Causal filter selection in microarray data

Gianluca Bontempi GBONTE@ULB.AC.BE
Patrick E. Meyer PMEYER@ULB.AC.BE
Machine Learning Group,
Computer Science Department, Faculty of Sciences
ULB, Université Libre de Bruxelles, Brussels, Belgium

Abstract

The importance of bringing causality into play when designing feature selection methods is more and more acknowledged in the machine learning community. This paper proposes a filter approach based on information theory which aims to prioritise direct causal relationships in feature selection problems where the ratio between the number of features and the number of samples is high. This approach is based on the notion of *interaction* which is shown to be informative about the relevance of an input subset as well as its causal relationship with the target. The resulting filter, called mIMR (min-Interaction Max-Relevance), is compared with state-of-the-art approaches. Classification results on 25 real microarray datasets show that the incorporation of causal aspects in the feature assessment is beneficial both for the resulting accuracy and stability. A toy example of causal discovery shows the effectiveness of the filter for identifying direct causal relationships.

1. Introduction

Feature selection is the domain of machine learning which studies data-driven methods to select, among a set of input variables, the ones that will lead to the most accurate predictive model (Guyon et al., 2006). *Causal inference* is the domain of artificial intelligence which aims to uncover causal relationships between variables from observational data (Spirtes et al., 1993). The importance of bringing causality into play when designing feature selection methods has been exhaustively discussed in the seminal paper (Guyon et al., 2007). According to the authors, *the benefits of incorporating causal discovery in feature selection include understanding more finely the data structure and making prediction possible under manipulations and some distribution changes.*

How to incorporate causal discovery issues in filter problems where the ratio between the number of features and the number of samples is high is still an open issue. This is typically relevant in microarray classification tasks where the goal is, for example, to distinguish between tumor classes or predict the effects of medical treatments on the basis of gene expression profiles (Xing et al., 2001). 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 inference of causal relationships between variables plays a major role since more and more biologists and medical doctors expect from data analysis not only accurate prediction models (e.g. for prognostic purposes) but also insights about causes of diseases (e.g. leukemia or diabete) and appropriate therapeutic targets.

The role of information-theoretic filters in feature selection has been largely discussed in the machine learning literature. Mutual information and related notions of information theory has been used in several filter algorithms like Ranking (Duch et al., 2003), Markov blanket filter (Koller & Sahami, 1996), Fast Correlation Based Filter (FCBF) (Yu & Liu, 2004), Relevance Filter (Battiti, 1994; Bell & Wang, 2000), Conditional Mutual Information Maximization (CMIM) filter (Fleuret, 2004), Minimum Redundancy Maximum Relevance (mRMR) filter (Peng et al., 2005) and Double Input Symmetrical Relevance (DISR) (Meyer et al., 2008) filter.

This paper proposes an information theoretic formulation of the feature selection problem which sheds

Appearing in *Proceedings of the 27th International Conference on Machine Learning*, Haifa, Israel, 2010.

light on the interplay between causality and information maximization. This is done by means of the notion of *feature interaction* which plays the role of missing link between causal discovery and feature selection. Qualitatively, feature interaction appears when we can model the dependencies between group of attributes only by considering them simultaneously (Freitas, 2001). Consider for instance two input features $\mathbf{x}_1$, $\mathbf{x}_2$ and a target class $\mathbf{y}$. It is said that $\mathbf{x}_1$ and $\mathbf{x}_2$ interact when the direction or magnitude of the relationship between $\mathbf{x}_1$ and $\mathbf{x}_2$ depends on the value of $\mathbf{y}$. Actually, this can be called a three-way interaction. Higher-order attribute interactions can be defined in a similar way. A nice aspect of the information theoretic formalism is that the interaction between attributes can be formalised on the basis of the notion of mutual information and conditional mutual information (McGill, 1954). At the same time the interaction between variables sheds a light on the possible causal patterns existing between them. The role of interaction in feature selection has already been discussed in the machine learning litterature. Jakulin (Jakulin, 2005) proposes an heuristic based on interaction for selecting attributes within the naive Bayesian classifier. The authors of (Meyer et al., 2008) proposed a filter algorithm which relies on the maximisation of an information theoretic criterion, denoted Double Input Symmetrical Relevance (DISR), which implicitly takes into account the interaction, or complementarity between variables, in the choice of the features. The paper of (Meyer et al., 2008) showed also that the maximisation of the DISR criterion is beneficial to the selection of the Markov blanket in a classification task. It is however known that the Markov blanket is a superset of the set of direct causes of a target $\mathbf{y}$, since it contains beyond direct causes, also the effects (direct descendants) and their parents (also known as spouses).

This paper proposes and assesses an original causal filter criterion which aims both to select a feature subset which is maximally informative and to prioritise direct causes. Our approach relies on the following considerations. The first consideration is that the maximization of the information of a subset of variables by forward selection can be shown to be equivalent to a problem of min-Interaction Max-Relevancy (mIMR), where the most informative variables are the one having both high mutual information with the target and high negative interaction (or high complementarity) with the others. The second consideration is that causal discovery differs from conventional feature selection since not all informative or strongly relevant variables are also direct causes. It follows that a causal filter algorithm should proceed by implementing a mIMR criterion but avoid to consider variables, like effects and spouses, which are not causally relevant. We propose here an approach which consists in restricting the selection to variables which have both positive relevance and negative interaction. Variables with positive interaction (i.e. effects) are penalised and variables with null relevance, even if interacting negatively (i.e. spouses), are discarded. An additional contribution of the paper is that the estimation of three-way interaction terms is sped up by conditioning on the values of the output class. The computational advantage is particularly evident in the Gaussian case, where the computational effort is limited to the computation of few correlation matrices of the inputs. The mIMR filter, was assessed in terms of accuracy on a set of 25 public microarray datasets and in terms of causal inference on a toy dataset. The real data experiment shows that mIMR outperforms most of the existing approaches, supporting the considerations in (Guyon et al., 2007) about the presumed robustness of causal feature selection methods. The causal inference experiment shows that the mIMR strategy leads to a more accurate retrieval of direct causes with respect to other filters.

