On the Use of Variable Complementarity for Feature Selection in Cancer Classification

Patrick E. Meyer and Gianluca Bontempi

Université Libre de Bruxelles, (CP 212), 1050 Bruxelles, Belgique,
(pmeyer,gbonte)@ulb.ac.be
home page: http://ulb.ac.be/di/mlg/

Abstract. The paper presents an original filter approach for effective feature selection in classification tasks with a very large number of input variables. The approach is based on the use of a new information theoretic selection criterion: the double input symmetrical relevance (DISR). The rationale of the criterion is that a set of variables can return an information on the output class that is higher than the sum of the informations of each variable taken individually. This property will be made explicit by defining the measure of *variable complementarity*. A feature selection filter based on the DISR criterion is compared in theoretical and experimental terms to recently proposed information theoretic criteria. Experimental results on a set of eleven microarray classification tasks show that the proposed technique is competitive with existing filter selection methods.

1 Introduction

Statisticians and data-miners are used to build predictive models and infer dependencies between variables on the basis of observed data. However, in a lot of emerging domains, like bioinformatics, they are facing datasets characterized by a very large number of features (up to several thousands), a large amount of noise, non-linear dependencies and, often, only several hundreds of samples. In this context, the detection of functional relationships as well as the design of effective classifiers appears to be a major challenge. Recent technological advances, like microarray technology, have made it possible to simultaneously interrogate thousands of genes in a biological specimen. It follows that two classification problems commonly encountered in bioinformatics are how to distinguish between tumor classes and how to predict the effects of medical treatments on the basis of microarray gene expression profiles. If we formalize this prediction task as a supervised classification problem, we realize that we are facing a problem where the number of input variables, represented by the number of genes, is huge (around several thousands) and the number of samples, represented by the clinical trials, is very limited (around several tens). As a consequence, the use of classification techniques in bioinformatics requires the capacity of managing datasets with many variables and few samples (also known as high *feature-to-sample ratio* datasets). Because of well-known numerical and statistical accuracy

issues, it is typically necessary to reduce the number of variables before starting a learning procedure. Furthermore, selecting features (i.e. genes) can increase the intelligibility of a model while at the same time decreasing measurements and storage requirements [1]. A number of experimental studies [2–4] have shown that irrelevant and redundant features can dramatically reduce the predictive accuracy of models builded from data.

Feature selection is a topic of machine learning whose objective is selecting, among a set of input variables, the ones that will lead to the best predictive model. We can formulate this problem as a stochastic optimization problem since the search for the best subset have to be conducted on the basis of statistical estimates of the accuracy associated to a given subset. Two well-known approaches in feature selection combine a search strategy with a stochastic evaluation function: the *filter approach* and the *wrapper approach* (see [3, 2]). 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. 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.

As a consequence of the increasing number of variables in datasets, search algorithms have evolved from exhaustive search [5] to heuristic search [2, 6–8] and evaluation functions are moving from learning algorithms (in the case of wrappers) [3, 2, 6, 7] to fast dependence measure such as mutual information (in the case of filters) [9–12]. Formal definitions of relevance [2, 8] and redundancy [10] have been developped in order to guide the search and render algorithms faster.

This paper will focus on the use of filter techniques for feature selection in supervised classification tasks. In particular, we present an original filter approach based on a new information theoretic selection criterion, called the *double input symmetrical relevance* (DISR). This criterion combines two well known intuitions 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 will be made explicit by defining the notion of *variable complementarity*. Secondly, in 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 intuition will be made formal by the computation of a lower-bound on the information of a subset of variables expressed as the average of information of all its sub-subsets.

