A New Perspective for Information Theoretic Feature Selection

Gavin Brown
School of Computer Science
University of Manchester
Oxford Road, UK
gavin.brown@manchester.ac.uk

Abstract

Feature Filters are among the simplest and fastest approaches to feature selection. A filter defines a statistical criterion, used to *rank* features on how useful they are expected to be for classification. The highest ranking features are retained, and the lowest ranking can be discarded. A common approach is to use the *Mutual Information* between the feature and class label. This area has seen a recent flurry of activity, resulting in a confusing variety of heuristic criteria all based on mutual information, and a lack of a principled way to understand or relate them. The contribution of this paper is a unifying theoretical understanding of such filters. In contrast to current methods which *manually construct* filter criteria with particular properties, we show how to naturally derive a *space* of *possible* ranking criteria. We will show that several recent contributions in the feature selection literature are points within this continuous space, and that there exist many points that have never been explored.

1 INTRODUCTION

High-dimensional datasets 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, or redundant in the context of others. Learning in this situation raises several issues, e.g. overfitting, computational burden, and interpretability of the final model. It is therefore an important research direction to automatically identify meaningful smaller subsets of these variables, i.e. *feature selection* (Guyon and Elisseeff, 2003). We focus on *filter criteria*—that is, model-independent criteria that provide a ranking of the features. In particular, we analyse approaches based on mutual information (Shannon, 1948), and note that there is a lack of a principled methodology for design/analysis of such criteria.

The key result of this research is a new expansion of the Shannon mutual information between the features and the class label, and the use of this expansion to naturally *derive* a space of possible filter criteria. We then “retrofit” existing criteria into the theoretical framework, showing that numerous published criteria are points within this space, and that there exist many points that have never been explored.

2 FEATURE SELECTION WITH MUTUAL INFORMATION

Mutual Information measures the amount of information *shared* by X and Y. As such it has found significant uptake in machine learning, where we imagine X as an input feature set, and Y as a target. It is not difficult to show that the Bayes error of predicting Y from X is lower-bounded by Fano's inequality (Fano, 1961), and upper-bounded by half the conditional entropy,

$$\frac{H(Y) - I(X;Y) - 1}{\log(|Y|)} \leq P(g(X) \neq Y) \leq \frac{1}{2} H(Y|X). \tag{1}$$

The first inequality states that for *any* function $g(X)$ of the inputs, the probability of error is lower bounded by an expression dependent on the mutual information. As the mutual information grows, the bound is minimized—whether or not the bound can be reached

Appearing in Proceedings of the 12^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2009, Clearwater Beach, Florida, USA. Volume 5 of JMLR: W&CP 5. Copyright 2009 by the authors.

depends on the ability of our classifier, i.e. the function $g(X)$. Our task is therefore to select features from a pool such that their *joint* mutual information $I(X_{1:n};Y)$ is maximised.

To know whether we should include a given candidate feature, we must be able to evaluate the mutual information—unfortunately, $I(X_{1:n};Y)$ involves high dimensional distributions, and thus is extremely difficult to reliably estimate. As a heuristic, we could assume the utility of each feature X_n is independent of all other features—and rank the features in descending order of the criterion $J_{mim} = I(X_n;Y)$. Under the assumption of Naive Bayes as our classification model, J_{mim} can in fact be derived as the optimal feature selection criterion in terms of maximising the log-likelihood of the dataset. However in the general case, where features are interdependent, this is *known* to be suboptimal.

In general, it is widely recognised that a good 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. Several criteria have been proposed that attempt to achieve this. For example, MIFS (Battiti, 1994), uses the ranking criterion,

$$J_{mifs} = I(X_n;Y) - \beta \sum_{k=1}^{n-1} I(X_n;X_k). \qquad (2)$$

This includes the objective $I(X_n;Y)$ term to ensure feature *relevance*, but introduces a penalty to enforce low inter-feature correlations. The β is a configurable parameter, which must be set experimentally. Using $\beta = 0$ would be equivalent to selecting features independently, while a larger value will place more emphasis on reducing inter-feature dependencies. The MIFS scheme was the first of many criteria that attempted to capture the relevance-redundancy tradeoff in feature sets with various heuristic terms (Yang and Moody, 1999; Kwak and Choi, 2002; Vidal-Naquet and Ullman, 2003; Fleuret, 2004; Peng et al., 2005; Lin and Tang, 2006). Current best practice has been to *hand-design* criteria, augmenting the individual feature relevance with various penalties to manage the redundancy. In the following section we offer a novel perspective on the problem.

