Journal of Machine Learning Research Submitted 12/10;

A Simple Filter Benchmark for Feature Selection

Athanasios Tsanas TSANAS@MATHS.OX.AC.UK, TSANASTHANASIS@GMAIL.COM
Oxford Centre for Industrial and Applied Mathematics
Mathematical Institute, University of Oxford
24-29 St. Giles', Oxford, OX1 3LB, UK

Max A. Little LITTLEM@PHYSICS.OX.AC.UK
Oxford Centre for Integrative Systems Biology
Department of Physics, University of Oxford
Parks Road, Oxford, OX1 3PU, UK

Patrick E. McSharry PATRICK@MCSHARRY.NET
Smith School of Enterprise and the Environment
University of Oxford, 75 George Street
Oxford, OX1 2BQ, UK

Editor:

Abstract

A new correlation-based filter approach for simple, fast, and effective feature selection (FS) is proposed. The association strength between each feature and the response variable (relevance) and between pairs of features (redundancy) is quantified via a simple nonlinear transformation of correlation coefficients inspired by information theoretic concepts. Furthermore, the association strength between a set of features and the response variable (feature complementarity) is explicitly addressed using a similar nonlinear transformation of partial correlation coefficients, where a feature is selected conditionally upon its additional information content when combined with the features already selected in the forward sequential process. The new filter scheme overcomes several major issues associated with competing FS algorithms, including computational complexity and difficulty in implementation, and can be used on both multi-class classification and regression problems. Experiments on five synthetic and twelve real datasets demonstrate that the proposed filter outperforms popular alternative filter approaches in terms of recovering the correct features. We envisage the proposed scheme setting a competitive benchmark against which more sophisticated FS algorithms can be compared. Documented Matlab source code is available on the first author's website.

Keywords: Benchmark, curse of dimensionality, feature selection, nonlinear transformation

1 Introduction

Data analysis is a ubiquitous problem in various disciplines, ranging from engineering and medical research to the social sciences. Typically, the researcher is faced with the problem of inferring a relationship between a set of *features* (characteristics of the examined dataset), and a measured quantity of interest known as the *response*; this is commonly referred to as the *supervised learning* setup (Hastie et al., 2009). However, the presence of a large number of features sometimes obstructs the interpretation of useful patterns in the data, and is often detrimental to the subsequent learning process of mapping the features to the response (Guyon and Eliseef, 2003; Liu and Yu, 2005; Gyon et al., 2006). This problem, widely known as the *curse of dimensionality* (Hastie et al., 2009), occurs because it is practically impossible to adequately populate the feature space with the available data (the number of required samples grows exponentially with the number of features). The problem is worse where the number of features is larger than the number of samples, for example, in microarray data analysis problems (Hastie et al., 2009).

To mitigate the curse of dimensionality, researchers often resort to either *feature transformation* or *feature selection*. Feature transformation aims to build a new feature space of reduced dimensionality, producing a compact representation of the information that may be distributed across several of the original features. Although it has shown promising results in many applications (Torkkola, 2003; Hastie et al., 2009), feature transformation is not easily interpretable because the physical meaning of the original features cannot be retrieved. In addition, it does not save on resources required during the data collection process since *all* the original features still need to be measured or computed. Moreover, in very high dimensional settings where the number of irrelevant features may exceed the number of relevant features, reliable feature transformation can be problematic (Torkkola 2003).

Feature selection (FS) aims to decide on a *feature subset*, discarding features that do not contribute towards predicting the response. FS algorithms can be broadly categorized into *wrappers* and *filters*. Wrappers incorporate the *learner* (classifier or regressor) in the process of selecting the feature subset, and may improve the overall machine learning algorithm performance (Tuv et al., 2009; Torkkola, 2003). However, there are at least four major issues with wrappers: a) increased *computational complexity* (compared to filters), which is exacerbated as the dataset grows larger, b) the selected feature subset for a specific learner may be suboptimal for a different learner, a problem known as *feature exportability* (that is, the selected feature subset is not exportable), c) *controlling internal parameters* (parameter fine-tuning) of the learner requires experimentation, expertise, and is time-consuming, and d) inherent *learner constraints*, for example some learners do not handle multi-class classification or regression problems. The problem with feature exportability arises because the features chosen in a wrapper scheme are tailored to optimize the performance of the *specific* learner irrespective of the general characteristics of the data. Hence, the selected feature subset may not reflect the global properties of the original dataset, which leads to the failure of generalization of wrapper-selected feature subsets in *alternative* learners (Hilario and Kalousis, 2008). Filters attempt to overcome these limitations of the wrapper methods and commonly evaluate feature subsets based on their information content (for example using statistical tests) instead of optimizing the performance of specific learners, and are computationally more efficient than wrappers. In the remainder of this study, FS is used to refer *exclusively* to filters.

There has been extensive research on filter schemes, often these schemes are demanding both in terms of computational effort and memory, whilst some require tuning of internal parameters to optimize performance. Moreover, some filters are limited in their application since they can only address binary classification problems, or cannot be generalized to regression settings. In addition, recent studies highlight the importance of using simple filters before experimenting with more sophisticated schemes, remarking that many promising but elementary concepts have been left unexplored (Guyon et al., 2007; Guyon, 2008; Brown, 2009).

In this study, we introduce a simple yet efficient filter, which we call *relevance*, *redundancy* and *complementarity trade-off* (RRCT). This correlation-based filter uses a simple nonlinear transformation of the correlation coefficients using *information theoretic concepts* to quantify the association of the features with the response, and the overlapping information between features. In addition, the RRCT explicitly takes into account the *additional* information content which is contributed by a new feature, conditional on the existing feature subset. We demonstrate that by generalizing point estimates of shared information content (quantified via correlation coefficients) and by accounting for multi-variable complementarity we improve the accuracy over classical filter schemes. The proposed algorithm is effective in multi-class classification and regression settings, is very fast, and does not require fine-tuning of parameters.

This paper is organized as follows: Section 2 describes the most widely used FS concepts. Section 3 reviews a selection of existing FS algorithms that will form the basis of the comparisons provided, and introduces the new FS scheme presented in this study. Section 4 shows experimental results using a range of synthetic and real datasets. Finally, Section 5 summarizes the results of this study, and outlines the proposed FS scheme's properties, strengths and limitations compared to the established filter schemes.

2 Terminology and main concepts of filter approaches

Feature selection algorithms abound in the literature (Guyon and Elisseeff, 2003; Guyon et al., 2006; Tuv et al., 2009; Sun, Todorovic and Goodison, 2010, Guyon et al., 2010). This section aims to review some of the most important concepts and algorithms, and motivate the need for the development of the proposed schemes. This review is necessarily brief, and we refer to Guyon et al. (2006; Guyon and Elisseeff 2003) as good starting points on the topic of feature selection. In the following, the terms *features*, *input variables*, *explanatory variables*, and *predictors* coincide and are used interchangeably throughout the text. Similarly, the terms *response*, *response variable* and *target* all refer to the outcome quantity of interest.

Given the input data matrix $\mathbf{X} \epsilon \mathbb{R}^{N\times M}$ and the response variable $\boldsymbol{y} \epsilon \mathbb{R}^{N\times 1}$ where N is the number of samples (instances) and M is the number of features, the FS algorithms aim to reduce the input feature space M into m features, where $m < M$ (m can be chosen based on prior knowledge and possible constraints of the application, or can be determined via cross validation). That is, we want to select a feature set S comprising m features $\{\boldsymbol{x}_i\}$ $i \epsilon (1 \dots M)$, where each $\boldsymbol{x}_i$ is a column vector in the data matrix $\mathbf{X}$. The optimal feature subset maximizes the *combined* information content of all features in the feature subset with respect to the response variable. However, this is a complex combinatorial problem, and the optimal solution can only be found by a brute force search. Since a brute force search is extremely computationally demanding, particularly for large datasets, sub-optimal alternatives must be sought. Although in principle combinatorial optimization methods (such as simulated annealing and genetic algorithms) can be applied to the FS problem, these techniques are also computationally expensive.

As an approximate solution to the combinatorial one, researchers often assess each feature *individually* in order to determine the overall information content of the feature subset from each individual feature in the subset. Then, there are two FS approaches: a) *sequential forward* process (features are sequentially added to the selected feature subset), and b) *backward elimination* (starting from the entire feature set and eliminating one feature at each step). Forward FS is often used in many filter applications (Peng, Long and Ding, 2005; Sun, Todorovic and Goodison, 2010), and is particularly suitable for those problems where we want to reduce a dataset comprising many features to a dataset with a fairly small number of features.

One of the simplest FS approaches is to use the features which are maximally related to the response, where the association strength of the features with the response can be quantified using a suitable *criterion* (or *metric*, not to be confused with a distance metric in the mathematical sense) $I(\cdot)$. One of the straightforward metrics is the Pearson correlation coefficient, which expresses the *linear* relationship between each feature $\boldsymbol{x}_i$ and $\boldsymbol{y}$. This assumes that the association strength between the response and each of the features can be characterized using the mean and standard deviations (first two statistical moments) alone, and that the higher order moments are zero, or at least sufficiently small that they can be neglected. Alternatively, the Spearman rank correlation coefficient, which is a more general *monotonic* metric, can be used to quantify the relationship between each feature and the response. More complicated criteria can also be used to characterize potentially *nonlinear* (and *non-monotonic*) relationships between the features and the response variable. One of the most important metrics to quantify the association strength between two random variables is the *mutual information* (MI), because it can be used to express arbitrary dependencies between the two quantities (Cover and Thomas, 2006). In fact, MI has attracted extensive and systematic interest in the feature selection literature (Battiti, 1994; Peng, Long and Ding, 2005; Meyer, Schretter and Bontempi, 2008; Estevez et al., 2009). However, the computation of MI is not trivial (particularly in domains with continuous variables), which hinders its widespread use (Torkkola, 2003).