The DISR criterion can be used to select among a finite number of alternative subsets the one expected to return the maximum amount of information on the output class. As we intend to benchmark its performance with respect to state-of-the-art information theoretic criteria we define an experimental session where several filter algorithms with different selection criteria but the same

search strategy are compared. In our experiments we compare the filter based on DISR with four state-of the art approaches: a Ranking algorithm [12] and three filters based on the same search strategy: the forward selection. The three state-of-the-art criteria are the Relevance criterion [8], the Minimum Redudancy Maximum Relevance criterion [9] and the Conditional Mutual Information Maximization criterion [11]. The assessment of the different filters is obtained by measuring the classification accuracy of several learning algorithms which adopt as inputs the set of variables returned by each of the filter methods. For our benchmark purposes, we use eleven public-domain multi-class microarray gene expression datasets. 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 notions of information theory and our definition of variable complementarity. Section 3 recalls four state-of-art information theoretic criteria which have been proposed in literature to deal with the feature selection problem. Section 4 provides the theoretical justification of the DISR information theoretic criterion. Section 5 illustrates the experimental results. The conclusions are summarized in Section 6.

2 Information theoretic notions for feature selection

This paper deals with supervised multi-class classification. We will assume either that all the variables are discrete or that they can be made discrete by a quantization step. Hereafter, we will denote by Y the discrete output random variable representing the class and by X the multi-dimensional discrete input random variable.

In qualitative terms, feature selection boils down to select, among a set of potential variables, the most *relevant* ones. At the same time it would be appealing that these selected variables are not *redundant*. The notions of relevance and redundancy can be made more formal thanks to the use of dependency measures [2, 10].

Let us first introduce some concepts of information theory:

Definition 1. *[13] The conditional entropy of Y given Z, denoted by $H(Y|Z)$ is defined as:*

$$H(Y|Z) = -\sum_{y\in\mathcal{Y}}\sum_{z\in\mathcal{Z}} p(y,z)\log p(y|z) \tag{1}$$

and the $I(X;Y|Z)$ is the conditional mutual information.

Definition 2. *[13] The conditional mutual information of the random variables X and Y given Z is defined as:*

$$I(X;Y|Z) = H(X|Z) - H(X|Z,Y) \tag{2}$$

These definitions allow us to introduce the following measure of relevance proposed by [8]:

Definition 3. Relevance.
Consider three random variables X,Y and Z and their joint probability distribution $p_{X,Y,Z}(x,y,z)$. If $H(Y|Z) = 0$, then the variable relevance of X to Y given Z, denoted by $r(X;Y|Z)$, is zero. Else if $H(Y|Z) \neq 0$, then the variable relevance of X to Y given Z is defined as:

$$r(X;Y|Z) = \frac{I(X;Y|Z)}{H(Y|Z)} \tag{3}$$

According to this definition the relevance is a function $0 \leq r(X;Y|Z) \leq 1$ that measures the relative reduction of uncertainty of Y provided by X once the value of Z is given.

In the following, we rather consider as measure of relevance the classical (non-normalized) mutual information, i.e. $I(X_i;Y|X_S)$, where X_i denotes an input variable and X_S a subset of variables not containing X_i.

The formalization of the notion of relevance makes explicit one of the major challenges in feature selection: the mutual information between an input variable X_i and the output class Y is *conditionally dependent*. This means that an input variable X_i having a significative relevance to Y given X_S, can return a null relevance conditioned on an other variable. The following two examples may serve to better illustrate this idea.

Example 1. Consider the four random variables Y, X_i, X_S and X_M such that $Y = X_i + X_S$ and $X_M = \frac{X_i}{2}$.

Given X_S, X_i has a large relevance ($I(X_i;Y|X_S) = H(Y|X_S)$), while its relevance goes to zero in the case where it is conditioned to X_M (i.e. $I(X_i;Y|X_M) = 0$). In this example, X_i and X_M have both a high mutual information with the output Y but a low conditional mutual information when conditioned to the other variable.

The next example shows that the relevance can also increase by conditioning.

Example 2. Consider the three random variables Y, X_i and X_S such that Y and X_i are independent and $Y = X_i + X_S$.

In this case the mutual information of X_i with Y is zero ($I(X_i;Y) = 0$) whereas the conditional mutual informatione increases up when conditioned to X_S (i.e. $I(X_i;Y|X_S) = H(Y|X_S)$).

