Journal of Machine Learning Research 13 (2012) 27-66 Submitted 12/10; Revised 6/11; Published 1/12

Conditional Likelihood Maximisation: A Unifying Framework for Information Theoretic Feature Selection

Gavin Brown GAVIN.BROWN@CS.MANCHESTER.AC.UK
Adam Pocock ADAM.POCOCK@CS.MANCHESTER.AC.UK
Ming-Jie Zhao MING-JIE.ZHAO@CS.MANCHESTER.AC.UK
Mikel Luján MIKEL.LUJAN@CS.MANCHESTER.AC.UK
School of Computer Science
University of Manchester
Manchester M13 9PL, UK

Editor: Isabelle Guyon

Abstract

We present a unifying framework for information theoretic feature selection, bringing almost two decades of research on heuristic filter criteria under a single theoretical interpretation. This is in response to the question: *"what are the implicit statistical assumptions of feature selection criteria based on mutual information?"*. To answer this, we adopt a different strategy than is usual in the feature selection literature—instead of trying to *define* a criterion, we *derive* one, directly from a clearly specified objective function: the conditional likelihood of the training labels. While many hand-designed heuristic criteria try to optimize a definition of feature 'relevancy' and 'redundancy', our approach leads to a probabilistic framework which naturally incorporates these concepts. As a result we can unify the numerous criteria published over the last two decades, and show them to be low-order approximations to the exact (but intractable) optimisation problem. The primary contribution is to show that *common heuristics for information based feature selection (including Markov Blanket algorithms as a special case) are approximate iterative maximisers of the conditional likelihood.* A large empirical study provides strong evidence to favour certain classes of criteria, in particular those that balance the relative size of the relevancy/redundancy terms. Overall we conclude that the JMI criterion (Yang and Moody, 1999; Meyer et al., 2008) provides the best tradeoff in terms of accuracy, stability, and flexibility with small data samples.

Keywords: feature selection, mutual information, conditional likelihood

1. Introduction

High dimensional data sets are a significant challenge for Machine Learning. Some of the most practically relevant and high-impact applications, such as *gene expression* data, may easily have more than 10,000 features. Many of these features may be completely *irrelevant* to the task at hand, or *redundant* in the context of others. Learning in this situation raises important issues, for example, over-fitting to irrelevant aspects of the data, and the computational burden of processing many similar features that provide redundant information. It is therefore an important research direction to automatically identify meaningful smaller subsets of these variables, that is, *feature selection*.

Feature selection techniques can be broadly grouped into approaches that are classifier-dependent ('wrapper' and 'embedded' methods), and classifier-independent ('filter' methods). Wrapper meth-

ods search the space of feature subsets, using the training/validation accuracy of a particular classifier as the measure of utility for a candidate subset. This may deliver significant advantages in generalisation, though has the disadvantage of a considerable computational expense, and may produce subsets that are overly specific to the classifier used. As a result, any change in the learning model is likely to render the feature set suboptimal. Embedded methods (Guyon et al., 2006, Chapter 3) exploit the structure of specific classes of learning models to *guide* the feature selection process. While the defining component of a wrapper method is simply the search procedure, the defining component of an embedded method is a criterion derived through fundamental knowledge of a specific class of functions. An example is the method introduced by Weston et al. (2001), selecting features to minimize a generalisation bound that holds for Support Vector Machines. These methods are less computationally expensive, and less prone to overfitting than wrappers, but still use quite strict model structure assumptions. In contrast, *filter* methods (Duch, 2006) separate the classification and feature selection components, and define a heuristic *scoring criterion* to act as a proxy measure of the classification accuracy. Filters evaluate statistics of the data *independently* of any particular classifier, thereby extracting features that are generic, having incorporated few assumptions.

Each of these three approaches has its advantages and disadvantages, the primary distinguishing factors being speed of computation, and the chance of overfitting. In general, in terms of speed, filters are faster than embedded methods which are in turn faster than wrappers. In terms of overfitting, wrappers have higher learning capacity so are more likely to overfit than embedded methods, which in turn are more likely to overfit than filter methods. All of this of course changes with extremes of data/feature availability—for example, embedded methods will likely outperform filter methods in generalisation error as the number of datapoints increases, and wrappers become more computationally unfeasible as the number of features increases. A primary advantage of filters is that they are relatively cheap in terms of computational expense, and are generally more amenable to a theoretical analysis of their design. Such theoretical analysis is the focus of this article.

The defining component of a filter method is the *relevance index* (also known as a *selection/scoring criterion*), quantifying the 'utility' of including a particular feature in the set. Numerous hand-designed heuristics have been suggested (Duch, 2006), all attempting to maximise feature 'relevancy' and minimise 'redundancy'. However, few of these are motivated from a solid theoretical foundation. It is preferable to start from a more principled perspective—the desired approach is outlined eloquently by Guyon:

> *"It is important to start with a clean mathematical statement of the problem addressed [...] It should be made clear how optimally the chosen approach addresses the problem stated. Finally, the eventual approximations made by the algorithm to solve the optimisation problem stated should be explained. An interesting topic of research would be to 'retrofit' successful heuristic algorithms in a theoretical framework."* (Guyon et al., 2006, pg. 21)

In this work we adopt this approach—instead of trying to *define* feature relevance indices, we *derive* them starting from a clearly specified objective function. The objective we choose is a well accepted statistical principle, *the conditional likelihood of the class labels given the features*. As a result we are able to provide deeper insight into the feature selection problem, and achieve precisely the goal above, to retrofit numerous hand-designed heuristics into a theoretical framework.

2. Background

In this section we give a brief introduction to information theoretic concepts, followed by a summary of how they have been used to tackle the feature selection problem.

2.1 Entropy and Mutual Information

The fundamental unit of information is the *entropy* of a random variable, discussed in several standard texts, most prominently (Cover and Thomas, 1991). The entropy, denoted $H(X)$, quantifies the uncertainty present in the distribution of X. It is defined as,

$$H(X) = -\sum_{x\in\mathcal{X}} p(x)\log p(x),$$

where the lower case x denotes a possible value that the variable X can adopt from the alphabet $\mathcal{X}$. To compute[1] this, we need an estimate of the distribution $p(X)$. When X is discrete this can be estimated by frequency counts from data, that is $\hat{p}(x) = \frac{\#x}{N}$, the fraction of observations taking on value x from the total N. We provide more discussion on this issue in Section 3.3. If the distribution is highly biased toward one particular event $x \in \mathcal{X}$, that is, little uncertainty over the outcome, then the entropy is low. If all events are equally likely, that is, maximum uncertainty over the outcome, then $H(X)$ is maximal.[2] Following the standard rules of probability theory, entropy can be *conditioned* on other events. The *conditional entropy* of X given Y is denoted,

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

This can be thought of as the amount of uncertainty remaining in X after we learn the outcome of Y. We can now define the *Mutual Information* (Shannon, 1948) between X and Y, that is, the amount of information *shared* by X and Y, as follows:

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(xy)\log\frac{p(xy)}{p(x)p(y)}. \end{aligned}$$

This is the difference of two entropies—the uncertainty *before* Y is known, $H(X)$, and the uncertainty *after* Y is known, $H(X|Y)$. This can also be interpreted as the amount of uncertainty in X which is removed by knowing Y, thus following the intuitive meaning of mutual information as the amount of information that one variable provides about another. It should be noted that the Mutual Information is symmetric, that is, $I(X;Y) = I(Y;X)$, and is zero if and only if the variables are statistically independent, that is $p(xy) = p(x)p(y)$. The relation between these quantities can be seen in Figure 1. The Mutual Information can also be conditioned—the *conditional information* is,

$$\begin{aligned} I(X;Y|Z) &= H(X|Z) - H(X|YZ) \\ &= \sum_{z\in\mathcal{Z}} p(z)\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(xy|z)\log\frac{p(xy|z)}{p(x|z)p(y|z)}. \end{aligned}$$

1. The base of the logarithm is arbitrary, but decides the 'units' of the entropy. When using base 2, the units are 'bits', when using base e, the units are 'nats.'
2. In general, $0 \le H(X) \le \log(|\mathcal{X}|)$.

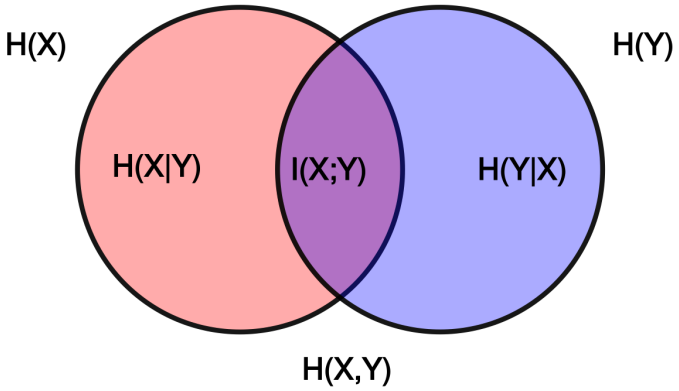


Figure 1: Illustration of various information theoretic quantities.

This can be thought of as the information still shared between X and Y after the value of a third variable, Z, is revealed. The conditional mutual information will emerge as a particularly important property in understanding the results of this work.

This section has briefly covered the principles of information theory; in the following section we discuss motivations for using it to solve the feature selection problem.

2.2 Filter Criteria Based on Mutual Information

Filter methods are defined by a criterion J, also referred to as a 'relevance index' or 'scoring' criterion (Duch, 2006), which is intended to measure how potentially useful a feature or feature subset may be when used in a classifier. An intuitive J would be some measure of correlation between the feature and the class label—the intuition being that a stronger correlation between these should imply a greater predictive ability when using the feature. For a class label Y, the *mutual information* score for a feature X_k is

$$J_{mim}(X_k) = I(X_k;Y). \tag{1}$$

This heuristic, which considers a score for each feature independently of others, has been used many times in the literature, for example, Lewis (1992). We refer to this feature scoring criterion as 'MIM', standing for *Mutual Information Maximisation*. To use this measure we simply rank the features in order of their MIM score, and select the top K features, where K is decided by some predefined need for a certain number of features or some other stopping criterion (Duch, 2006). A commonly cited justification for this measure is that the mutual information can be used to write both an upper and lower bound on the Bayes error rate (Fano, 1961; Hellman and Raviv, 1970). An important limitation is that this assumes that each feature is independent of all other features—and effectively ranks the features in descending order of their individual mutual information content. However, where features may be interdependent, this is known to be suboptimal. In general, it is widely accepted that a useful and parsimonious set of features should not only be individually *relevant*, but also should not be *redundant* with respect to each other—features should not be highly correlated. The reader is warned that while this statement seems appealingly intuitive, it is *not strictly correct*, as will be expanded upon in later sections. In spite of this, several criteria have

been proposed that attempt to pursue this 'relevancy-redundancy' goal. For example, Battiti (1994) presents the *Mutual Information Feature Selection* (MIFS) criterion:

$$J_{mifs}(X_k) = I(X_k;Y) - \beta \sum_{X_j \in S} I(X_k;X_j),$$

where S is the set of currently selected features. This includes the $I(X_k;Y)$ term to ensure feature *relevance*, but introduces a penalty to enforce low correlations with features already selected in S. Note that this assumes we are selecting features *sequentially*, iteratively constructing our final feature subset. For a survey of other search methods than simple sequential selection, the reader is referred to Duch (2006); however it should be noted that all theoretical results presented in this paper will be generally applicable to any search procedure, and based solely on properties of the criteria themselves. The β in the MIFS criterion is a configurable parameter, which must be set experimentally. Using $\beta = 0$ would be equivalent to $J_{mim}(X_k)$, selecting features independently, while a larger value will place more emphasis on reducing inter-feature dependencies. In experiments, Battiti found that $\beta = 1$ is often optimal, though with no strong theory to explain why. The MIFS criterion focuses on reducing *redundancy*; an alternative approach was proposed by Yang and Moody (1999), and also later by Meyer et al. (2008) using the *Joint Mutual Information* (JMI), to focus on increasing *complementary* information between features. The JMI score for feature X_k is

$$J_{jmi}(X_k) = \sum_{X_j \in S} I(X_k X_j;Y).$$

This is the information between the targets and a *joint* random variable $X_k X_j$, defined by pairing the candidate X_k with each feature previously selected. The idea is if the candidate feature is 'complementary' with existing features, we should include it.

The MIFS and JMI schemes were the first of many criteria that attempted to manage the relevance-redundancy tradeoff with various heuristic terms, however it is clear they have very different motivations. The criteria identified in the literature 1992-2011 are listed in Table 1. The practice in this research problem has been to *hand-design* criteria, piecing criteria together as a jigsaw of information theoretic terms—the overall aim to manage the relevance-redundancy trade-off, with each new criterion motivated from a different direction. Several questions arise here: Which criterion should we believe? What do they assume about the data? Are there other useful criteria, as yet undiscovered? In the following section we offer a novel perspective on this problem.

3. A Novel Approach

In the following sections we formulate the feature selection task as a conditional likelihood problem. We will demonstrate that precise links can be drawn between the well-accepted statistical framework of likelihood functions, and the current feature selection heuristics of mutual information criteria.

3.1 A Conditional Likelihood Problem

We assume an underlying i.i.d. process $p : X \to Y$, from which we have a sample of N observations. Each observation is a pair $(\mathbf{x}, y)$, consisting of a d-dimensional feature vector $\mathbf{x} = [x_1, ..., x_d]^T$, and a target class y, drawn from the underlying random variables $X = \{X_1, ..., X_d\}$ and Y. Furthermore, we assume that $p(y|\mathbf{x})$ is defined by a *subset* of the d features in $\mathbf{x}$, while the remaining features are

Criterion	*Full name*	*Authors*
MIM	Mutual Information Maximisation	Lewis (1992)
MIFS	Mutual Information Feature Selection	Battiti (1994)
KS	Koller-Sahami metric	Koller and Sahami (1996)
JMI	Joint Mutual Information	Yang and Moody (1999)
MIFS-U	MIFS-'Uniform'	Kwak and Choi (2002)
IF	Informative Fragments	Vidal-Naquet and Ullman (2003)
FCBF	Fast Correlation Based Filter	Yu and Liu (2004)
AMIFS	Adaptive MIFS	Tesmer and Estevez (2004)
CMIM	Conditional Mutual Info Maximisation	Fleuret (2004)
MRMR	Max-Relevance Min-Redundancy	Peng et al. (2005)
ICAP	Interaction Capping	Jakulin (2005)
CIFE	Conditional Infomax Feature Extraction	Lin and Tang (2006)
DISR	Double Input Symmetrical Relevance	Meyer and Bontempi (2006)
MINRED	Minimum Redundancy	Duch (2006)
IGFS	Interaction Gain Feature Selection	El Akadi et al. (2008)
SOA	Second Order Approximation	Guo and Nixon (2009)
CMIFS	Conditional MIFS	Cheng et al. (2011)

Table 1: Various information-based criteria from the literature. Sections 3 and 4 will show how these can all be interpreted in a single theoretical framework.

irrelevant. Our modeling task is therefore two-fold: firstly to identify the features that play a functional role, and secondly to use these features to perform predictions. In this work we concentrate on the first stage, that of selecting the relevant features.

