Informative Priors for Markov Blanket Discovery

Adam Pocock
School of Computer Science
University of Manchester
Manchester, M13 9PL
apocock@cs.manchester.ac.uk

Mikel Luján
School of Computer Science
University of Manchester
Manchester, M13 9PL
mlujan@cs.manchester.ac.uk

Gavin Brown
School of Computer Science
University of Manchester
Manchester, M13 9PL
gbrown@cs.manchester.ac.uk

Abstract

We present a novel interpretation of information theoretic feature selection as optimization of a discriminative model. We show that this formulation coincides with a group of mutual information based filter heuristics in the literature, and show how our probabilistic framework gives a well-founded extension for informative priors. We then derive a particular sparsity prior that recovers the well-known IAMB algorithm (Tsamardinos & Aliferis, 2003) and extend it to create a novel algorithm, IAMB-IP, that includes domain knowledge priors. In empirical evaluations, we find the new algorithm to improve Markov Blanket recovery even when a misspecified prior was used, in which half the prior knowledge was incorrect.

1 Introduction

Classifying examples is one of the most fundamental tasks in Machine Learning. *Feature selection* (FS) is the process of determining which inputs or *features* should be presented to the classification algorithm. In the past this was performed by domain experts, who chose features they thought relevant to the classification task. Many modern machine learning domains collect as many features as possible and use a *feature selection algorithm* to find a *relevant* subset of features to use in classification. An even more recent trend has been the incorporation of domain experts back into the feature selection process, to guide it in complex spaces. The question of how to integrate such *domain knowledge* into an algorithm is therefore important.

Appearing in Proceedings of the 15^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2012, La Palma, Canary Islands. Volume 22 of JMLR: W&CP 22.

Helleputte & Dupont (2009) present a good example of much recent work on informative priors for regularized linear models. They develop a prior for an approximate zero-norm minimization, constraining some of the dimensions according to knowledge gained from the biological literature. Krupka et al. (2008) define *meta-features*, where each feature is qualified by additional information, and a mapping is learned from meta-features to some measure of feature 'quality'. Knowledge can then be transferred between tasks by learning the feature-quality mapping; however the challenge remains to define a good quality measure and reliably learn the mapping. In the special case where the features can be assumed to be faithful to an unknown Bayesian network, the feature selection problem is equivalent to learning the local structure of the network around the target node. Mukherjee & Speed (2008) present a method of learning global structures with priors on network properties such as sparsity and degree, as well as specific arc constraints. The generated network contains the relationships between all of the nodes (features), hence the feature selection problem can be solved by inspecting the Markov Blanket (the set of child, parent and spouse nodes) for any given target node. However, the possible network structures grow super-exponentially with the number of nodes, and so the procedure becomes impractical with large datasets.

In this paper we begin from a clear probabilistic model of the data, and derive an iterative filter feature selection algorithm which optimizes our model. We show how our probabilistic interpretation gives a well-founded extension in the form of informative priors for sparsity and domain knowledge. We show how a particular choice of sparsity prior recovers the well-known IAMB algorithm (Tsamardinos & Aliferis, 2003) and create a novel algorithm IAMB-IP which extends IAMB to include domain knowledge priors.

2 The feature selection problem

There are three main types of feature selection algorithm: filters, wrappers and embedded methods (Guyon et al., 2006). In this work we concentrate on filters, which are commonly specified as a correlation measure (or filter "criterion") between features and class labels. Features exhibiting a higher value of the criterion are favored for inclusion in the feature set, with the assumption that higher correlation implies more predictive power. However the exact relationship between the criterion and the classification error is usually unclear. An alternative perspective is to define a probabilistic model, and then *derive* the appropriate criterion to maximize the posterior of that model. Therefore we begin by defining a discriminative model and deriving an error term for the feature selection process; this error term is a root criterion from which many mutual information based selection criteria can be derived.

