation Based Filter is a ranking method combined with a redundancy analysis ([45]). The FCBF starts by selecting the variable with the highest mutual information, denoted by $X_{(1)}$. Then, all the features which are less relevant to Y than redundant to $X_{(1)}$ are eliminated from the list. For example X_i is removed if $I(X_i; X_{(1)}) > I(X_i; Y)$. The algorithms proceeds by focusing on $X_{(2)}$, the variable with the second highest mutual information in the list of remaining variables, and repeating the elimination step. The procedure stops when no more variables remains to be taken into consideration.

This method is fast because a few (less than n^2) evaluations of bivariate mutual information are computed. However, although the method addresses redundancy, it presents the risk of eliminating relevant variables. Indeed, it is possible to have $I(X_i; X_{(1)}) > I(X_i; Y)$ and still have X_i strongly relevant, as shown in the Example 9. Note also that this method cannot return a total ranking of the variables of the dataset. The effectiveness of the approach was shown in [45] by comparing it with two well-known filters: Relief ([24]) and

Focus ([1]). Note that in ([45]), a normalized measure of mutual information called the symmetrical uncertainty, is used, i.e. $SU(X,Y) = \frac{2\times I(X;Y)}{H(X)+H(Y)}$. The latter often improves the performances of the selection by penalizing inputs with large entropies.

4.4. **Forward Selection and Minimum Redundancy - Maximum Relevance criterion (MRMR).** The minimum redundancy-maximum relevance (MRMR) criterion has been proposed in ([32, 31]) in combination with a forward selection search strategy. Given a set X_S of selected variables, the criterion consists of updating X_S with the variable $X_i \in X_{-S}$ that maximizes $u_i - z_i$, where u_i is a relevance term and z_i is a redundancy term. More precisely, u_i is the relevance of X_i to the output Y alone, and z_i is the average redundancy of X_i to the selected variables $X_j \in X_S$.

$$u_i = I(X_i;Y) \tag{4.3}$$

$$z_i = \frac{1}{d} \sum_{X_j \in X_S} I(X_i;X_j) \tag{4.4}$$

$$X_i^{MRMR} = \arg \max_{X_i \in X_{-S}} \{u_i - z_i\} \tag{4.5}$$

At each step, this method selects the variable which has the best trade-off between relevance and redundancy. This selection criterion is fast and efficient. At step d of the forward search, the search algorithm computes $n-d$ evaluations where each evaluation requires the estimation of $(d+1)$ bivariate densities (one for each already selected variables plus one with the output). It has been shown in ([32]) that the MRMR criterion is an optimal first order approximation of the max-dependency criterion (see 4.1) if we select (add) a feature at the time. Furthermore, MRMR avoids the estimation of multivariate densities by using multiple bivariate densities.

Note that, although the method addresses the issue of bivariate redundancy through the term z_i, it is not able to take into account the complementarities between variables. This could be ineffective in situations like the one in Example 9 where, although the set $\{X_i, X_S\}$ is highly relevant, we have

(1) the redundancy term z_i is large due to the redundancy of X_i and X_S,
(2) the relevance term u_i is small since X_i "alone" is not relevant to Y.

4.5. **Forward Selection and Conditional Mutual Information Maximization criterion (CMIM).** This approach ([17]) proposes to select the feature $X_i \in X_{-S}$ whose minimal relevance $I(X_i;Y|X_j)$ conditioned to the selected features $X_j \in X_S$, is maximal. This requires the computation of the mutual information of X_i and the output Y, conditioned on each feature $X_j \in X_S$ previously selected. Then, the minimal value is retained and the feature that has a maximal minimal conditional relevance is selected.

In formal notation, the variable returned according to the CMIM is

$$X_i^{CMIM} = \arg \max_{X_i \in X_{-S}} \{ \min_{X_j \in X_S} I(X_i; Y|X_j) \} \tag{4.6}$$

[17] showed that CMIM is competitive with AdaBoost and FCBF ([45]) in selecting binary features for a pattern recognition task. This selection criterion selects relevant variables, avoids redundancy, avoids estimating high dimensional multivariate densities and, unlike the previous method, it does not ignore variable complementarity. However, it does not necessarily select a variable complementary with the already selected variables. Indeed, a variable that has a high complementarity to the already selected variable will be characterized by a high conditional mutual information with that variable but not necessarily by a high minimal conditional information (see Examples 18 and 9).

In the following section, we propose a new criterion that deals more explicitly with complementary variables.

5. The k-Average Sub-Subset Information criterion (kASSI)

This section presents and discusses an original criterion (kASSI) for feature selection which aims at maximizing the mutual information avoiding, meanwhile, large multivariate density estimation:

$$X_S^{kASSI} = \arg \max_{X_S \subset X} \{ \sum_{V \subset S : |V| = k} I(X_V; Y) \} \tag{5.1}$$

The rationale behind the k-Average Sub-Subset Information criterion is that a good approximation of the information of a set of variables is proportional to the average of the information of its subsets. For example, if we want to estimate $I(X_{1,2,3,4}; Y)$ the criterion proposes to use the average of the information gains of the trivariate subsets: $I(X_{1,2,3}; Y)$, $I(X_{2,3,4}; Y)$, $I(X_{1,3,4}; Y)$ and $I(X_{1,2,4}; Y)$.