We adopt a d-dimensional binary vector θ: a 1 indicating the feature is selected, a 0 indicating it is discarded. Notation $\mathbf{x}_\theta$ indicates the vector of selected features, that is, the full vector $\mathbf{x}$ projected onto the dimensions specified by θ. Notation $\mathbf{x}_{\tilde{\theta}}$ is the complement, that is, the unselected features. The full feature vector can therefore be expressed as $\mathbf{x} = \{\mathbf{x}_\theta, \mathbf{x}_{\tilde{\theta}}\}$. As mentioned, we assume the process p is defined by a subset of the features, so for some unknown optimal vector θ^*, we have that $p(y|\mathbf{x}) = p(y|\mathbf{x}_{\theta^*})$. We approximate p using a hypothetical predictive model q, with two layers of parameters: θ representing which features are selected, and τ representing parameters used to predict y. Our problem statement is to identify the minimal subset of features such that we *maximize the conditional likelihood of the training labels, with respect to these parameters.* For i.i.d. data $\mathcal{D} = \{(\mathbf{x}^i, y^i); i = 1..N\}$ the conditional likelihood of the labels given parameters $\{\theta, \tau\}$ is

$$\mathcal{L}(\theta, \tau|\mathcal{D}) = \prod_{i=1}^{N} q(y^i|\mathbf{x}^i_\theta, \tau).$$

The (scaled) conditional *log*-likelihood is

$$\ell = \frac{1}{N}\sum_{i=1}^{N} \log q(y^i|\mathbf{x}^i_\theta, \tau). \tag{2}$$

This is the error function we wish to optimize with respect to the parameters $\{\tau, \theta\}$; the scaling term has no effect on the optima, but simplifies exposition later. Using conditional likelihood has

become popular in so-called *discriminative* modelling applications, where we are interested only in the classification performance; for example Grossman and Domingos (2004) used it to learn Bayesian Network classifiers. We will expand upon this link to discriminative models in Section 9.3. Maximising conditional likelihood corresponds to minimising KL-divergence between the true and predicted class posterior probabilities—for classification, we often only require the *correct* class, and not precise estimates of the posteriors, hence Equation (2) is a proxy lower bound for classification accuracy.

We now introduce the quantity $p(y|\mathbf{x}_\theta)$: this is the true distribution of the class labels given the selected features $\mathbf{x}_\theta$. It is important to note the distinction from $p(y|\mathbf{x})$, the true distribution given *all* features. Multiplying and dividing q by $p(y|\mathbf{x}_\theta)$, we can re-write the above as,

$$\ell = \frac{1}{N}\sum_{i=1}^{N}\log\frac{q(y^i|\mathbf{x}^i_\theta,\tau)}{p(y^i|\mathbf{x}^i_\theta)} + \frac{1}{N}\sum_{i=1}^{N}\log p(y^i|\mathbf{x}^i_\theta). \tag{3}$$

The second term in (3) can be similarly expanded, introducing the probability $p(y|\mathbf{x})$:

$$\ell = \frac{1}{N}\sum_{i=1}^{N}\log\frac{q(y^i|\mathbf{x}^i_\theta,\tau)}{p(y^i|\mathbf{x}^i_\theta)} + \frac{1}{N}\sum_{i=1}^{N}\log\frac{p(y^i|\mathbf{x}^i_\theta)}{p(y^i|\mathbf{x}^i)} + \frac{1}{N}\sum_{i=1}^{N}\log p(y^i|\mathbf{x}^i).$$

These are finite sample approximations, drawing datapoints i.i.d. with respect to the distribution $p(\mathbf{x}y)$. We use $E_{\mathbf{x}y}\{\cdot\}$ to denote statistical expectation, and for convenience we negate the above, turning our maximisation problem into a minimisation. This gives us,

$$-\ell \approx E_{\mathbf{x}y}\Big\{\log\frac{p(y|\mathbf{x}_\theta)}{q(y|\mathbf{x}_\theta,\tau)}\Big\} + E_{\mathbf{x}y}\Big\{\log\frac{p(y|\mathbf{x})}{p(y|\mathbf{x}_\theta)}\Big\} - E_{\mathbf{x}y}\Big\{\log p(y|\mathbf{x})\Big\}. \tag{4}$$

These three terms have interesting properties which together define the feature selection problem. It is particularly interesting to note that the second term is *precisely* that introduced by Koller and Sahami (1996) in their definitions of optimal feature selection. In their work, the term was adopted ad-hoc as a sensible objective to follow—here we have shown it to be a direct and natural consequence of adopting the conditional likelihood as an objective function. Remembering $\mathbf{x} = \{\mathbf{x}_\theta, \mathbf{x}_{\tilde{\theta}}\}$, this second term can be developed:

$$\begin{aligned}
\Delta_{KS} &= E_{\mathbf{x}y}\Big\{\log\frac{p(y|\mathbf{x})}{p(y|\mathbf{x}_\theta)}\Big\} \\
&= \sum_{\mathbf{x}y} p(\mathbf{x}y)\log\frac{p(y|\mathbf{x}_\theta\mathbf{x}_{\tilde{\theta}})}{p(y|\mathbf{x}_\theta)} \\
&= \sum_{\mathbf{x}y} p(\mathbf{x}y)\log\frac{p(y|\mathbf{x}_\theta\mathbf{x}_{\tilde{\theta}})}{p(y|\mathbf{x}_\theta)}\frac{p(\mathbf{x}_{\tilde{\theta}}|\mathbf{x}_\theta)}{p(\mathbf{x}_{\tilde{\theta}}|\mathbf{x}_\theta)} \\
&= \sum_{\mathbf{x}y} p(\mathbf{x}y)\log\frac{p(\mathbf{x}_{\tilde{\theta}}y|\mathbf{x}_\theta)}{p(\mathbf{x}_{\tilde{\theta}}|\mathbf{x}_\theta)p(y|\mathbf{x}_\theta)} \\
&= I(X_{\tilde{\theta}};Y|X_\theta).
\end{aligned} \tag{5}$$

This is the conditional mutual information between the class label and the remaining features, given the selected features. We can note also that the third term in (4) is another information theoretic

quantity, the conditional entropy $H(Y|X)$. In summary, we see that our objective function can be decomposed into three distinct terms, each with its own interpretation:

$$\lim_{N\to\infty} -\ell \;=\; E_{\mathbf{x}y}\Big\{ \log \frac{p(y|\mathbf{x}_\theta)}{q(y|\mathbf{x}_\theta,\tau)} \Big\} + I(X_{\tilde{\theta}};Y|X_\theta) + H(Y|X). \tag{6}$$

The first term is a likelihood ratio between the true and the predicted class distributions given the selected features, averaged over the input space. The size of this term will depend on how well the model q can approximate p, given the supplied features.[3] When θ takes on the true value θ^* (or consists of a superset of θ^*) this becomes a KL-divergence $p||q$. The second term is $I(X_{\tilde{\theta}};Y|X_\theta)$, the conditional mutual information between the class label and the unselected features, given the selected features. The size of this term depends solely on the choice of features, and will decrease as the selected feature set X_θ explains more about Y, until eventually becoming zero when the remaining features $X_{\tilde{\theta}}$ contain no additional information about Y in the context of X_θ. It can be noted that due to the chain rule, we have

$$I(X;Y) = I(X_\theta;Y) + I(X_{\tilde{\theta}};Y|X_\theta),$$

hence minimizing $I(X_{\tilde{\theta}};Y|X_\theta)$ is equivalent to maximising $I(X_\theta;Y)$. The final term is $H(Y|X)$, the conditional entropy of the labels given *all features*. This term quantifies the uncertainty still remaining in the label even when we know *all possible* features; it is an irreducible constant, independent of all parameters, and in fact forms a bound on the Bayes error (Fano, 1961).

These three terms make explicit the effect of the feature selection parameters θ, separating them from the effect of the parameters τ in the model that *uses* those features. If we somehow had the optimal feature subset θ^*, which perfectly captured the underlying process p, then $I(X_{\tilde{\theta}};Y|X_\theta)$ would be zero. The remaining (reducible) error is then down to the KL divergence $p||q$, expressing how well the predictive model q can *make use* of the provided features. Of course, different models q will have different predictive ability: a good feature subset will not necessarily be put to good use if the model is too simple to express the underlying function. This perspective was also considered by Tsamardinos and Aliferis (2003), and earlier by Kohavi and John (1997)—the above results place these in the context of a precise objective function, the conditional likelihood. For the remainder of the paper we will use the same assumption as that made implicitly by *all* filter selection methods. For completeness, here we make the assumption explicit:

Definition 1 : Filter assumption
Given an objective function for a classifier, we can address the problems of optimizing the feature set and optimizing the classifier in two stages: first picking good features, then building the classifier to use them.

This implies that the second term in (6) can be optimized independently of the first. In this section we have formulated the feature selection task as a conditional likelihood problem. In the following, we consider how this problem statement relates to the existing literature, and discuss how to solve it in practice: including how to optimize the feature selection parameters, and the estimation of the necessary distributions.

3. In fact, if q is a *consistent* estimator, this term will approach zero with large N.

3.2 Optimizing the Feature Selection Parameters

Under the filter assumption in Definition 1, Equation (6) demonstrates that the optima of the conditional likelihood coincide with that of the conditional mutual information:

$$\arg\max_{\theta} \mathcal{L}(\theta|\mathcal{D}) = \arg\min_{\theta} I(X_{\tilde{\theta}}; Y|X_\theta). \tag{7}$$

There may of course be multiple global optima, in addition to the trivial minimum of selecting all features. With this in mind, we can introduce a minimality constraint on the size of the feature set, and define our problem:

$$\theta^* = \arg\min_{\theta'} \{|\theta'| : \theta' = \arg\min_{\theta} I(X_{\tilde{\theta}}; Y|X_\theta)\}. \tag{8}$$

This is the smallest feature set X_θ, such that the mutual information $I(X_{\tilde{\theta}}; Y|X_\theta)$ is minimal, and thus the conditional likelihood is maximal. It should be remembered that the likelihood is only our proxy for classification error, and the minimal feature set in terms of classification could be smaller than that which optimises likelihood. In the following paragraphs, we consider how this problem is implicitly tackled by methods already in the literature.

A common heuristic approach is a sequential search considering features one-by-one for addition/removal; this is used for example in Markov Blanket learning algorithms such as IAMB (Tsamardinos et al., 2003). We will now demonstrate that this sequential search heuristic is in fact equivalent to a greedy iterative optimisation of Equation (8). To understand this we must time-index the feature sets. Notation $X_{\theta^t}/X_{\tilde{\theta}^t}$ indicates the selected and unselected feature sets at timestep t—with a slight abuse of notation treating these interchangeably as sets and random variables.

Definition 2 : Forward Selection Step with Mutual Information
The forward selection step adds the feature with the maximum mutual information in the context of the currently selected set X_{θ^t}. The operations performed are:

$$\begin{aligned} X_k &= \arg\max_{X_k \in X_{\tilde{\theta}^t}} I(X_k; Y|X_{\theta^t}), \\ X_{\theta^{t+1}} &\leftarrow X_{\theta^t} \cup X_k, \\ X_{\tilde{\theta}^{t+1}} &\leftarrow X_{\tilde{\theta}^t} \setminus X_k. \end{aligned}$$

A subtle (but important) implementation point for this selection heuristic is that it should *not* add another feature if $\forall X_k,\ I(X_k; Y|X_\theta) = 0$. This ensures we will not unnecessarily increase the size of the feature set.

Theorem 3 *The forward selection mutual information heuristic adds the feature that generates the largest possible increase in the conditional likelihood—a greedy iterative maximisation.*

Proof With the definitions above and the chain rule of mutual information, we have that:

$$I(X_{\tilde{\theta}^{t+1}}; Y|X_{\theta^{t+1}}) = I(X_{\tilde{\theta}^t}; Y|X_{\theta^t}) - I(X_k; Y|X_{\theta^t}).$$

The feature X_k that *maximises* $I(X_k; Y|X_{\theta^t})$ is the same that *minimizes* $I(X_{\tilde{\theta}^{t+1}}; Y|X_{\theta^{t+1}})$; therefore the forward step is a greedy *minimization* of our objective $I(X_{\tilde{\theta}}; Y|X_\theta)$, and therefore maximises the conditional likelihood. ■

Definition 4 : Backward Elimination Step with Mutual Information
In a backward step, a feature is removed—the utility of a feature X_k is considered as its mutual information with the target, conditioned on all other elements of the selected set without X_k. The operations performed are:

$$\begin{aligned} X_k &= \arg\min_{X_k \in X_{\theta^t}} I(X_k; Y | \{X_{\theta^t} \setminus X_k\}). \\ X_{\theta^{t+1}} &\leftarrow X_{\theta^t} \setminus X_k \\ X_{\widetilde{\theta}^{t+1}} &\leftarrow X_{\widetilde{\theta}^t} \cup X_k \end{aligned}$$

Theorem 5 *The backward elimination mutual information heuristic removes the feature that causes the minimum possible decrease in the conditional likelihood.*

Proof With these definitions and the chain rule of mutual information, we have that:

$$I(X_{\widetilde{\theta}^{t+1}}; Y | X_{\theta^{t+1}}) = I(X_{\widetilde{\theta}^t}; Y | X_{\theta^t}) + I(X_k; Y | X_{\theta^{t+1}}).$$

The feature X_k that *minimizes* $I(X_k; Y | X_{\theta^{t+1}})$ is that which keeps $I(X_{\widetilde{\theta}^{t+1}}; Y | X_{\theta^{t+1}})$ as close as possible to $I(X_{\widetilde{\theta}^t}; Y | X_{\theta^t})$; therefore the backward elimination step removes a feature while attempting to maintain the likelihood as close as possible to its current value. ■

To strictly achieve our optimization goal, a backward step should *only* remove a feature if $I(X_k; Y | \{X_{\theta^t} \setminus X_k\}) = 0$. In practice, working with real data, there will likely be estimation errors (see the following section) and thus very rarely the strict zero will be observed. This brings us to an interesting corollary regarding IAMB (Tsamardinos and Aliferis, 2003).