2. Mutual information and interaction

Let us consider three random[1] variables $\mathbf{x}_1$, $\mathbf{x}_2$ and $\mathbf{y}$ where $\mathbf{x}_1 \in \mathcal{X}_1, \mathbf{x}_2 \in \mathcal{X}_2$ are continuous and $\mathbf{y} \in \mathcal{Y} = \{y_1, \ldots, y_C\}$. The mutual information $I(\mathbf{x}_1; \mathbf{x}_2)$ (Cover & Thomas, 1990) measures the amount of stochastic dependence between $\mathbf{x}_1$ and $\mathbf{x}_2$ and is also called two-way interaction (Jakulin, 2005). Note that, if $\mathbf{x}_1$ and $\mathbf{x}_2$ are Gaussian distributed the following relation holds

$$I(\mathbf{x}_1; \mathbf{x}_2) = -\frac{1}{2}\log(1-\rho^2) \tag{1}$$

where ρ is the Pearson correlation coefficient.

Let us now consider the target $\mathbf{y}$, too. The *conditional mutual information* $I(\mathbf{x}_1; \mathbf{x}_2|\mathbf{y})$ (Cover & Thomas, 1990) between $\mathbf{x}_1$ and $\mathbf{x}_2$ once $\mathbf{y}$ is given is defined by

$$\int\int\int p(x_1, x_2, y) \log \frac{p(x_1, x_2|y)}{p(x_1|y)p(x_2|y)} dx_1 dx_2 dy.$$

The conditional mutual information is null iff $\mathbf{x}_1$ and $\mathbf{x}_2$ are conditionally independent given $\mathbf{y}$. The change of dependence betwen $\mathbf{x}_1$ and $\mathbf{x}_2$ due to the knowledge of $\mathbf{y}$ is measured by the three-way *interaction information* defined in (McGill, 1954) as

$$I(\mathbf{x}_1; \mathbf{x}_2; \mathbf{y}) = I(\mathbf{x}_1; \mathbf{y}) - I(\mathbf{x}_1; \mathbf{y}|\mathbf{x}_2). \tag{2}$$

[1] Boldface denotes random variables.

This measure quantifies the amount of mutual dependence that cannot be explained by bivariate interactions. When it is different from zero, we say that $\mathbf{x}_1$, $\mathbf{x}_2$ and $\mathbf{y}$ 3-interact. A non-zero interaction can be either negative, and in this case we say that there is a synergy or complementarity betwen the variables, or positive, and we say that there is redundancy. Because of the symmetry we have

$$I(\mathbf{x}_1;\mathbf{x}_2;\mathbf{y}) = I(\mathbf{x}_1;\mathbf{y}) - I(\mathbf{x}_1;\mathbf{y}|\mathbf{x}_2) = \\ = I(\mathbf{x}_2;\mathbf{y}) - I(\mathbf{x}_2;\mathbf{y}|\mathbf{x}_1) = I(\mathbf{x}_1;\mathbf{x}_2) - I(\mathbf{x}_1;\mathbf{x}_2|\mathbf{y}) \tag{3}$$

Since by (3) we derive $I(\mathbf{x}_1;\mathbf{y}|\mathbf{x}_2) = I(\mathbf{x}_1;\mathbf{y}) - I(\mathbf{x}_1;\mathbf{x}_2;\mathbf{y})$, it follows that by adding $I(\mathbf{x}_2;\mathbf{y})$ to both sides we obtain

$$I((\mathbf{x}_1,\mathbf{x}_2);\mathbf{y}) = I(\mathbf{x}_1;\mathbf{y}) + I(\mathbf{x}_2;\mathbf{y}) - I(\mathbf{x}_1;\mathbf{x}_2;\mathbf{y}) = \\ = I(\mathbf{x}_1;\mathbf{y}) + I(\mathbf{x}_2;\mathbf{y}) + I(\mathbf{x}_1;\mathbf{x}_2|\mathbf{y}) - I(\mathbf{x}_1;\mathbf{x}_2) \tag{4}$$

Note that the above relationships hold also when either $\mathbf{x}_1$ or $\mathbf{x}_2$ are vectorial random variables.

2.1. Interaction and optimal feature selection

Consider a multi-class classification problem (Duda & Hart, 1976) where $\mathbf{x} \in \mathbf{X} \subset \mathbb{R}^n$ is the n-variate input and $\mathbf{y} \in \mathcal{Y}$ is the target variable. Let $A = \{1,\ldots,n\}$ be the set of indices of the n inputs. Let us formulate the feature selection problem as the problem of finding the subset $\mathbf{X}^*$ of v variables such that

$$\mathbf{X}^* = \arg \max_{S \subset A:|S|=v} I(\mathbf{X}_S;\mathbf{y}) \tag{5}$$

In other terms for a given number v of variables the optimal feature set is the one which maximizes the information about the target. Note that this formulation of the feature selection problem, also known as Max-Dependency (Peng et al., 2005; Meyer et al., 2008), is classifier-independent.