A theoretical justification of the criterion can be found in the following theorem which states that, under specific assumptions, the term maximized in (5.1) appears in the lower bound of the quantity $I(X_S; Y)$.

Theorem 19. *[Lower bound on mutual information] (see proof in Appendix).*

$$\frac{1}{\binom{d}{k}} \sum_{V \subseteq S : |V| = k} I(X_V; Y) \leq I(X_S; Y) \tag{5.2}$$

where X_V denotes subsets of variables (of size k) of X_S (of size d)

In other words, if there is no subset of complementary variables of size $p > k$, then

$$K_1 \times X_S^{kASSI} \leq I(X_S; Y) \tag{5.3}$$

Note that the higher the k, the tighter the bound but also the higher is the computational cost.

Another theoretical justification of the selection criterion (5.1) as an example of the *averaging estimators* approach is worthy to be provided. Suppose we aim to estimate the information per variable of a subset X_S, i.e. the quantity $I_V(X_S;Y) = \frac{I(X_S;Y)}{d}$. The kASSi criterion boils down to an estimator $\hat{I}_V$ of I_V which is nothing but the average of estimates of the target quantity computed for all the subsets of S of size k , i.e

$$\hat{I}_V = \frac{1}{k! \begin{pmatrix} d \\ k \end{pmatrix}} \sum_{V \subseteq S : |V|=k} \frac{I(X_V;Y)}{k} \tag{5.4}$$

The criterion (5.1) can be then interpreted as an averaging approach to the estimation of the unknown quantity I_V which relies on the average of existing estimates of lower order.

5.1. **The $k = 2$ case.** Note that the kASSI criterion for $k = d$ is equivalent to the REL approach and for $k = 1$ boils down to the ranking algorithm. Here, we are mainly interested in the $k = 2$ case:

$$X_S^{2ASSI} = \arg\max_{X_S \in X} \{ \sum_{X_i \in X_S} \sum_{X_j \in X_S} I(X_{i,j};Y) \} \tag{5.5}$$

This formulation is particularly interesting since it can deal with complementarities up to order 2 (like those of the examples 9 and 18) by preserving the same computational complexity of the MRMR and CMIM criteria (Table 2). Indeed, we can see that optimizing $I(X_{i,j};Y)$,

$$I(X_{i,j};Y) = I(X_i;Y) + I(X_j;Y) - C(X_i;X_j;Y) \tag{5.6}$$

means optimizing the relevance of each variable alone minus an interaction term. The lower the interaction term, the less redundant are X_i and X_j and the higher is their complementarity (if the interaction term is negative).

A variant of the criterion (5.5) is obtained by replacing the mutual information, with a normalized relevance measure defined by [44] as

$$\mathrm{SR}(X_S;Y) = \frac{I(X_S;Y)}{H(X_S,Y)} \tag{5.7}$$

[28] denote this measure as the *symmetrical relevance*.

This normalization aims to improve the selection strategy by penalizing inputs with large entropies. The criterion

$$X_S^{DISR} = \arg\max_{X_S \in X} \{ \sum_{X_i \in X_S} \sum_{X_j \in X_S} SR(X_{i,j};Y) \} \tag{5.8}$$

takes the name of *double input symmetrical relevance* (DISR).

Note that this criterion is an evaluation function of a subset of variables. As a result, when DISR is combined with a forward selection (as the criteria MRMR, CMIM and REL), X_S is updated at each step by the variable X_i^{DISR},

$$X_i^{DISR} = \arg \max_{X_i \in X_{-S}} \{ \sum_{X_j \in X_S} SR(X_{i,j}; Y) \} \tag{5.9}$$

5.2. **A comparison with other criteria.** It is interesting to situate the proposed criterion with respect to the state-of-the-art methods of Section 4. The presented criteria can be analyzed under different perspectives. We stress here i) which issues, among relevance, redundancy and complementarity, are taken into account, ii) the ability of a criterion to avoid the estimation of large multivariate densities and iii) whether it returns a ranking of variables. A concise summary of the presented techniques is available in Table 1. Note that variable complementarity is a desirable feature in biological data analysis where combination of genes or proteins has to be detected.

methods:	RANK	FCBF	REL	CMIM	MRMR	DISR
Select Relevance	V	V	V	V	V	V
Eliminate Redundancy	-	V	V	V	V	V
2th order Complementarity	-	-	V	-	-	V
Avoid Multivariate Density	V	V	-	V	V	V
Return Ranking	V	-	V	V	V	V

TABLE 1. Comparison of the properties (relevance, redundancy and complementarity, ability to avoid estimation of large multivariate densities, ability to rank the variables) that are taken into account in each selection criterion.