Conceptually, the simple approach discussed thus far that relies solely on the association strength between individual features and the response variable works well in the presence of *independent* (orthogonal) features (no correlations amongst features). It is now well established that in most practical applications a good feature subset needs to account for *overlapping information* amongst features for predicting the response variable. That is, the *relevance* (association strength of a feature with the response variable) needs to be counter-weighted with *redundancy* (overlapping information amongst features in the feature subset towards predicting the response) (Battiti, 1994; Yu and Liu, 2004; Guyon et al. 2006). Battiti (1994) proposed a compromising setup between relevance and redundancy:

$$\mathrm{FS}_{\mathrm{Battiti}}(\beta) = \max_{i\,\in\,Q-S}\left[\underbrace{I(\boldsymbol{x}_i;\boldsymbol{y})}_{relevance} - \beta\underbrace{\sum_{s\,\in\,S} I(\boldsymbol{x}_i;\boldsymbol{x}_s)}_{\text{redundancy}}\right] \tag{1}$$

where $\boldsymbol{x}_i$ denotes the i^{th} variable in the initial M-dimensional feature space, $\boldsymbol{x}_s$ is a variable that has been already selected in the feature index subset S (s is an integer, Q contains the indices of all the features in the initial feature space, that is $1\ldots M$, S contains the indices of selected features and $Q-S$ denotes the indices of the features not in the selected subset), $\boldsymbol{y}$ is the response, β is a parameter to compromise between the relevance term and the redundancy term, and $I(\cdot)$ is the metric used to quantify the relevance or redundancy. Battiti's (1994) algorithm is a heuristic incremental (*greedy*) search solution, which consists of the following steps: 1) (Selecting the first feature index) include the feature index i: $\max_{i\,\in\,Q}\big(I(\boldsymbol{x}_i;\boldsymbol{y})\big)$ in the initially empty set S, that is $\{i\}\rightarrow S$, 2) (Selecting the next $m-1$ features, one at each step, by repeating the following) apply the criterion in Equation 1 to incrementally select the next feature index i, and include it in the set: $S\cup\{i\}\rightarrow S$, 3) obtain the feature subset by selecting the features $\{\boldsymbol{x}_i\}_{i=1}^{m}$, $i \in S$ from the original data matrix $\mathbf{X}$.

A major problem with the approach formalized by Equation 1 is that it requires the specification of the free parameter β (which can be achieved using grid search and cross validation). Moreover, the optimal value of β may vary with the size of the feature subset. Peng, Long and Ding (2005) modified the criterion in Equation 1 to avoid the fine tuning of the free parameter, proposing the *minimum redundancy maximum relevance* (mRMR) (see Equation 2):

$$\mathrm{mRMR} \overset{\text{def}}{=} \max_{i\,\in\,Q-S}\left[I(\boldsymbol{x}_i;\boldsymbol{y}) - \frac{1}{|S|}\sum_{s\,\in\,S} I(\boldsymbol{x}_i;\boldsymbol{x}_s)\right] \tag{2}$$

where $|S|$ is the cardinality of the selected subset. As in Battiti's (1994) study, Peng, Long and Ding (2005) used MI to express the relevance and redundancy, and the greedy search solution follows the same steps described above. In practice the mRMR filter approach is highly successful in many applications, (Peng, Long and Ding, 2005; Meyer, Schretter and Bontempi, 2008), thereby justifying the intuitive concept that selecting features based on the compromise between relevance and redundancy may be more appropriate than relying solely on the naïve idea of selecting features only on the basis of strong association with the response.

More recently, Estevez et al. (2009) refined the criterion used in mRMR by dividing through the redundancy term with the minimum of the entropy $H(\cdot)$ of the two features (see Equations 3 and 4). Their argument is founded on the fact that the MI is bounded ($0 \leq MI(\boldsymbol{x_i}; \boldsymbol{x_s}) \leq \min\{H(\boldsymbol{x_i}), H(\boldsymbol{x_s})\}$), and the use of the normalized version of the redundancy term compensates for the MI bias, which is a common problem in MI estimation (Quinlan, 1986):

$$\text{mRMR}_{\text{normalized}} \overset{\text{def}}{=} \max_{i \in Q-S} \left[I(\boldsymbol{x}_i; \boldsymbol{y}) - \frac{1}{|S|} \sum_{s \in S} NI(\boldsymbol{x}_i; \boldsymbol{x}_s) \right] \tag{3}$$

$$NI(\boldsymbol{x}_i; \boldsymbol{x}_s) = I(\boldsymbol{x}_i; \boldsymbol{x}_s)/\min\{H(\boldsymbol{x}_i), \mathrm{H}(\boldsymbol{x}_s)\} \tag{4}$$

A further aspect of FS that is often underestimated or ignored is *variable complementarity*. Variable complementarity (also known as *conditional relevance*) is the property of two or more features being strongly associated with the response variable when they are *combined*, whilst the same features may be only moderately associated with the response *individually*. This issue has been topical lately, and has been explicitly addressed in a number of recent studies, for example (Meyer, Schretter and Bontempi, 2008; Brown 2009; Zhao and Liu, 2009). Meyer, Schretter and Bontempi (2008) extended mRMR to include up to second order interactions because in general this keeps algorithm complexity low, although in principle the interactions could be generalized to higher order. They demonstrated that their algorithm has the potential to outperform mRMR in some datasets, although their scheme was not universally superior. This suggests that second order complementarity proves quite useful in some datasets, and their results may indicate that including higher order interactions could further improve the performance of the FS filter scheme. However, the computation of high order interactions is both computationally expensive and difficult to be accurately computed generalizing criteria such as MI (for example using total correlation); in Section 3.2 we propose one way to tackle the computation of high order interactions very efficiently with the proposed FS algorithm.

There are many FS algorithms making use of the concepts outlined briefly here (relevance, redundancy, complementarity); specific algorithms will be introduced in the next Section.

3 Simple feature selection schemes and the new algorithm

Having outlined the main conceptual approaches of FS, we now focus on the actual schemes used in this study. We study simple sequential forward filters that can be used for multi-class classification and regression problems, are computationally efficient, and do not require the estimation of any internal parameters which rely on user expertise or experimentation.

3.1 Simple feature selection algorithms

The algorithmic description of mRMR (Peng, Long and Ding, 2005) was outlined in the previous section. We have used the mRMR source code from Peng (the MI computation relies on density estimation using histograms)[1]. In addition, we use a computationally cheap alternative to the original mRMR approach that used MI; here the relevance and redundancy are computed using the nonparametric Spearman rank correlation coefficient. We refer to the mRMR algorithm of Peng, Long and Ding (2005) by $mRMR_{MI}$, and to the alternative using the Spearman correlation coefficient by $mRMR_{Spearman}$.

An alternative FS algorithm where features are selected on the basis of being correlated to the target and minimally correlated to the existing feature subset is the Gram-Schmidt orthogonalization (GSO) (Stoppiglia et al., 2003). The GSO algorithm projects the original features onto the *null space* of those features already selected, and the feature that is maximally correlated with the target in that projection is selected next. The procedure iterates until the number of desired features has been selected. Further details of the GSO algorithm used for FS

[1] The Matlab source code for mRMR is available at http://penglab.janelia.org/software/Hanchuan_Peng_Software/software.html

can be found in Stoppiglia et al. (2003) and in Guyon et al. (2006). We have used the implementation of Guyon (2008).

A further very successful FS algorithm is the *least absolute shrinkage and selection operator* (LASSO) (Equation 5), which has generated major interest particularly in the statistics literature (Tibshirani, 1996; Efron et al., 2004; Donoho, 2006; Meinshausen and Yu, 2009), and which we have used in our previous studies (Tsanas et al., 2010a, Tsanas et al., 2010b, Tsanas et al., 2010c). The LASSO is a principled *shrinkage* method, enforcing the sum of absolute coefficient value penalty in a standard linear regression setting, formalized as follows:

$$\hat{\boldsymbol{b}}_{LASSO} = \arg\min_{b} \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{M} x_{ij} b_j \right)^2 + \lambda \sum_{j=1}^{M} |b_j| \quad (5)$$

where $\boldsymbol{b}$ is a vector containing the linear regression coefficients, and λ is the LASSO regularization parameter. Tibshirani (1996) has shown that this L_1-norm penalty promotes *sparsity* (some coefficients in $\boldsymbol{b}$ become zero), and therefore the LASSO can be used as an FS tool. Efron et al. (2004) have designed an efficient algorithm to determine the entire LASSO *regularization path* (that is, the values of the variables as λ is varied), increasing the popularity of the method, since this obviates the need for the user to search manually for the best λ by varying across the entire range of the regularization parameter. The LASSO has been shown extremely effective in environments where the features are not highly correlated (Donoho, 2006), and more recent research endorses its use even under those circumstances (Meinshausen and Yu, 2009). We have used K. Skoglund's implementation to determine the entire LASSO regularization path[2].

3.2 The proposed feature selection algorithm

The new FS scheme attempts to address the major issues outlined above: relevance, redundancy and complementarity. Initially, it relies on the computation of correlation coefficients, which are subsequently transformed using a function inspired by *information theoretic* (IT) *concepts* appropriate when the underlying distribution is Gaussian. This simplifying approach, that assumes normality of the features, is common in diverse machine learning applications and often works well in practice (Bishop, 2007). One reason for the success of the normal assumption is that the central limit theorem states that the distribution of the sum of an increasingly large number of non-Gaussian random variables tends to the Gaussian (under mild assumptions).

This starting assumption of normality greatly facilitates analysis since important IT concepts applied to the Gaussian distribution that are of central importance to this new algorithm are simple to compute and manipulate analytically. Before any processing of the dataset, the features and the response variable are standardized to have zero mean and unit standard deviation. This is also a common pre-processing step in machine learning applications, facilitating subsequent analysis: for example, it finds use in the LASSO algorithm (Hastie et al., 2009) and in mRMR (Peng, Long and Ding, 2005).

First, we compute the Spearman correlation coefficient between the features and the response variable to obtain the vector of rank correlations $\boldsymbol{r} = [r_1, r_2 \dots r_M]$, where each entry denotes the correlation of each feature with the response. Then, we compute the covariance matrix $\boldsymbol{\Sigma}$, and denote its entries with ρ_{ij}: these entries are the linear (Pearson) correlation coefficients computed between the features $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, where $i, j \in (1 \dots M)$. The fact that we choose a different metric to quantify relevance (with rank correlation coefficients) and redundancy (with linear correlation coefficients) may seem counter-intuitive; we address this in the Discussion.