Corollary 6 *Since the IAMB algorithm uses precisely these forward/backward selection heuristics, it is a greedy iterative maximisation of the conditional likelihood. In IAMB, a backward elimination step is only accepted if $I(X_k; Y | \{X_{\theta^t} \setminus X_k\}) \approx 0$, and otherwise the procedure terminates.*

In Tsamardinos and Aliferis (2003) it is shown that IAMB returns the Markov Blanket of any target node in a Bayesian network, and that this set coincides with the strongly relevant features in the definitions from Kohavi and John (1997). The precise links to this literature are explored further in Section 7. The IAMB family of algorithms adopt a common assumption, that the data is *faithful* to some unknown Bayesian Network. In the cases where this assumption holds, the procedure was proven to identify the unique Markov Blanket. Since IAMB uses precisely the forward/backward steps we have derived, we can conclude that *the Markov Blanket coincides with the (unique) maximum of the conditional likelihood function.* A more recent variation of the IAMB algorithm, called MMMB (Min-Max Markov Blanket) uses a series of optimisations to mitigate the requirement of exponential amounts of data to estimate the relevant statistical quantities. These optimisations do not change the underlying behaviour of the algorithm, as it still maximises the conditional likelihood for the selected feature set, however they do slightly obscure the strong link to our framework.

3.3 Estimation of the Mutual Information Terms

In considering the forward/backward heuristics, we must take account of the fact that we do not have perfect knowledge of the mutual information. This is because we have implicitly assumed we have access to the true distributions $p(\mathbf{x}y),\ p(y|\mathbf{x}_\theta)$, etc. In practice we have to estimate these from data. The problem calculating mutual information reduces to that of *entropy estimation*, and is fundamental in statistics (Paninski, 2003). The mutual information is defined as the expected logarithm of a ratio:

$$I(X;Y) = E_{xy}\Big\{ \log \frac{p(xy)}{p(x)p(y)} \Big\}.$$

We can estimate this, since the Strong Law of Large Numbers assures us that the sample estimate using $\hat{p}$ converges *almost surely* to the expected value—for a dataset of N i.i.d. observations (x^i, y^i),

$$I(X;Y) \approx \hat{I}(X;Y) = \frac{1}{N}\sum_{i=1}^{N} \log \frac{\hat{p}(x^i y^i)}{\hat{p}(x^i)\hat{p}(y^i)}.$$

In order to calculate this we need the estimated distributions $\hat{p}(xy)$, $\hat{p}(x)$, and $\hat{p}(y)$. The computation of entropies for continuous or ordinal data is highly non-trivial, and requires an assumed model of the underlying distributions—to simplify experiments throughout this article, we use discrete data, and estimate distributions with *histogram estimators* using fixed-width bins. The probability of any particular event $p(X = x)$ is estimated by maximum likelihood, the frequency of occurrence of the event $X = x$ divided by the total number of events (i.e., datapoints). For more information on alternative entropy estimation procedures, we refer the reader to Paninski (2003).

At this point we must note that the approximation above holds *only* if N is large *relative to the dimension of the distributions over x and y.* For example if x,y are binary, $N \approx 100$ should be more than sufficient to get reliable estimates; however if x,y are multinomial, this will likely be insufficient. In the context of the sequential selection heuristics we have discussed, we are approximating $I(X_k;Y|X_\theta)$ as,

$$I(X_k;Y|X_\theta) \approx \hat{I}(X_k;Y|X_\theta) = \frac{1}{N}\sum_{i=1}^{N} \log \frac{\hat{p}(x_k^i y^i|\mathbf{x}_\theta^i)}{\hat{p}(x_k^i|\mathbf{x}_\theta^i)\hat{p}(y^i|\mathbf{x}_\theta^i)}. \tag{9}$$

As the dimension of the variable X_θ grows (i.e., as we add more features) then the necessary probability distributions become more high dimensional, and hence our estimate of the mutual information becomes less reliable. This in turn causes increasingly poor judgements for the inclusion/exclusion of features. For precisely this reason, the research community have developed various low-dimensional approximations to (9). In the following sections, we will investigate the implicit statistical assumptions and empirical effects of these approximations.

In the remainder of this paper, we use $I(X;Y)$ to denote the ideal case of being able to compute the mutual information, though in practice on real data we use the finite sample estimate $\hat{I}(X;Y)$.

3.4 Summary

In these sections we have in effect *reverse-engineered* a mutual information-based selection scheme, starting from a clearly defined conditional likelihood problem, and discussed estimation of the various quantities involved. In the following sections we will show that we can retrofit numerous existing relevancy-redundancy heuristics from the feature selection literature into this probabilistic framework.

4. Retrofitting Successful Heuristics

In the previous section, starting from a clearly defined conditional likelihood problem, we derived a greedy optimization process which assesses features based on a simple scoring criterion on the utility of including a feature $X_k \in X_{\widetilde{\theta}}$. The score for a feature X_k is,

$$J_{cmi}(X_k) = I(X_k;Y|S), \tag{10}$$

where *cmi* stands for conditional mutual information, and for notational brevity we now use $S = X_\theta$ for the currently selected set. An important question is, how does (10) relate to existing heuristics in the literature, such as MIFS? We will see that MIFS, and certain other criteria, can be phrased cleanly as *linear combinations* of Shannon entropy terms, while some are non-linear combinations, involving *max* or *min* operations.

4.1 Criteria as Linear Combinations of Shannon Information Terms

Repeating the MIFS criterion for clarity,

$$J_{mifs}(X_k) = I(X_k;Y) - \beta \sum_{X_j \in S} I(X_k;X_j). \tag{11}$$

We can see that we first need to rearrange (10) into the form of a simple relevancy term between X_k and Y, plus some additional terms, before we can compare it to MIFS. Using the identity $I(A;B|C) - I(A;B) = I(A;C|B) - I(A;C)$, we can re-express (10) as,

$$J_{cmi}(X_k) = I(X_k;Y|S) = I(X_k;Y) - I(X_k;S) + I(X_k;S|Y). \tag{12}$$

It is interesting to see terms in this expression corresponding to the concepts of 'relevancy' and 'redundancy', that is, $I(X_k;Y)$ and $I(X_k;S)$. The score will be increased if the relevancy of X_k is large and the redundancy with existing features is small. This is in accordance with a common view in the feature selection literature, observing that we wish to avoid redundant variables. However, we can also see an important additional term $I(X_k;S|Y)$, which is not traditionally accounted for in the feature selection literature—we call this the *conditional redundancy*. This term has the opposite sign to the redundancy $I(X_k;S)$, hence J_{cmi} will be increased when this is large, that is, a strong class-conditional dependence of X_k with the existing set S. Thus, we come to the important conclusion that *the inclusion of correlated features can be useful*, provided the correlation *within classes* is stronger than the overall correlation. We note that this is a similar observation to that of Guyon et al. (2006), that "correlation does not imply redundancy"—Equation (12) effectively embodies this statement in information theoretic terms.

The sum of the last two terms in (12) represents the three-way interaction between the existing feature set S, the target Y, and the candidate feature X_k being considered for inclusion in S. To further understand this, we can note the following property:

$$I(X_kS;Y) = I(S;Y) + I(X_k;Y|S) = I(S;Y) + I(X_k;Y) - I(X_k;S) + I(X_k;S|Y).$$

We see that if $I(X_k;S) > I(X_k;S|Y)$, then the total utility when including X_k, that is $I(X_kS;Y)$, is *less* than the sum of the individual relevancies $I(S;Y) + I(X_k;Y)$. This can be interpreted as X_k having unnecessary duplicated information. In the opposite case, when $I(X_k;S) < I(X_k;S|Y)$, then X_k and

S combine well and provide more information *together* than by the sum of their parts, $I(S;Y)$, and $I(X_k;Y)$.

The important point to take away from this expression is that the terms are in a *trade-off*—we do not require a feature with low redundancy for its own sake, but instead require a feature that best trades off the three terms so as to maximise the score overall. Much like the bias-variance dilemma, attempting to decrease one term is likely to increase another.

The relation of (10) and (11) can be seen with assumptions on the underlying distribution $p(\mathbf{x}y)$. Writing the latter two terms of (12) as entropies:

$$\begin{aligned} J_{cmi}(X_k) &= I(X_k;Y) \\ &\quad - H(S) + H(S|X_k) \\ &\quad + H(S|Y) - H(S|X_kY). \end{aligned} \tag{13}$$

To develop this further, we require an assumption.

Assumption 1 *For all unselected features $X_k \in X_{\tilde{\theta}}$, assume the following,*

$$\begin{aligned} p(\mathbf{x}_\theta|x_k) &= \prod_{j\in S} p(x_j|x_k) \\ p(\mathbf{x}_\theta|x_ky) &= \prod_{j\in S} p(x_j|x_ky). \end{aligned}$$

This states that the selected features X_θ are independent and class-conditionally independent given the unselected feature X_k under consideration.

Using this, Equation (13) becomes,

$$\begin{aligned} J'_{cmi}(X_k) &= I(X_k;Y) \\ &\quad - H(S) + \sum_{j\in S} H(X_j|X_k) \\ &\quad + H(S|Y) - \sum_{j\in S} H(X_j|X_kY). \end{aligned}$$

where the prime on J indicates we are making assumptions on the distribution. Now, if we introduce $\sum_{j\in S} H(X_j) - \sum_{j\in S} H(X_j)$, and $\sum_{j\in S} H(X_j|Y) - \sum_{j\in S} H(X_j|Y)$, we recover mutual information terms, between the candidate feature and each member of the set S, plus some additional terms,

$$\begin{aligned} J'_{cmi}(X_k) &= I(X_k;Y) \\ &\quad - \sum_{j\in S} I(X_j;X_k) + \sum_{j\in S} H(X_j) - H(S) \\ &\quad + \sum_{j\in S} I(X_j;X_k|Y) - \sum_{j\in S} H(X_j|Y) + H(S|Y). \end{aligned} \tag{14}$$

Several of the terms in (14) are constant with respect to X_k—as such, removing them will have *no effect on the choice of feature*. Removing these terms, we have an equivalent criterion,

$$J'_{cmi}(X_k) = I(X_k;Y) - \sum_{j\in S} I(X_j;X_k) + \sum_{j\in S} I(X_j;X_k|Y). \tag{15}$$

This has in fact already appeared in the literature as a filter criterion, originally proposed by Lin and Tang (2006), as Conditional Infomax Feature Extraction (CIFE), though it has been repeatedly rediscovered by other authors (El Akadi et al., 2008; Guo and Nixon, 2009). It is particularly interesting as it represents a sort of 'root' criterion, from which several others can be derived. For example, the link to MIFS can be seen with one further assumption, that the features are pairwise class-conditionally independent.

Assumption 2 *For all features i, j, assume $p(x_i x_j|y) = p(x_i|y)p(x_j|y)$. This states that the features are pairwise class-conditionally independent.*

With this assumption, the term $\sum I(X_j;X_k|Y)$ will be zero, and (15) becomes (11), the MIFS criterion, with $\beta = 1$. The β parameter in MIFS can be interpreted as encoding a strength of belief in another assumption, that of unconditional independence.

Assumption 3 *For all features i, j, assume $p(x_i x_j) = p(x_i)p(x_j)$. This states that the features are pairwise independent.*

A β close to zero implies very strong belief in the independence statement, indicating that any measured association $I(X_j;X_k)$ is in fact spurious, possibly due to noise in the data. A β value closer to 1 implies a lesser belief, that any measured dependency $I(X_j;X_k)$ should be incorporated into the feature score exactly as observed. Since MIM is produced by setting $\beta = 0$, we can see that MIM also adopts Assumption 3. The same line of reasoning can be applied to a very similar criterion proposed by Peng et al. (2005), the *Minimum-Redundancy Maximum-Relevance* criterion,

$$J_{mrmr}(X_k) = I(X_k;Y) - \frac{1}{|S|}\sum_{j\in S} I(X_k;X_j).$$

Since mRMR omits the conditional redundancy term entirely, it is implicitly using Assumption 2. The β coefficient has been set inversely proportional to the size of the current feature set. If we have a large set S, then β will be extremely small. The interpretation is then that as the set S grows, mRMR adopts a stronger belief in Assumption 3. In the original paper, (Peng et al., 2005, Section 2.3) it was claimed that mRMR is equivalent to (10). In this section, through making explicit the intrinsic assumptions of the criterion, we have clearly illustrated that this claim is incorrect.

Balagani and Phoha (2010) present an analysis of the three criteria mRMR, MIFS and CIFE, arriving at similar results to our own: that these criteria make highly restrictive assumptions on the underlying data distributions. Though the conclusions are similar, our approach includes their results as a special case, and makes explicit the link to a likelihood function.

The relation of the MIFS/mRMR to Equation (15) is relatively straightforward. It is more challenging to consider how closely other criteria might be re-expressed in this form. Yang and Moody (1999) propose using *Joint Mutual Information* (JMI),

$$J_{jmi}(X_k) = \sum_{j\in S} I(X_k X_j;Y). \tag{16}$$

Using some relatively simple manipulations (see appendix) this can be re-written as,

$$J_{jmi}(X_k) = I(X_k;Y) - \frac{1}{|S|}\sum_{j\in S}\Big[I(X_k;X_j) - I(X_k;X_j|Y)\Big]. \tag{17}$$

This criterion (17) returns *exactly* the same set of features as the JMI criterion (16); however in this form, we can see the relation to our proposed framework. The JMI criterion, like mRMR, has a stronger belief in the pairwise independence assumptions as the feature set S grows. Similarities can of course be observed between JMI, MIFS and mRMR—the differences being the scaling factor and the conditional term—and their subsequent relation to Equation (15). It is in fact possible to identify numerous criteria from the literature that can all be re-written into a common form, corresponding to variations upon (15). A *space* of potential criteria can be imagined, where we parameterize criterion (15) as so:

$$J'_{cmi} = I(X_k;Y) - \beta \sum_{j \in S} I(X_j;X_k) + \gamma \sum_{j \in S} I(X_j;X_k|Y). \tag{18}$$

Figure 2 shows how the criteria we have discussed so far can all be fitted inside this unit square corresponding to β/γ parameters. MIFS sits on the left hand axis of the square—with $\gamma = 0$ and $\beta \in [0,1]$. The MIM criterion, Equation (1), which simply assesses each feature individually without any regard of others, sits at the bottom left, with $\gamma = 0, \beta = 0$. The top right of the square corresponds to $\gamma = 1, \beta = 1$, which is the CIFE criterion (Lin and Tang, 2006), also suggested by El Akadi et al. (2008) and Guo and Nixon (2009). A very similar criterion, using an assumption to approximate the terms, was proposed by Cheng et al. (2011).