5.3. **The MASSIVE implementation.** An interesting property of the DISR maximization step is that it can be expressed as a weighted Dense Subgraph Problem (DSP) ([2]) or equivalently as a Dispersion Sum Problem ([33]). In the dense subgraph problem, we have a complete undirected graph

on the node set $A = \{1, ..., n\}$ where each edge (i, j) takes a weight $w_{ij} \geq 0$, $w_{ii} = 0$. The goal is

$$maximize \sum_{i \in A} \sum_{j \in A} w_{ij} v_i v_j \tag{5.10}$$

$$subject\ to \sum_{i \in A} v_i = d \tag{5.11}$$

$$v_i \in \{0, 1\}, \quad i \in A \tag{5.12}$$

which means to select a node subset $S \subseteq A$ of fixed size $|S| = d$ such that the total edge weight in the induced subgraph is maximum.

We can put the DISR problem in a DSP framework by assuming that the ith node represents the variable X_i, the binary variable $v_i, i = 1, \ldots, n$ takes the value 1 if the ith variable is selected and 0 otherwise and the weight $w_{ij} = \frac{I(X_{i,j};Y)}{H(X_{i,j})}$ is the symmetrical relevance of the two variables linked by the edge.

The DSP is a NP-hard problem since it can be reduced to the CLIQUE problem ([33]). However, there exist a branch and bound algorithm able to deal up to 90 variables ([33]) and several promising results on the performance of greedy searches. [6] showed that backward elimination combined with a sequential search (BESR) performs well on binary quadratic problems. The BESR method starts with a set containing all the variables (i.e. $v_j = 1$ for all $j \in A$) and then selects the variable i whose removal (i.e. $v_i = 1 \leftarrow 0$) induces the lowest decrease of the objective function till the adequate number of variable is reached (i.e $\sum_{i \in A} v_i = d$). The procedure is enhanced by an iterative sequential replacement which, at each step, swaps the status of a selected and a non selected variable (i.e. swapping $v_i = 1$ and $v_j = 0$) such that the largest increase in the objective function is achieved. The sequential replacement is stopped when no further improvement is met.

5.3.1. *Computational Complexity.* The proposed implementation (called MASSIVE from Matrix of Average Sub-Subset Information for Variable Elimination) works as follows: at first the DISR-matrix is computed. This demands $\frac{n(n-1)}{2}$ weight evaluations $w_{ij} = \frac{I(X_{i,j};Y)}{H(X_{i,j})}$ since the average sub-subset information is symmetric. Then a backward elimination as well as the sequential replacement are applied to the DISR-matrix. This has a complexity of $O(n^2)$ if the implementation for binary quadratic problems of ([27]) is adopted.

Table 2 reports a comparative analysis of the different techniques in terms of computational complexity of the evaluation step. Note that the table reports a naive implementation of CMIM. A more efficient algorithm with CMIM is given in [17]. The total complexity of the MASSIVE is $O(C \times n^2)$ where C is the cost of an estimation of the mutual information involving N samples and three variables (two inputs and one output). A total ranking can be returned by MASSIVE by selecting a subset composed of all the n

variables (with no increase in the asymptotic comptutational cost). In that case, the ranking is the one given by the backward elimination since there are no remaining variables to be used for sequential replacement. Note that a conventional forward selection (up to d variables) based on an information criterion (e.g., MRMR) demands $O(n \times d)$ evaluations, each having a complexity depending on d and on the number of samples N. A conventional backward selection, where the evaluation is done inside the loop and not precomputed as in MASSIVE, demands $O(n^2)$ evaluations.

The MASSIVE implementation makes possible the adoption of a BESR strategy at a cost lying between a conventional forward and a conventional backward approach.

methods:	REL	CMIM	MRMR	DISR	MASS
calls of evaluation function	$\frac{(n\times d)}{2}$	$\frac{(n\times d)}{2}$	$\frac{(n\times d)}{2}$	$\frac{(n\times d)}{2}$	$\frac{(n\times n)}{2}$
calls of MI by evaluation	1	$d-1$	d	$d-1$	1
k-variate density	$d+1$	3	2	3	3

TABLE 2. The number of calls of the evaluation function is $n \times d$ in a forward selection strategy. Note that $d = n$ for a backward elimination instead of a forward selection or for a complete ranking of the n features. The computational cost of the criteria REL, CMIM, MRMR, DISR and MASSIVE is the number of calls of mutual information (MI) times the cost of an estimation of the mutual information involving a k-variate density and N samples.

The C++ code for this method can be found at http://www.ulb.ac.be/di/map/pmeyer/links.html. Note that on a 3.4GHz Intel Pentium4 processor with 2GB RAM, the MASSIVE computation with 5327 variables and 72 samples (see Leukemia1 dataset in Table 3) takes less than 4 minutes. Running a backward elimination on that matrix, in order to eliminate 5300 variables, took less than 10 seconds.

6. EXPERIMENTS

The validation of the proposed approach relies on two feature selection experiments. The first one uses 11 public domain microarray expression datasets available from (Plymouth University, http://www.tech.plym.ac.uk/

[1]) and commonly used for comparing cancer classification techniques. The second experiment uses synthetically generated data to show a practical situation where the notion of complementarity can be exploited to obtain improved selection.