If we want to carry out the maximization (5), both an estimation of I and a search strategy in the space of subsets of $\mathbf{X}$ are required. Section 2.3 will discuss the estimation issues. As far as the search is concerned, according to the Cover and Van Campenhout theorem (Devroye et al., 1996), to be assured of finding the optimal feature set of size v, all feature subsets should be assessed. Given the infeasibility of exhaustive approaches for large n, we will limit to consider here only forward selection search approaches. Forward selection starts with an empty set of variables and incrementally updates the solution by adding the variable that is expected to bring the best improvement (according to a given criterion). The hill-climbing search selects a subset of $v < n$ variables in v steps by exploring only $\sum_{i=1}^{v}(n-i+1)$ configurations. For this reason the forward approach is commonly adopted in filter approaches for classification problems with high dimensionality (Fleuret, 2004; Peng et al., 2005). If $v = 1$ the optimal set returned by (5) is composed of the most relevant variable, i.e. the one carrying the highest mutual information to $\mathbf{y}$. For $v > 1$, we need to provide an incremental solution to (5) in order to obtain, given a set of d variables, the $d+1$th feature which maximizes the increase of the dependency. We propose an incremental step based on the relation (4). Let $\mathbf{X}_S$ be the set of the $d < v$ variables selected in the first d steps. In a forward perspective, the optimal variable $\mathbf{x}^*_{d+1} \in \mathbf{X} - \mathbf{X}_S$ to be added to the set $\mathbf{X}_S$ is the one satisfying

$$\mathbf{x}^*_{d+1} = \arg \max_{\mathbf{x}_k \in \mathbf{X}-\mathbf{X}_S} I((\mathbf{X}_S;\mathbf{x}_k);\mathbf{y}) = \\ = \arg \max_{\mathbf{x}_k \in \mathbf{X}-\mathbf{X}_S} \Big[I(\mathbf{X}_S;\mathbf{y}) + I(\mathbf{x}_k;\mathbf{y}) - I(\mathbf{X}_S;\mathbf{x}_k;\mathbf{y})\Big] = \\ = \arg \max_{\mathbf{x}_k \in \mathbf{X}-\mathbf{X}_S} \Big[I(\mathbf{x}_k;\mathbf{y}) - I(\mathbf{X}_S;\mathbf{x}_k;\mathbf{y})\Big] \tag{6}$$

that is the variable minimizing the interaction $I(\mathbf{X}_S;\mathbf{x}_k;\mathbf{y})$ with $\mathbf{X}_S$ and maximizing the relevance $I(\mathbf{x}_k;\mathbf{y})$.

2.2. Interaction and causal patterns

This section will discuss how the interaction measure sheds light about the potential causal patterns existing between variables. We will limit to consider here patterns of three variables only since, for estimation and computational reasons, we will not consider interactions of degree higher than three. According to the definition (3), a negative interaction between $\mathbf{x}_1$ and $\mathbf{x}_2$ means that the knowledge of the value $\mathbf{y}$ increases the amount of dependence. This situation can occur in two cases: i) the *common effect* configuration (Figure 1a) (this is also known as the *explaining-away* effect and ii) the *spouse* configuration where $\mathbf{x}_1$ is the common descendant of $\mathbf{y}$ and $\mathbf{x}_2$ (Figure 1b).

On the contrary a positive interaction between $\mathbf{x}_1$ and $\mathbf{x}_2$ means that the knowledge of the value $\mathbf{y}$ decreases the amount of dependence. This situation can occur in four cases: i) the *common cause* configuration (Figure 1c) where two dependent effects $\mathbf{x}_1$ and $\mathbf{x}_2$ become independent once the value of the common cause $\mathbf{y}$ is known, ii)-iii) the brotherhood configurations where $\mathbf{y}$ is a brother of $\mathbf{x}_1$ ($\mathbf{x}_2$) (Figure 1d) and iv) the *causal chain* configuration (Figure 1e) where one of the variables (e.g. $\mathbf{x}_1$) is the cause and the other (e.g. $\mathbf{x}_2$) is the effect of $\mathbf{y}$. Note that positive interaction can also be considered as synonimous of redundancy.

It is interesting to remark that if we are interested to identify the set of direct causes of $\mathbf{y}$, the interaction term is able to disambiguate only partially the situation. If on one side, a positive three-way interaction term implies that at least one variable is not a direct cause (Figures 1cde), on the other side negative interaction could be induced by a spouse configuration (Figures 1b). Note that, it is well-known that, though spouses are not causal, they belong to the Markov blanket set and as such they are strongly relevant for performing an accurate prediction (Tsamardinos & Aliferis, 2003). This is confirmed by Equation (6) which shows that all variables having negative interaction with the selected set $\mathbf{X}_s$ (including spouses) bring non-redundant information about the target. How is it then possible to discriminate spouses from direct causes? A possible solution comes from the fact that spouses are independent of the target and as a consequence their mutual information with $\mathbf{y}$ is null. A filter algorithm which aims to perform a forward selection of direct causes should therefore implement the minimum Interaction Maximum Relevance strategy (6) but discard variables with a null relevance (even if their interaction is negative). This is the idea that will be implemented in our causal filter mIMR.