The JMI and mRMR criteria are unique in that they *move linearly* within the space as the feature set S grows. As the size of the set S increases they move closer towards the origin and the MIM criterion. The particularly interesting point about this property is that the *relative magnitude* of the relevancy term to the redundancy terms stays approximately constant as S grows, whereas with MIFS, the redundancy term will in general be $|S|$ times bigger than the relevancy term. The consequences of this will be explored in the experimental section of this paper. Any criterion expressible in the unit square has made independence Assumption 1. In addition, any criteria that sit at points other than $\beta = 1, \gamma = 1$ have adopted varying degrees of belief in Assumptions 2 and 3.

A further interesting point about this square is simply that it is sparsely populated. An obvious unexplored region is the bottom right, the corner corresponding to $\beta = 0, \gamma = 1$; though there is no clear intuitive justification for this point, for completeness in the experimental section we will evaluate it, as the *conditional redundancy* or 'condred' criterion. In previous work (Brown, 2009) we explored this unit square, though derived from an expansion of the mutual information function rather than directly from the conditional likelihood. While this resulted in an identical expression to (18), the probabilistic framework we present here is far more expressive, allowing exact specification of the underlying assumptions.

The unit square of Figure 2 describes *linear* criteria, named as so since they are linear combinations of the relevance/redundancy terms. There exist other criteria that follow a similar form, but involving other operations, making them *non-linear*.

4.2 Criteria as Non-Linear Combinations of Shannon Information Terms

Fleuret (2004) proposed the *Conditional Mutual Information Maximization* criterion,

$$J_{cmim}(X_k) = \min_{X_j \in S} \Big[I(X_k;Y|X_j) \Big].$$

This can be re-written,

$$J_{cmim}(X_k) = I(X_k;Y) - \max_{X_j \in S} \Big[I(X_k;X_j) - I(X_k;X_j|Y) \Big]. \tag{19}$$

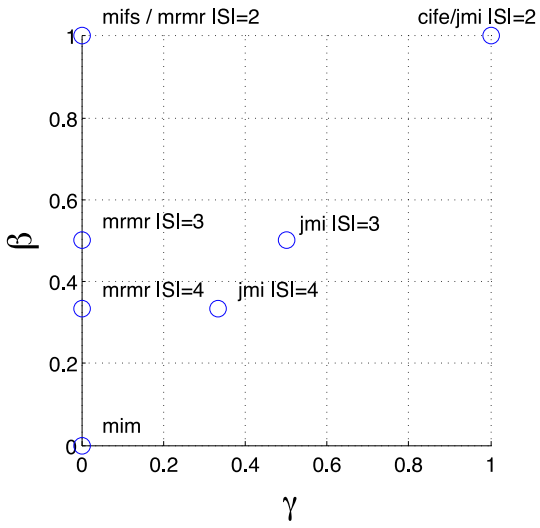


Figure 2: The full space of *linear* filter criteria, describing several examples from Table 1. Note that *all* criteria in this space adopt Assumption 1. Additionally, the γ and β axes represent the criteria belief in Assumptions 2 and 3, respectively. The left hand axis is where the mRMR and MIFS algorithms sit. The bottom left corner, MIM, is the assumption of completely independent features, using just marginal mutual information. Note that some criteria are equivalent at particular sizes of the current feature set $|S|$.

The proof is again available in the appendix. Due to the *max* operator, the probabilistic interpretation is a little less straightforward. It is clear however that CMIM adopts Assumption 1, since it evaluates only pairwise feature statistics.

Vidal-Naquet and Ullman (2003) propose another criterion used in Computer Vision, which we refer to as *Informative Fragments*,

$$J_{if}(X_k) = \min_{X_j \in S} \Big[I(X_k X_j; Y) - I(X_j; Y) \Big].$$

The authors motivate this criterion by noting that it measures the gain of combining a new feature X_k with each existing feature X_j, over simply using X_j by itself. The X_j with the least 'gain' from being paired with X_k is taken as the score for X_k. Interestingly, using the chain rule $I(X_k X_j; Y) = I(X_j; Y) + I(X_k; Y|X_j)$, therefore IF is equivalent to CMIM, that is, $J_{if}(X_k) = J_{cmim}(X_k)$, making the same assumptions. Jakulin (2005) proposed the criterion,

$$J_{icap}(X_k) = I(X_k; Y) - \sum_{X_j \in S} \max \Big[0, \{ I(X_k; X_j) - I(X_k; X_j|Y) \} \Big].$$

Again, this adopts Assumption 1, using the same redundancy and *conditional* redundancy terms, yet the exact probabilistic interpretation is unclear.

An interesting class of criteria use a normalisation term on the mutual information to offset the inherent bias toward high arity features (Duch, 2006). An example of this is *Double Input*

Symmetrical Relevance (Meyer and Bontempi, 2006), a modification of the JMI criterion:

$$J_{disr}(X_k) = \sum_{X_j \in S} \frac{I(X_k X_j; Y)}{H(X_k X_j Y)}.$$

The inclusion of this normalisation term breaks the strong theoretical link to a likelihood function, but again for completeness we will include this in our empirical investigations. While the criteria in the unit square can have their probabilistic assumptions made explicit, the nonlinearity in the CMIM, ICAP and DISR criteria make such an interpretation far more difficult.

4.3 Summary of Theoretical Findings

In this section we have shown that numerous criteria published over the past two decades of research can be 'retro-fitted' into the framework we have proposed—the criteria are approximations to (10), each making different assumptions on the underlying distributions. Since in the previous section we saw that accepting the top ranked feature according to (10) provides the maximum possible increase in the likelihood, we see now that the criteria are *approximate* maximisers of the likelihood. Whether or not they indeed provide the maximum increase at each step will depend on how well the implicit assumptions on the data can be trusted. Also, it should be remembered that even if we used (10), it is not guaranteed to find the global optimum of the likelihood, since (a) it is a greedy search, and (b) finite data will mean distributions cannot be accurately modelled. In this case, we have reached the limit of what a theoretical analysis can tell us about the criteria, and we must close the remaining 'gaps' in our understanding with an experimental study.

5. Experiments

In this section we empirically evaluate some of the criteria in the literature against one another. Note that we are not pursuing an exhaustive analysis, attempting to identify the 'winning' criterion that provides best performance overall[4]—rather, we primarily observe how the theoretical properties of criteria relate to the similarity of the returned feature sets. While these properties are interesting, we of course must acknowledge that classification performance is the ultimate evaluation of a criterion—hence we also include here classification results on UCI data sets and in Section 6 on the well-known benchmark NIPS Feature Selection Challenge.

In the following sections, we ask the questions: "how stable is a criterion to small changes in the training data set?", "how similar are the criteria to each other?", "how do the different criteria behave in limited and extreme small-sample situations?", and finally, "what is the relation between stability and accuracy?".

To address these questions, we use the 15 data sets detailed in Table 2. These are chosen to have a wide variety of example-feature ratios, and a range of multi-class problems. The features within each data set have a variety of characteristics—some binary/discrete, and some continuous. Continuous features were discretized, using an equal-width strategy into 5 bins, while features already with a categorical range were left untouched. The 'ratio' statistic quoted in the final column is an indicator of the difficulty of the feature selection for each data set. This uses the number of datapoints (N), the median arity of the features (m), and the number of classes (c)—the ratio quoted in

4. In any case, the No Free Lunch Theorem applies here also (Tsamardinos and Aliferis, 2003).

the table for each data set is $\frac{N}{mc}$, hence a smaller value indicates a more challenging feature selection problem.

A key point of this work is to understand the statistical assumptions on the data imposed by the feature selection criteria—if our classification model were to make even more assumptions, this is likely to obscure the experimental observations relating performance to theoretical properties. For this reason, in all experiments we use a simple nearest neighbour classifier ($k = 3$), this is chosen as it makes few (if any) assumptions about the data, and we avoid the need for parameter tuning. For the feature selection search procedure, the filter criteria are applied using a simple forward selection, to select a fixed number of features, specified in each experiment, before being used with the classifier.

Data	*Features*	*Examples*	*Classes*	Ratio
breast	30	569	2	57
congress	16	435	2	72
heart	13	270	2	34
ionosphere	34	351	2	35
krvskp	36	3196	2	799
landsat	36	6435	6	214
lungcancer	56	32	3	4
parkinsons	22	195	2	20
semeion	256	1593	10	80
sonar	60	208	2	21
soybeansmall	35	47	4	6
spect	22	267	2	67
splice	60	3175	3	265
waveform	40	5000	3	333
wine	13	178	3	12

Table 2: Data sets used in experiments. The final column indicates the difficulty of the data in feature selection, a smaller value indicating a more challenging problem.

5.1 How Stable are the Criteria to Small Changes in the Data?

The set of features selected by any procedure will of course depend on the data provided. It is a plausible complaint if the set of returned features varies wildly with only slight variations in the supplied data. This is an issue reminiscent of the *bias-variance dilemma*, where the sensitivity of a classifier to its initial conditions causes high variance responses. However, while the bias-variance decomposition is well-defined and understood, the corresponding issue for feature selection, the 'stability', has only recently been studied. The stability of a feature selection criterion requires a measure to quantify the 'similarity' between two selected feature sets. This was first discussed by Kalousis et al. (2007), who investigated several measures, with the final recommendation being the Tanimoto distance between sets. Such set-intersection measures seem appropriate, but have limitations; for example, if two criteria selected identical feature sets of size 10, we might be less surprised if we knew the overall pool of features was of size 12, than if it was size 12,000. To account

for this, Kuncheva (2007) presents a *consistency index*, based on the hypergeometric distribution with a correction for chance.

Definition 7 *The consistency for two subsets $A, B \subset X$, such that $|A| = |B| = k$, and $r = |A \cap B|$, where $0 < k < |X| = n$, is*

$$C(A,B) = \frac{rn - k^2}{k(n-k)}.$$

The consistency takes values in the range $[-1,+1]$, with a positive value indicating similar sets, a zero value indicating a purely random relation, and a negative value indicating a strong anti-correlation between the features sets.

One problem with the consistency index is that it does not take feature *redundancy* into account. That is, two procedures could select features which have different array indices, so are identified as 'different', but in fact are so highly correlated that they are effectively identical. A method to deal with this situation was proposed by Yu et al. (2008). This method constructs a weighted complete bipartite graph, where the two node sets correspond to two different feature sets, and weights are assigned to the arcs are the normalized mutual information between the features at the nodes, also sometimes referred to as the symmetrical uncertainty. The weight between node i in set A, and node j in set B, is

$$w(A(i),B(j)) = \frac{I(X_{A(i)};X_{B(j)})}{H(X_{A(i)}) + H(X_{B(j)})}.$$

The Hungarian algorithm is then applied to identify the maximum weighted matching between the two node sets, and the overall similarity between sets A and B is the final matching cost. This is the *information consistency* of the two sets. For more details, we refer to Yu et al. (2008).

We now compare these two measures on the criteria from the previous sections. For each data set, we take a bootstrap sample and select a set of features using each feature selection criterion. The (information) stability of a single criterion is quantified as the average pairwise (information) consistency across 50 bootstraps from the training data.

Figure 3 shows Kuncheva's stability measure on average over 15 data sets, selecting feature sets of size 10; note that the criteria have been displayed ordered left-to-right by their median value of stability over the 15 data sets. The marginal mutual information, MIM, is as expected the most stable, given that it has the lowest dimensional distribution to approximate. The next most stable is JMI which includes the relevancy/redundancy terms, but *averages* over the current feature set; this averaging process might therefore be interpreted empirically as a form of 'smoothing', enabling the criteria overall to be resistant to poor estimation of probability distributions. It can be noted that the far right of Figure 3 consists of the MIFS, ICAP and CIFE criteria, all of which do not attempt to average the redundancy terms.

Figure 4 shows the same data sets, but instead the *information stability* is computed; as mentioned, this should take into account the fact that some features are highly correlated. Interestingly, the two box-plots show broadly similar results. MIM is the most stable, and CIFE is the least stable, though here we see that JMI, DISR, and MRMR are actually more stable than Kuncheva's stability index can reflect. An interesting line of future research might be to combine the best of these two stability measures—one that can take into account both feature redundancy and a correction for random chance.

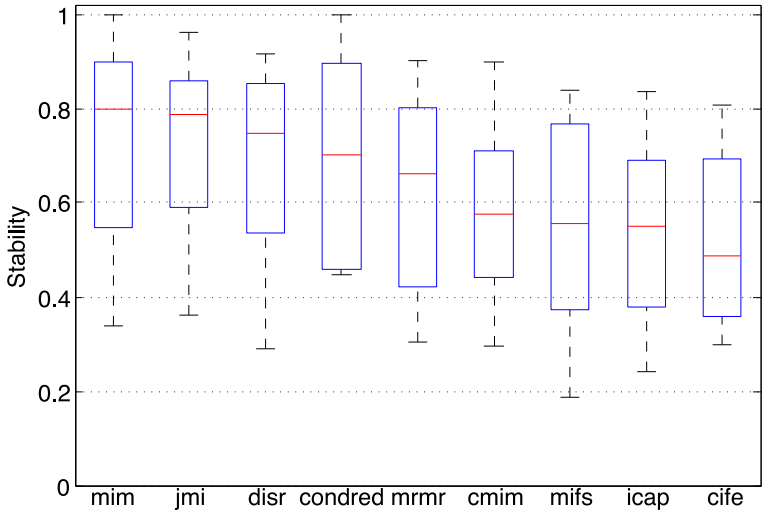


Figure 3: Kuncheva's Stability Index across 15 data sets. The box indicates the upper/lower quartiles, the horizontal line within each shows the median value, while the dotted crossbars indicate the maximum/minimum values. For convenience of interpretation, criteria on the x-axis are ordered by their median value.