[2] The Matlab source code for computing the LASSO path is available at http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=3897

$$\mathbf{\Sigma} = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1M} \\ \rho_{12} & 1 & \cdots & \rho_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1M} & \rho_{2M} & \cdots & 1 \end{bmatrix} \tag{6}$$

Then, for Gaussian distributions, there is an analytic expression for MI relying only on the linear correlation coefficient ρ (Cover and Thomas, 2006) (note that the MI also relies on the variance, but this is 1 due to the standardization pre-processing step):

$$MI = -0.5 \cdot \log\left(1 - \rho^2\right) \tag{7}$$

Throughout this study, we use the natural logarithm. For the purpose of this work, we are using Equation 7 as an IT quantity that is obtained using either the linear correlation coefficient or the rank correlation coefficient. For convenience, we will use the notation $r_{\mathrm{IT}}(X,Y) = -0.5 \cdot \log[1 - r_{XY}^2]$ to refer to the non-linearly transformed linear or rank correlation coefficient r_{XY} between two random variables X, Y. Now, we can write in compact vector form all the *relevance* terms using the IT inspired transform in Equation 7:

$$\boldsymbol{r}_{\mathrm{ITL}} = -0.5 \cdot \log[1 - r_1^2 \quad \cdots \quad 1 - r_M^2] \tag{8}$$

Similarly, using the covariance matrix $\mathbf{\Sigma}$ (Equation 6) and Equation 7, the *redundancy* between pairs of features can be conveniently expressed as a matrix, where each (*i,j*) entry denotes the information that two features share towards predicting the response:

$$\mathbf{\Sigma}_{\mathrm{IT}} = -0.5 \cdot \log \begin{bmatrix} 1 & 1-\rho_{12}^2 & \cdots & 1-\rho_{1M}^2 \\ 1-\rho_{12}^2 & 1 & \cdots & 1-\rho_{2M}^2 \\ \vdots & \vdots & \ddots & \vdots \\ 1-\rho_{1M}^2 & 1-\rho_{2M}^2 & \cdots & 1 \end{bmatrix} \tag{9}$$

Now, using the relevance terms in Equation 8 across the main diagonal of $\mathbf{\Sigma}_{\mathrm{IT}}$ in Equation 9, we obtain a matrix which will be used to compute the compromise between relevance and redundancy:

$$\mathbf{D} = -0.5 \cdot \log \begin{bmatrix} 1-r_1^2 & 1-\rho_{12}^2 & \cdots & 1-\rho_{1M}^2 \\ 1-\rho_{12}^2 & 1-r_2^2 & \cdots & 1-\rho_{2M}^2 \\ \vdots & \vdots & \ddots & \vdots \\ 1-\rho_{1M}^2 & 1-\rho_{2M}^2 & \cdots & 1-r_M^2 \end{bmatrix} \tag{10}$$

The matrix $\mathbf{D}$ is essentially a compact form of mRMR relying on the IT quantity of Equation 7 which alleviates the need for repeated computation of the relevance and complementarity terms in the iterative steps (therefore this expedites the incremental FS process in large datasets). Conceptually, the IT transformation of the (linear or rank) correlation coefficient assigns greater weight to coefficients above the absolute value 0.5 (see Figure 1). The fundamental idea is that weak associations (between a feature and the target or between features) are penalized; conversely strong associations (large absolute correlation coefficients) are enhanced.

Now, using the notion that MI needs to be normalized for the redundancy term as described in Estevez et al. (2009), we divide the redundancy by the minimum of the entropies between the two features. For the Gaussian distribution, the entropy is simply a scalar value (there is no dependency on the variance because of the pre-processing step):

$$H(\boldsymbol{x}_i) = H(\boldsymbol{y}) = -0.5 \cdot \log\left(2\pi e\right) \tag{11}$$

Therefore, all the terms not in the main diagonal in Equation 10, are divided by the scalar quantity in Equation 11 giving rise to **D2**. We introduce the subscript (n) in the form $r_{\mathrm{IT,n}}$ to denote the normalization of the term r_{IT} $\left(r_{\mathrm{IT,n}} = r_{\mathrm{IT}}/H(\boldsymbol{y})\right)$.

The proposed algorithm developed thus far can be seen as an extension of the classical mRMR using an *information theoretic* inspired transformation, and for this reason we call it mRMR$_{\mathrm{ITL}}$. Thus, the mRMR$_{\mathrm{ITL}}$ is conveniently calculated in terms of the matrix **D2**, where for the computation of the new candidate feature $\boldsymbol{x_i}$ (which corresponds to a feature not in the existing feature subset) we focus on the i^{th} row. The relevance of the feature $\boldsymbol{x_i}$ lies on the main diagonal of the matrix **D2**, and the redundancy is computed from the average of the terms that appear in the column s (the $\text{D2}_{i,s}$ entries) where s corresponds to features in the already selected subset ($s \in S$).

We introduce the concept of quantifying the *conditional relevance* (complementarity) of a feature as the usefulness of that feature in predicting the response *conditional upon the existing feature subset*. This is achieved using the rank partial correlation coefficient $r_{\text{p}}(\boldsymbol{x_i}; \boldsymbol{y}|S)$. That is, the partial correlation coefficient r_{p} is defined as the rank correlation coefficient between a new candidate feature $\boldsymbol{x_i}$ and the response $\boldsymbol{y}$, controlling for the existing features in the subset.

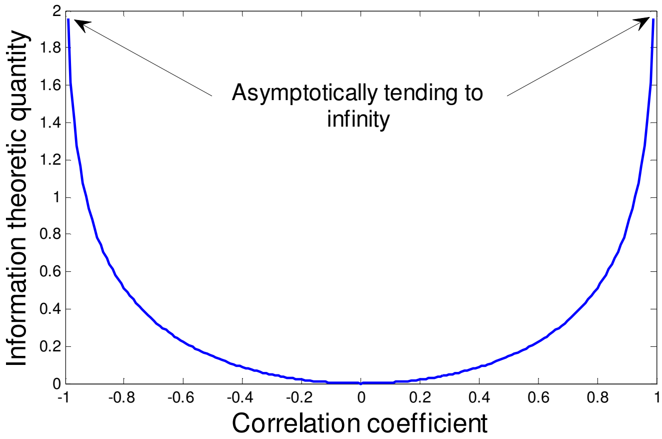

Figure 1: Information theoretic (IT) quantity (relevance or redundancy) as a function of the linear (Pearson) or rank (Spearman) correlation coefficient ρ, computed as $I(\rho) = -0.5 \cdot \log(1 - \rho^2)$. Asymptotically, as the absolute value of the correlation coefficient tends to ± 1, the IT quantity becomes infinite (in practice we set this to a very large value). We demonstrate that this IT nonlinear transformation of the correlation coefficients is valuable in feature selection.

This approach aims to incorporate how well the candidate feature pairs up with the existing features that have already been chosen. Then, we transform the computed partial correlation coefficient using the IT inspired transformation in Equation 7, which gives:

$$r_{\text{p,IT}} = -0.5 \cdot \log\left[1 - r_{\text{p}}^2\right] \tag{12}$$

Since the controlling variables S (whose effect needs to be removed to compute the partial correlation coefficient) are not known and will vary at each step, it is not possible to express this quantity in vector or matrix form as we did above for **D2**.

This additional term in Equation 12 is added to mRMR_{ITL}, and we therefore obtain the new FS algorithm which we call *relevance, redundancy and complementarity trade-off* (RRCT):

$$\begin{aligned}\text{RRCT} \overset{\text{def}}{=} \max_{i \,\in\, Q-S} \Bigg[& r_{\text{IT}}(\boldsymbol{x_i}; \boldsymbol{y}) - \frac{1}{|S|} \sum_{s \,\in\, S} r_{\text{IT,n}}(\boldsymbol{x_i}; \boldsymbol{x_s}) + \text{sign}\left(r_{\text{p}}(\boldsymbol{x_i}; \boldsymbol{y}|S)\right) \\ & \cdot \text{sign}\left(r_{\text{p}}(\boldsymbol{x_i}; \boldsymbol{y}|S) - r(\boldsymbol{x_i}; \boldsymbol{y})\right) \cdot r_{\text{p,IT}} \Bigg]\end{aligned} \tag{13}$$

sign(·) returns +1 if the quantity (·) is positive and -1 if (·) negative, and is used to determine whether $r_{\mathrm{p,IT}}$ is added or subtracted in Equation 13. RRCT follows Battiti's (1994) algorithmic steps (see Section 2) using Equation 13 instead of Equation 1 to select features. Care needs to be exercised in the RRCT expression when including the $r_{\mathrm{p,IT}}$ term. Given that this term is non-negative due to the IT transformation, we need to determine whether the inclusion of the candidate feature to the existing subset actually contributes *additional* information conditional on the features in the selected subset (conditionally relevant). Consideration must be made of both the sign of the partial correlation coefficient, and the sign of the difference in magnitudes between $r_{\mathrm{p}}(\boldsymbol{x_i};\boldsymbol{y}|S)$ and $r(\boldsymbol{x_i};\boldsymbol{y})$. The $\mathrm{sign}\left(r_{\mathrm{p}}(\boldsymbol{x_i};\boldsymbol{y}|S)-r(\boldsymbol{x_i};\boldsymbol{y})\right)$ term in Equation 13 is used to determine whether the conditional relevance term $r_{\mathrm{p}}(\boldsymbol{x_i};\boldsymbol{y}|S)$ is larger than $r(\boldsymbol{x_i};\boldsymbol{y})$ in terms of magnitude; that would mean that including the candidate feature has additional (conditional) relevance given the features in the selected subset. The $\mathrm{sign}\left(r_{\mathrm{p}}(\boldsymbol{x_i};\boldsymbol{y}|S)\right)$ term is used to make the overall complementarity contribution positive in the case that $r(\boldsymbol{x_i};\boldsymbol{y})<0$, $r_{\mathrm{p}}(\boldsymbol{x_i};\boldsymbol{y}|S)<0$ and $\left(r_{\mathrm{p}}(\boldsymbol{x_i};\boldsymbol{y}|S)-r(\boldsymbol{x_i};\boldsymbol{y})\right)<0$, because then the term $\mathrm{sign}\left(r_{\mathrm{p}}(\boldsymbol{x_i};\boldsymbol{y}|S)-r(\boldsymbol{x_i};\boldsymbol{y})\right)$ would indicate the additional contribution offered by the complementarity term is negative.

To isolate the advantages of using the partial correlation coefficient from the advantages of using the IT transformation in mRMR$_{\mathrm{ITL}}$, we define an alternative FS scheme, *RRCT*$_0$. RRCT$_0$ is identical to Equation 13 except that all the terms (relevance, redundancy, and complementarity) have not undergone IT transformation. That is, we use the raw correlation coefficients and the partial correlation coefficient instead.

We aim to demonstrate that the simple nonlinear transformation of the correlation coefficients using IT concepts derived under the assumption of Gaussianity, brings a tangible advantage in FS over alternative approaches (for example, over the mRMR$_{\mathrm{Spearman}}$ scheme). Moreover, introducing the conditional relevance term that controls for the existing features in the selected subset at each iteration, combined with IT transformation, brings additional power in selecting a parsimonious feature subset rich in information content.

So far, the IT approach has assumed that all the distributions of the features and the response are Gaussian. Because this may be substantially inaccurate in some circumstances, we use the *Box-Cox transform*, which aims to normalize non-Gaussian random variables (Box and Cox, 1964). The Box-Cox transformation (see Equation 14) belongs to a family of *power transformations*, and takes the form:

$$f(x,\lambda)=\begin{cases}\dfrac{(x^{\lambda}-1)}{\lambda},\lambda\neq 0\\ \log(x),\lambda=0\end{cases}\qquad(14)$$

where λ is determined via optimization to maximize the associated log likelihood function.