6.1. **Data.** The eleven datasets are detailed in Table 3 where n is the number of features corresponding to the number of expressed gene probes, m is the number of observations corresponding to the tumor samples and c is the number of cancer classes.

	datasets	n	m	c	ts (min)
1	SRBCT	2308	83	4	0,7
2	Leukemia1	5327	72	3	3
3	DLBCL	7129	77	2	5,5
4	9_Tumors	7129	60	9	4,6
5	Brain_Tumor1	7129	60	2	4,5
6	Brain_Tumor2	12625	50	2	11,4
7	Prostate_Tumor	12600	102	2	23
8	Leukemia2	12582	72	3	16,5
9	11_Tumors	12533	174	11	43
10	Lung_Cancer	12600	203	5	46,4
11	14_Tumors	15009	308	26	198

TABLE 3. The 11 datasets of microarray cancer from http://www.tech.plym.ac.uk/spmc/. The column n contains the number of different sequences measured in the microarray, m the number of clinical samples and c the number of cancer classes. The column ts reports the time needed to select 15 features with the C++ massive toolbox on a 3.4GHz Intel Pentium4 processor with 2GB RAM.

6.2. **Entropy Estimation .** Each continuous variable is quantized according to two different methods: equal sized interval and equal frequency interval (see [15]). The number of intervals of each input is based on the Scott criterion. The latter is based on the estimate of the standard error of a normal distribution ([35]). The entropy estimation method used on the discretized dataset is the simple empirical entropy. This entropy estimation method is known to be biased ([30]). However, this estimator is fast to compute, i.e., $O(N \times d)$. Note also that all the feature selection algorithms are compared with the same entropy estimation method (on the same number N of samples) and at each step of a forward/backward search, $n - d$ subsets of variables of exactly the same size are compared. As a result, the impact of the bias of the estimator on the selection of the best subset (in each step

[1]http://www.tech.plym.ac.uk/spmc/links/bioinformatics/microarray/microarray_cancers.html

of each algorithm) should be weak. A similar remark can be found in ([25]). However, further research on this topic is required

6.3. **Selection of Variable.** We compare the filter based on DISR with the five state-of-the-art approaches discussed in Section 4: the Ranking algorithm, the FCBF and three filters based on the Relevance criterion, the MRMR criterion, and the CMIM criterion.

A two-fold cross-validation is used to partition each dataset into separate datasets for feature selection and validation. Each selection method stops after 15 variables have been selected. It appeared in the experimental session that the classifiers required less than ten features in order to reach their optimal accuracy (for most datasets).

6.4. **Validation of the Selection.** The learning procedure uses a 3-nearest neighbor classifier and a SVM learning algorithm (with a gaussian kernel). The assessment of each selection criterion is done as follows: once the ranking of $d = 15$ variables is done on the selection set, we perform, on the validation set, a 10-fold cross-validation for a number of variables ranging from 1 to 15 (according to the ranking returned by the filter). In other words, the dataset is first split into two equal parts. The latter half-dataset is then split into ten parts in order to select the best number of variables. The best classification accuracy, i.e., $\frac{\#\, good\ classifications}{\#\ samples}$, is presented in the following section. Also, an adjusted p-value (using the method in [5]) using a statistical paired permutation test on the test set is reported. A bold-faced notation is used when the p-value is lower than 0.05.

As far as the implementation of the learning methods is concerned, we used the algorithms made available by the R statistical language.

6.5. **Results.** The experimental results show that the MASSIVE method is competitive with the state-of the-art information-theoretic filters. In particular, MASSIVE is significatively better than all other information-theoretic criteria except MRMR. Note that the results for the FCBF can be poor on given datasets because of the internal stop criterion. Indeed, on several datasets, the FCBF stopped after having selected one variable.

6.6. **Experiments with synthetic datasets.** The results of the previous experiments showed that MASSIVE is competitive with state-of-the-art approaches although it is not able to outperform the MRMR approach. A more detailed comparison between MASSIVE and MRMR is then worthy to be done. This section considers two synthetic experiments designed on purpose a) to show a situation where the notion of complementarity may be advantageous and b) to assess the capability of the MASSIVE algorithm to select the optimal feature set for a large number of samples.

Classification accuracy in a Bayesian network setting

This experiment aims to show that the notion of complementarity used by the DISR brings a significant improvement with respect to MRMR in terms

	MASS	CMIM	MRMR	FCBF	RANK	REL
1	83.13	**42.17**	78.31	**57.83**	77.11	**53.01**
2	87.5	80.56	91.67	**70.83**	86.11	**68.06**
3	81.82	80.52	90.91	76.62	84.42	75.32
4	6.67	5	10	0	8.33	6.67
5	63.33	60	63.33	55	65	61.67
6	74	58	64	58	72	**54**
7	91.18	91.18	90.2	83.33	90.2	**78.43**
8	88.89	79.17	87.5	**43.06**	80.56	**54.17**
9	60.92	**42.53**	54.02	**22.99**	**49.43**	**34.48**
10	79.8	80.3	81.28	74.88	77.83	80.3
11	19.48	**12.99**	18.51	**12.01**	19.48	**12.99**
Avg	67	**57**	66	**50**	**65**	**53**

TABLE 4. **SVM classifier and equal sized quantization**: Accuracy with 10-fold cross-validation on the test set. Bold notation is used to identify which techniques are significantly different (pval $< .05$) from MASSIVE in terms of accuracy.