3 A NOVEL PERSPECTIVE

As discussed in the previous section, our ultimate objective is to pick features $\{X_1, ..., X_n\}$ that maximise $I(X_{1:n};Y)$. That is, we want to maximise the mutual information between label and the joint variable of all the features. The current practice to designing ranking criteria can be viewed as heuristic "*bottom-up*" approaches—manually construct a criterion composed of various terms, that attempt to balance relevance against redundancy at each step, with the expectation that it should have a desirable effect on our overall objective. In this work we take a principled "*top-down*" perspective. We begin with the objective $I(X_{1:n};Y)$, and *analytically expand it* into all possible correlations that exist within the feature set. Then, rather than starting with the marginal information $I(X_i;Y)$, and adding arbitrary terms, we start from the expanded information, and *discard* terms. Note that we are **not** offering a *prescriptive* framework here, that is we do not claim any better reason for discarding a term than including it, given particular data. Instead, we offer a *descriptive* framework, showing that several heuristic criteria in the literature (Battiti, 1994; Yang and Moody, 1999; Kwak and Choi, 2002; Vidal-Naquet and Ullman, 2003; Fleuret, 2004; Peng et al., 2005; Lin and Tang, 2006), can be expressed in a common functional form.

A key component of our approach is to use *multivariate* mutual information. While Shannon's mutual information $I(X;Y)$ measures dependence between a *pair* of variables, the multivariate form, known as *Interaction Information* (McGill, 1954), can account for dependencies among *multiple* variables: $I(\{X,Y,Z\})$. For a set of size 2, the Interaction Information reduces to Shannon's definition. An important and nonintuitive property is that the interaction information can take *negative* values. See Appendix A for the definition and properties.

Theorem 1
Given a set of input features $S = \{X_1, ..., X_n\}$, and a target Y, their Shannon mutual information can be expanded as

$$I(X_{1:n};Y) = \sum_{T \subseteq S} I(\{T \cup Y\}), \qquad |T| \geq 1. \qquad (3)$$

That is, the Shannon Mutual Information between $X_{1:n}$ and Y expands into a sum of Interaction Information terms. Note that $\sum_{T \subseteq S}$ should be read, "sum over all possible subsets T drawn from S".

Example: As an illustrative example for the 4 variable case, the Shannon information between a joint variable $X_{1:3}$ and a target Y can be re-written as

$$\begin{aligned} I(X_{1:3};Y) \;\; = & \\ & I(\{X_1,Y\}) + I(\{X_2,Y\}) + I(\{X_3,Y\}) \\ & +I(\{X_1,X_2,Y\}) + I(\{X_1,X_3,Y\}) + I(\{X_2,X_3,Y\}) \\ & +I(\{X_1,X_2,X_3,Y\}). \qquad (4) \end{aligned}$$

This result explicitly separates the interactions of every possible pair, triple, quadruple (etc) of variables.

We will now show how this can be exploited to better understand several criteria that have appeared in the literature to date.

What would it mean if we *truncated* the expansion at a certain order? For example, keeping only terms with $|T| = 2$? In this case, we would have terms $I(X_i;Y)$ for all i, plus terms $I(\{X_j, X_k, Y\})$ for all j, k, but no further terms. When Interaction Information is calculated for sets of size 3, it represents *pairwise* and *conditional pairwise* interactions. When the set is larger than 3, the information represents properties of higher-order interactions—above pairwise interactions. In summary, if we truncate terms beyond $|T| = 2$, we assume there exist only conditional and unconditional pairwise relations, and no higher order relations. In the following we will show that *all currently used filter criteria assume exactly this.*

We proceed and retain only terms involving *pairs* of features; so all terms where $|T| \leq 2$. This gives us,

$$I(X_{1:n};Y) \approx \sum_{i=1}^{n} I(X_i;Y) + \sum_{j=1}^{n}\sum_{k=j+1}^{n} I(\{X_j, X_k, Y\}). \tag{5}$$