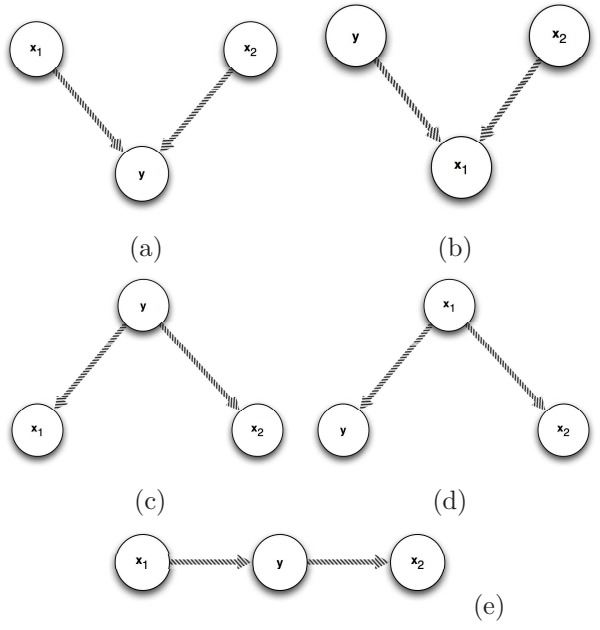

Figure 1. a) Common effect pattern, b) spouse pattern c) common cause pattern, d) brotherhood pattern and, e) causal chain pattern .

2.3. Estimation of the interaction

For a given set $\mathbf{X}_S$ of selected variables, the optimal forward selection step (6) requires the estimation of the interaction term $I(\mathbf{X}_S;\mathbf{x}_k;\mathbf{y}) = I(\mathbf{X}_S;\mathbf{x}_k) - I(\mathbf{X}_S;\mathbf{x}_k|\mathbf{y})$. The high feature-to-sample ratio nature of the microarray datasets demands a specific approach to the estimation of this quantity. A large amount of litterature on microarray classification converges on the consideration that only simple, constrained and low-variate estimation techniques are robust enough to cope with the noise and the high dimensionality of the problem (Dougherty, 2001). This is confirmed by the success of univariate ranking approaches which are widely used in explorative genomic studies despite of their evident limitations(Saeys et al., 2007). Although the term $I(\mathbf{X}_S;\mathbf{x}_k;\mathbf{y})$ is multivariate what we propose here is a biased but robust estimator based only on bivariate terms. In order to design such estimator we can take advantage of the following simplifications. Since $\mathbf{y}$ is a class and takes value in a finite set of C values $\{y_1,\ldots,y_C\}$, we can write the term $I(\mathbf{X}_S;\mathbf{x}_k|\mathbf{y})$ as follows:

$$\int\int \log\frac{p(X_S,x_k|y)}{p(X_S|y)p(x_k|y)}p(X_S,x_k,y)dX_S dx_k dy = \\ = \sum_{c=1}^{C} I(\mathbf{X}_S;\mathbf{x}_k|\mathbf{y}=y_c)\mathrm{Prob}\{\mathbf{y}=y_c\} \quad (7)$$

where $\mathrm{Prob}\{\mathbf{y}=y_c\}$ is the a-priori probability of the cth class. After this transformation it appears that the estimation of the interaction term requires the estimation of $C+1$ terms: the mutual information $I(\mathbf{X}_S;\mathbf{x}_k)$ and the C conditional mutual information terms $I(\mathbf{X}_S;\mathbf{x}_k|\mathbf{y}=y_c)$, $c=1,\ldots,C$ obtained by controlling the value of the target. In practical terms this boils down to estimate the terms $I(\mathbf{X}_S;\mathbf{x}_k|\mathbf{y}=y_c)$ by considering only the portion of the training set for which the target $\mathbf{y}=y_c$. Unfortunately each of these terms is highly dimensional because of the term $\mathbf{X}_S$. Most of the information theoretic filters proposed in litterature so far proposed several ways to decompose multivariate mutual information terms in a linear combination of low-variate mutual information terms. We recall here the approximations underlying some of the most effective state-of-the-art filter approaches:

Max-Relevance (Peng et al., 2005): it uses

$$I(\mathbf{X}_S;\mathbf{y}) \approx \frac{1}{d}\sum_{\mathbf{x}_i\in\mathbf{X}_S} I(\mathbf{x}_i;\mathbf{y})$$

where d is the size of the set $\mathbf{X}_S$.

Conditional Mutual Information Maximisation (CMIM) (Fleuret, 2004): it makes the approximation

$$I(\mathbf{x}_k;\mathbf{y}|\mathbf{X}_S) \approx \min_{\mathbf{x}_i \in X_S} I(\mathbf{x}_k;\mathbf{y}|\mathbf{x}_i)$$

Minimum Redundancy Maximum Relevance (mRMR) (Peng et al., 2005): it approximates the total interaction term (i.e. the interaction term when all the components of $\mathbf{X}_S$ are independent) (Watanabe, 1960):

$$J(\mathbf{X}_S;\mathbf{x}_k) \approx \frac{1}{d}\sum_{\mathbf{x}_i \in \mathbf{X}_S} I(\mathbf{x}_i;\mathbf{x}_k)$$