	MASS	CMIM	MRMR	FCBF	RANK	REL
1	89.16	**49.4**	90.36	**49.4**	79.52	**49.4**
2	91.67	81.94	97.22	**75**	90.28	**69.44**
3	85.71	85.71	87.01	**71.43**	87.01	**72.73**
4	16.67	16.67	16.67	6.67	10	6.67
5	60	56.67	61.67	55	58.33	65
6	64	62	66	58	66	58
7	85.29	81.37	88.24	80.39	89.22	83.33
8	87.5	81.94	83.33	**50**	73.61	**58.33**
9	48.85	**36.78**	50.57	**33.33**	40.8-0.225	**27.59**
10	83.74	79.8	80.3	**77.34**	81.77	**77.34**
11	16.56	13.31	15.26	**11.04**	16.56	15.58
Avg	66	**59**	67	**52**	**63**1	**53**

TABLE 5. **3NN classifier and equal sized quantization**: Accuracy with 10-fold cross-validation on the test set. Bold notation is used to identify which techniques are significantly different (pval $< .05$) from MASSIVE in terms of accuracy.

of accuracy if variables dependency may be described by a graphical model . Let us consider a set of random variables A, B, Y, S, H whose statistical dependencies are described by the Bayesian network of Figure 6.1 and the conditional probability of Table 8. The interpretation of the Bayesian network is the following: an abnormal cellular activity (A) can cause a cancer

	MASS	CMIM	MRMR	FCBF	RANK	REL
1	79.52	**38.55**	79.52	**49.4**	72.29	**43.37**
2	87.5	81.94	88.89	81.94	83.33	81.94
3	94.81	**80.52**	92.2	**77.92**	**81.82**	**80.52**
4	15	8.33	8.33	**5**	8.33	6.67
5	65	65	65	63.33	65	65
6	76	**54**	68	60	60	**52**
7	91.18	84.31	94.12	**81.37**	92.16	84.31
8	93.06	**61.11**	90.28	**51.39**	90.28	**63.89**
9	46.55	**33.33**	53.45	**19.54**	52.3	**34.48**
10	82.27	**74.88**	83.25	**74.88**	78.33	**71.43**
11	17.53	15.58	22.4	**8.44**	19.16	**11.04**
Avg	68	**54**	68	**52**	**64**	**54**

TABLE 6. **SVM classifier and equal frequency quantization**: Accuracy with 10-fold cross-validation on the test set. Bold notation is used to identify which techniques are significantly different (pval $< .05$) from MASSIVE in terms of accuracy.

	MASS	CMIM	MRMR	FCBF	RANK	REL
1	79.52	**50.6**	84.34	**46.99**	72.29	**55.42**
2	88.89	77.78	90.28	**76.39**	86.11	77.78
3	93.51	83.12	92.21	**83.12**	92.21	**81.82**
4	23.33	18.33	21.67	11.67	21.67	13.33
5	58.33	66.67	66.67	60	60	60
6	62	68	68	60-1	68	68
7	89.22	81.37	90.2	**75.49**	90.2	84.31
8	83.33	72.22	81.94	**62.5**	90.28	72.22
9	50.57	**38.51**	55.17	**17.24**	51.15	**32.18**
10	82.76	78.33	82.76	**74.88**	78.82	**71.43**
11	18.83	15.26	**29.87**	**10.06**	19.81	**12.01**
Avg	66	**59**	**69**	**53**	66	**57**

TABLE 7. **3NN classifier and equal frequency quantization**: Accuracy with 10-fold cross-validation on the test set. Bold notation is used to identify which techniques are significantly different (pval $< .05$) from MASSIVE in terms of accuracy.

(Y) and activate a blood marker (B). A possible symptom of that cancer is a headache (H). Luckily, headaches may have also innocuous causes, like sinus

infection (S).

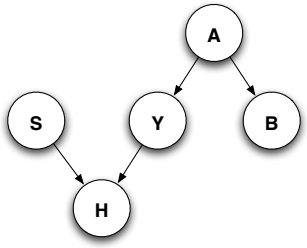

FIGURE 6.1. Artificial Bayesian network connecting variables : abnormal cellular activity (A), cancer (Y), blood marker (B), headache (H) and sinus infection (S).