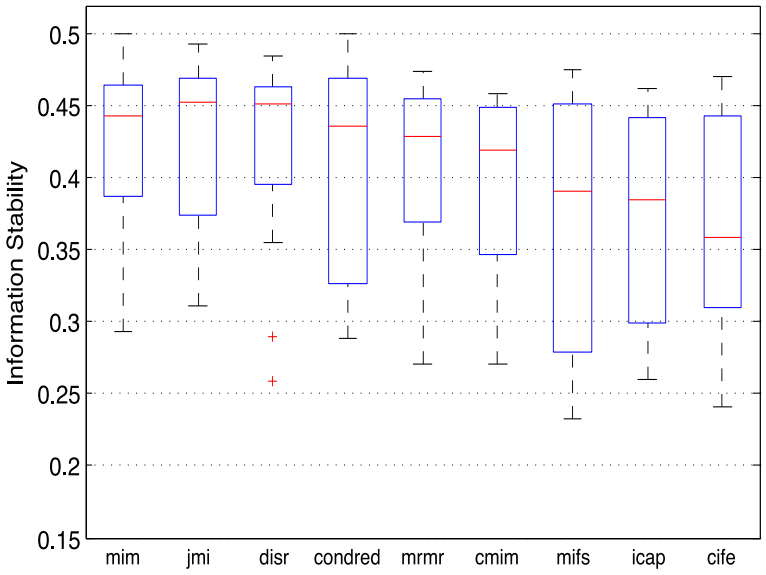
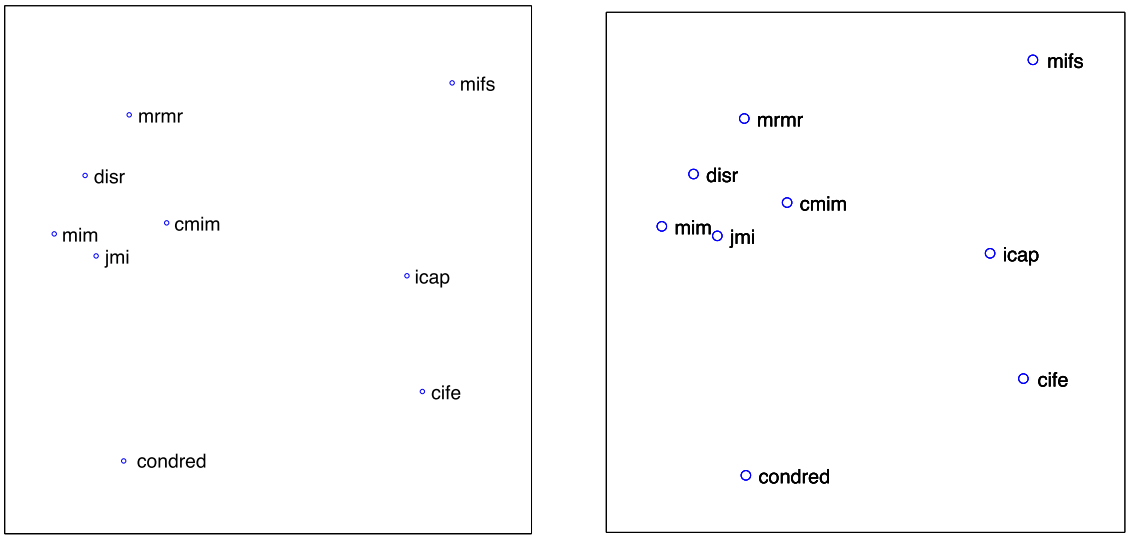


Figure 4: Yu et al's Information Stability Index across 15 data sets. For comparison, criteria on the x-axis are ordered identically to Figure 3. The general picture emerges similarly, though the information stability index is able to take feature redundancy into account, showing that some criteria are slightly more stable than expected.



(a) Kuncheva's Consistency Index.

(b) Yu et al's Information Stability Index.

Figure 5: Relations between feature sets generated by different criteria, on average over 15 data sets. 2-D visualisation generated by classical multi-dimensional scaling.

5.2 How Similar are the Criteria?

Two criteria can be directly compared with the same methodology: by measuring the consistency and information consistency between selected feature subsets on a common set of data. We calculate the mean consistencies between two feature sets of size 10, repeatedly selected over 50 bootstraps from the original data. This is then arranged in a similarity matrix, and we use classical multi-dimensional scaling to visualise this as a 2-d map, shown in Figures 5a and 5b. Note again that while the indices may return different absolute values (one is a normalized mean of a hypergeometric distribution and the other is a pairwise sum of mutual information terms) they show very similar relative 'distances' between criteria.

Both diagrams show a cluster of several criteria, and 4 clear outliers: MIFS, CIFE, ICAP and CondRed. The 5 criteria clustering in the upper left of the space appear to return relatively similar feature sets. The 4 outliers appear to return quite significantly different feature sets, both from the clustered set, and from each other. A common characteristic of these 4 outliers is that they do not scale the redundancy or conditional redundancy information terms. In these criteria, the upper bound on the redundancy term $\sum_{j \in S} I(X_k; X_j)$ grows linearly with the number of selected features, whilst the upper bound on the relevancy term $I(X_k; Y)$ remains constant. When this happens the relevancy term is overwhelmed by the redundancy term and thus the criterion selects features with minimal redundancy, rather than trading off between the two terms. This leads to strongly divergent feature sets being selected, which is reflected in the stability of the criteria. Each of the outliers are different from each other as they have different combinations of redundancy and conditional redundancy. We will see this 'balance' between relevancy and redundancy emerge as a common theme in the experiments over the next few sections.

5.3 How do Criteria Behave in Limited and Extreme Small-sample Situations?

To assess how criteria behave in data poor situations, we vary the number of datapoints supplied to perform the feature selection. The procedure was to randomly select 140 datapoints, then use the remaining data as a hold-out set. From this 140, the number provided to each criterion was increased in steps of 10, from a minimal set of size 20. To allow a reasonable testing set size, we limited this assessment to only data sets with at least 200 datapoints total; this gives us 11 data sets from the 15, omitting *lungcancer*, *parkinsons*, *soybeansmall*, and *wine*. For each data set we select 10 features and apply the 3-nn classifier, recording the rank-order of the criteria in terms of their generalisation error. This process was repeated and averaged over 50 trials, giving the results in Figure 6.

To aid interpretation we label MIM with a simple point marker, MIFS, CIFE, CondRed, and ICAP with a circle, and the remaining criteria (DISR, JMI, mRMR and CMIM) with a star. The criteria labelled with a star balance the relative magnitude of the relevancy and redundancy terms, those with a circle do not attempt to balance them, and MIM contains no redundancy term. There is a clear separation between those criteria with a star outperforming those with a circle, and MIM varying in performance between the two groups as we allow more training datapoints.

Notice that the highest ranked criteria coincide with those in the cluster at the top left of Figures 5a and 5b. We suggest that the relative difference in performance is due to the same reason noted in Section 5.2, that the redundancy term grows with the size of the selected feature set. In this case, the redundancy term eventually grows to outweigh the relevancy by a large degree, and the new features are selected solely on the basis of redundancy, ignoring the relevance, thus leading to poor classification performance.

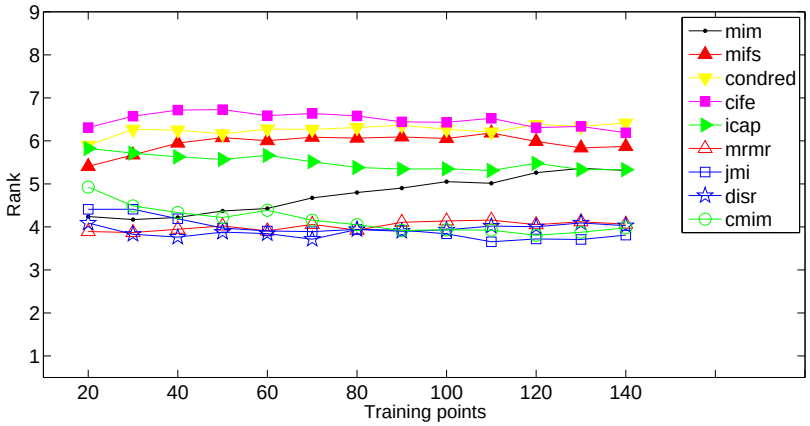


Figure 6: Average ranks of criteria in terms of test error, selecting 10 features, across 11 data sets. Note the clear dominance of criteria which do not allow the redundancy term to overwhelm the relevancy term (unfilled markers) over those that allow redundancy to grow with the size of the feature set (filled markers).

Data	*Features*	*Examples*	*Classes*
Colon	2000	62	2
Leukemia	7070	72	2
Lung	325	73	7
Lymph	4026	96	9
NCI9	9712	60	9

Table 3: Data sets from Peng et al. (2005), used in experiments.

5.4 Extreme Small-Sample Experiments

In the previous sections we discussed two theoretical properties of information-based feature selection criteria: whether it balances the relative magnitude of relevancy against redundancy, and whether it includes a class-conditional redundancy term. Empirically on the UCI data sets, we see that the balancing is far more important than the inclusion of the conditional redundancy term—for example, MRMR succeeds in many cases, while MIFS performs poorly. Now, we consider whether same property may hold in extreme small-sample situations, when the number of examples is so low that reliable estimation of distributions becomes extremely difficult. We use data sourced from Peng et al. (2005), detailed in Table 3. Results are shown in Figure 7, selecting 50 features from each data set and plotting leave-one-out classification error. It should of course be remembered that on such small data sets, making just one additional datapoint error can result in seemingly large changes in accuracy. For example, the difference between the best and worst criteria on Leukemia was just 3 datapoints. In contrast to the UCI results, the picture is less clear. On Colon, the criteria all perform similarly; this is the least complex of all the data sets, having the smallest number of classes with a (relatively) small number of features. As we move through the data sets with increasing numbers of features/classes, we see that MIFS, CONDRED, CIFE and ICAP start to break away, performing poorly compared to the others. Again, we note that these do not attempt to balance relevancy/redundancy. This difference is clearest on the NCI9 data, the most complex with 9 classes and 9712 features. However, as we may expect with such high dimensional and challenging problems, there are some exceptions—the Colon data as mentioned, and also the Lung data where ICAP/MIFS perform well.

5.5 What is the Relation Between Stability and Accuracy?

An important question is whether we can find a good balance between the stability of a criterion and the classification accuracy. This was considered by Gulgezen et al. (2009), who studied the stability/accuracy trade-off for the MRMR criterion. In the following, we consider this trade-off in the context of *Pareto-optimality*, across the 9 criteria, and the 15 data sets from Table 2. Experimental protocol was to take 50 bootstraps from the data set, each time calculating the out-of-bag error using the 3-nn. The stability measure was Kuncheva's stability index calculated from the 50 feature sets, and the accuracy was the mean out-of-bag accuracy across the 50 bootstraps. The experiments were also repeated using the Information Stability measure, revealing almost identical results. Results using Kuncheva's stability index are shown in Figure 8.

The *Pareto-optimal set* is defined as the set of criteria for which no other criterion has both a higher accuracy and a higher stability, hence the members of the Pareto-optimal set are said to be *non-dominated* (Fonseca and Fleming, 1996). Thus, each of the subfigures of Figure 8, criteria

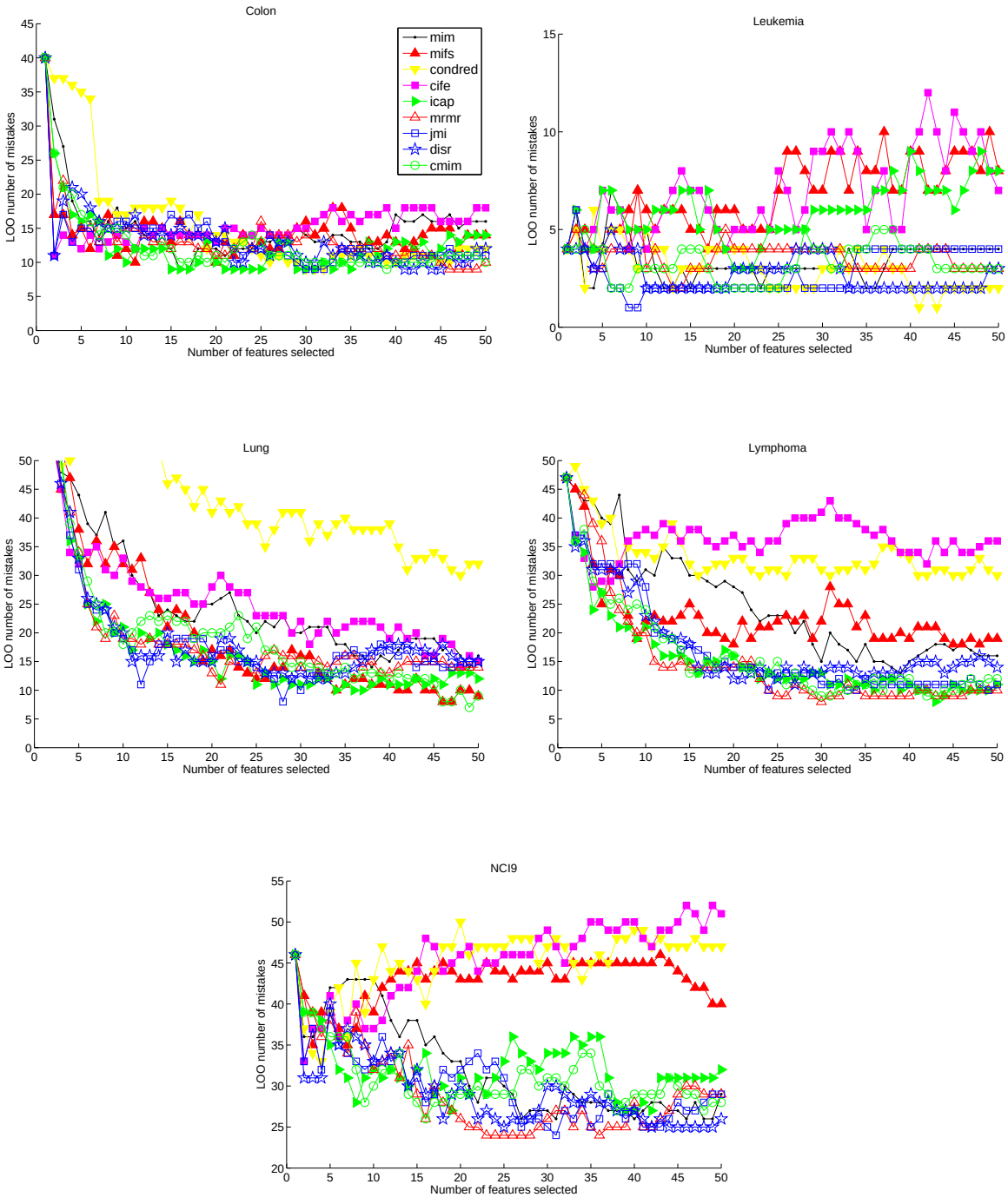


Figure 7: LOO results on Peng's data sets : Colon, Lymphoma, Leukemia, Lung, NCI9.