This is an approximation to the Shannon information, assuming pairwise feature interactions, for a feature set of size n. We now take the next important step to understand the filtering process. If we already had $n-1$ features, then the *utility* of X_n, i.e. the information gain when it is included, is quantified as $I(X_n;Y|X_{1:n-1}) = I(X_{1:n};Y) - I(X_{1:n-1};Y)$. Using the approximation in (5) for these terms, we obtain an estimate for this as

$$I(X_n;Y|X_{1:n-1}) \approx I(X_n;Y) + \sum_{k=1}^{n-1} I(\{X_n, X_k, Y\}). \tag{6}$$

which, using the definition of interaction information, can be re-written as,

$$J_{fou} = I(X_n;Y) - \sum_{k=1}^{n-1}\Big[I(X_n;X_k) - I(X_n;X_k|Y)\Big]. \tag{7}$$

We call this, the *first-order utility* (FOU) of including feature X_n. It is *first*-order because it includes only first-order (∼pairwise) interactions. The FOU for X_n is composed of three parts: its own mutual information, subtract a positive term penalising high correlations between itself and the existing features, *plus* another positive term dependent on the class-conditional probabilities. This tells us that the best feature is a trade-off between these components: the individual predictive power of the feature, the unconditional correlations, and the class-conditional correlations. Understanding that we need a trade-off between these components, we could parameterize it, as so,

$$J = I(X_n;Y) - \beta\sum_{k=1}^{n-1} I(X_n;X_k) + \gamma\sum_{k=1}^{n-1} I(X_n;X_k|Y). \tag{8}$$

A remarkable similarity to MIFS is self-evident. With $\gamma = 0$, this is exactly the MIFS criteria. If we allow β and γ to vary in $[0, 1]$, we have a *unit square* describing a space of possibilities. In the following section, we will see that many heuristic criteria already in the literature can be reproduced from eq(8), and are in fact points within the space.

4 SUBSUMING PREVIOUS CRITERIA

In a wide survey of the feature selection literature, we have to date identified 12 separate criteria that can all be described within this framework. In this section, due to space limitations, we present a selection of the most well-known criteria.

Each of the following have justified their use with different arguments, all with the central aspiration to "increase feature relevancy" and "decrease feature redundancy". Where a re-writing of the heuristic is necessary, proofs are in the appendix.

Battiti (1994) proposed the *Mutual Information-Based Feature Selection* (MIFS) criterion,

$$J_{mifs} = I(X_n;Y) - \beta\sum_{k=1}^{n-1} I(X_n;X_k). \tag{9}$$

The MIFS scheme shows a clear link to eq (8) is seen—it includes relevance and redundancy, but omits the conditional term.

Peng et al. (2005) propose the *Maximum-Relevance Minimum-Redundancy* criterion,

$$J_{mrmr} = I(X_n;Y) - \frac{1}{n-1}\sum_{k=1}^{n-1} I(X_n;X_k). \tag{10}$$

It can clearly be seen that MRMR is equivalent to MIFS with $\beta = \frac{1}{n-1}$, and that it takes the mean of the redundancy term, but again omits the conditional term.

Yang and Moody (1999) propose using *Joint Mutual Information* (JMI),

$$J_{jmi} = \sum_{k=1}^{n-1} I(X_n X_k;Y) \tag{11}$$

This is the information between the targets and a joint random variable, defined by pairing the candidate X_n

with each current feature. This can be re-written as,

$$= I(X_n;Y) - \frac{1}{n-1}\sum_{k=1}^{n-1}\Big[I(X_n,X_k) - I(X_n;X_k|Y)\Big]. \tag{12}$$

Intermediate steps for this re-writing are given in Appendix C. JMI fully captures the redundancy term, but takes the mean value. We can also see a close relationship between MRMR and JMI, i.e. the JMI criterion is the MRMR criterion plus $\frac{1}{n-1}\sum_{k=1}^{n-1} I(X_n;X_k|Y)$.

Kwak and Choi (2002) propose an improvement to MIFS, called MIFS-U. This claimed to be more suited to problems where information is distributed *uniformly* in the space, by using the criterion,

$$J_{mifsu} = I(X_n;Y) - \beta\sum_{k=1}^{n-1}\frac{I(X_k;Y)}{H(X_k)}I(X_n;X_k). \tag{13}$$

in practice, the authors found $\beta = 1$ optimal. In this case, the equivalence of eq(13) and eq(7) can be proven; that is, $J_{mifsu} = J_{fou}$.

Lin and Tang (2006) propose a criterion for the Computer Vision literature, called *Conditional Infomax Feature Extraction*. This turns out to be exactly equivalent to our own proposal, eq(7). The key difference between this and our result is that we have derived the general case, allowing arbitrary orders of expansion.