Double Input Symmetrical Relevance (DISR) (Meyer et al., 2008): it adopts the approximation

$$I((\mathbf{X}_S,\mathbf{x}_k);\mathbf{y}) \approx \frac{1}{d}\sum_{\mathbf{x}_i \in \mathbf{X}_S} I((\mathbf{x}_i,\mathbf{x}_k);\mathbf{y})$$

Similarly to what is done in DISR, we adopt a fast approximation of $I(\mathbf{X}_S;\mathbf{x}_k)$ which consists in replacing the multivariate term by a linear combination of bivariate terms. The two terms of interaction are then approximated as follows

$$I(\mathbf{X}_S;\mathbf{x}_k) \approx \frac{1}{d}\sum_{\mathbf{x}_i \in \mathbf{X}_S} I(\mathbf{x}_i;\mathbf{x}_k) \quad (8)$$

$$I(\mathbf{X}_S;\mathbf{x}_k|\mathbf{y}=y_c) \approx \frac{1}{d}\sum_{\mathbf{x}_i \in \mathbf{X}_S} I(\mathbf{x}_i;\mathbf{x}_k|\mathbf{y}=y_c) \quad (9)$$

This means that we restrict to consider only the interaction terms involving both the candidate feature and at most one of the previously selected features.

2.4. The min-Interaction Max-Relevancy (mIMR) filter algorithm

Let $\mathbf{X}_+ = \{\mathbf{x}_i \in \mathbf{X} : I(\mathbf{x}_i;\mathbf{y}) > 0\}$ the subset of $\mathbf{X}$ containing all variables having non null mutual information (i.e. non null relevance) with $\mathbf{y}$. Once the approximations (8) and (9) are done, the forward step (6) can be written as follows

$$\begin{aligned}\mathbf{x}^*_{d+1} &= \arg\max_{\mathbf{x}_k \in \mathbf{X}_+ - \mathbf{X}_S} [I(\mathbf{x}_k;\mathbf{y}) - I(\mathbf{X}_S;\mathbf{x}_k;\mathbf{y})] = \\ &= \arg\max_{\mathbf{x}_k \in \mathbf{X}_+ - \mathbf{X}_S} \left[I(\mathbf{x}_k;\mathbf{y}) - \frac{1}{d}\sum_{\mathbf{x}_i \in \mathbf{X}_S} (I(\mathbf{x}_i;\mathbf{x}_k;\mathbf{y})\right] = \\ &= \arg\max_{\mathbf{x}_k \in \mathbf{X}_+ - \mathbf{X}_S} \Big[I(\mathbf{x}_k;\mathbf{y}) + \\ &+ \frac{1}{d}\sum_{\mathbf{x}_i \in \mathbf{X}_S}\sum_{c=1}^{C} p_c(I(\mathbf{x}_i;\mathbf{x}_k|\mathbf{y}=y_c) - I(\mathbf{x}_i;\mathbf{x}_k))\Big] \quad (10)\end{aligned}$$

where $p_c = \text{Prob}\{\mathbf{y} = y_c\}$.

The resulting algorithm, called mIMR (minimum Interaction Maximum Relevance), relies on four main ideas: i) avoid to select spouses by restricting the selection to the subset of relevant variables $\mathbf{X}_+$, ii) select incrementally, among the set of variables $\mathbf{X}_+$ the ones which minimise the interaction (mI) and maximise the relevancy (MR), ii) simplify the computation of the interaction term by limiting to 3-way interactions and, iv) speed up the three-way interaction computation by conditioning the interaction compution on the value of the target (i.e. restraining the training set to the set of observations with the same output class). Note that in causal terms the mIMR algorithm tends to select features which are both relevant and on average take part to common effect patterns (Figure 1a) with previously selected variables. In order to initialise the mIMR algorithm (10) with a pair of direct causes, we put

$$\mathbf{x}^*_1, \mathbf{x}^*_2 = \arg\max_{\mathbf{x}_i, \mathbf{x}_k \in \mathbf{X}_+} I((\mathbf{x}_i,\mathbf{x}_k);\mathbf{y}).$$

Table 1 summarizes a set of causal patterns, involving the candidate feature $\mathbf{x}_k \in \mathbf{X}_+$, the selected feature $\mathbf{x}_i \in \mathbf{X}_S$ and the target $\mathbf{y}$ and the related values of relevancy and interaction. Since mIMR maximizes relevance and minimize interaction it appears that such algorithm prioritises the selection of causal variables $\mathbf{x}_k$. The only case when interaction is negative but $\mathbf{x}_k$ is not causal is discarded a-priori since $I(\mathbf{x}_k;\mathbf{y}) = 0 \Rightarrow \mathbf{x}_k \notin \mathbf{X}_+$. According to Table 1 it is then reasonable to expect that the mIMR algorithm proceeds by prioritizing the direct causes, penalizing the effects (because of their positive interaction) and neglecting the spouses.

Let us now discuss the relation between the mIMR filter and two related state-of-the-art approaches: DISR and mRMR.

mIMR vs. DISR: the incremental step of the DISR algorithm is

$$\mathbf{x}^*_{d+1} = \arg\max_{\mathbf{x}_k \in \mathbf{X} - \mathbf{X}_S} \sum_{\mathbf{x}_i \in \mathbf{X}_S} I(\mathbf{x}_k,\mathbf{x}_i;\mathbf{y}) \quad (11)$$

In analytical terms, because of (4), if $\mathbf{X}_+$ coincides with $\mathbf{X}$ the solution of (11) coincides with (10). What makes a difference between mIMR and DISR is that mIMR i) removes variables with no mutual information with the target and ii) decomposes the mIMR criterion into the sum of a relevance and interaction terms. In the Gaussian case (1), the mIMR algorithm requires the computation of the input-output correlation vector and $C+1$ input correlation matrices.