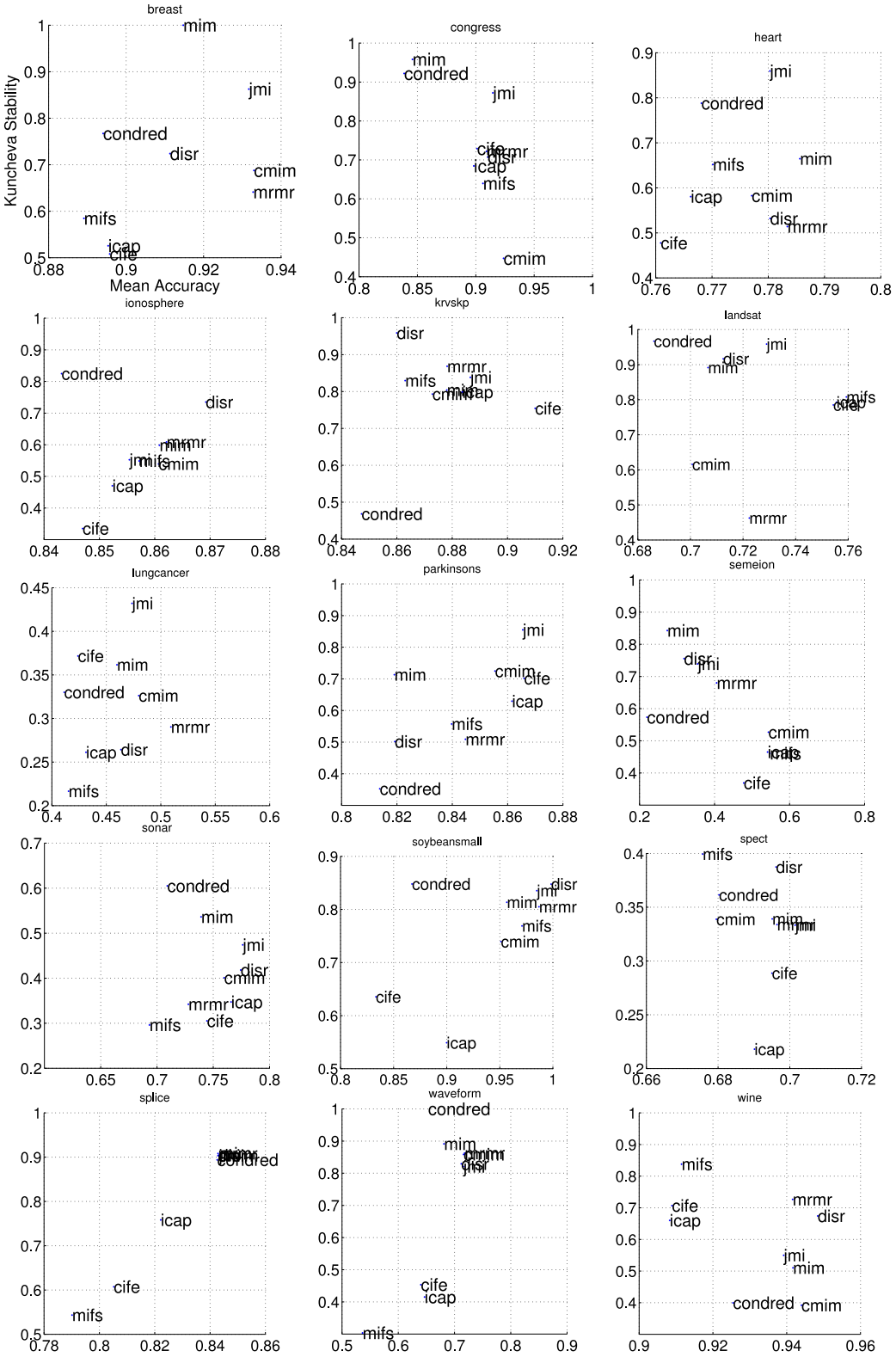


Figure 8: Stabilty (y-axes) versus Accuracy (x-axes) over 50 bootstraps for each of the UCI data sets. The pareto-optimal rankings are summarised in Table 4.

Accuracy/Stability(Yu)	Accuracy/Stability(Kuncheva)	Accuracy
JMI (1.6)	JMI (1.5)	JMI (2.6)
DISR (2.3)	DISR (2.2)	MRMR (3.6)
MIM (2.4)	MIM (2.3)	DISR (3.7)
MRMR (2.5)	MRMR (2.5)	CMIM (4.5)
CMIM (3.3)	CONDRED (3.2)	ICAP (5.3)
ICAP (3.6)	CMIM (3.4)	MIM (5.4)
CONDRED (3.7)	ICAP (4.3)	CIFE (5.9)
CIFE (4.3)	CIFE (4.8)	MIFS (6.5)
MIFS (4.5)	MIFS (4.9)	CONDRED (7.4)

Table 4: *Column 1:* Non-dominated Rank of different criteria for the trade-off of accuracy/stability. Criteria with a higher rank (closer to 1.0) provide a better tradeoff than those with a lower rank. *Column 2:* As column 1 but using Kuncheva's Stability Index. *Column 3:* Average ranks for accuracy alone.

that appear further to the top-right of the space *dominate* those toward the bottom left—in such a situation there is no reason to choose those at the bottom left, since they are dominated on both objectives by other criteria.

A summary (for both stability and information stability) is provided in the first two columns of Table 4, showing the *non-dominated rank* of the different criteria. This is computed per data set as the number of other criteria which dominate a given criterion, in the Pareto-optimal sense, then averaged over the 15 data sets. We can see that these rankings are similar to the results earlier, with MIFS, ICAP, CIFE and CondRed performing poorly. We note that JMI, (which both balances the relevancy and redundancy terms and includes the conditional redundancy) outperforms all other criteria.

We present the average accuracy ranks across the 50 bootstraps in column 3. These are similar to the results from Figure 6 but use a bootstrap of the full data set, rather than a small sample from it. Following Demšar (2006) we analysed these ranks using a Friedman test to determine which criteria are statistically significantly different from each other. We then used a Nemenyi post-hoc test to determine which criteria differed, with statistical significances at 90%, 95%, and 99% confidences. These give a partial ordering for the criteria which we present in Figure 9, showing a *Significant Dominance Partial Order* diagram. Note that this style of diagram encapsulates the same information as a Critical Difference diagram (Demšar, 2006), but allows us to display multiple levels of statistical significance. A bold line connecting two criteria signifies a difference at the 99% confidence level, a dashed line at the 95% level, and a dotted line at the 90% level. Absence of a link signifies that we do not have the statistical power to determine the difference one way or another. Reading Figure 9, we see that with 99% confidence JMI is significantly superior to CondRed, and MIFS, but not statistically significantly different from the other criteria. As we lower our confidence level, more differences appear, for example MRMR and MIFS are only significantly different at the 90% confidence level.

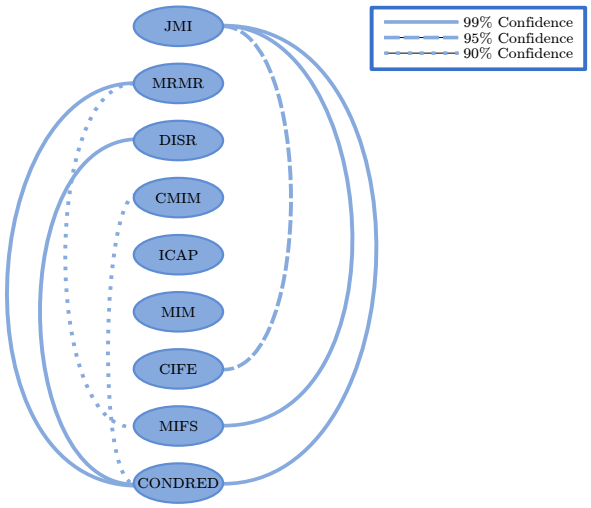


Figure 9: Significant dominance partial-order diagram. Criteria are placed top to bottom in the diagram by their rank taken from column 3 of Table 4. A link joining two criteria means a statistically significant difference is observed with a Nemenyi post-hoc test at the specified confidence level. For example JMI is significantly superior to MIFS ($\beta = 1$) at the 99% confidence level. Note that the absence of a link does not signify the lack of a statistically significant difference, but that the Nemenyi test does not have sufficient power (in terms of number of data sets) to determine the outcome (Demšar, 2006). It is interesting to note that the four bottom ranked criteria correspond to the corners of the unit square in Figure 2; while the top three (JMI/MRMR/DISR) are all very similar, scaling the redundancy terms by the size of the feature set. The middle ranks belong to CMIM/ICAP, which are similar in that they use the min/max strategy instead of a linear combination of terms.

5.6 Summary of Empirical Findings

From experiments in this section, we conclude that the balance of relevancy/redundancy terms is extremely important, while the inclusion of a class conditional term seems to matter less. We find that some criteria are inherently more *stable* than others, and that the trade-off between accuracy (using a simple k-nn classifier) and stability of the feature sets differs between criteria. The best overall trade-off for accuracy/stability was found in the JMI and MRMR criteria. In the following section we re-assess these findings, in the context of two problems posed for the NIPS Feature Selection Challenge.

6. Performance on the NIPS Feature Selection Challenge

In this section we investigate performance of the criteria on data sets taken from the NIPS Feature Selection Challenge (Guyon, 2003).

6.1 Experimental Protocols

We present results using GISETTE (a handwriting recognition task), and MADELON (an artificially generated data set).

Data	*Features*	*Examples (Tr/Val)*	*Classes*
GISETTE	5000	6000/1000	2
MADELON	500	2000/600	2

Table 5: Data sets from the NIPS challenge, used in experiments.

To apply the mutual information criteria, we estimate the necessary distributions using histogram estimators: features were discretized independently into 10 equal width bins, with bin boundaries determined from training data. After the feature selection process the original (undiscretised) data sets were used to classify the validation data. Each criterion was used to generate a ranking for the top 200 features in each data set. We show results using the full top 200 for GISETTE, but only the top 20 for MADELON as after this point all criteria demonstrated severe overfitting. We use the Balanced Error Rate, for fair comparison with previously published work on the NIPS data sets. We accept that this does not necessarily share the same optima as the classification error (to which the conditional likelihood relates), and leave investigations of this to future work.

Validation data results are presented in Figure 10 (GISETTE) and Figure 11 (MADELON). The minimum of the validation error was used to select the best performing feature set size, the training data alone used to classify the testing data, and finally test labels were submitted to the challenge website. Test results are provided in Table 6 for GISETTE, and Table 7 for MADELON.[5]

Unlike in Section 5, the data sets we have used from the NIPS Feature Selection Challenge have a greater number of datapoints (GISETTE has 6000 training examples, MADELON has 2000) and thus we can present results using a direct implementation of Equation (10) as a criterion. We refer to this criterion as CMI, as it is using the conditional mutual information to score features. Unfortunately there are still estimation errors in this calculation when selecting a large number of

5. We do not provide classification confidences as we used a nearest neighbour classifier and thus the AUC is equal to 1− BER.

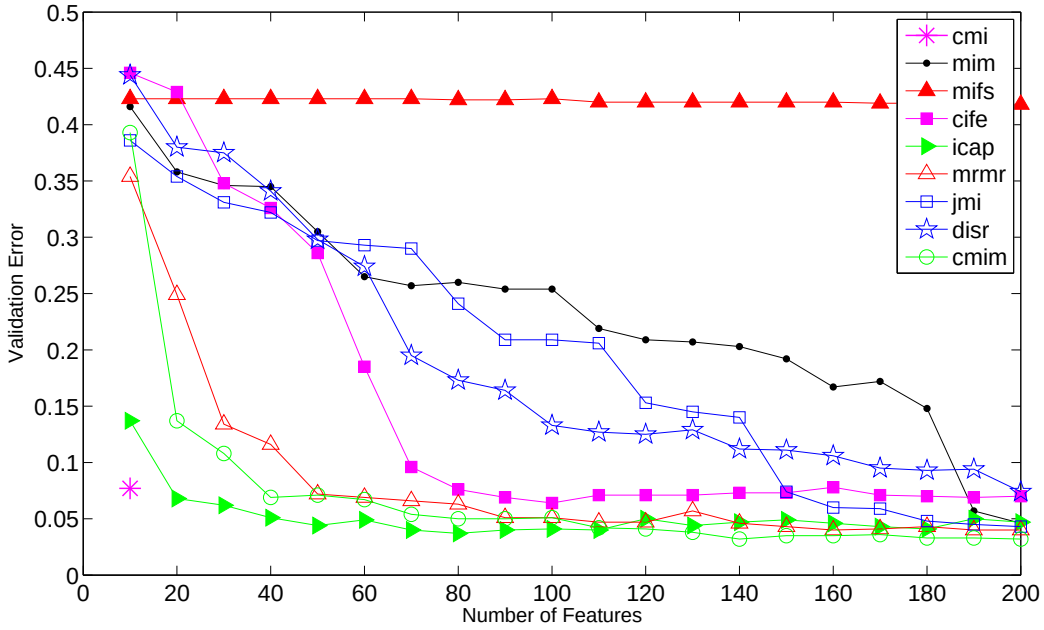


Figure 10: Validation Error curve using GISETTE.