The criteria we have described so far can all be rearranged into a common functional form, such that they can be exactly reproduced from various parameterizations of eq(8). Consequently, they all fit neatly into a unit square, illustrated in figure 1. These are *linear* criteria, as they take linear combinations of the relevance/redundancy terms. We will now cover two criteria that follow a similar form, with the same relevance/redundancy terms, though they involve a *max* operation, making them *nonlinear*.

Fleuret (2004) is probably the most well-known recent criterion, based on *Conditional Mutual Information Maximization*,

$$J_{cmim} = min_k\Big[I(X_n;Y|X_k)\Big]. \tag{14}$$

This can be re-written,

$$= I(X_n;Y) - max_k\Big[I(X_n;X_k) - I(X_n;X_k|Y)\Big]. \tag{15}$$

The proof is again available in the appendix. CMIM examines the information between a feature and the target, *conditioned on* each current feature. From the re-writing, it is clearer why the CMIM criterion succeeds. The term in square brackets is the interaction information—which can be both positive and negative. A negative value indicates that the shared information between X_k and Y has *decreased* as a result of including X_n. CMIM takes the smallest value, therefore identifying that X_n interacts badly with at least one of the existing features. CMIM therefore takes a “pessimistic” view of X_n if it interacts badly with the existing set.

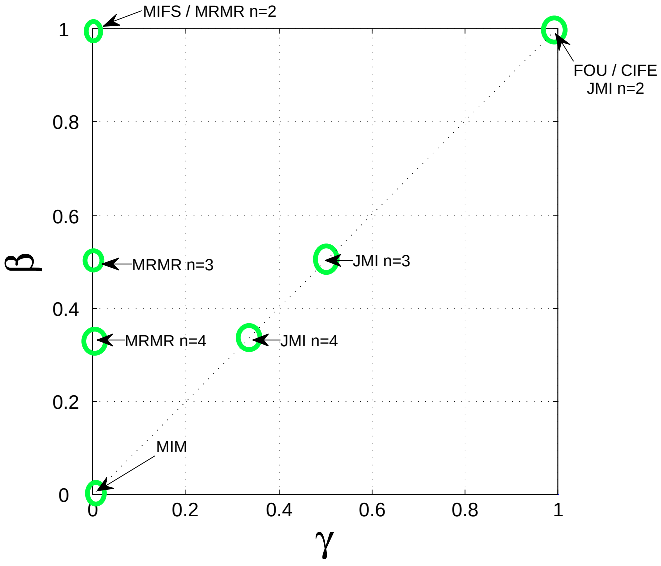

Figure 1: The full space of first-order linear filter criteria, derived from equation (8). The left hand axis of this graph is where the MRMR and MIFS algorithms sit. The bottom right corner is the assumption of independent features, using just the marginal mutual information.

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

$$J_{if} = min_k\Big[I(X_nX_k;Y) - I(X_k;Y)\Big]. \tag{16}$$

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

5 EXPERIMENTS

An interesting question is how the criteria perform relative to each other in practice. In this section we examine the overfitting behaviours, with the hypothesis that criteria involving both the conditional and uncon-

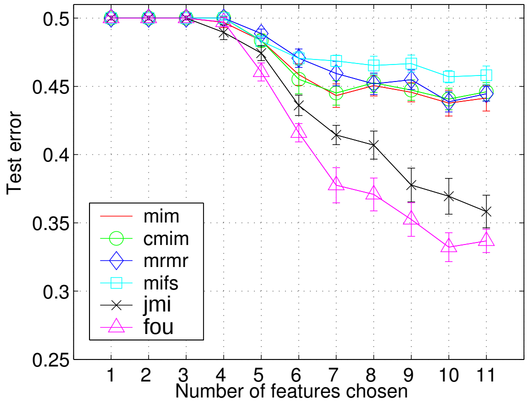

Figure 2: Noisy multiplexer problem: FOU and JMI show significant benefit compared to criteria measuring only pairwise dependencies.

ditional redundancy terms should overfit to a greater extent, considering the larger number of parameters.