Table 1. Causal patterns, relevance and interactions terms.

CAUSAL PATTERN	$I(\mathbf{x}_k;\mathbf{y})$	$I(\mathbf{x}_i;\mathbf{x}_k;\mathbf{y})$
$\mathbf{x}_i \rightarrow \mathbf{y} \leftarrow \mathbf{x}_k$	>0	<0
$\mathbf{x}_i \rightarrow \mathbf{y} \rightarrow \mathbf{x}_k$	>0	>0
$\mathbf{x}_k \leftarrow \mathbf{x}_i \rightarrow \mathbf{y}$	>0	>0
$\mathbf{x}_k \rightarrow \mathbf{x}_i \rightarrow \mathbf{y}$	>0	>0
$\mathbf{x}_i \leftarrow \mathbf{y} \rightarrow \mathbf{x}_k$	>0	>0
$\mathbf{x}_k \rightarrow \mathbf{x}_i \leftarrow \mathbf{y}$	0	<0

mIMR vs. mRMR: the incremental step of the mRMR algorithm is

$$\mathbf{x}^*_{d+1} = \arg \max_{\mathbf{x}_k \in \mathbf{X}-\mathbf{X}_S} \left[I(\mathbf{x}_k;\mathbf{y}) - \frac{1}{d} \sum_{\mathbf{x}_i \in \mathbf{X}_S} I(\mathbf{x}_k;\mathbf{x}_i)) \right] \tag{12}$$

The mIMR algorithm differentiates from the mRMR algorithh since it replaces the redundancy term based on 2-way interaction with a causal term based on 3-way interaction. The experimental setting will show that this could be important in some situations.

3. Experiments

The experiment uses 25 public domain microarray expression (Table 2[2]) to compare the performance of the mIMR approach with 4 state-of-the-art filters: DISR, mRMR, CMIM, MB and RANK. For all these filters the bivariate mutual information terms are computed by making the Gaussian assumption (1). Note that this assumption is simplistic but has the merit of returning a low-variance estimation of the mutual information and of making possible the creation of $\mathbf{X}_+$ by a statistical test on the correlation. The implementation of the MB filter is based on the sequential version proposed in (Xing et al., 2001) and uses a linear SVM classifier to assess the redundancy term Δ. A first dimensionality reduction step is carried out by hierarchically clustering the variables into 1000 compressed features obtained by averaging the probes of the same cluster (Park et al., 2007). Table 3 reports the average balanced cross-validated misclassification error of a set of classifiers $\mathcal{C}$ composed of a linear support vector machine, a random forest and a three KNN with different number of neighbours ($K = 1, 3, 5$). The use of more than a single classifier is expected to reduce the bias of a feature assessment based on a specific classification strategy. An external cross-validation scheme (Ambroise & McLachlan, 2002) is used to prevent feature selection bias in our assessment. For each step of the 10-fold cross-validation, for each selection approach and for each classifier, once selected features are returned, the generalization accuracy is assessed by (i) training the classifier on the same dataset used for feature selection and (ii) testing the trained classifier on the remaining tenth. Note that because of the scarcity of the data and to avoid the bias related to the selection of the feature set size, we average the performance over all the classifiers and over all the feature sets whose size is ranging from $d = 1$ to $d = 20$. Table 3 reports the multi-class balanced error measure proposed in (Melvin et al., 2007) in order to account for the unbalanced nature of the samples. An error misclassification percentage takes the bold notation when it is significantly different (Benjamini-Hochberg adjusted p-value < 0.05) from the accuracy of the mIMR strategy. Together with accuracy, another important issue in the use of feature selection techniques for microarray data is the stability of the resulting selected set. In order to assess this property, we consider how the selected set varies during the different cross-validation steps. Table 4 reports a measure of the stability of the different selection procedures obtained by averaging over all possible pairs of cross-validation folds the percentage size of the intersection of the selected features. It follows that the higher this quantity, the higher is the stability of the feature algorithm

[2]For reasons of limited space the complete reference list of the datasets is available in a supplementary file that can be accessed in `http://www.ulb.ac.be/di/map/gbonte/Papers.html`

Table 2. Microarray data sets: name of the dataset, number N of samples, number n of features and number C of classes.

#	Name	N	n	C
1	Gordon	240	12533	2
2	Golub	72	7129	2
3	Alon	62	2000	2
4	Notterman	36	7457	2
5	Nutt	50	12625	2
6	Shipp	77	7129	2
7	Singh	102	12600	2
8	Sorlie	76	7937	2
9	Wang	286	22283	2
10	Van't Veer	65	24481	2
11	VandeVijver	295	24496	2
12	Sotiriou	99	7650	2
13	Pomeroy	60	7129	2
14	LYM	47	4026	2
15	Beer	96	7129	2
16	Petricoin	96	7129	2
17	Khan		83 2308	4
18	Novartis	103	1000	4
19	West	49	7129	4
20	Staunton	60	7129	9
21	Su	174	12533	11
22	Bhattacharjee	203	12600	5
23	Armstrong	72	12582	3
24	Ma	60	22575	3
25	Hedenfalk	22	3226	3

Table 3. Balanced misclassification error (ten-fold cross-validation) averaged over the classifiers of the family $\mathcal{C}$. The bold notation stands for significantly different from the accuracy of the mIMR strategy in terms of an adjusted (BH criterion) p-value (pv < 0.05). The Avg line returns the average of the balanced misclassification percentages. The W/B line returns the number of times that the technique is worse/better than mIMR.