	Relevance	Redundancy	Complementarity	Information theoretic transformation	Box-Cox transformation
mRMR$_{\mathrm{MI}}$	X	X	-	-	-
mRMR$_{\mathrm{Spearman}}$	X	X	-	-	-
GSO	X	-	-	-	-
LASSO	X	X	-	-	-
mRMR$_{\mathrm{ITL}}$	X	X	-	X	-
mRMR$_{\mathrm{ITL,Box\text{-}Cox}}$	X	X	-	X	X
RRCT$_0$	X	X	X	-	-
RRCT	X	X	X	X	-
RRCT$_{\mathrm{Box\text{-}Cox}}$	X	X	X	X	X

Table 1: Summary of the properties for each feature selection algorithm used in this study.

There is active research into the optimal determination of λ (Marazzi and Yohai, 2006) beyond the scope of this work and here we will use the standard reference approach with the maximum likelihood estimate. We apply the Box-Cox transform to the raw data prior to standardization, and compute the RRCT on this transformed data, in addition to RRCT for the non-transformed data. This is indicated as $\text{RRCT}_{\text{Box-Cox}}$ for convenience. Table 1 summarizes the main properties of the FS algorithms used in this study.

4 Datasets

Table 2 summarizes the data used in this study. All the selected datasets are publicly available, and most have been previously used in the FS literature. In cases of missing entries in a dataset, the corresponding row in the data matrix was deleted.

Dataset	Design matrix	Associated task	Attributes
Artificial 1	1000×100	Regression	C (100)
TIED (Artificial 2)[3] (Statnikov and Aliferis, 2009)	750×999	Classification (4 classes)	D (999)
Friedman (4 sets) (Artificial 3) (Friedman, 1999)	500×50 (2) 1000×50 (2)	Regression	C (50)
Hepatitis[4] (Diaconis and Efron, 1983)	155×19	Classification (2 classes)	C (17), D (2)
Acute inflammations (urinary bladder)[4] (Czerniak and Zarzycki, 2003)	120×6	Classification (2 classes)	C (1), D (5)
Breast cancer (diagnostic)[4] (Wolberg and Mangasarian, 1990)	569×30	Classification (2 classes)	C (30)
Breast cancer (prognostic)[4] (Wolberg and Mangasarian, 1990)	198×33	Classification (2 classes)	C (31), D (2)
Statloat heart[4]	270×13	Classification (2 classes)	C (1), D (12)
Parkinson's[4] (Little et al., 2009)	195×22	Classification (2 classes)	C (22)
Liver[4]	345×6	Classification (2 classes)	C (1), D (5)
Heart disease[4]	303×13	Classification (5 classes)	C (1), D (12)
Los Angeles Ozone[5]	330×9	Classification (35 classes)	D (9)
Concrete compressive strength[4]	1030×8	Regression	C (7), D (1)
Boston Housing[4]	506×13	Regression	C (10), D (3)
Prostate[5]	97×8	Regression	C (4), D (4)

Table 2: Summary of datasets used in this study. The size of the design matrix is $N \times M$, where N denotes the number of instances (samples), and M denotes the number of features. The 'attributes' denote the type of the design matrices' variables: continuous (C) or discrete (D). In cases of missing entries, the entire row in the data matrix was deleted.

Artificial Dataset 1 comprises 1000 samples and 100 features (that is, the design matrix $\mathbf{X}$ is 1000×100) where the 100 continuous-valued, independent features were generated using the

[3] Downloaded from the *Causality Workbench Repository*: http://www.causality.inf.ethz.ch/repository.php

[4] Downloaded from the *UCI Machine Learning Repository*: http://archive.ics.uci.edu/ml/datasets.html

[5] Downloaded from the website of the *Elements of Statistical Learning* book: http://www-stat.stanford.edu/~tibs/ElemStatLearn/

standard normal distribution. The response was determined as a linear combination of 10 inputs $\{\boldsymbol{x_1} \dots \boldsymbol{x_{10}}\}$, where each feature was assigned a random coefficient $\{a_1 \dots a_{10}\}$ in the range 10 to 100, and no two features were allowed to have the same coefficient. All weights were positive to avoid any masking issues where one feature might be effectively cancelled by another. Thus, the target variable has the form:

$$\boldsymbol{y} = a_1\boldsymbol{x_1} + a_2\boldsymbol{x_2} + \dots + a_{10}\boldsymbol{x_{10}}$$

To simulate a real world scenario, after obtaining the response variable $\boldsymbol{y}$ we added 10% independent and identically distributed (i.i.d.) Gaussian noise to each of the features. Therefore, the first dataset is a standard regression problem with 10 *true* predictors (features that contribute towards predicting the response) and 90 *false* (redundant, irrelevant, or noisy) predictors.

The second artificial dataset (TIED) was generated by Statnikov and Aliferis (2009). It is obtained from a discrete Bayesian network where there are 750 instances with 999 variables, and the response has four classes. The 13 relevant variables are known *a priori*: $\{X_1, X_2, X_3, X_4, X_8, X_9, X_{10}, X_{11}, X_{12}, X_{13}, X_{18}, X_{19}, X_{20}\}$. These features are organized into the following five subsets contributing towards the 72 Markov boundaries (smallest subset of features for which the response variable is conditionally independent of other features) of the TIED dataset (one variable from each subset): (a) $\{X_9\}$, (b) $\{X_4, X_8\}$, (c) $\{X_{11}, X_{12}, X_{13}\}$, (d) $\{X_{18}, X_{19}, X_{20}\}$, and (e) $\{X_1, X_2, X_3, X_{10}\}$. A separate set to test the effectiveness of feature selection is also provided, which has 3,000 samples. The pilot study illustrated that this is a very challenging problem, where some advanced FS algorithms failed to detect all the variables in the Markov boundaries. This dataset was recently used in Tuv et al. (2009) who demonstrated that their FS scheme based on ensembles (relying on random forests' importance score) can identify all the true features including three false features ($\{X_{15}, X_{29}, X_{14}\}$).

The third artificial dataset uses the well-known Friedman data generator. The model has multiple nonlinear interactions amongst the explanatory variables, and includes relevant, redundant, and noisy input variables. We used four realisations of this generator[6]. All the datasets used here have 50 continuous valued variables, where only 5 are true. We used two dataset sizes: 500 and 1000 samples. In addition, we used two *collinearity degrees* (number of variables which depend on other variables): two and four.

One real dataset widely used in FS scheme comparisons is the *hepatitis* dataset (Diaconis and Efron, 1983). It includes 155 patients and the binary outcome (healthy control subject versus subject with hepatitis disease) depends on 19 features. This dataset has been subject to close scrutiny by Breiman (2001), who concluded that features X_{17} and X_{12} were highly indicative of the response (and highly correlated with each other). Breiman suggested that either of those two explanatory variables individually carries almost as much information as the entire feature set. The features X_{19} and X_{11} were also identified as conveying some additional information towards predicting the response. More recently Tuv et al. (2009) identified the following feature subset using a scheme based on random forests: $\{X_6, X_{17}, X_{14}, X_{19}, X_{11}\}$. That study contrasted their proposed FS scheme with three alternative FS algorithms, which also unanimously identified variable X_2.

We refer to the original studies and the publicly available repositories cited in Table 2 for further details regarding each dataset.

5 Results

In general, there are two approaches to evaluate FS algorithms: the first aims to determine whether the optimal feature subset was selected (optimal being the combination of features maximally associated with the response), and the second aims to improve a performance metric in

[6] The four realizations of the Friedman generator used in this study are available at http://www.cs.waikato.ac.nz/~ml/weka/index_datasets.html.

the subsequent learning phase where the feature subsets are input to a learner. This later approach is a surrogate to validate the efficiency of FS schemes, although it does not necessarily correspond to selecting the optimal feature subset, and it is possible that different learners might lead to different conclusions. Moreover, in practice some weakly relevant or redundant features could improve the learners' performance; conversely, the benefit of discarding relevant features may outweigh loss in information content (Guyon et al., 2007). Therefore, in this study we focus almost exclusively on the first approach, trying to determine whether the correct set of features has been selected. This aim is actually infeasible for real world datasets, because we do not know *a priori* the best feature subset; in Section 5.2 we propose an approach to tackle this problem.

In the subsequent analysis, all datasets are standardized to have zero mean and unit standard deviation. With the exception of the TIED dataset, the features are always chosen using 10% hold out data with 100 repetitions (effectively this is the same approach as in 10-fold cross-validation). That is, we randomly permute the initial design matrix and select 90% of the data to select a feature subset; the process is repeated 100 times where each time the initial dataset is randomly permuted. We do this to determine which of the features are robustly selected and are true (intrinsically useful towards predicting the response in the given task); although the feature subsets are not identical in the 100 repetitions, this approach gives us confidence in the selected features.

5.1 Focusing on selecting the optimal feature subset: two artificial benchmarks

The first artificial dataset serves as a very simple benchmark, to test how accurately the FS algorithms select features. This test demonstrated that GSO, LASSO, RRCT_0, RRCT and $\text{RRCT}_{\text{Box-Cox}}$ accurately identified the true features in all 100 repetitions of the cross validation process with no false alarms (false features appearing in the place of true features). The remaining filters had a few false alarms.

mRMR_{MI}	$\text{mRMR}_{\text{Spearman}}$	GSO	LASSO	mRMR_{ITL}	$\text{mRMR}_{\text{ITL,Box-Cox}}$	RRCT_0	RRCT	$\text{RRCT}_{\text{Box-Cox}}$
10	**10**	**11**	**11**	**10**	**10**	**10**	**10**	**10**
11	14	**10**	**13**	**12**	**12**	240	**12**	**2**
18	**11**	**18**	**12**	**18**	**2**	**2**	**18**	**3**
4	**18**	14	**10**	15	**3**	**1**	**19**	**1**
12	**3**	**13**	**18**	**11**	**18**	**3**	15	**18**
13	**13**	**19**	14	**2**	**1**	58	**2**	**11**
19	**2**	**12**	15	**3**	15	302	**11**	**19**
29	15	**8**	**8**	**13**	**19**	388	**3**	15
20	**12**	**4**	228	29	14	810	**13**	14
8	**1**	772	351	**19**	**8**	29	29	**8**
9	**19**	417	362	**1**	6	15	**1**	6
15	29	501	120	**8**	42	24	**8**	**4**
2	**8**	228	417	14	288	**8**	14	**9**
14	20	31	772	6	364	6	4	42
3	6	498	77	288	235	19	6	305
6	473	305	730	42	109	18	9	288
(2)	(3)	(5)	(7)	(3)	(5)	(8)	(3)	(3)

Table 3: Feature selection results for the TIED dataset. The TIED dataset is artificially constructed from a discrete Bayesian network, where 13 features are relevant (see text for details). All the feature selection (FS) algorithms in this study are greedy, and rely on forward sequential selection. Each row in the Table shows the feature that maximizes the criterion used in each FS algorithm in the iterative (greedy) processing steps. The correctly detected features appear in bold, and the final row summarizes the number of falsely detected features. This number is computed for the 13 features ($m = 13$); the Table presents the first 16 features ($m = 16$) to show whether some algorithms might have closely missed the true features.