5.1 ARTIFICIAL DATA

In a first experiment we wish to evaluate the criteria on data *known* to exhibit very strong feature interdependencies. The *multiplexer problem* is a boolean function used in signal processing. This is defined on binary strings of length 11, and treats the string as being composed of an index segment (the first 3 bits) and a data segment (the remaining 8 bits). The output of the function is the value of the indexed bit in the data segment; so, for example, the correct output for `10100000100` is 1, since the first three (index) bits `101` point to data bit 5. This function is particularly interesting as the features interact heavily—as such we expect only the criteria that assess higher order interactions will perform well. To further test the criteria, 10 random boolean features were added as noise, giving a total 21 features, 2048 examples. We used 10% for training, 90% for testing, over ten trials, with simple forward selection. The selected features were used in a 1-nearest neighbour classifier. Results clearly show that FOU and JMI are the top performers, explained by the fact that they include the multi-way interaction between two features and the target.

5.2 REAL DATA

The Lung and SRBCT datasets, are freely available gene expression datasets. Figures 3 and 4 show experiments with these. We used a 1-nn classifier, and performed a leave-one-out cross validation, at each step, we also restricted the amount of training data that was made available to the classifier, thus giving us a *learning curve*. The far left of the graphs represent the data-poor training environments—both Lung and SRBCT show the FOU criterion severely overfitting, while the less complex criteria manage to resist until the extreme data limit. At this extreme, note that *all criteria eventually overfit.* In the data-rich environments, it appears the optimal criteria that incorporate the right dependency assumptions are MRMR/JMI.

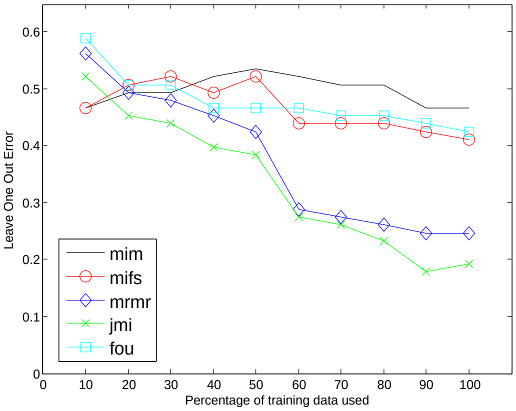

Figure 3: Lung data. Learning curve with 10 features, showing that all criteria overfit, and become essentially equivalent when data is limited. At the extreme (10%) the more complex the criterion, the more it overfits.

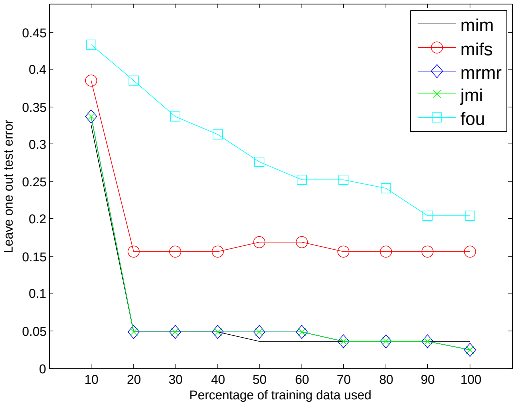

Figure 4: SRBCT data. Learning curve with 10 features. Simpler criteria resist overfitting until the extreme data limit.

The experiments we have performed so far are all with previous studied criteria. From figure 1 it is evident that there *exist many points within the space that have never been explored.* Figures 5 and 6 explore the space, illustrating errors (a smaller square means lower error), when picking 10 features for a 1-nn classifier. On the Lung data the minimum in the space is $\{\beta = 0.8, \gamma = 0.2\}$. On the SRBCT data, it is at $\{\beta = 0.6, \gamma = 0.8\}$. Neither of these points correspond exactly to any previously studied criteria. Whilst a full empirical study is outside the scope of this paper, this serves as proof-of-concept that this space warrants investigation.

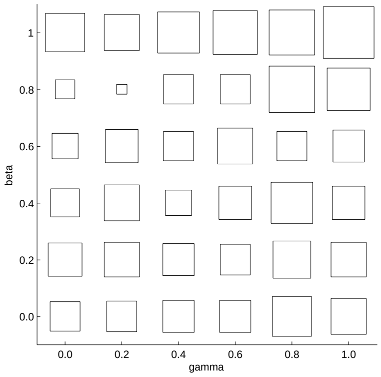

Figure 5: Lung data. Error rates for a 1-nn classifier, using 10 features. Each coordinate in the space corresponds to a different criterion, some previously investigated, some not.