Several definitions of a relevant variable and a redundant variable have been well defined in the literature, see [2, 3, 10].

These examples show that it is hard to predict, in terms of relevance, the joint effect of several input variables on an output variable Y.

As shown in Example 1 the mutual information $I(X_{i,j};Y)$ of a set $\{X_i, X_j\}$ (aka joint mutual information [14]) can be smaller than the sum of each relevance taken separately. In generic terms, we could describe these two variables as *redundant* for the task of classifying Y. For a formal definition of redundancy we refer the reader to [10, 9]. Also, as shown in Example 2, it could happen that two variables have jointly a larger mutual information with Y than when they

are considered separately. In this case we say that the two variables are *complementary*. Note that *variable complementarity* should warn us against eliminating variables with null mutual information with the output (i.e. $I(X_i;Y) = 0$) since the joint information of two random variables $I(X_{i,j};Y)$ can be higher than the sum of their individual informations $I(X_i;Y)$ and $I(X_j;Y)$.

Variable complementarity was underlined experimentally in [1] and explained in [15] as a second order term of the Möbius representation of the mutual information. It can be useful to define the notion of complementarity between two variables with respect to an output Y as the difference between the joint mutual information and the sum of the "individual" mutual informations. We introduce then the following measure:

Definition 4. *The* complementarity *of two random variables X_i and X_j with respect to an output Y is defined as,*

$$C_Y(X_i, X_j) = I(X_{i,j};Y) - I(X_i;Y) - I(X_j;Y) \tag{4}$$

where $X_{i,j} = \{X_i, X_j\}$.

We define two variables as complementary if their measure of complementarity with respect to Y is positive. We remark also that a negative value of the complementarity can be taken as a measure of the redundancy of a pair of variables for the task of predicting Y.

The example 2 is an illustration of complementarity between X_i and X_S since in that case:

$$I(X_{i,S};Y) > \underbrace{I(X_i;Y)}_{0} + I(X_S;Y) \tag{5}$$

Another illustration of complementarity is given by the well-known XOR problem [2, 1]:

Example 3. Xor problem:

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

We see that X_1 and X_2 have a null mutual information with the output, once they are taken individually (i.e. $I(X_1;Y) = 0$, $I(X_2;Y) = 0$). However, when they are taken together the mutual information $I(X_{1,2};Y) = H(Y) > 0$ of the subset is positive.

Complementarity explains why an apparently irrelevant combination of variables can eventually perform efficiently in a learning task. In the following section, we will proceed to a critical survey of information theoretic approaches existing in literature, by stressing when and where the notion of complementarity is taken into account.

3 State of the Art

As mutual information can measure relevance, this quantity is currently used in literature for performing feature selection. One of the main reasons for adopting

it is its low complexity computational cost ($O(d \times N)$ where d is the number of variables and N is the number of samples) in the case of discrete variables. The following sections will sketch four state-of-the-art filter approaches that use this quantity.

3.1 Variable Ranking (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 [12].

The main advantage of the method is its rapidity of execution. Indeed, only n computations of mutual information are required for a resulting complexity $O(n \times 2 \times N)$. The main drawback derives from the fact that possible redundancies between variables is not taken into account. Indeed, two redundant variables, yet highly relevant taken individually, will be both well ranked. As a result, a model that uses these two variables is dangerously prone to an increased variance without any gain in terms of bias reduction. 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 2 or Example 3). As a consequence, these variables could be badly ranked, or worse eliminated, by the ranking filter.

3.2 Filters Combining Relevance and Redundancy Analysis

Although the variable ranking algorithm is reputed to be fast, it may be poorly efficient as it only relies on individual relevance. Recently, new algorithms that combine relevance and redundancy analysis offer a good trade-off between accuracy and computational load as the Fast Correlation Based Filter [10]. Also, some heuristic search methods such as the best first search (also known as the forward selection) can be combined efficiently with information theoretic criteria in order to select the best variable given a previously selected subset.