The selected features for the TIED dataset appear in Table 3. We note that mRMR$_{\mathrm{MI}}$ had the lowest number of false alarms. The false alarms come mainly from $\{X_{15}, X_{29}, X_{14}\}$ (for example these are the only three false alarms in the RRCT algorithm), features which were also incorrectly selected in Tuv et al. (2009) who used a much more sophisticated FS scheme. Interestingly, if we decide to use the best feature subset out of 16 features (and not out of 13, that is allowing for three false alarms), RRCT recovers two out of the three remaining relevant variables (and mRMR$_{\mathrm{MI}}$ recovers one out of the remaining two). Thus, RRCT is very competitive with a considerably more sophisticated FS algorithm (Tuv et al., 2009) in terms of finding the relevant features in the TIED dataset. Alternative FS algorithms that worked well for the first artificial example, and in particular LASSO, do not perform well on this dataset.

5.2 False discovery rate

Assuming the number of true and false (collectively referring to redundant, irrelevant and noisy) features in a dataset is known (ground truth), we define the *False Discovery Rate* (FDR) as the number of false features erroneously identified by the FS algorithm as true. For artificial datasets the optimal feature subset is known *a priori*, information which is typically unknown for real datasets. Therefore, for the artificial datasets it is easy to quantify the performance of each FS algorithm at selecting the optimal feature subset. To quantify the performance of feature selection algorithms for real datasets we need a different strategy.

First, we assume that *most* of the features in the original design matrix are in some way related to the response (this implicitly assumes that the researchers who collected the data in the first place had a reasonable idea of the relevant explanatory variables for their application). The features in the original dataset are assumed to be true (useful in generating the response). Then, by appending a large number of *irrelevant* features we test the ability of the FS scheme to discard those artificial *probes* (false features). For each real dataset, we appended to the original design matrix 100 irrelevant features. Each irrelevant feature was randomly chosen to belong each time to one of eight possible different distributions (normal, extreme value, uniform, beta, chi-square, gamma, generalized extreme value, and Weibull) and was independently sampled. The use of widely different distributions ensures that the resulting dataset will resemble more closely real domain applications, where the distributions of the features could vary widely. In addition, this also tests the versatility of the FS algorithms proposed in this study, which exploit the Gaussian assumptions for the features, in recovering the true feature subset.

Next, we generate another 100 irrelevant features to append to the data matrix (in addition to the 100 irrelevant features generated as described above). Each of those new irrelevant features is independently generated by randomly permuting a randomly chosen feature from the original data matrix. Thus, each of the new 100 irrelevant features will have the same empirical probability distribution as the randomly chosen (true) feature which was used to generate it. The aim of this step is to investigate whether the FS algorithms are misled by features with the same distributions as the original features. Now we focus on applying the FDR methodology.

In the case of the four datasets obtained from the Friedman data generator, we set $m = 5$ (we check for five true features), and examine the ability of the FS algorithms to detect the true features at each of the iterative $1 \dots m$ steps. We used 10-fold cross validation with 100 repetitions for confidence and report the average FDR in Figure 2. These results are averaged over the 100 repetitions and should be interpreted sequentially: each step in the x-axis denotes the iterative step in the FS algorithms, and the values in the y-axis denote whether each FS algorithm's choice identified true features in the subset (or whether it selected a probe). For example, a value of 0.1 in the y-axis for the first iterative step (x-axis=1) for one of the iterative FS algorithms would denote that in 10 out of the 100 repetitions the first feature that is selected for the given FS scheme is a probe. Similarly, for the second iterative step (x-axis=2) a value of 0.1 in the y-axis would denote that 10 times out of the 100 repetitions one out of the first two selected features is a probe, and so on. The results in Figure 2 clearly indicate that the $\text{RRCT}_{\text{Box-Cox}}$ works particularly well in this problem.

We report the FDR results for the real datasets used in this study using the same methodology we used to obtain the findings in Figure 2. In Figure 3 we use each dataset with 100 additional irrelevant appended features in the design matrix (the 100 irrelevant features sampled from the eight possible distributions). In Figure 4 we use each dataset with 200 additional irrelevant appended features in the design matrix (100 irrelevant features sampled from the eight possible distributions, and 100 features where each feature is generated by randomly permuting a randomly chosen feature from the original data matrix.) In both Figure 3 and Figure 4 we focus on how many true features the FS algorithms recover when we set $m = \min(20, M)$, presenting the results for all the iterative steps $1 \dots m$.

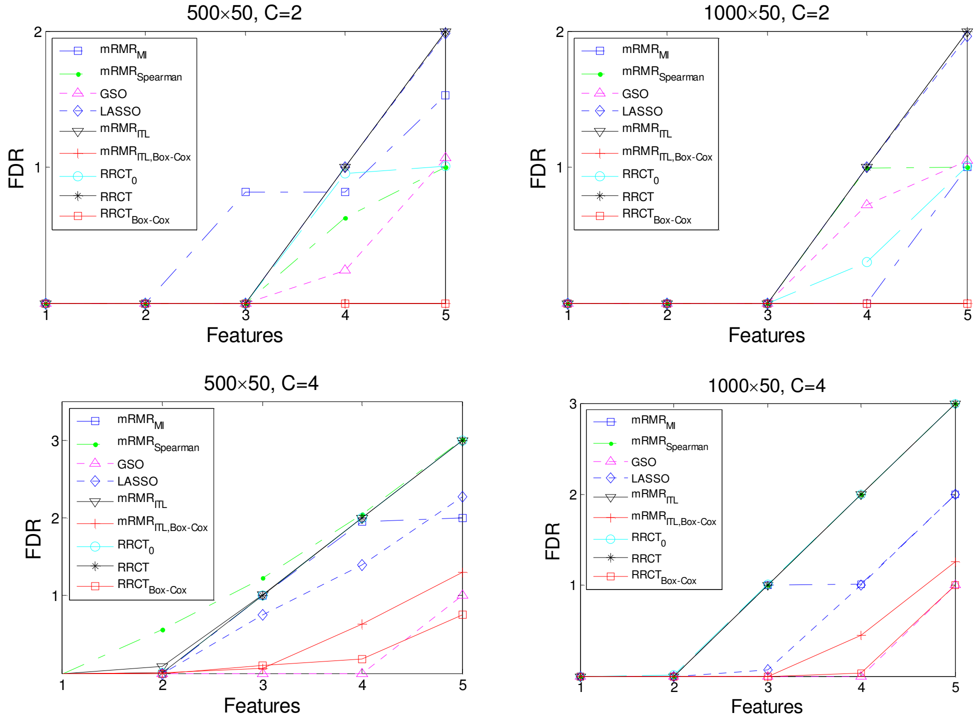

Figure 2: False Discovery Rate (FDR) (number of artificially added, false features identified by the algorithm as true), as a function of including features (the iterative steps in the FS algorithms were set to five). The smaller the FDR the better the feature selection algorithm. The design matrix is given in the form $N \times M$, where N denotes the number of instances (samples), and M denotes the number of features. Each of the four datasets was developed using the Friedman data generator, and has 50 features (5 are true). The degree of collinearity (C) denotes the presence of two or four collinear features. The results are the average FDR computed using 10-fold cross validation with 100 repetitions.