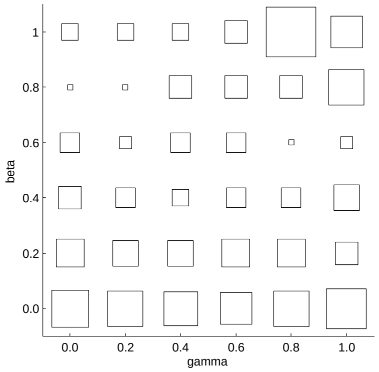

Figure 6: SRBCT data. Error rates for a 1-nn classifier, using 10 features.

6 CONCLUSIONS

We have presented a unifying viewpoint on the existing information theoretic feature ranking literature. The core of this is a 'root' criterion, which explicitly quantifies the *utility* of a new feature. We showed how several published heuristics are reproduced by parameterizations of this root criterion. All the heuristics assume first-order feature interactions, and some additionally omit the conditional term $I(X_n, X_k|Y)$. Criteria that omit this conditional redundancy term are MIFS and MRMR. Criteria that incorporate both redundancy and conditional redundancy are CMIM, IF, JMI and MIFS-U. Since $I(X_n, X_k|Y)$ is positive, these latter criteria could be seen as more "aggressive" than the former, assuming a higher utility for the same candidate feature. The consequence of this was demonstrated in the overfitting behaviour of the criteria. This approach has resulted in theoretical explanations for the success and relations of several recently published heuristics (Battiti, 1994; Yang and Moody, 1999; Kwak and Choi, 2002; Vidal-Naquet and Ullman, 2003; Fleuret, 2004; Peng et al., 2005; Lin and Tang, 2006). This also opens the door for a *data-driven* framework for feature selection. If we have some a-priori belief of the dependencies within the dataset, we may be able to build these into our choice of β and γ. Further work will also address the relative merits of using *higher-order* feature interactions in the criteria.

It is important to note that, we do not claim eq(8) as a universally superior criterion. The success of any filter criterion will depend on the *true* dependencies in the data, on the dependence structure assumed (explicitly or implicitly) by the classifier, and also on the search strategy employed. Instead, we hope that this work is a step toward a longer term goal, a solid mathematical foundation for feature selection methodology.

Appendix A : Multivariate Information

Several authors have offered what they claim to be "natural" multi-variate extensions of Shannon's mutual information. Each definition has its own properties, some of which are desirable and some undesirable. It is widely acknowledged that McGill (1954) was the first to propose a multi-variate mutual information measure, the *Interaction Information.* We use the 2-arity function $I(\cdot;\cdot)$ to denote the Shannon Mutual Information between two variables, while the 1-arity function $I(\cdot)$ denotes McGill's Interaction Information among all variables in the supplied set argument. For *three* random variables, the *Interaction Information* is

$$I(\{X_1, X_2, X_3\}) = I(X_1; X_2|X_3) - I(X_1; X_2), \quad (17)$$

that is, a difference of the conditional mutual information and the simple mutual information. The case for n variables is defined recursively,

$$I(\{X_1, ..., X_n\}) = \\ I(\{X_1, ..., X_{n-1}\}|X_n) - I(\{X_1, ..., X_{n-1}\}). \quad (18)$$

The conditional form is defined by simply marginalising over the distribution of X_n. A general closed-form definition can be shown by re-writing it as entropies, with $S = \{X_1, ..., X_n\}$,

$$I(S) = -\sum_{T \subseteq S} (-1)^{|S \setminus T|} H(T), \quad (19)$$

where $\sum_{T \subseteq S}$ denotes a sum over all possible subsets drawn from S. The case for $n = 2$ reduces to Shannon's definition. An important property to note is that the interaction information *can be negative*—it can be understood as the gain (or loss) of information between two variables, due to additional knowledge of a third. A positive value indicates a synergistic interaction among the variables; that is, the shared information between A and B has increased as a result of observing C. A negative value indicates the opposite, that shared information is decreased by observing C. This is a highly non-intuitive property of information content. A reader less practised with information theory might assume since conditioning *reduces* entropy, $I(A;B|C) < I(A;B)$, therefore (17) is *always* negative. However, as these are *sums* of entropies, and it can easily be the opposite case, $I(A;B|C) > I(A;B)$.