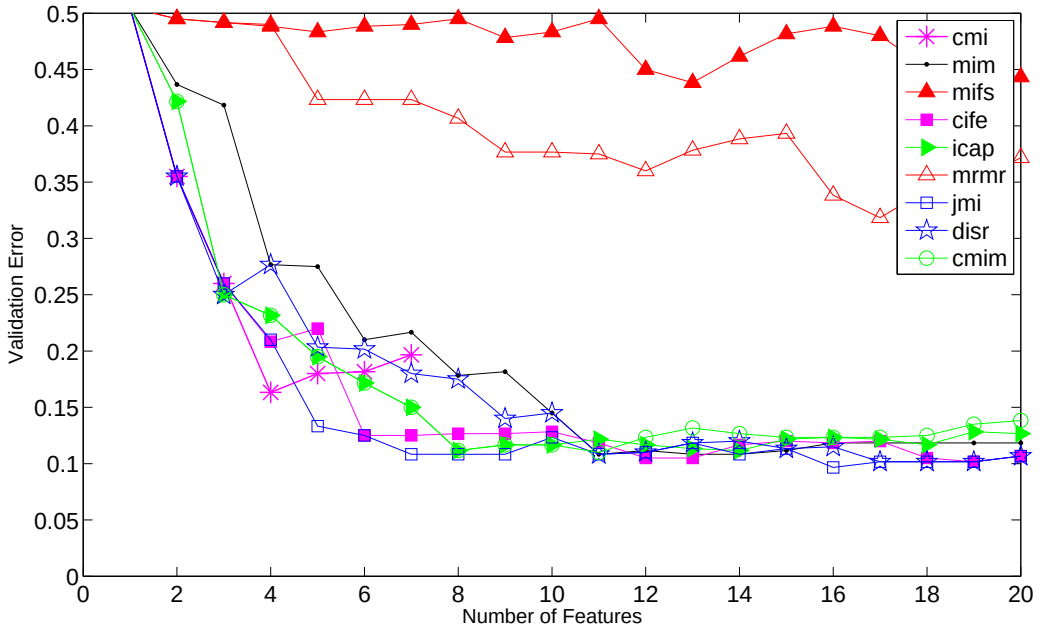


Figure 11: Validation Error curve using MADELON.

features, even given the large number of datapoints and so the criterion fails to select features after a certain point, as each feature appears equally irrelevant. In GISETTE, CMI selected 13 features, and so the top 10 features were used and thus one result is shown. In MADELON, CMI selected 7 features and so 7 results are shown.

6.2 Results on Test Data

In Table 6 there are several distinctions between the criteria, the most striking of which is the failure of MIFS to select an informative feature set. The importance of balancing the magnitude of the relevancy and the redundancy can be seen whilst looking at the other criteria in this test. Those criteria which balance the magnitudes, (CMIM, JMI, & mRMR) perform better than those which do not (ICAP,CIFE). The DISR criterion forms an outlier here as it performs poorly when compared to JMI. The only difference between these two criteria is the normalization in DISR—as such, this is the likely cause of the observed poor performance, the introduction of more variance by estimating the normalization $H(X_k X_j Y)$.

We can also see how important the low dimensional approximation is, as even with 6000 training examples CMI cannot estimate the required joint distribution to avoid selecting probes, despite being a direct iterative maximisation of the conditional likelihood in the limit of datapoints.

Criterion	*BER*	*AUC*	*Features (%)*	*Probes (%)*
MIM	4.18	95.82	4.00	0.00
MIFS	42.00	58.00	4.00	58.50
CIFE	6.85	93.15	2.00	0.00
ICAP	4.17	95.83	1.60	0.00
CMIM	**2.86**	**97.14**	**2.80**	**0.00**
CMI	8.06	91.94	0.20	20.00
mRMR	2.94	97.06	3.20	0.00
JMI	3.51	96.49	4.00	0.00
DISR	8.03	91.97	4.00	0.00
Winning Challenge Entry	**1.35**	**98.71**	**18.3**	**0.0**

Table 6: NIPS FS Challenge Results: GISETTE.

The MADELON results (Table 7) show a particularly interesting point—the top performers (in terms of BER) are JMI and CIFE. Both these criteria include the class-conditional redundancy term, but CIFE does not balance the influence of relevancy against redundancy. In this case, it appears the 'balancing' issue, so important in our previous experiments seems to have little importance—instead, the presence of the conditional redundancy term is the differentiating factor between criteria (note the poor performance of MIFS/MRMR). This is perhaps not surprising given the nature of the MADELON data, constructed precisely to require features to be evaluated jointly.

It is interesting to note that the challenge organisers benchmarked a 3-NN using the optimal feature set, achieving a 10% test error (Guyon, 2003). Many of the criteria managed to select feature sets which achieved a similar error rate using a 3-NN, and it is likely that a more sophisticated classifier is required to further improve performance.

This concludes our experimental study—in the following, we make further links to the literature for the theoretical framework, and discuss implications for future work.

Criterion	*BER*	*AUC*	*Features (%)*	*Probes (%)*
MIM	10.78	89.22	2.20	0.00
MIFS	46.06	53.94	2.60	92.31
CIFE	**9.50**	**90.50**	**3.80**	**0.00**
ICAP	11.11	88.89	1.60	0.00
CMIM	11.83	88.17	2.20	0.00
CMI	21.39	78.61	0.80	0.00
mRMR	35.83	64.17	3.40	82.35
JMI	**9.50**	**90.50**	**3.20**	**0.00**
DISR	9.56	90.44	3.40	0.00
Winning Challenge Entry	**7.11**	**96.95**	**1.6**	**0.0**

Table 7: NIPS FS Challenge Results: MADELON.

7. Related Work: Strong and Weak Relevance

Kohavi and John (1997) proposed definitions of *strong* and *weak* feature relevance. The definitions are formed from statements about the conditional probability distributions of the variables involved. We can re-state the definitions of Kohavi and John (hereafter KJ) in terms of mutual information, and see how they can fit into our conditional likelihood maximisation framework. In the notation below, notation X_i indicates the ith feature in the overall set X, and notation $X_{\setminus i}$ indicates the set $\{X \setminus X_i\}$, all features *except* the ith.

Definition 8 : Strongly Relevant Feature (Kohavi and John, 1997)
Feature X_i is strongly relevant *to Y iff there exists an assignment of values x_i, y, $x_{\setminus i}$ for which $p(X_i = x_i, X_{\setminus i} = x_{\setminus i}) > 0$ and $p(Y = y|X_i = x_i, X_{\setminus i} = x_{\setminus i}) \neq p(Y = y|X_{\setminus i} = x_{\setminus i})$.*

Corollary 9 *A feature X_i is* strongly relevant *iff $I(X_i;Y|X_{\setminus i}) > 0$.*

Proof The KL divergence $D_{KL}(p(y|xz) \;||\; p(y|z)) > 0$ iff $p(y|xz) \neq p(y|z)$ for some assignment of values x, y, z. A simple re-application of the manipulations leading to Equation (5) demonstrates that the expected KL-divergence $E_{xz}\{p(y|xz)||p(y|z)\}$ is equal to the mutual information $I(X;Y|Z)$. In the definition of strong relevance, if there exists a single assignment of values x_i, y, $x_{\setminus i}$ that satisfies the inequality, then $E_x\{p(y|x_i x_{\setminus i})||p(y|x_{\setminus i})\} > 0$ and therefore $I(X_i;Y|X_{\setminus i}) > 0$. ■

Given the framework we have presented, we can note that this strong relevance comes from a combination of *three terms*,

$$I(X_i;Y|X_{\setminus i}) = I(X_i;Y) - I(X_i;X_{\setminus i}) + I(X_i;X_{\setminus i}|Y).$$

This view of strong relevance demonstrates explicitly that a feature may be individually irrelevant (i.e., $p(y|x_i) = p(y)$ and thus $I(X_i;Y) = 0$), but still strongly relevant if $I(X_i;X_{\setminus i}|Y) - I(X_i;X_{\setminus i}) > 0$.

Definition 10 : Weakly Relevant Feature (Kohavi and John, 1997)
Feature X_i is weakly relevant *to Y iff it is not strongly relevant and there exists a subset $Z \subset X_{\setminus i}$, and an assignment of values x_i, y, z for which $p(X_i = x_i, Z = z) > 0$ such that $p(Y = y|X_i = x_i, Z = z) \neq p(Y = y|Z = z)$.*

Corollary 11 *A feature X_i is weakly relevant to Y iff it is not strongly relevant and $I(X_i;Y|Z) > 0$ for some $Z \subset X_{\setminus i}$.*

Proof This follows immediately from the proof for the strong relevance above. ■

It is interesting, and somewhat non-intuitive, that there can be cases where there are *no* strongly relevant features, but *all* are weakly relevant. This will occur for example in a data set where all features have exact duplicates: we have $2M$ features and $\forall i,\ X_{M+i} = X_i$. In this case, for any X_k (such that $k < M$) we will have $I(X_k;Y|X_{\setminus i}) = 0$ since its duplicate feature X_{M+k} will carry the same information. In this case, for any feature X_k (such that $k < M$) that is strongly relevant in the data set $\{X_1,...,X_M\}$, it is *weakly* relevant in the data set $\{X_1,...,X_{2M}\}$.

This issue can be dealt with by refining our definition of relevance with respect to a subset of the full feature space. A particular subset about which we have some information is the currently selected set S. We can relate our framework to KJ's definitions in this context. Following KJ's formulations,

Definition 12 : Relevance with respect to the current set S.
Feature X_i is relevant *to Y with respect to S iff there exists an assignment of values x_i, y, s for which $p(X_i = x_i, S = s) > 0$ and $p(Y = y|X_i = x_i, S = s) \neq p(Y = y|S = s)$.*

Corollary 13 *Feature X_i is* relevant *to Y with respect to S, iff $I(X_i;Y|S) > 0$.*

A feature that is relevant with respect to S is either strongly or weakly relevant (in the KJ sense) but it is not possible to determine in which class it lies, as we have not conditioned on $X_{\setminus i}$. Notice that the definition coincides exactly with the forward selection heuristic (Definition 2), which we have shown is a hill-climber on the conditional likelihood. As a result, we see *that hill-climbing on the conditional likelihood corresponds to adding the* most *relevant feature with respect to the current set S.* Again we re-emphasize, that the resultant gain in the likelihood comes from a combination of *three sources*:

$$I(X_i;Y|S) = I(X_i;Y) - I(X_i;S) + I(X_i;S|Y).$$

It could easily be the case that $I(X_i;Y) = 0$, that is a feature is entirely irrelevant when considered on its own—but the sum of the two redundancy terms results in a positive value for $I(X_i;Y|S)$. We see that if a criterion does not attempt to model both of the redundancy terms, even if only using low dimensional approximations, it runs the risk of evaluating the relevance of X_i incorrectly.

Definition 14 : Irrelevance with respect to the current set S.
Feature X_i is irrelevant *to Y with respect to S iff $\forall$ x_i, y, s for which $p(X_i = x_i, S = s) > 0$ and $p(Y = y|X_i = x_i, S = s) = p(Y = y|S = s)$.*

Corollary 15 *Feature X_i is* irrelevant *to Y with respect to S, iff $I(X_i;Y|S) = 0$.*

In a forward step, if a feature X_i is irrelevant with respect to S, adding it alone to S *will not increase the conditional likelihood.* However, there may be further additions to S in the future, giving us a selected set S'; we may then find that X_i is then *relevant* with respect to S'. In a backward step we check whether a feature is irrelevant with respect to $\{S \setminus X_i\}$, using the test $I(X_i;Y|\{S \setminus X_i\}) = 0$. In this case, removing this feature *will not decrease the conditional likelihood.*

8. Related Work: Structure Learning in Bayesian Networks

The framework we have described also serves to highlight a number of important links to the literature on structure learning of directed acyclic graphical (DAG) models (Korb, 2011). The problem of DAG learning from observed data is known to be NP-hard (Chickering et al., 2004), and as such there exist two main families of approximate algorithms. *Metric* or *Score-and-Search* learners construct a graph by searching the space of DAGs directly, assigning a score to each based on properties of the graph in relation to the observed data; probably the most well-known score is the BIC measure (Korb, 2011). However, the space of DAGs is superexponential in the number of variables, and hence an exhaustive search rapidly becomes computationally infeasible. Grossman and Domingos (2004) proposed a greedy hill-climbing search over structures, using conditional likelihood as a scoring criterion. Their work found significant advantage from using this 'discriminative' learning objective, as opposed to the traditional 'generative' joint likelihood. The potential of this discriminative model perspective will be expanded upon in Section 9.3.

Constraint learners approach the problem from a constructivist point of view, adding and removing arcs from a single DAG according to conditional independence tests given the data. When the candidate DAG passes all conditional independence statements observed in the data, it is considered to be a good model. In the current paper, for a feature to be eligible for inclusion, we required that $I(X_k;Y|S) > 0$. This is equivalent to a conditional independence test $X_k \not\perp\!\!\!\perp Y \mid S$. One well-known problem with constraint learners is that if a test gives an incorrect result, the error can 'cascade', causing the algorithm to draw further incorrect conclusions on the network structure. This problem is also true of the popular greedy-search heuristics that we have described in this work.

In Section 3.2, we showed that Markov Blanket algorithms (Tsamardinos et al., 2003) are an example of the framework we propose. Specifically, the solution to Equation (7) is a (possibly non-unique) Markov Blanket, and the solution to Equation (8) is exactly the Markov *boundary*, that is, a minimal, unique blanket. It is interesting to note that these algorithms, which are a restricted class of structure learners, assume *faithfulness* of the data distribution. We can see straightforwardly that all criteria we have considered, when combined with a greedy forward selection, also make this assumption.

9. Conclusion

This work has presented a unifying framework for information theoretic feature selection, bringing almost two decades of research on heuristic scoring criteria under a single theoretical interpretation. This is achieved via a novel interpretation of information theoretic feature selection as *an optimization of the conditional likelihood*—this is in contrast to the current view of mutual information, as a heuristic measure of feature relevancy.

9.1 Summary of Contributions

In Section 3 we showed how to decompose the conditional likelihood into three terms, each with their own interpretation in relation to the feature selection problem. One of these emerges as a *conditional mutual information*. This observation allows us to answer the following question:

What are the implicit statistical assumptions of mutual information criteria? The investigations have revealed that the various criteria published over the past two decades are all *approximate iterative maximisers of the conditional likelihood.* The approximations are due to implicit assumptions

on the data distribution: some are more restrictive than others, and are detailed in Section 4. The approximations, while heuristic, are necessary due to the need to estimate high dimensional probability distributions. The popular Markov Blanket learning algorithm IAMB is included in this class of procedures, hence can also bee seen as an iterative maximiser of the conditional likelihood.

The main differences between criteria are whether they include a *class-conditional* term, and whether they provide a mechanism to *balance* the relative size of the redundancy terms against the relevancy term. To ascertain how these differences impact the criteria in practice, we conducted an empirical study of 9 different heuristic mutual information criteria across 22 data sets. We analyzed how the criteria behave in large/small sample situations, how the stability of returned feature sets varies between criteria, and how similar criteria are in the feature sets they return. In particular, the following questions were investigated:

How do the theoretical properties translate to classifier accuracy? Summarising the performance of the criteria under the above conditions, including the class-conditional term is *not* always necessary. Various criteria, for example MRMR, are successful without this term. However, without this term criteria are blind to certain classes of problems, for example, the MADELON data set, and will perform poorly in these cases. Balancing the relevancy and redundancy terms is however *extremely* important—criteria like MIFS, or CIFE, that allow redundancy to swamp relevancy, are ranked lowest for accuracy in almost all experiments. In addition, this imbalance tends to cause large instability in the returned feature sets—being highly sensitive to the supplied data.