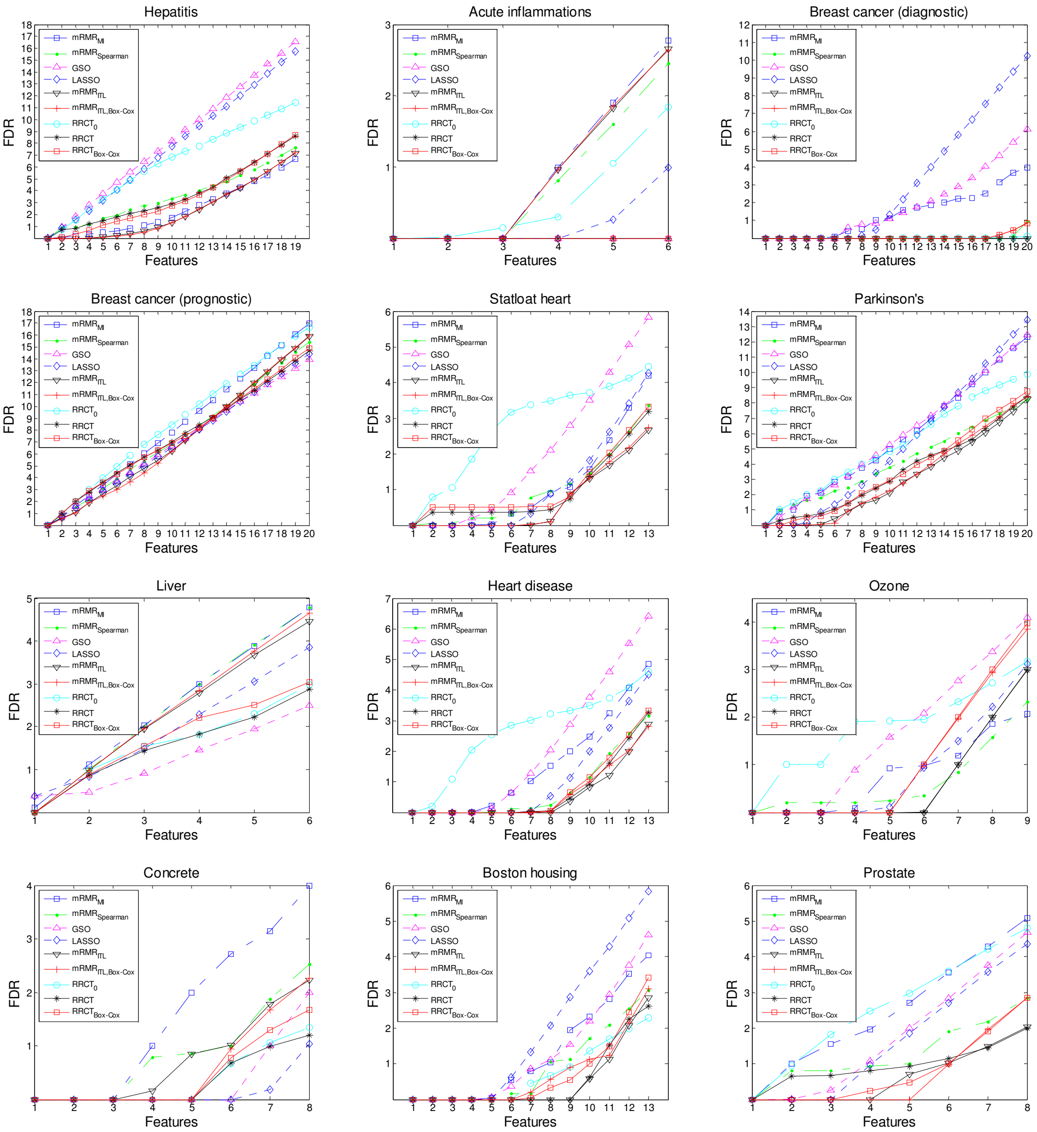

Figure 3: False discovery rate (FDR) (number of artificially added, false features identified by the algorithm as true), as a function of including features. The iterative steps in the FS algorithms were set to $m = \min(20, M)$, where M is the number of features in each dataset. The smaller the FDR the better the feature selection algorithm. For each dataset, we appended to the original design matrix 100 irrelevant features, independently sampled and randomly chosen each time from eight different distributions (see text for details). The results are the average FDR computed using 10-fold cross validation with 100 repetitions.

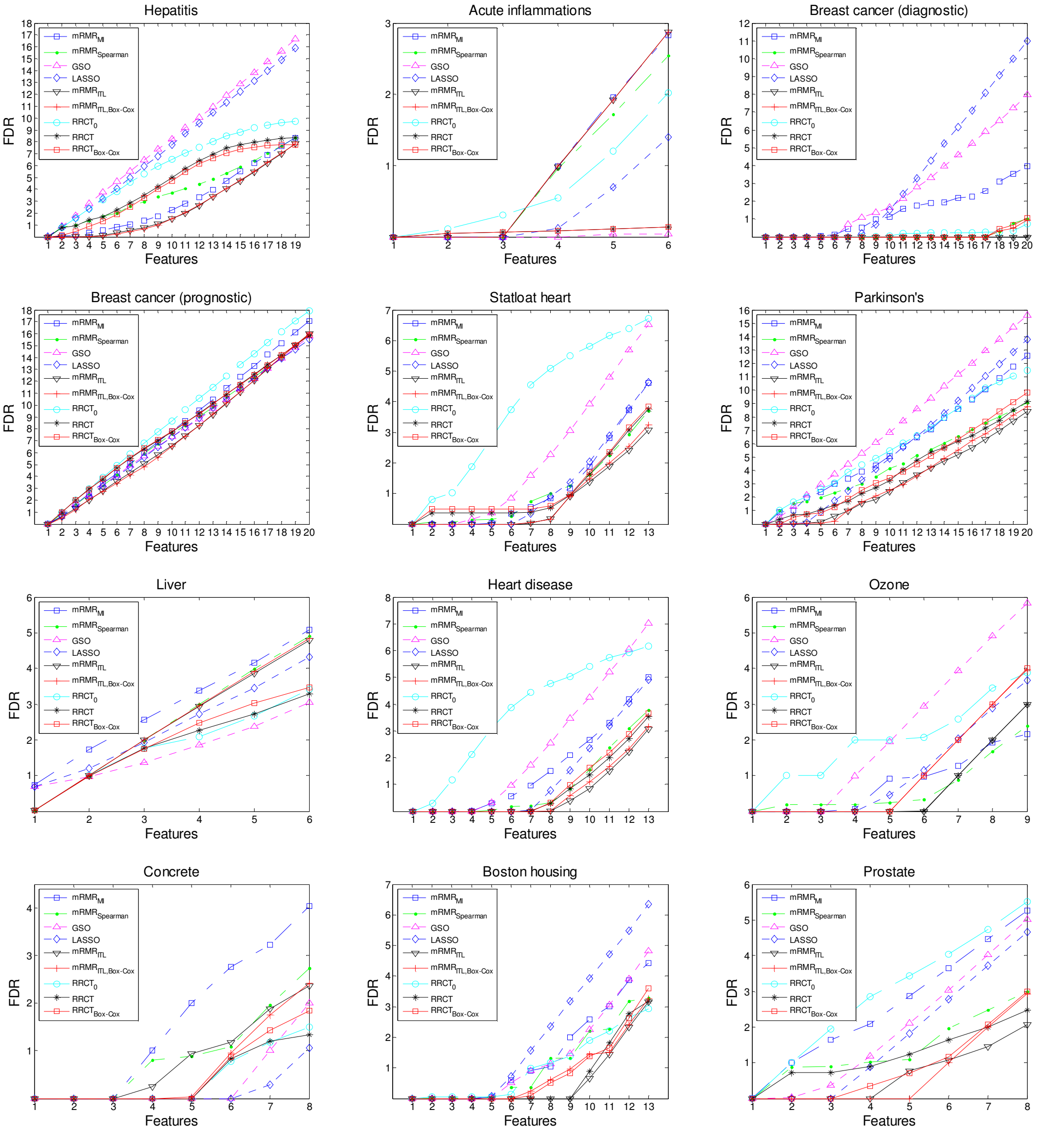

Figure 4: False discovery rate (FDR) (number of artificially added, false features identified by the algorithm as true), as a function of including features. The iterative steps in the FS algorithms were set to $m = \min(20, M)$, where M is the number of features in each dataset. The smaller the FDR the better the feature selection algorithm. For each dataset, we appended to the original design matrix 100 irrelevant features, independently sampled and randomly chosen each time from eight different distributions. In addition, we appended 100 irrelevant features, randomly generated by permuting the entries of a randomly chosen feature in the original data matrix each time (see text for details). The results are the average FDR computed using 10-fold cross validation with 100 repetitions.

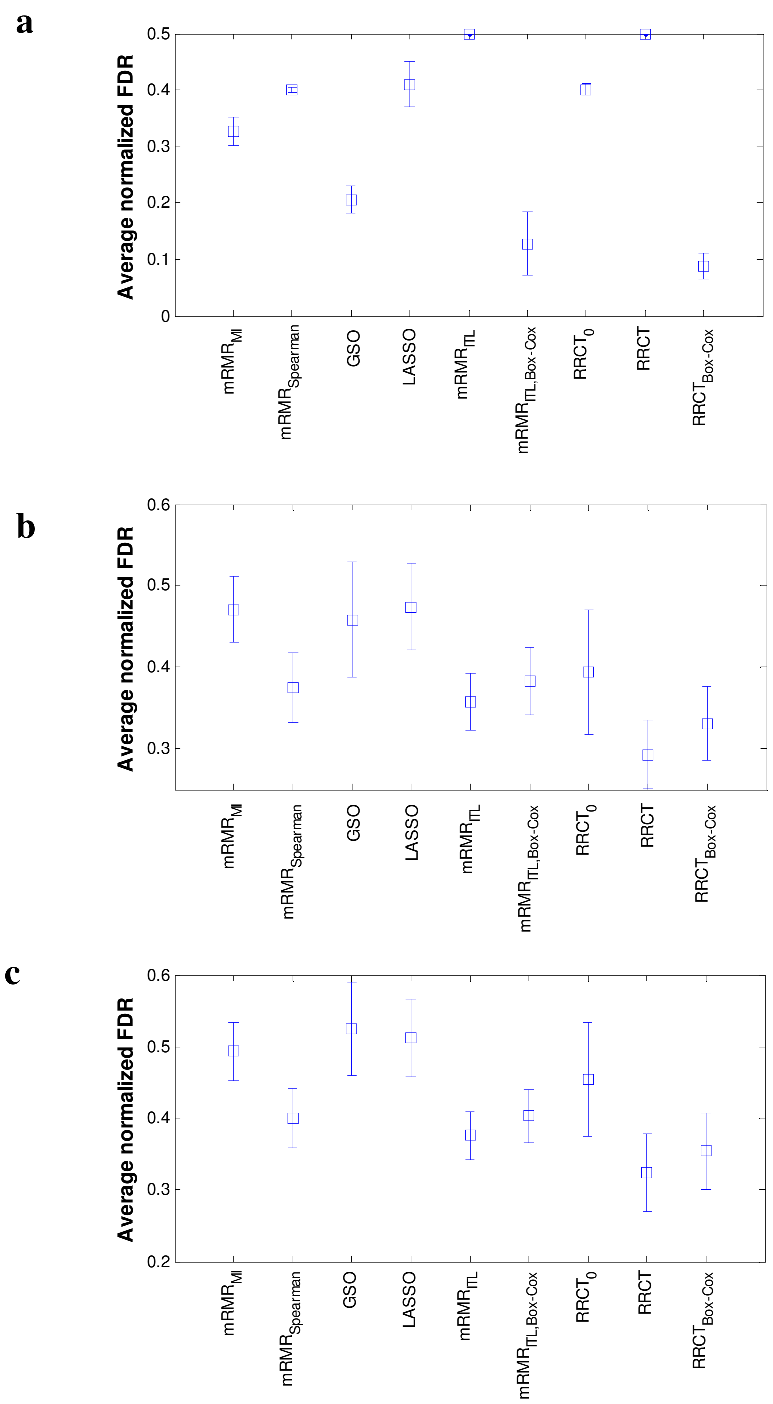

Figure 5: Summarizing the results of Figures 2-4: a) includes the four datasets from Friedman's generator, b) includes the twelve real datasets using 100 probes, c) includes the twelve real datasets using 200 probes. The normalized FDR for each dataset was computed out of m=min(20, M), where M is the number of features in each dataset, and then averaged across the datasets in each of the three cases in Figures 2-4. The boxes denote the average normalized FDR and the lines denote the standard deviation.

Figure 5 summarizes the findings reported in Figures 2-4, presenting box plots of the average normalized FDR. For each dataset we computed the ratio of $\text{FDR}(m)$ and m, where $m = \min(20, M)$. $\text{FDR}(m)$ denotes the m^{th} (rightmost) average FDR scalar value in each dataset (see Figures 2-4). The $\text{FDR}(m)/m$ ratio provides the normalized FDR (a scalar) for each dataset. Then, we computed the mean and standard deviation normalized FDR scores across three clusters of datasets: a) for the four datasets from Friedman's generator, b) for the twelve datasets where 100 artificial probes had been inserted using the methodology outlined to obtain the results for Figure 3, and c) for the twelve datasets where 200 artificial probes had been inserted using the methodology outlined to obtain the results for Figure 4. The results in Figure 5 suggest that RRCT is, on average, outperforming the competing FS schemes for the real datasets. Interestingly, RRCT with prior Box-Cox transformation is remarkably accurate for the Friedman datasets. We reflect further on these findings in the Discussion.