Forward Selection is a search method that starts with an empty set of variables. At each step, it selects the variable that brings the best improvement (according to the selection criterion). As a consequence, each selected variable does influence the evaluations of the following steps. This hill-climbing search selects a subset of $d < n$ variables in d steps and explores only $\sum_{i=0}^{d}(n - i)$ evaluations.

In the following sections, several information theoretic criteria existing in the literature and that can be easily combined with the forward selection, are presented.

Relevance Criterion (REL) The relevance criterion is a well-known criterion which is used together with the forward selection search strategy [8]. The

approach consists in 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_{REL} returned by the relevance criterion is,

$$X_{REL} = \arg \max_{X_i \in X_{-S}} \{I(X_i; Y|X_S)\} \tag{6}$$

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[1].

Although this method is appealing, it presents some major drawbacks. The estimation of the relevance requires the estimation of several 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 two reasons we recently assisted to the adoption of selection criteria based on bi- and tri-variate densities only.

Minimum Redundancy - Maximum Relevance criterion (MRMR) The minimum redundancy - maximum relevance criterion [9] consists in selecting the variable X_i among the not yet selected features 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 mean redundancy of X_i to each variables $X_j \in X_S$ already selected.

$$u_i = I(X_i; Y) \tag{7}$$

$$z_i = \frac{1}{d} \sum_{X_j \in X_S} I(X_i; X_j) \tag{8}$$

$$X_{MRMR} = \arg \max_{X_i \in X_{-S}} \{u_i - z_i\} \tag{9}$$

At each step, this method selects the variable which has the best trade-off relevance-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 $(d + 1)$ bi-variate densities (one for each already selected variables plus one with the output). It has been shown in [9] that the MRMR criterion is an optimal first order approximation of the conditional relevance criterion. Furthermore, MRMR avoids the estimation of multivariate densities by using multiple bivariate densities.

Note that, although the method aims to address the issue of redundancy between variables through the term z_i, it is not able to take into account the

[1] Note that in [8] a normalized version of relevance (Eq. 3) is used.

complementarities between variables. This could be ineffective in situations like the one of Example 2 where, although the set $\{X_i, X_S\}$ has a large relevance to Y, we observe that

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 is not relevant to Y.

Conditional Mutual Information Maximization Criterion (CMIM) This approach [11] proposes to select the feature $X_i \in X_{-S}$ whose minimal conditional relevance $I(X_i; Y|X_j)$ among 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, conditional 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 [2] criterion is,

$$X_{CMIM} = \arg \max_{X_i \in X_{-S}} \{ \min_{X_j \in X_S} I(X_i; Y|X_j)\} \tag{10}$$

This selection criterion is powerful. It selects relevant variables, it avoids redundancy, it avoids estimating high dimensional multivariate densities and unlike the previous method, it does not ignore variable complementarity. However, it will not necessary select a variable complementary with the already selected variables. Indeed, a variable that has a high complementarity with an 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 example 3).

In terms of complexity, note that at the dth step of the forward search, the algorithm computes $n - d$ evaluations where each evaluation following CMIM requires the estimation of d tri-variate densities (one for each previously selected variable).

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

4 Double input symmetrical relevance (DISR) criterion

A lower bound on mutual information In this section, we derive a lower bound on the the mutual information between a subset X_S and a target variable Y. It is shown [16] that this quantity is lower bounded by the average of the same quantity computed for all the sub-subsets $X_{S-i} = X_S \setminus X_i$ of X_S.

Theorem 1. *: Let $X_S = \{X_1, ..., X_d\}$ be a subset of d variables of X and $X_{S-i} = X_S \setminus X_i$, $i \in 1, \ldots, d$ a subset of X_S that does not contain the variable X_i. Then,*

[2] Note that in [11] this method was applied to select binary features in a pattern recognition task.

$$I(X_S;Y) \geq \frac{1}{d} \sum_{i \in S} I(X_{S-i};Y) \tag{11}$$

The theorem expresses that the mutual information of a subset S and a target variable Y is lower bounded by the quantity $\mathcal{L}\left(I(X_S;Y)\right)$, that is the average of the same quantity computed for all the sub-subsets X_{S-i} of X_S.