$p(A)$	$p(B\|$ $A=1)$	$p(Y\|$ $A=1)$	$p(H\|S=1,$ $Y=0)$	$p(H\|S=1,$ $Y=1)$
(0.8, 0.2)	(0.3, 0.7)	(0.3, 0.7)	(0.2, 0.8)	(0.1, 0.9)
$p(S)$	$p(B\|$ $A=0)$	$p(Y\|$ $A=0)$	$p(H\|S=0,$ $Y=1)$	$p(H\|S=0,$ $Y=0)$
(0.3, 0.7)	(0.7, 0.3)	(0.7, 0.3)	(0.2, 0.8)	(0.9, 0.1)

TABLE 8. All variables are binary, p(X)=(P(X=0),P(X=1)). For example first column indicates that 20% patients coming to the consultation have an abnormal cellular activities and 70% have a sinus infection, the 4th column indicates that there is 80% chance having some headache either if you have a sinus infection or cancer.

Let us consider a classification task where the goal is to predict the variable Y using both the set $\{A, B, H, S\}$ of variables and a set of ten irrelevant features. The training set is composed of 300 i.i.d. samples and the test set is composed of 700 i.i.d. samples. A 3-Nearest Neighbor and Support Vector Machine algorithm were used to assess the performances of both filters using the five first variables selected.

	MASSIVE	MRMR
3NN	62	**56**
SVM	75	**70**

TABLE 9. Accuracy with 10-fold cross-validation on the test set, boldfaced if p-value by a paired permutation test <0.05. The learning algorithm consider the five first variables selected by both filters.

Let us note that in this example only the three variables $\{A, S, H\}$ are relevant for predicting Y. Note that they form the Markov blanket of Y (see Definition 13) composed of the target's parents, the target's children, and the target's children's parents.

The experiment shows that MASSIVE is able to select $\{A, S, H\}$. Indeed, variable S and H are complementary in order to predict Y since $C(S; H; Y) = I(S; Y|H) - I(S; Y) > 0$. It worth noting that the information-theoretic notion of complementarity corresponds to *the explaining away effect* of the Bayesian network litterature ([20]). MRMR, instead, selects only the variables A and H since they are relevant to the output and discards all the others since the latter are either irrelevant or are redundant with A or H. The different selections of MASSIVE and MRMR have a significant impact on the classification accuracy (Table 9) coherent with the fact that the Markov blanket of a target variable is the minimum set that achieves optimal classification accuracy under certain general conditions ([40]). The dataset and the R code used to generate it are available with the MASSIVE C++ toolbox.

Optimal feature set selection

The potential difficulties of performing feature selection in small sample size situations, due to the curse of dimensionality, have been first raised in ([39]) and later discussed in [19]. Although the goal of the paper is to assess the potentiality of the complementarity notion in small sample and real feature selection problem, it is worthy investigating how the selection accuracy scales up with the number of samples. In order to carry out this analysis we had recourse to the *SynTReN* generator ([13]), a simulator designed to generate microarray data by selecting subnetworks from *E. coli* and *S. cerevisiae* source networks. Here the mRNA expression levels for the genes in the network are obtained by simulating equations based on Michaelis-Menten and Hill kinetics under different conditions.

The SynTRen simulator is used to generate 14 different classification tasks where the target is the simulated expression of a target gene and the set of relevant inputs is made of seven genes which are connected to the target in the network. For each of the simulated tasks 500 irrelevant variables were added to the seven genes that regulate the target and training sets with increasing numbers of samples (m=100, 1000, 2000) were generated. The resulting dataset is then discretized using the equal frequency binning algorithm (see Sec. 6.2). A MRMR and a MASSIVE feature selection (d=7) was performed. For each dataset, the number of relevant variables selected by MRMR and MASSIVE is reported in the table 10. We observe that i) for both algorithms the quality of the selection increases with the number of samples coherently with the conclusion reached in ([19]) and ii) the quality of the MASSIVE selection converges more rapidly toward the optimum.

MRMR	MASS	MRMR	MASS	MRMR	MASS
2	3	3	4	3	6
5	6	6	7	6	7
5	7	6	7	6	7
2	2	3	2	3	6
2	3	3	3	3	6
2	2	3	3	3	6
1	1	2	1	2	6
2	2	3	3	3	6
3	4	3	3	3	7
2	2	3	3	3	2
2	3	3	6	3	7
1	1	2	1	2	6
2	3	3	4	3	7
3	3	4	5	4	7
35%	43%	48%	53%	48%	86%

TABLE 10. Table reporting the number of relevant variables selected by MRMR and MASSIVE on 14 datasets of 507 input variables and 100 (column 2,3), 1000 (column 4,5) and 2000 (column 5,6) samples, respectively. The expression of the target is generated by a synthetic microarray data simulator based on Michaelis-Menten and Hill kinetics equations. The last line reports the average covering of the selection.

7. Conclusions

The paper introduced MASSIVE, a new information-theoretic filter approach for mining microarray data and compared it with state-of-the-art approaches. This filter relies on a criterion which consists in maximizing a term appearing in the lower bound of the mutual information of a subset. The normalized second-order approximation of the criterion, called DISR, is well suited to microarray analysis for three main reasons: (i) the low computational cost $O(n)$ allows to deal with a large number of variables, (ii) highly multivariate mutual informations are approximated by trivariate mutual informations, (iii) the potential two-by-two gene complementarity is taken into account at the same computational cost of existing methods.