Figure 6 presents an illustrative example of the trade-offs involved for $\text{RRCT}_{\text{Box-Cox}}$ and $\text{mRMR}_{\text{ITL,Box-Cox}}$ for one of the real datasets (Acute Inflammations). In this example we used all the data samples (120) and the additional 200 false variables generated as explained above (100 irrelevant features generated from the eight distributions, and 100 irrelevant features where each feature is generated by randomly selecting each time a feature and randomly permuting it), setting $m = 6$ ($M = 6$ in the Acute Inflammations dataset). The results in Figure 6 illustrate nicely how the complementarity term works in the forward incremental FS process and provide intuitive insight for understanding why conditional relevance may be valuable in many FS applications (promoting or relegating the need to include a feature). For example, in the third iterative step of the $\text{RRCT}_{\text{Box-Cox}}$ scheme a particular feature was promoted because the combination with the two features already selected contributes markedly towards predicting the response. Similarly, in the fourth iterative step of the $\text{RRCT}_{\text{Box-Cox}}$ scheme a feature is almost irrelevant *individually* (relevance term is almost zero). However, this particular feature is recovered because the complementarity term suggests it combines well with the features selected in the preceding steps of the algorithm. Although it is not obvious in Figure 6 whether the selected features were recovered accurately, we get a visual insight for how complementarity works. In this example, the $\text{RRCT}_{\text{Box-Cox}}$ method selected all the six true features, whereas $\text{mRMR}_{\text{ITL,Box-Cox}}$ selected only the first three true features correctly (subsequently three false features were selected).

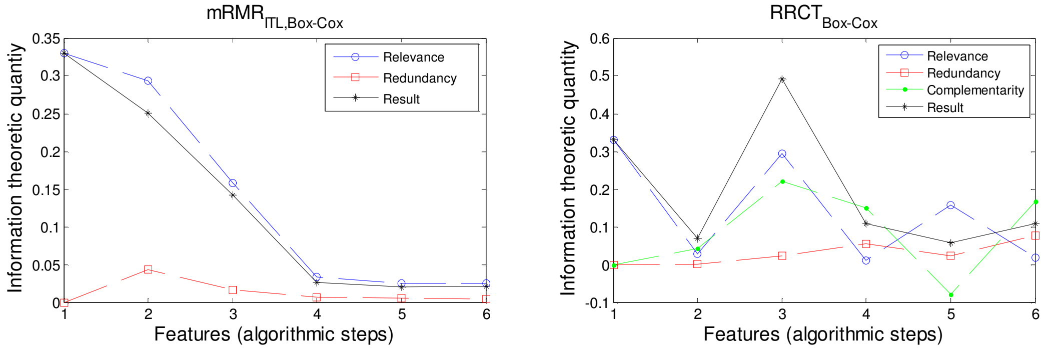

Figure 6: Illustrative example presenting the values of the relevance, redundancy and complementarity terms in the sequential forward feature selection process contrasting $\text{mRMR}_{\text{ITL,Box-Cox}}$ and $\text{RRCT}_{\text{Box-Cox}}$ for the Acute Inflammations dataset. In this experiment we appended to the original design matrix (120×6) 200 additional irrelevant features (see text for details). The information theoretic quantity in the y axis is obtained by the nonlinear transformation of the correlation coefficients (see Equation 7).

$mRMR_{MI}$	$mRMR_{Spearman}$	GSO	LASSO	$mRMR_{ITL}$	$mRMR_{ITL,Box\text{-}Cox}$	$RRCT_0$	RRCT	$RRCT_{Box\text{-}Cox}$
17	17	19	19	17	17	17	17	17
2	2	1	1	11	11	2	2	2
11	11	2	2	12	12	6	18	18
12	12	3	3	18	18	8	11	11
18	18	4	4	14	14	18	14	14
14	13	5	5	13	13	7	12	12
6	1	6	6	6	6	5	13	13

Table 4: Feature selection results for the hepatitis dataset. The hepatitis dataset has 155 samples and 19 features and the outcome is a binary response. All the feature selection (FS) algorithms in this study are greedy, and rely on forward sequential selection (we set $m = 7$ to show the results for the first seven features selected in the iterative process). Each row in the Table shows the feature that maximizes the criterion used in each FS algorithm in the iterative (greedy) processing steps. We used 10-fold cross validation with 100 repetitions to obtain statistical confidence. The Table reports the feature subset that was selected most often in the 100 repetitions.

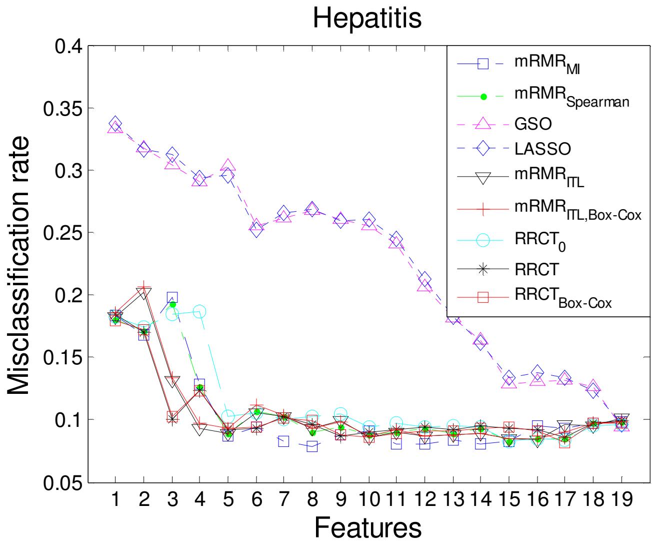

Figure 7: Out of sample performance (using 10-fold cross validation with 1000 repetitions) as a function of including the features iteratively selected by each feature selection algorithm for the hepatitis dataset (see Table 4 for each feature selection algorithm's choices as the first seven iterations proceed). For illustration in this classification problem we have used the Naïve Bayes classifier.

5.3 Focusing on prediction performance

We have already discussed prediction performance of a learner as a proxy solution to determine the accuracy of the FS algorithms, and have explained the perennial pitfalls in that approach. For illustrative purposes we present here results for the performance of the algorithms for the widely used hepatitis dataset. In Table 4 we report the feature subsets selected by the FS algorithms for the hepatitis dataset. The feature subsets were selected using 10-fold cross validation (we used 100 repetitions to obtain confidence and selected the subset that occurred most often). In our experiments, X_{12} or X_{17} was almost always chosen by most of the FS schemes in the first step in the 100 repetitions, thus confirming Breiman's (2001) observations. Moreover, the strong masking effect between the two variables was verified in this study in accordance to the findings of Breiman (2001) and Tuv et al. (2009). Subsequently, we chose to use the Naïve Bayes classifier as the learner for this classification problem. Although this learner inherently assumes that the input variables are uncorrelated, in practice it works well even if this condition is violated (Hastie et al., 2009). For each FS scheme we repeated the training process with the Naïve Bayes classifier 19 times ($m = 19$, the number of features in the hepatitis dataset), where each time we used as input into the learner the $1 \dots 19$ features in the order they were selected (see Table 4 for the first seven iterative steps of the FS algorithms). Figure 7 presents the out of sample results following 10-fold cross validation with 1000 repetitions. With the exception of the LASSO and GSO that do not perform well on this dataset, the remaining FS schemes are not statistically significantly different with respect to misclassifying the response. Interestingly, if we use as input to the Naïve Bayes learner the features selected in Tuv et al. (2009) the resulting performance of the learner is still considerably worse compared to RRCT (it is similar to the results reported when the features are selected using the LASSO).

6 Discussion

We have developed a simple, yet efficient filter feature selection scheme which can be used as a benchmark against which more complicated algorithms can be compared. The proposed scheme works in both multi-class classification and regression settings, is extremely fast, and outperforms alternative simple feature selection schemes which are widely used.

The new algorithm relies on the computation of correlation coefficients, which are subsequently transformed using a nonlinear relationship inspired from the analytic expression linking linear correlation coefficients and mutual information when the underlying distribution is normal. Prior transformation of the datasets using the Box-Cox transform appears to be particularly constructive in some datasets, since then the densities tend to become more Gaussian. Nevertheless, in some cases the Box-Cox transformation leads to degraded performance in recovering the true variables in a dataset. This is an inherent problem of the maximum likelihood approach used for the computation of λ in Equation 14 in the presence of outliers; hence the application of more robust approaches to power transformations might be better in these cases (Marazzi and Yohai, 2006).

There are two approaches to validate FS algorithms in the literature: i) detect the correct feature subset, ii) improve a performance metric (for example classification accuracy) in the subsequent learning phase where the selected feature subsets are fed into a learner. Detecting the correct feature subset assumes knowledge of the ground truth, and typically requires the use of artificially generated datasets. Alternatively, as a proxy solution we can introduce artificial probes into real datasets, and test the FS algorithm on its ability to recover the explanatory variables and discard the probes. Evaluating an FS scheme on the basis of out of sample performance involves a learner, which complicates the assessment, since different learners could promote different FS schemes. Therefore, in this study we focused primarily on the former approach aiming to quantify the ability of the FS schemes at detecting the true feature subset.

We used some classical artificial datasets which have been widely used in FS studies. For those artificial datasets the ground truth (true features) is known, and therefore it is easy to evaluate the performance of each FS algorithm in terms of discarding probes. We have shown that mRMR_{ITL} and especially RRCT are very promising FS techniques, and may be enhanced in some datasets when they are combined with the Box-Cox transformation. Interestingly, they consistently outperformed the popular LASSO algorithm in both the TIED dataset and the four datasets created using Friedman's data generator. The mRMR_{MI} algorithm had one false alarm less than RRCT for the TIED dataset, but as we have seen RRCT quickly recovers the remaining two out of the three relevant explanatory variables in subsequent FS steps. Moreover, RRCT with Box-Cox is consistently better than mRMR_{MI} in the Friedman datasets.