Data	mIMR	DISR	mRMR	CMIM	MB	RANK
1	2.44	**2.71**	2.3	**2.04**	**5.3**	**2.94**
2	5.73	6.15	5.77	**6.48**	**10.76**	5.87
3	18.34	**19.4**	**19.98**	**21.59**	**22.99**	**20.55**
4	6.78	6.63	**7.14**	**7.59**	**11.39**	**7.78**
5	30.53	**23.11**	30.33	30.2	**27.03**	**24.94**
6	17.63	**16.21**	17.88	16.93	**24.08**	17.76
7	11.51	**10.99**	11.34	11.27	**15.35**	**12.59**
8	35.68	**37.04**	**36.43**	**38.75**	**40.74**	**40.67**
9	44.21	**44.89**	**45.91**	**46.32**	**45.16**	**45.66**
10	29.21	28.85	**30.98**	**32.75**	29.33	**27.66**
11	44.69	44.52	45.28	44.24	**46.54**	44.83
12	49.8	**51.44**	49.68	50.67	**51.6**	**53.37**
13	49.12	49.95	49.82	47.99	**55.78**	**50.64**
14	6.65	**5.7**	**6.01**	7.21	**24**	6.13
15	3.75	**2.88**	3.7	**7.4**	**1.3**	**2.75**
16	0.36	**0.5**	**0.58**	**0.69**	**4.07**	**2.84**
17	7.95	8.49	**7.52**	7.67	**12.45**	**8.68**
18	8.47	**16.32**	**8.03**	**11.71**	**35.55**	**36.98**
19	57.95	**54.13**	**59.35**	**60.48**	58.59	**54.77**
20	67.08	**65.32**	67.12	**70.08**	**77.56**	**65.44**
21	34.31	**40.11**	**34.78**	**34.98**	**48.51**	**42.95**
22	16.33	**18.54**	**15.04**	**14.37**	**30.02**	**24.78**
23	6.32	6.19	6.17	**7.43**	**27.62**	**8.59**
24	66.14	**67.11**	**66.86**	66.22	**67.15**	66.75
25	41.65	**45.36**	**38.76**	**37.65**	**60.93**	**38.43**
Avg	25.49	**26.27**	**25.77**	**26.18**	**31.62**	**28.43**
W/B		11/7	9/5	13/3	19/1	14/6

The classification results show that mIMR outperforms the state-of-the-art informaton-theoretic filters. The comparison with DISR shows that in several circumstances the removal of spouses may have a beneficial effect in terms of accuracy. A possible interpretation is that datasets where mIMR is significantly better than DISR are datasets where the predictive role of direct causes is strong. The comparison with mRMR quantifies the accuracy improvement due to the introduction of the interaction term. It is interesting also to remark that the mIMR accuracy improvement with respect to mRMR, CMIM and MB is obtained by increasing at the same time the stability of the selected features: as shown in Table 4 the mIMR filter selection is 14 (resp. 20 and 21) times more stable than the one of mRMR (resp. CMIM and MB). A final consideration raises from the fact that RANK, though definitely less accurate than mIMR, outperforms the mIMR filter in stability terms. This result should be interpreted by considering that good accuracy derives from a suitable trade-off between stability (low variance) and relevancy (low bias). The incorporation of causal terms appears to be advantageous in terms of the overall trade-off, therefore suggesting that the mIMR gain in terms of lower bias dominates the loss due to the increased variability of the selection.

Table 4. Stability measure in terms of percentage size of the intersection. The Avg line returns the average of the intersection percentage sizes. The L/H line returns the number of times that the stability is lower/higher than mIMR.

Name	mIMR	DISR	mRMR	CMIM	MB	RANK
1	62.2	59.9	61.8	47.0	25.8	65.2
2	43.6	46.1	43.6	35.1	18.8	51.5
3	63.6	68.7	67.8	56.8	27.4	65.5
4	6.0	6.5	6.0	4.3	3.8	6.9
5	28.1	28.3	28.3	25.6	13.4	31.5
6	40.1	38.2	42.2	40.7	16.5	39.1
7	68.0	67.3	69.2	62.7	28.5	67.9
8	27.4	27.3	26.4	24.0	15.6	29.0
9	17.9	18.5	16.5	15.5	32.9	21.7
10	17.2	22.8	16.4	14.7	15.7	24.9
11	24.9	29.8	23.4	21.0	11.0	32.5
12	16.3	15.8	15.7	16.8	12.7	19.2
13	16.8	15.2	16.8	15.9	10.6	13.9
14	61.5	63.4	61.4	52.8	19.3	65.4
15	71.9	72.5	73.3	46.7	28.6	75.2
16	7.3	10.9	7.4	4.6	3.2	11.1
17	58.5	64.4	58.0	52.4	28.9	61.8
18	61.8	74.9	65.4	51.0	52.8	88.3
19	25.6	35.8	26.1	27.3	14.9	46.4
20	28.3	49.4	28.6	27.9	47.5	24.5
21	50.1	62.7	48.6	50.5	39.9	40.0
22	54.3	70.9	51.1	46.7	30.5	60.5
23	39.2	40.4	38.2	32.7	42.8	41.4
24	16.6	17.5	16.2	14.6	9.8	14.2
25	4.7	5.9	4.9	4.7	6.7	6.3
Avg	36.5	40.5	36.5	31.7	22.3	40.2
L/H		6/19	14/11	20/5	21/4	6/19