Another appealing aspect of the MASSIVE method is that the resulting selection procedure can be formulated as a well-known quadratic optimization problem, the DSP, and solved by a search strategy like the backward elimination combined with a sequential replacement. This method, typically expensive in terms of the number of subset evaluations, is made computationally affordable by the computation and the storage of the DISR-matrix.

The experimental results show that the proposed method is competitive with five state-of-the-art approaches and significantly better than four of them. A word of caution about the cross-validated assessment if however necessary here: given the small sample size nature of the microarray problem a cross-validation assessment runs the risk of being too variant ([9]) and not precise enough for distinguishing between very close performance. Notwithstanding, we are confident that the identical experimental conditions (estimator, quantization) as well as the large number of datasets provide sufficient statistical evidence to our conclusions.

Further research is expected to focus on (i) the application of the MASSIVE method to other classification tasks, (ii) the study of the role of the entropy estimator on the quality of the selection, (iii) the definition of optimality conditions for the DISR criterion.

Appendix

Theorem 19 states that

$$\frac{1}{\left(\begin{array}{c} d \\ k \end{array}\right)} \sum_{V \subseteq S : |V| = k} I(X_V; Y) \leq I(X_S; Y)$$

Proof. Since mutual information can only increase by adding variables, we have

$$\max_i I(X_{S-i}; Y) \leq I(X_S; Y) \tag{7.1}$$

By the definition of average:

$$\frac{1}{d} \sum_{i=1}^{d} I(X_{S-i}; Y) \leq \max_i I(X_{S-i}; Y) \leq I(X_S; Y) \tag{7.2}$$

Applying it recursively gives:

$$\begin{aligned} \frac{1}{d} \sum_{i \in S} I(X_{S-i}; Y) &\geq \frac{1}{d(d-1)} \sum_{i \in S} \sum_{j \in S-i} I(X_{S-i,j}; Y) \\ &\geq \frac{\frac{(d-k)!}{d!} \sum_{V \subseteq S : |V| = k} I(X_V; Y)}{\frac{1}{k! \left(\begin{array}{c} d \\ k \end{array}\right)} \sum_{V \subseteq S : |V| = k} I(X_V; Y)} \end{aligned}$$

multiplying by $k!$ for all ordering, we have

$$\frac{1}{\left(\begin{array}{c} d \\ k \end{array}\right)} \sum_{V \subseteq S : |V| = k} I(X_V; Y) \leq I(X_S; Y)$$

□

Acknowledgments

We are grateful to Jean Cardinal, Martine Labbe, Yann-Ael Le Borgne and Abhilash Miranda from the Université Libre de Bruxelles, for their precious time and advices concerning this work. We would like also to thank the editor and the anonymous reviewers for their useful suggestions and comments.

References

[1] H. Almuallim and T. G. Dietterich. Learning with many irrelevant features. In *Proceedings of the Ninth National Conference on Artificial Intelligence (AAAI91)*, pages 547–552. AAAI Press, 1991.

[2] Y. Asahiro, K. Iwama, H. Tamaki, and T. Tokuyama. Greedily finding a dense subgraph. *Journal of Algorithms*, 34(1):203–221, 2000.

[3] R. Battiti. Using mutual information for selecting features in supervised neural net learning. In *IEEE Transactions on Neural Networks*, 1994.

[4] D. A. Bell and H. Wang. A formalism for relevance and its application in feature subset selection. *Machine Learning*, 41(2):175–195, 2000.

[5] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, Series B(57):289–300, 1995.

[6] A. Billionnet and F. Calmels. Linear programming for the 0-1 quadratic knapsack problem. *European Journal of Operational Research*, 92:310–325, 1996.

[7] A. Blum and P. Langley. Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97:245–271, 1997.

[8] A. Blum and R. L. Rivest. Training a 3-node neural network is NP-complete. In *Machine Learning: From Theory to Applications*, volume 661 of *Lecture Notes in Computer Science*. Springer, 1993.

[9] U. M. Braga-Neto. Fads and fallacies in the name of small-sample microarray classification. *IEEE Signal Processing Magazine, Special Issue on Signal Processing Methods in Genomics and Proteomics*, 24(1):91–99, 2007.

[10] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley, New York, 1990.

[11] M. Dash and H. Liu. Feature selection for classification. *Intelligent Data Analysis*, IOS Press 1997.

[12] S. Davies and S. Russell. Np-completeness of searches for smallest possible feature sets. In *Proceedings of the AAAI Fall Symposium on Relevance*, 1994.

[13] T. Van den Bulcke, K. Van Leemput, B. Naudts, P. van Remortel, H. Ma, A. Verschoren, B. De Moor, and K. Marchal. Syntren: a generator of synthetic gene expression data for design and analysis of structure learning algorithms. *BMC Bioinformatics*, 7(1):43, 2006.

[14] C. Ding and H. Peng. Minimum redundancy feature selection from microarray gene expression data. *Journal of Bioinformatics and Computational Biology*, 3(2):185–205, 2005.