In the following, we will use this theorem as a theoretical support to the following heuristic: *without any additional knowledge on how subsets of d variables should combine, the most promising subset is a combination of the best performing subsets of* $(d-1)$ *variables).*

Criterion Given a fixed number d of variables, we can write the problem of feature selection in the following form:

$$S_{best} = \arg \max_{S \,:\, |S|=d} I(X_S;Y) \tag{12}$$

In other words, the goal of feature selection is to find the subset of d variables which maximizes the mutual information with the output Y.

Our idea consists in replacing the maximization of the quantity $I(X_S;Y)$ by the maximization of its lower bound $\mathcal{L}\left(I(X_S;Y)\right)$:

$$\max_{S \,:\, |S|=d} I(X_S;Y) \geq \max_{S \,:\, |S|=d} \sum_{i \in S} I(X_{S-i};Y) \tag{13}$$

Replacing again the right-hand term by its lower bound and recursively until we have subsets of two variables:

$$\geq \max_{S} \sum_{i \in S} \sum_{j \in S} I(X_{S-(i,j)};Y) \geq \max_{S} \sum_{i \in S} \sum_{j \in S} I(X_{i,j};Y) \tag{14}$$

In other words, without any information on how to combine subsets made of more than two variables, the most promising subset (the best bound) is the one with the highest sum of mutual information on all the combinations of two variables. We choose to stop the recursivity at two variables because it is the minimal size of subset that can capture variable complementarity (i.e. $I(X_{i,j};Y) = I(X_i;Y) + I(X_j;Y) + C_Y(X_i;X_j)$). Note that this strategy, when we stop the recursion at one variable, boils down to the ranking approach.

A similar approach has been developped in [15] based on the Möbius representation of mutual information. However, in order to improve the selection procedure we use here a normalized measure of mutual information very close to the symmetrical uncertainty presented in [10]: the *symmetrical relevance*.

Definition 5. *Given two random variables X,Y a joint probability distribution* $p(x,y)$*, the symmetrical relevance* $SR(X,Y)$ *is defined as:*

$$SR(X;Y) = \frac{I(X,Y)}{H(X,Y)} \tag{15}$$

This definition expresses that symmetrical relevance is a function $0 \leq SR(X;Y) \leq 1$ that indicates the "concentration" of mutual information "contained" in $p(x,y)$.

As a consequence, our resulting criterion is the following:

$$X_{DISR} = \arg \max_{X_i \in X_{-S}} \{ \sum_{X_j \in X_S} SR(X_{i,j};Y) \} \tag{16}$$

The main advantage in using this criterion for selecting variables is that a complementary variable of an already selected one has a much higher probability to be selected than with other criteria. As this criterion measures symmetrical relevance on all the combination of two variables (double input) of a subset, we have called the criterion: the *double input symmetrical relevance* (DISR). At the dth step of the forward search, the search algorithm computes $n-d$ evaluations where each evaluation requires the estimation of d tri-variate densities (one for each previously selected variable). In the next section, the DISR criterion is assessed and compared with the other heuristic search filters discussed in the Section 3.

Table 1 summarizes the methods discussed so far in terms of some peculiar aspects: the capacity of selecting relevant variables, of avoiding redundancy, of selecting complementary features and of avoiding the computation of multivariate densities.

5 Experiments

A major topic in bioinformatics is how to build accurate classifiers for cancer diagnostic and prognostic purposes on the basis of microarray genomic signatures. This task can be considered as a challenging benchmark for feature selection algorithms [9] given the high feature to sample ratio.

We use eleven public domain multi-class datasets from [17] (Table 2) in order to assess and compare our technique with the state-of-the-art approaches.