As an example, imagine A and B are two independently sampled binary values, and that $C = \neg(A \ \mathit{xor} \ B)$, i.e. $C = 1$ when both are true, or both false. Without knowledge of C, the information between A and B is zero, however when C is revealed, A and B become completely dependent, therefore $I(A;B|C) > I(A;B)$. The converse can also be shown. Imagine A and B are now noisy observations of another variable C, i.e. $A = (C \ \mathit{xor} \ \phi)$ and $B = (C \ \mathit{xor} \ \phi)$, where ϕ is a function returning true/false with equal probability when called. Now when C is revealed, the only component left is ϕ, a completely random quantity, therefore the information between A and B is again 0 and $I(A;B|C) < I(A;B)$. It is important to note that the quantity $I(ABC;D)$ has a very different meaning to $I(\{A,B,C,D\})$. The former is Shannon's mutual information between the joint random variable ABC and the variable D; the latter is McGill's interaction information among all four variables $\{A,B,C,D\}$. Theorem 1 in the main body of this paper shows a deep connection between these two definitions.

Appendix B : Proof of Theorem 1

To prove theorem 1, we use a classic technique from number theory, Möbius inversion. This is a method for inverting finite sums over partially ordered sets — for our use, the poset is the set of all subsets of $S = \{X_1, ..., X_n\}$, ordered by inclusion. For a set S, if we define a function

$$f(S) = \sum_{T \subseteq S} f'(T), \qquad (20)$$

where $\sum_{T \subseteq S}$ should be read, "sum over all possible subsets T drawn from S". Then, we can write $f'(\cdot)$ as,

$$f'(S) = \sum_{T \subseteq S} \mu(S,T) f(T), \qquad \mu(S,T) = (-1)^{|S \setminus T|} \qquad (21)$$

Here, $\mu(S,T)$ is the classical Möbius function from number theory (Rota, 1964). Hence $f'(\cdot)$ is the *inversion* of $f(\cdot)$, and vice versa. The formula allows us to define either $f(\cdot)$ or $f'(\cdot)$, and derive the other without ambiguity. For further details, we refer the reader to an excellent survey (Bender and Goldman, 1975), discussing generalisations of the formula and its use in combinatorial analysis.

For our own needs, with $S = \{X_1, ..., X_n\}$, we define $f(S) = I(X_{1:n};Y)$, the Shannon information. Then from (21) we have

$$f'(S) = \sum_{k=1}^{|S|} \sum_{\substack{T \subseteq S, \\ |T|=k}} (-1)^{|S \setminus T|} \Big[H(Y) + H(T) - H(TY) \Big] \qquad (22)$$

where we note that for the subset $T = \emptyset$, we have $I(T;Y) = 0$. The $H(Y)$ in eq (22) is independent of the sum, so can be separated and simplified thus

$$\begin{aligned} \sum_{k=1}^{|S|} \sum_{\substack{T \subseteq S, \\ |T|=k}} (-1)^{|S \setminus T|} H(Y) &= \sum_{k=1}^{|S|} \binom{|S|}{k} (-1)^{|S|-k} H(Y) \\ &= (-1)^{|S|-1} H(Y). \qquad (23) \end{aligned}$$

The remaining entropy terms in eq (22) can be rewritten in a more compact notation, using T_k to indicate a set T of size k, and so $|S \setminus T_k| = |S| - k$, giving us

$$\phi = \sum_{k=1}^{|S|} (-1)^{|S|-k} \Big[\sum_{T_k \subseteq S} H(T_k) - \sum_{T_k \subseteq S} H(T_k Y) \Big]$$

We note that the following property holds,

$$\sum_{T_k \subseteq S} H(T_k) + \sum_{T_{k-1} \subseteq S} H(T_{k-1} Y) = \sum_{T_k \subseteq \{S \cup Y\}} H(T_k) \qquad (24)$$

For example, (a function of) all subsets of size 3 drawn from S, plus (a function of) Y with all subsets of size 2 drawn from S, gives us all subsets of size 3 drawn from $S \cup Y$. If $S = \{A,B,C,D\}$, then all size 3 subsets is $\{\{ABC\},\{ABD\},\{BCD\},\{ACD\}\}$. All size 2 subsets, each unioned with Y, is $\{\{ABY\},\{ACY\},\{ADY\},\{BCY\},\{BDY\},\{CDY\}\}$. Finally the union of these two sets gives us all subsets of size 3 from $\{S \cup Y\}$. This property allows us to rearrange our expression for ϕ,

$$\begin{aligned} \phi &= \sum_{k=1}^{|S|} (-1)^{|S|-k} \Big[\sum_{T_k \subseteq S} H(T_k) - \sum_{T_k \subseteq S} H(T_k Y) \Big] \\ &= (-1)^{|S|-1} \Big[\sum_{T_1 \subseteq S} H(T_k) \Big] + \\ &\qquad \sum_{k=2}^{|S \cup Y|} (-1)^{|S|-k} \Big[\sum_{T_k \subseteq \{S \cup Y\}} H(T_k) \Big] \qquad (25) \end{aligned}$$