Table 5. Average (over 200 runs) position of the two direct classes ($\mathbf{x}_1$ and $\mathbf{x}_5$ in Figure 2) in the rankings returned by the filters. N is the number of samples

N	mIMR	mRMR	CMIM	RANK
50	3.92	4.34	3.96	53.67
100	3.36	3.93	3.32	3.27
200	2.31	2.64	2.46	2.60
500	2.00	2.18	2.13	2.42
1500	1.95	2.16	2.19	2.42
2000	1.93	2.06	2.08	2.38

The second experiment assesses the capacity of the mIMR filter of prioritizing the selection of the direct causes in a toy example inspired to the LUCAS dataset[3]. In order to reuse the continuous Gaussian estimator used in the previous experiment, we generated a set of artificial datasets made of 11 continuous inputs and 1 class target $\mathbf{y}$ where all the dependencies are linear (Figure 2). We compared the ranking of the two direct causes $\mathbf{x}_1$ and $\mathbf{x}_5$ returned by the mIMR, mRMR, CMIM and RANK for 6 different sizes of datasets. Table 5 reports the average (over 200 runs) position of the two direct causes in the ranking returned by the considered filters. Note that the lowest values obtained with mIMR show that this filter is the most effective one in ranking the two direct causes in high position.

[3] `http://www.causality.inf.ethz.ch/data/LUCAS.html`

4. Conclusions

The bioinformatics community is demanding of learning algorithms able to detect in a fast and reliable manner subsets of informative and causal features. An open issue is to understand whether the strive for causal patterns is in contradiction with the objective of maximising the generalisation accuracy. This paper shows that the two objectives are synergetic and that their link is represented by the notion of information interaction. By taking into account this term it is possible to address both accuracy and causal explanation.

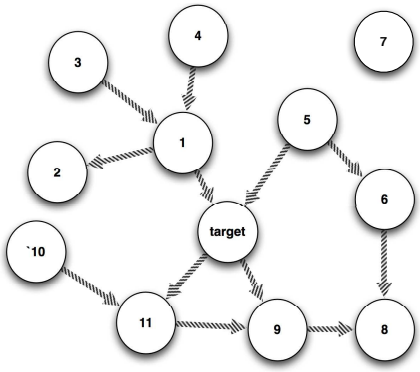

Figure 2. Causal network

References

Ambroise, C. and McLachlan, G.J. Selection bias in gene extraction on the basis of microarray gene-expression data. *PNAS*, 99:6562–6566, 2002.

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

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

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

Devroye, L., Györfi, L., and Lugosi, G. *A Probabilistic Theory of Pattern Recognition*. Springer Verlag, 1996.

Dougherty, E.R. Small sample issues for microarray-based classification. *Comp. Funct. Genomics*, 2:28–34, 2001.

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

Duda, R. O. and Hart, P. E. *Pattern Classification and Scene Analysis*. Wiley, 1976.

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

Freitas, A. A. Understanding the crucial role of attribute interaction in data mining. *Artficial Intelligence Review*, 6:177–199, 2001.

Guyon, I., Aliferis, C., and Elisseeff, A. *Computational Methods of Feature Selection*, chapter Causal Feature Selection, pp. 63–86. Chapman and Hall, 2007.

Guyon, Isabelle, Gunn, Steve, Nikravesh, Masoud, and Zadeh, Lotfi A. *Feature Extraction: Foundations and Applications*. Springer-Verlag New York, Inc., 2006.

Jakulin, A. *Machine Learning Based on Attribute Interactions*. PhD thesis, University of Ljubliana, Faculty of Computer and Information Science, 2005.

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

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

Melvin, I., Ie, E., Weston, J., Noble, W. S., and Leslie, C. Multi-class protein classification using adaptive codes. *Journal of Machine Learning Research*, 8:1557–1581, 2007.

Meyer, P.E., Schretter, C., and Bontempi, G. Information-theoretic feature selection in microarray data using variable complementarity. *IEEE Journal of Selected Topics in Signal Processing*, 2:261–274, 2008.

Park, M., Hastie, T., and Tibshirani, R. Averaged gene expression for regression. *Biostatistics*, 8(2):212–227, 2007.

Peng, H., Long, F., and Ding, C. 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.

Saeys, Y., Inza, I., and Larranaga, P. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23:2507–2517, 2007.

Spirtes, P., Glymour, C., and Scheines, R. *Causation, Prediction and Search*. Springer Verlag, Berlin, 1993.

Tsamardinos, I. and Aliferis, C. Towards principled feature selection: Relevancy. In *Proceedings of the 9th International Workshop on Artificial Intelligence and Statistics*, 2003.

Watanabe, S. Information theoretical analysis of multivariate correlation. *IBM Journal of Research and Development*, 4:66–82, 1960.

Xing, E. P., Jordan, M. I., and Karp, R. M. Feature selection for high-dimensional genomic microarray data. In *Proceedings of the 18th International Conf. on Machine Learning*, pp. 601–608. Morgan Kaufmann, San Francisco, CA, 2001.

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