How stable are the criteria to small changes in the data? Several criteria return wildly different feature sets with just small changes in the data, while others return similar sets each time, hence are 'stable' procedures. The most stable was the univariate mutual information, followed closely by JMI (Yang and Moody, 1999; Meyer et al., 2008); while among the least stable are MIFS (Battiti, 1994) and ICAP (Jakulin, 2005). As visualised by multi-dimensional scaling in Figure 5, several criteria appear to return quite similar sets, while there are some outliers.

How do criteria behave in limited and extreme small-sample situations? In extreme small-sample situations, it appears the above rules (regarding the conditional term and the balancing of relevancy-redundancy) can be broken—the poor estimation of distributions means the theoretical properties do not translate immediately to performance.

9.2 Advice for the Practitioner

From our investigations we have identified three desirable characteristics of an information based selection criterion. The first is whether it includes reference to a conditional redundancy term—criteria that do not incorporate it are effectively blind to an entire class of problems, those with strong class-conditional dependencies. The second is whether it keeps the relative size of the redundancy term from swamping the relevancy term. We find this to be *essential*—without this control, the relevancy of the kth feature can easily be ignored in the selection process due to the $k-1$ redundancy terms. The third is simply whether the criterion is a low-dimensional approximation, hence making it usable with small sample sizes. On GISETTE with 6000 examples, we were unable to select more than 13 features with any kind of reliability. Therefore, low dimensional approximations, the focus of this article, are essential.

A summary of the criteria is shown in Table 8. Overall we find only 3 criteria that satisfy these properties: CMIM, JMI and DISR. We recommend the JMI criterion, as from empirical investigations it has the best trade-off (in the Pareto-optimal sense) of accuracy and stability. DISR is

a normalised variant of JMI—in practice we found little need for this normalisation and the extra computation involved. If higher stability is required—the MIM criterion, as expected, displayed the highest stability with respect to variations in the data—therefore in extreme data-poor situations we would recommend this as a first step. If speed is required, the CMIM criterion admits an fast exact implementation giving orders of magnitude speed-up over a straightforward implementation—refer to Fleuret (2004) for details.

To aid replicability of this work, implementations of all criteria we have discussed are provided at: `http://www.cs.man.ac.uk/~gbrown/fstoolbox/`

	MIM	mRMR	MIFS	CMIM	JMI	DISR	ICAP	CIFE	CMI
Cond Redund term?	✗	✗	✗	✔	✔	✔	✔	✔	✔
Balances rel/red?	✔	✔	✗	✔	✔	✔	✗	✗	✔
Estimable?	✔	✔	✔	✔	✔	✔	✔	✔	✗

Table 8: Summary of criteria. They have been arranged left to right in order of ascending estimation difficulty. *Cond Redund term*: does it include the conditional redundancy term? *Balances rel/red*: does it balance the relevance and redundancy terms? *Estimable*: does it use a low dimensional approximation, making it usable with small samples?

9.3 Future Work

While advice on the suitability of existing criteria is of course useful, perhaps a more interesting result of this work is the perspective it brings to the feature selection problem. We were able to *explicitly* state an objective function, and derive an appropriate information-based criterion to maximise it. This begs the question, what selection criteria would result from different objective functions? Dmochowski et al. (2010) study a weighted conditional likelihood, and its suitability for cost-sensitive problems—it is possible (though outside the scope of this paper) to derive information-based criteria in this context. The reverse question is equally interesting, what objective functions are implied by other existing criteria, such as the Gini Index? The KL-divergence (which defines the mutual information) is a special case of a wider family of measures, based on the f-divergence—could we obtain similar efficient criteria that pursue these measures, and what overall objectives do they imply?

In this work we explored criteria that use pairwise (i.e., $I(X_k;X_j)$) approximations to the derived objective. These approximations are commonly used as they provide a reasonable heuristic while still being (relatively) simple to estimate. There has been work which suggests relaxing this pairwise approximation, and thus increasing the number of terms (Brown, 2009; Meyer et al., 2008), but there is little exploration of how much data is required to estimate these multivariate information terms. A theoretical analysis of the tradeoff between estimation accuracy and additional information provided by these more complex terms could provide interesting directions for improving the power of filter feature selection techniques.

A very interesting direction concerns the motivation behind the conditional likelihood as an objective. It can be noted that the conditional likelihood, though a well-accepted objective function in its own right, can be derived from a probabilistic discriminative model, as follows. We approximate the true distribution p with our model q, with three distinct parameter sets: θ for feature selection,

τ for classification, and λ modelling the input distribution $p(\mathbf{x})$. Following Minka (2005), in the construction of a discriminative model, our joint likelihood is

$$\mathcal{L}(\mathcal{D},\theta,\tau,\lambda) = p(\theta,\tau)p(\lambda)\prod_{i=1}^{N} q(y^i|\mathbf{x}^i,\theta,\tau)q(\mathbf{x}^i|\lambda).$$

In this type of model, we wish to maximize $\mathcal{L}$ with respect to θ (our feature selection parameters) and τ (our model parameters), and are not concerned with the generative parameters λ. Excluding the generative terms gives

$$\mathcal{L}(\mathcal{D},\theta,\tau,\lambda) \propto p(\theta,\tau)\prod_{i=1}^{N} q(y^i|\mathbf{x}^i,\theta,\tau).$$

When we have no particular bias or prior knowledge over which subset of features or parameters are more likely (i.e., a flat prior $p(\theta,\tau)$), this reduces to the conditional likelihood:

$$\mathcal{L}(\mathcal{D},\theta,\tau,\lambda) \propto \prod_{i=1}^{N} q(y^i|\mathbf{x}^i,\theta,\tau),$$

which was exactly our starting point for the current paper. An obvious extension here is to take a non-uniform prior over features. An important direction for machine learning is to incorporate *domain knowledge*. A non-uniform prior would mean influencing the search procedure to incorporate our background knowledge of the features. This is applicable for example in gene expression data, when we may have information about the metabolic pathways in which genes participate, and therefore which genes are likely to influence certain biological functions. This is outside the scope of this paper but is the focus of our current research.

Acknowledgments

This research was conducted with support from the UK Engineering and Physical Sciences Research Council, on grants EP/G000662/1 and EP/F023855/1. Mikel Luján is supported by a Royal Society University Research Fellowship. Gavin Brown would like to thank James Neil, Sohan Seth, and Fabio Roli for invaluable commentary on drafts of this work.

Appendix A.

The following proofs make use of the identity, $I(A;B|C) - I(A;B) = I(A;C|B) - I(A;C)$.

A.1 Proof of Equation (17)

The *Joint Mutual Information* criterion (Yang and Moody, 1999) can be written,

$$\begin{aligned} J_{jmi}(X_k) = & \sum_{X_j \in S} I(X_k X_j; Y), \\ = & \sum_{X_j \in S} \Big[I(X_j;Y) + I(X_k;Y|X_j) \Big]. \end{aligned}$$

The term $\sum_{X_j \in S} I(X_j;Y)$ in the above is constant with respect to the X_k argument that we are interested in, so can be omitted. The criterion therefore reduces to (17) as follows,

$$\begin{aligned} J_{jmi}(X_k) &= \sum_{X_j \in S} \Big[I(X_k;Y|X_j) \Big] \\ &= \sum_{X_j \in S} \Big[I(X_k;Y) - I(X_k;X_j) + I(X_k;X_j|Y) \Big] \\ &= |S| \times I(X_k;Y) - \sum_{X_j \in S} \Big[I(X_k;X_j) - I(X_k;X_j|Y) \Big] \\ &\propto I(X_k;Y) - \frac{1}{|S|} \sum_{X_j \in S} \Big[I(X_k;X_j) - I(X_k;X_j|Y) \Big]. \end{aligned}$$

A.2 Proof of Equation (19)

The rearrangement of the Conditional Mutual Information criterion (Fleuret, 2004) follows a very similar procedure. The original, and its rewriting are,

$$\begin{aligned} J_{cmim}(X_k) &= \min_{X_j \in S} \Big[I(X_k;Y|X_j) \Big] \\ &= \min_{X_j \in S} \Big[I(X_k;Y) - I(X_k;X_j) + I(X_k;X_j|Y) \Big] \\ &= I(X_k;Y) + \min_{X_j \in S} \Big[I(X_k;X_j|Y) - I(X_k;X_j) \Big] \\ &= I(X_k;Y) - \max_{X_j \in S} \Big[I(X_k;X_j) - I(X_k;X_j|Y) \Big], \end{aligned}$$

which is exactly Equation (19).

References

K. S. Balagani and V. V. Phoha. On the feature selection criterion based on an approximation of multidimensional mutual information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(7):1342–1343, 2010. ISSN 0162-8828.

R. Battiti. Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on Neural Networks*, 5(4):537–550, 1994.

G. Brown. A new perspective for information theoretic feature selection. In *International Conference on Artificial Intelligence and Statistics*, volume 5, pages 49–56, 2009.

H. Cheng, Z. Qin, C. Feng, Y. Wang, and F. Li. Conditional mutual information-based feature selection analyzing for synergy and redundancy. *Electronics and Telecommunications Research Institute (ETRI) Journal*, 33(2), 2011.

D. M. Chickering, D. Heckerman, and C. Meek. Large-sample learning of bayesian networks is np-hard. *Journal of Machine Learning Research*, 5:1287–1330, 2004.

T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience New York, 1991.

J. Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

J. P. Dmochowski, P. Sajda, and L. C. Parra. Maximum likelihood in cost-sensitive learning: model specification, approximations, and upper bounds. *Journal of Machine Learning Research*, 11: 3313–3332, 2010.

W. Duch. *Feature Extraction: Foundations and Applications*, chapter 3, pages 89–117. Studies in Fuzziness & Soft Computing. Springer, 2006. ISBN 3-540-35487-5.

A. El Akadi, A. El Ouardighi, and D. Aboutajdine. A powerful feature selection approach based on mutual information. *International Journal of Computer Science and Network Security*, 8(4):116, 2008.

R. M. Fano. *Transmission of Information: Statistical Theory of Communications*. New York: Wiley, 1961.

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

C. Fonseca and P. Fleming. On the performance assessment and comparison of stochastic multiobjective optimizers. *Parallel Problem Solving from Nature*, pages 584–593, 1996.

D. Grossman and P. Domingos. Learning bayesian network classifiers by maximizing conditional likelihood. In *International Conference on Machine Learning*. ACM, 2004.

G. Gulgezen, Z. Cataltepe, and L. Yu. Stable and accurate feature selection. *Machine Learning and Knowledge Discovery in Databases*, pages 455–468, 2009.

B. Guo and M. S. Nixon. Gait feature subset selection by mutual information. *IEEE Trans Systems, Man and Cybernetics*, 39(1):36–46, January 2009.

I. Guyon. *Design of experiments for the NIPS 2003 variable selection benchmark*. http://www.nipsfsc.ecs.soton.ac.uk/papers/NIPS2003-Datasets.pdf, 2003.

I. Guyon, S. Gunn, M. Nikravesh, and L. Zadeh, editors. *Feature Extraction: Foundations and Applications*. Springer, 2006. ISBN 3-540-35487-5.

M. Hellman and J. Raviv. Probability of error, equivocation, and the chernoff bound. *IEEE Transactions on Information Theory*, 16(4):368–372, 1970.

A. Jakulin. *Machine Learning Based on Attribute Interactions*. PhD thesis, University of Ljubljana, Slovenia, 2005.

A. Kalousis, J. Prados, and M. Hilario. Stability of feature selection algorithms: a study on high-dimensional spaces. *Knowledge and information systems*, 12(1):95–116, 2007. ISSN 0219-1377.

R. Kohavi and G. H. John. Wrappers for feature subset selection. *Artificial intelligence*, 97(1-2): 273–324, 1997. ISSN 0004-3702.

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

K. Korb. *Encyclopedia of Machine Learning*, chapter Learning Graphical Models, page 584. Springer, 2011.

L. I. Kuncheva. A stability index for feature selection. In *IASTED International Multi-Conference: Artificial Intelligence and Applications*, pages 390–395, 2007.

N. Kwak and C. H. Choi. Input feature selection for classification problems. *IEEE Transactions on Neural Networks*, 13(1):143–159, 2002.

D. D. Lewis. Feature selection and feature extraction for text categorization. In *Proceedings of the workshop on Speech and Natural Language*, pages 212–217. Association for Computational Linguistics Morristown, NJ, USA, 1992.

D. Lin and X. Tang. Conditional infomax learning: An integrated framework for feature extraction and fusion. In *European Conference on Computer Vision*, 2006.

P. Meyer and G. Bontempi. On the use of variable complementarity for feature selection in cancer classification. In *Evolutionary Computation and Machine Learning in Bioinformatics*, pages 91–102, 2006.

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*, 2(3): 261–274, 2008.

T. Minka. Discriminative models, not discriminative training. *Microsoft Research Cambridge, Tech. Rep. TR-2005-144*, 2005.

L. Paninski. Estimation of entropy and mutual information. *Neural Computation*, 15(6):1191–1253, 2003. ISSN 0899-7667.

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

C. E. Shannon. A mathematical theory of communication. *Bell Systems Technical Journal*, 27(3): 379–423, 1948.

M. Tesmer and P. A. Estevez. Amifs: Adaptive feature selection by using mutual information. In *IEEE International Joint Conference on Neural Networks*, volume 1, 2004.

I. Tsamardinos and C. F. Aliferis. Towards principled feature selection: Relevancy, filters and wrappers. In *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics (AISTATS)*, 2003.

I. Tsamardinos, C. F. Aliferis, and A. Statnikov. Algorithms for large scale markov blanket discovery. In *16th International FLAIRS Conference*, volume 103, 2003.

M. Vidal-Naquet and S. Ullman. Object recognition with informative features and linear classification. *IEEE Conference on Computer Vision and Pattern Recognition*, 2003.

J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, and V. Vapnik. Feature selection for svms. *Advances in Neural Information Processing Systems*, pages 668–674, 2001. ISSN 1049-5258.

H. Yang and J. Moody. Data visualization and feature selection: New algorithms for non-gaussian data. *Advances in Neural Information Processing Systems*, 12, 1999.

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

L. Yu, C. Ding, and S. Loscalzo. Stable feature selection via dense feature groups. In *Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 803–811, 2008.