In addition to the artificial datasets, we have used twelve diverse real datasets, with few instances and a large number of features (for example the hepatitis dataset), as well as a large number of instances and low number of features (for example the concrete compressive strength dataset). For the real datasets we appended 100 irrelevant features independently sampled from eight different distribution functions to perform a first series of tests (results shown in Figure 3). Some studies have sampled irrelevant features from the Gaussian distribution only, for example (Sun, Todorovic and Goodison, 2010). We have found that the use of only Gaussian probes gives better FDR results for all the algorithms, but is not as challenging for FS schemes. We believe that using a more generalized pool of distribution functions is a more accurate reflection of the performance of the new schemes in actual applications. Furthermore, we introduced 100 additional irrelevant predictors with identical distributions with the true predictors, randomly and independently sampling from each true predictor to build these new probes. The results for feature recovery are presented in Figure 4. Collectively, the results in Figures 3 and 4 indicate that the new schemes, mRMR_{ITL} and in particular RRCT are very competitive with widely used filters, showing consistently good performance in terms of recovering the true features.

The results in Figure 5 summarize the findings of this study and indicate that, on average, RRCT has an edge in detecting more true features compared to the competing FS algorithms. Whilst RRCT is not universally best in *all* the datasets we examined, it almost always ranks amongst the best approaches at recovering most of the explanatory variables in the original datasets, and rejecting the artificial probes. We believe this finding is particularly compelling, since the ability of alternative FS schemes to discard probes varies widely. Comparing the results in Figure 5b and 5c, we see that almost all the FS algorithms exhibit slight degradation in recovering the true features when presented with identical empirical probability distributions as the true features.

The success of RRCT can be attributed to two main factors: a) the nonlinear transformation of the rank and linear correlation coefficients inspired by IT considerations, and b) the integration of the concept of complementarity that quantifies the additional information content a feature exhibits, conditional upon an existing feature subset. The former assertion is verified by comparing RRCT with RRCT_0, whilst the latter claim is backed up evidentially by comparisons of RRCT with mRMR_{ITL}. Overall, the empirical findings of this study indicate that promoting or relegating features using the IT quantity (see Figure 1) by transforming the correlation coefficients is generally useful. Furthermore, the present study's findings agree with the recent research literature that emphasizes the need to account for feature complementarity (Meyer, Schretter and Bontempi, 2008; Brown 2009; Zhao and Liu, 2009). The use of the partial correlation coefficients to account for complementarity is a convenient way to overcome the problems associated with well known but often very difficult to assess metrics such as total correlation or conditional mutual information. Similarly to the correlation coefficients, the partial correlation coefficients use the IT inspired transformation which suppresses very low correlations amongst features towards predicting the response. This is in general useful because relatively low correlations may not actually reflect true structure in the dataset in terms of joint correlation of the features with the response. Conversely, the IT transformation further promotes relatively high joint correlations of the features.

A somewhat surprising finding is that we have obtained empirically better FDR results when we used the Spearman rank correlation coefficient to compute the relevance and conditional relevance terms whilst computing the redundancy using the linear correlation coefficient, as opposed to using either the linear or rank correlation coefficients exclusively. However, the empirical findings of this study (on the basis of five artificial datasets and twelve real datasets) suggest that this heuristic combination of the two correlation coefficients (rank for the relevance and conditional relevance terms, and linear for the redundancy term) appears to work surprisingly well in practice. We remark that by definition the mRMR scheme is a *heuristic* approach trying to balance the benefit of including a feature (relevance) against the disadvantage of including a feature that has overlapping information with the existing feature subset. This had already been suggested by Battiti (1994) and lately Brown has urged for the exploration of different trade-offs of relevance and redundancy in mRMR-type approaches. This can be achieved using "regularization terms" (free parameters) such as β in Equation 1 (Brown, 2009). Hence, it is conceivable that quantifying the relevance and conditional relevance using a simple nonlinear monotonic correlation and a linear correlation for redundancy followed by IT transformation could work well in the form of Equation 13. Indeed, the trade-off of relevance, redundancy, and complementarity could be parameterized by introducing "regularization parameters" that introduce some extra degrees of freedom, as suggested by Brown (2009).

RRCT has the desirable characteristic that it explicitly considers complementarities of the order equivalent to the number of features selected until that iterative step as part of its selection process, and is very easy to compute relying on correlation coefficients and partial correlation coefficients. The price to pay for this is that RRCT cannot quantify arbitrary relationships between features and the response variable as some complicated FS schemes relying on MI do. It would be particularly interesting to compare the performance of RRCT against more complicated filters that compute the MI using Parzen windows or Frazer's algorithm, as for example in Estevez et al. (2009). Preliminary results suggest that mRMR where MI is computed using kernel density estimates may be superior to the FS algorithms proposed here (as expected), but this is computationally extremely demanding, particularly for large datasets. In sharp contrast, the algorithms in this study are all computationally extremely efficient (each takes a few seconds of computational time compared to many hours with Parzen-window based mRMR-type schemes).

We envisage the proposed FS method finding use as a fast, simple, *off the shelf* feature selection algorithm in both multi-class classification and regression application problems.

Acknowledgements

A. Tsanas gratefully acknowledges the financial support of Intel Corporation and the Engineering and Physical Sciences Research Council (EPSRC).

References

R. Battiti. Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on Neural Networks*, 5:537–550, 1994

C.M. Bishop, *Pattern recognition and machine learning*, Springer, 2007

G. E. P. Box, D. R. Cox. An Analysis of Transformation. *Journal of the Royal Statistical Society. Series B (Methodological)*, 26:211-252, 1964

L. Breiman. Statistical modelling: the two cultures. *Statistical Science*, 16:199-231 (with comments and discussion), 2001

G. Brown. A New Perspective for Information Theoretic Feature Selection. In *12th International Conference on Artificial Intelligence and Statistics,* pages 49-54, Florida, June 2009

T. Cover and J. Thomas, *Elements of information theory*, Wiley-Blackwell, 2nd edition, 2006

J. Czerniak, H. Zarzycki. Application of rough sets in the presumptive diagnosis of urinary system diseases. Artifical Inteligence and Security in Computing Systems, *9th International Conference Proceedings*, Kluwer Academic Publishers, pages 41-51, 2003

P. Diaconis and B. Efron. Computer intensive methods in statics. *Scientific American*, 248:116-131, 1983

D. Donoho. For most large underdetermined systems of equations, the minimal L1-norm near-solution approximates the sparsest near-solution. *Communications Pure and Applied Mathematics*, 59:904-934, 2006

B. Efron, T. Hastie, I. Johnstone, R. Tibshirani. Least angle regression. *Annals of Statistics*, 32:407-499, 2004

P. Estevez, M. Tesmer, C. A. Perez, J. M. Zurada. Normalized mutual information feature selection. *IEEE Transactions on Neural Networks*, 20:189-201, 2009

J. Friedman. Greedy function approximation: a gradient boosting machine. Technical report, Department of Statistics, Stanford University, 1999

I. Guyon and A. Eliseeff. An introduction to variable and feature selection, *Journal of Machine Learning Research*, 3:1157-1182, 2003

I. Guyon, S. Gunn, M. Nikravesh, L. A. Zadeh (Eds.), *Feature Extraction: Foundations and Applications*, Springer, 2006

I. Guyon, J. Li, T. Mader, P. A. Pletscher, G. Schneider, M. Uhr. Competitive baseline methods set new standards for the NIPS 2003 feature selection benchmark. *Pattern Recognition Letters*, 28:1438-1444, 2007

I. Guyon. Practical feature selection: from correlation to causality. In *Mining Massive Data Sets for Security*. IOS Press, 2008. Available online at http://eprints.pascal-network.org/archive/00004038/01/PracticalFS.pdf

I. Guyon, A. Saffari, G. Dror, G. Cawley. Model Selection: Beyond the Bayesian/Frequentist Divide. *Journal of Machine Learning Research*, 11:61-87, 2010

T. Hastie, R. Tibshirani, J. Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2nd ed., 2009

M. Hilario and A. Kalousis. Approaches to Dimensionality Reduction in Proteomic Biomarker Studies. *Briefings in Bioinformatics*, 9:102-118, 2008.

H. Liu and L. Yu. Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions Knowledge and Data Engineering*, 17:491-502, 2005

A. Marazzi, V. J. Yohai. Robust Box–Cox transformations based on minimum residual autocorrelation. *Computational Statistics & Data Analysis*, 50:2752-2768, 2006

N. Meinshausen, and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics*, 37:246-270, 2009

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

H. Peng, F. Long, 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*, 27:1226-1238, 2005

J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1:81-106, 1986

A. Statnikov, and C.F. Aliferis. TIED: An Artificially Simulated Dataset with Multiple Markov Boundaries. In *Journal of Machine Learning Research Workshop and Conference Proceedings*, Vol. 6: Causality: Objectives and Assessment (NIPS 2008), pages 249-256, 2009

H. Stoppiglia, G. Dreyfus, R. Dubois, and Y. Oussar. Ranking a random feature for variable and feature selection. *Journal of Machine Learning Research*, 3:1399–1414, 2003

Y. Sun, S. Todorovic and S. Goodison, Local learning based feature selection for high dimensional data analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32:1610-1626, 2010

K. Torkkola. Feature extraction by non-parametric mutual information maximization. *Journal of Machine Learning Research*, 3:1415-1438, 2003

R. Tibshirani. Regression Shrinkage and Selection via the LASSO, *J. R. Statist. Soc. B*, 58:267-288, 1996

A. Tsanas, M.A. Little, P.E. McSharry, L.O. Ramig. Accurate telemonitoring of Parkinson's disease progression using non-invasive speech tests, *IEEE Transactions Biomedical Engineering*, 57:884-893, 2010a

A. Tsanas, M.A. Little, P.E. McSharry, L.O. Ramig. Enhanced classical dysphonia measures and sparse regression for telemonitoring of Parkinson's disease progression. *IEEE Signal Processing Society, International Conference on Acoustics, Speech and Signal Processing* (ICASSP '10), pages 594-597, Dallas, Texas, US, 2010b

A. Tsanas, M.A. Little, P.E. McSharry, L.O. Ramig. Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson's disease symptom severity, *Journal of the Royal Society Interface*, 2010c forthcoming (doi:10.1098/rsif.2010.0456)

E. Tuv, A. Borisov, G. Runger and K. Torkkola. Feature Selection with Ensembles, Artificial Variables, and Redundancy Elimination. *Journal of Machine Learning Research*, 10:1341-1366, 2009

W. H. Wolberg and O.L. Mangasarian. Multisurface method of pattern separation for medical diagnosis applied to breast cytology. *Proceedings of the National Academy of Sciences*, 87:9193-9196, 1990

L. Yu and H. Liu. Efficient Feature Selection via Analysis of Relevance and Redundancy. *Journal of Machine Learning Research*, 5:1205–1224, 2004

Z. Zhao and H. Liu. Searching for interacting features in subset selection. *Intelligent Data Analysis*, 13:207-228, 2009