In our experimental framework, each continuous variable has been discretized in equal sized interval. The number of intervals of each input is based on the Scott criterion, see [18]. All the datasets are partitioned into two parts: a selection set and a test set (each having size equal to $N/2$). We compare the filter based on DISR with the four state-of the art approaches discussed above: a Ranking algorithm and three filters based on the Relevance criterion, the Minimum Redudancy Maximum Relevance criterion and the Conditional Mutual Information Maximization criterion, respectively. Each selection method stops after that 15 variables have been selected. Then, the evaluation of the selection is done on the test set, by using a ten-fold cross validation with a 3-nearest neighbor, a naive Bayes and a SVM learning algorithm with a radial kernel. Each learning technique led to the choice of a different number of variables in a range from 2 to 15. Then for each of the eleven datasets and for each selection method, the best number of variables and the classification accuracy is computed. A set of statistical paired t-test on the set of classification errors are reported in Table 3.

As far as the implementation of the three learning methods is concerned, we used the algorithms made available by the R statistical language [19].

According to Table 2, the DISR criterion outperforms slightly all the other methods in terms of average accuracy. Furthermore, our method is one of the two best methods for 7 out of 11 datasets.

Table 3 reports the significant wins, ties or losses (at 0.1 and 0.2 significance levels of a paired two-tailed t-test, respectively) of the DISR criterion against all the other. We remark that in the case a 3-Nearest Neighbor, the DISR criterion is equivalent to MRMR and better than all the other methods. For a naive Bayes classifier, the performances of the DISR are slightly lower. This is not surprising because the benefits of the DISR criterion are related to variable complementarity whereas the success of a naive Bayes classifier typically relies on the opposite, that is variable independence. As far as the SVM classifier is concerned, at the 0.1 significance level, DISR appears to be slightly better than both REL and CMIM, and slightly worse than RANK and MRMR. However, at 0.2 significance level the DISR outperforms all the other methods except MRMR.

criterion	rank	REL	MRMR	CMIM	DISR
relevance selection	V	V	V	V	V
redundancy avoidance	–	V	V	V	V
complementarity selection	–	V	–	–	V
multivariate density avoidance	V	–	V	V	V

Table 1. Qualitative comparison of different information theoretic filters, according to different aspects: relevance selection, redundancy avoidance, complementarity selection and multivariate densities avoidance.

6 Conclusion and future work

This paper formalized an original notion in feature selection: variable complementarity. Also, a lower bound on the mutual information of a subset of variables with the output was demonstrated. On the basis of these considerations, we proposed a new selection criterion: the double input symmetrical relevance (DISR). The experimental session shows that this criterion is promising in high feature-to-sample ratio classifaction tasks like gene expression microarray datasets. Note that in gene selection, variable complementarity can be biologically meaningful since it is common to observe combination of genes acting together.

Further experiments will focus on (i) datasets with more samples and/or less features, (ii) other search strategies than the forward selection in order to validate the criterion in a wider range of domains, (iii) the impact of the discretization method to the efficiency of the feature selection algorithms.

Dataset (DN)	n	N	c	Rank	REL	CMIM	MRMR	DISR
11_Tumors	12534	87	11	**49%**	46%	48%	42%	**49%**
14_Tumors	15010	308	26	22%	**25%**	20%	**26%**	19%
9_Tumors	5727	60	9	19%	**36%**	20%	23%	**28%**
Leukemia1	5328	72	3	71%	**74%**	71%	69%	**78%**
Leukemia2	11226	72	3	**68%**	57%	62%	65%	**67%**
Prostate_Tumor	10510	102	2	**76%**	66%	62%	**78%**	73%
Brain_Tumor1	5921	90	5	**73%**	70%	70%	70%	**71%**
Brain_Tumor2	10368	50	4	47%	47%	49%	**59%**	**60%**
Lung_Cancer	12601	203	5	**84%**	74%	**82%**	77%	74%
SRBCT	2309	83	4	70%	53%	**75%**	**75%**	67%
DLBCL	5470	77	2	77%	**88%**	70%	71%	**88%**

Table 2. The 11 datasets of microarray cancer from `http://www.tech.plym.ac.uk`. The column n represents the number of probes in the microarray, the column N the number of samples and the column c the number of classes. The remaining columns contain the average accuracy of each selection method averaged over the three classifiers (SVM, 3-NN, naive Bayes). The accuracy of the two best methods for each dataset is typed in bold face.