Recombining this with eq (23) and rearranging,

$$\begin{aligned} f'(S) &= \sum_{k=1}^{|S|} \sum_{T_k \subseteq \{S \cup Y\}} (-1)^{|S \setminus T_k|} H(T_k) \\ &= -\sum_{k=1}^{|S \cup Y|} \sum_{T_k \subseteq \{S \cup Y\}} (-1)^{|\{S \cup Y\} \setminus T_k|} H(T_k) \\ &= I(\{S \cup Y\}) \qquad (26) \end{aligned}$$

where the final step uses the closed-form definition of interaction information, eq(19). We now reinsert this into (20) and we have the desired result.

Appendix C : Proof of eqs (12) and (15)

The following proofs make use of the information identity, $I(A;B|C) - I(A;B) = I(A;C|B) - I(A;C)$. The *Joint Mutual Information* criterion (Yang and Moody, 1999) can be written,

$$\begin{aligned} J_{jmi} = & \sum_{k=1}^{n-1} I(X_n X_k; Y) \\ = & \sum_{k=1}^{n-1} \Big[I(X_k;Y) + I(X_n;Y|X_k) \Big] \end{aligned}$$

The term $\sum_{k=1}^{n-1} I(X_k;Y)$ in the above is constant with respect to the X_n argument that we are interested in, so can be omitted. The criterion therefore reduces to,

$$\begin{aligned} J_{jmi} \; &= \sum_{k=1}^{n-1} \Big[I(X_n;Y|X_k) \Big] \\ = \sum_{k=1}^{n-1} &\Big[I(X_n;Y) - I(X_n;X_k) + I(X_n;X_k|Y) \Big] \\ = (n-1)I(X_n;Y) &- \sum_{k=1}^{n-1} \Big[I(X_n;X_k) - I(X_n;X_k|Y) \Big] \end{aligned}$$

Multiplying the above by the constant $\frac{1}{n-1}$ gives us eq(12). The rearrangement of the Conditional Mutual Information criterion (Fleuret, 2004) follows a very similar procedure. The original, and its rewriting are,

$$\begin{aligned} J_{cmim} &= min_k \Big[I(X_n;Y|X_k) \Big] \\ &= min_k \Big[I(X_n;Y) - I(X_n;X_k) + I(X_n;X_k|Y) \Big] \\ &= I(X_n;Y) + min_k \Big[I(X_n;X_k|Y) - I(X_n;X_k) \Big] \\ &= I(X_n;Y) + min_k \Big[I_m(\{X_n, X_k, Y\}) \Big]. \end{aligned}$$

ACKNOWLEDGEMENTS

The author would like to thank Mikel Lujan and James Neil for many useful comments on this work.

References

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

Bender, E. A. and Goldman, J. R. (1975). On the Applications of Mobius Inversion in Combinatorial Analysis. *Amer. Math. Monthly*, 82:789–803.

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

Fleuret, F. (2004). Fast Binary Feature Selection with Conditional Mutual Information. *The Journal of Machine Learning Research*, 5:1531–1555.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3(7-8):1157–1182.

Kwak, N. and Choi, C. (2002). Input Feature Selection for Classification Problems. *Neural Networks, IEEE Transactions on*, 13(1):143–159.

Lin, D. and Tang, X. (2006). Conditional Infomax Learning: An Integrated Framework for Feature Extraction and Fusion. In *European Conference on Computer Vision.*

McGill, W. (1954). Multivariate information transmission. *IEEE Trans. Inf. Theory*, 4(4):93–111.

Peng, H., Long, F., and Ding, C. (2005). 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.

Rota, G. (1964). On the Foundations of Combinatorial Theory I. Theory of Mobius Functions. *Probability Theory and Related Fields*, 2:340–368.

Shannon, C. (1948). A mathematical theory of communication, Bell Syst. *Tech. J*, 27(3):379–423.

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

Yang, H. and Moody, J. (1999). Data Visualization and Feature Selection: New Algorithms for Nongaussian Data. *Advances in Neural Information Processing Systems*, 12.