[15] J. Dougherty, R. Kohavi, and M. Sahami. Supervised and unsupervised discretization of continuous features. In *International Conference on Machine Learning*, pages 194–202, 1995.

[16] W. Duch, T. Winiarski, J. Biesiada, and A. Kachel. Feature selection and ranking filters. In *International Conference on Artificial Neural Networks (ICANN) and International Conference on Neural Information Processing (ICONIP)*, pages 251–254, June 2003.

[17] F. Fleuret. Fast binary feature selection with conditional mutual information. *Journal of Machine Learning Research*, 5:1531–1555, 2004.

[18] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[19] A. Jain and D. Zongker. Feature selection: evaluation, application, and small sampleperformance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19, 1997.

[20] A. Jakulin and I. Bratko. Quantifying and visualizing attribute interactions, 2003.

[21] R. Kohavi and G. H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2):273–324, 1997.

[22] I. Kojadinovic. Relevance measures for subset variable selection in regression problems based on k-additive mutual information. *Computational Statistics and Data Analysis*, 49, 2005.

[23] D. Koller and M. Sahami. Toward optimal feature selection. In *International Conference on Machine Learning*, pages 284–292, 1996.

[24] I. Kononenko. Estimating attributes: Analysis and extensions of RELIEF. In *European Conference on Machine Learning*, pages 171–182, 1994.

[25] A. A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. Dalla Favera, and A. Califano. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, 7, 2006.

[26] W. J. McGill. Multivariate information transmission. *Psychometrika*, 19, 1954.

[27] P. Merz and B. Freisleben. Greedy and local search heuristics for unconstrained binary quadratic programming. *Journal of Heuristics*, 8(2):1381–1231, 2002.

[28] P. E. Meyer and G. Bontempi. On the use of variable complementarity for feature selection in cancer classification. In F. Rothlauf et al., editor, *Applications of Evolutionary Computing: EvoWorkshops*, volume 3907 of *Lecture Notes in Computer Science*, pages 91–102. Springer, 2006.

[29] P. E. Meyer, K. Kontos, F. Lafitte, and G. Bontempi. Information-theoretic inference of large transcriptional regulatory networks. *EURASIP Journal on Bioinformatics and Systems Biology*, Special Issue on Information-Theoretic Methods for Bioinformatics, 2007.

[30] L. Paninski. Estimation of entropy and mutual information. *Neural Computation*, 15(6):1191–1253, 2003.

[31] H. Peng and F. Long. An efficient max-dependency algorithm for gene selection. In *36th Symposium on the Interface: Computational Biology and Bioinformatics*, may 2004.

[32] H. Peng, F. Long, and C. Ding. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238, 2005.

[33] D. Pisinger. Upper bounds and exact algorithms for dispersion problems. *Computers & OR*, 33:1380–1398, 2006.

[34] G. Provan and M. Singh. Learning bayesian networks using feature selection. In *in Fifth International Workshop on Artificial Intelligence and Statistics*, pages 450–456, 1995.

[35] D. W. Scott. *Multivariate Density Estimation. Theory,*. Wiley, 1992.

[36] M. Studený and J. Vejnarová. The multiinformation function as a tool for measuring stochastic dependence. In *Proceedings of the NATO Advanced Study Institute on Learning in graphical models*, pages 261–297, 1998.

[37] N. Tishby, F. Pereira, and W. Bialek. The information bottleneck method. In *Proceedings of the 37-th Annual Allerton Conference on Communication, Control and Computing*, 1999.

[38] G. D. Tourassi, E. D. Frederick, M. K. Markey, and Jr. C. E. Floyd. Application of the mutual information criterion for feature selection in computer-aided diagnosis. *Medical Physics*, 28(12):2394–2402, 2001.

[39] G. Trunk. A problem of dimensionality: A simple example. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1979.

[40] I. Tsamardinos and C. Aliferis. Towards principled feature selection: Relevancy, filters, and wrappers. *Artificial Intelligence and Statistics*, 2003.

[41] M. J. van de Vijver, Y. D. He, L. J. van't Veer, H. Dai, A. A. Hart, D. W. Voskuil, G. J. Schreiber, J. L. Peterse, C. Roberts, M. J. Marton, M. Parrish, D. Atsma, A. Witteveen, A. Glas, L. Delahaye, T. van der Velde T, H. Bartelink H, S. Rodenhuis, E. T. Rutgers, S. H. Friend, and R. Bernards. A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medecine*, 347, 2002.

[42] L. J. van 't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 406, 2002.

[43] W. Wienholt and B. Sendhoff. How to determine the redundancy of noisy chaotic time series. *International Journal of Bifurcation and Chaos*, 6(1):101–117, 1996.

[44] Y. Y. Yao, S. K. Michael Wong, and Cory J. Butz. On information-theoretic measures of attribute importance. In *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 1999.

[45] L. Yu and H. Liu. Efficient feature selection via analysis of relevance and redundancy. *Journal of Machine Learning Research*, 5:1205–1224, 2004.

[46] Z. Zhao and H. Liu. Searching for interacting features. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI-07)*, 2007.