W/T/L VS DISR	Rank	REL	CMIM	MRMR	Rank	REL	CMIM	MRMR
3-NN	0/9/2	1/8/2	1/7/3	2/7/2	1/5/5	1/7/3	1/7/3	2/7/2
Naive Bayes	3/8/0	2/7/2	1/8/2	1/9/1	4/7/0	3/6/2	1/8/2	1/9/1
SVM	2/9/0	2/6/3	2/6/3	2/8/1	2/6/3	3/4/4	2/5/4	3/5/3

Table 3. Statistically (0.1 level and 0.2 level of a paired two-tailed t-test) significant wins, ties or losses over best first search combined with DISR criterion.

References

1. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. Journal of Machine Learning Research **3** (2003) 1157–1182
2. Kohavi, R., John, G.H.: Wrappers for feature subset selection. Artificial Intelligence **97**(1-2) (1997) 273–324
3. Blum, A., Langley, P.: Selection of relevant features and examples in machine learning. Artificial Intelligence **97** (1997) 245–271
4. Provan, G., Singh, M.: Learning bayesian networks using feature selection. In: in Fifth International Workshop on Artificial Intelligence and Statistics. (1995) 450–456
5. Almuallim, H., Dietterich, T.G.: Learning with many irrelevant features. In: Proceedings of the Ninth National Conference on Artificial Intelligence (AAAI91), AAAI Press (1991) 547–552
6. Caruana, R., Freitag, D.: Greedy attribute selection. In: International Conference on Machine Learning. (1994) 28–36
7. Aha, D.W., Bankert, R.L.: A comparative evaluation of sequential feature selection algorithms. Artificial Intelligence and Statistics **5** (1996)
8. Bell, D.A., Wang, H.: A formalism for relevance and its application in feature subset selection. Machine Learning **41**(2) (2000) 175–195
9. Peng, H., Long, F.: An efficient max-dependency algorithm for gene selection. In: 36th Symposium on the Interface: Computational Biology and Bioinformatics. (2004)
10. Yu, L., Liu, H.: Efficient feature selection via analysis of relevance and redundancy. Journal of Machine Learning Research **5** (2004) 1205–1224
11. Fleuret, F.: Fast binary feature selection with conditional mutual information. Journal of Machine Learning Research **5** (2004) 1531–1555
12. Duch, W., Winiarski, T., Biesiada, J., Kachel, A.: Feature selection and ranking filters. In: International Conference on Artificial Neural Networks (ICANN) and International Conference on Neural Information Processing (ICONIP). (2003) 251–254
13. Cover, T.M., Thomas, J.A.: Elements of Information Theory. John Wiley, New York (1990)
14. Yang, H., Moody, J.: Feature selection based on joint mutual information. In: In Advances in Intelligent Data Analysis (AIDA), Computational Intelligence Methods and Applications (CIMA), Rochester New York, ICSC (1999)
15. Kojadinovic, I.: Relevance measures for subset variable selection in regression problems based on k-additive mutual information. Computational Statistics and Data Analysis **49** (2005)
16. Meyer, P.: Information theoretic filters for feature selection. Technical report, Universite Libre de Bruxelles ((548) 2005)
17. web: (http://www.tech.plym.ac.uk/spmc/bioinformatics/microarray_cancers.html)
18. Scott, D.W.: Multivariate Density Estimation. Theory,. Wiley (1992)
19. R-project: (www.r-project.org)