2.1 Notation

We assume an underlying true distribution $p(\mathbf{x}, y)$ from which we have an i.i.d. sample of N observations, denoted as $\mathcal{D} = \{(\mathbf{x}^i, y^i); i = 1...N\}$. Each observation is a pair $(\mathbf{x}, y)$, consisting of a d-dimensional feature vector $\mathbf{x} = \{x_1, ..., x_d\}$, and a target class y, drawn from the underlying variables $X = \{X_1, ..., X_d\}$ and Y. We further assume that $p(y|\mathbf{x})$ is defined by a *subset*, X^*, of the features X, while the remaining features are redundant or irrelevant. We adopt a d-dimensional binary vector $\boldsymbol{\theta}$, specifying the selected features: a 1 indicates the feature is selected, and a 0 indicates it is discarded. We use $\neg\boldsymbol{\theta}$ for the negation of $\boldsymbol{\theta}$, i.e. the unselected features. We then define $X_{\boldsymbol{\theta}}$ as the set of selected features, and $X_{\neg\boldsymbol{\theta}}$ as the set complement of $X_{\boldsymbol{\theta}}$, the set of unselected features. Therefore $X = X_{\boldsymbol{\theta}} \cup X_{\neg\boldsymbol{\theta}}$, as $X_{\boldsymbol{\theta}}$ and $X_{\neg\boldsymbol{\theta}}$ form a partition. We use $\mathbf{x}_{\boldsymbol{\theta}}$ for an observation of the selected features $X_{\boldsymbol{\theta}}$, and similarly for $\mathbf{x}_{\neg\boldsymbol{\theta}}$. We define $p(y|\mathbf{x}, \boldsymbol{\theta})$ as $p(y|\mathbf{x}_{\boldsymbol{\theta}})$, and use the latter when specifically talking about feature selection. We then formally define X^* as the minimal feature set s.t. $\forall\ \mathbf{x}, y\ p(y|\mathbf{x}_{\boldsymbol{\theta}^*}) = p(y|\mathbf{x})$ and use $\boldsymbol{\theta}^*$ as the vector indicating this feature set. The feature selection problem is to identify this vector. We define τ as the other model parameters involved in the generation of class labels, and λ as the generative parameters for the observations $\mathbf{x}$. We use $||$ to denote the KL-Divergence between two distributions.

2.2 A discriminative model for FS

In this section we decompose the likelihood of a discriminative model into a sum of terms, each with a natural interpretation in the context of the feature selection problem. When making the commonly adopted filter assumption that the feature selection and model fitting processes can be performed separately, this leads naturally to the choice of the mutual information as a filter criterion.

We approximate the true distribution p with a hypothetical model q, with separate parameters for feature selection, $\boldsymbol{\theta}$, and classification, τ. Following Minka (2005) and Lasserre et al. (2006), in the construction of a discriminative model, the joint likelihood is

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

As this is a discriminative model we wish to maximize $\mathcal{L}$ with respect to $\boldsymbol{\theta}$ (our feature selection parameters) and τ (our model parameters), thus we are not concerned with the generative parameters λ. Excluding the generative terms gives

$$\mathcal{L}(\mathcal{D}, \boldsymbol{\theta}, \tau, \lambda) \propto p(\boldsymbol{\theta}, \tau)\prod_{i=1}^{N} q(y^i|\mathbf{x}^i, \boldsymbol{\theta}, \tau). \tag{2}$$

We wish to find the Maximum a Posteriori (MAP) solution, with respect to the parameters $\{\boldsymbol{\theta}, \tau\}$. We choose to work with the scaled negative log-likelihood, $-\ell$, converting our maximization problem into a minimization problem, without changing the position of the optima. This gives

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

which is the function we will minimize with respect to $\{\boldsymbol{\theta}, \tau\}$; the scaling term is to simplify exposition later.

We are interested in decomposing this likelihood to extract terms related to feature selection and to classification. We begin by introducing the ratio $\frac{p(y^i|\mathbf{x}^i)}{p(y^i|\mathbf{x}^i)}$ into the logarithm. This is the probability of the correct class given all the features. As this ratio is unity it does not change the value of the log likelihood, nor the positions of its optima. We can then expand the resulting logarithm to give several terms,

$$\begin{aligned} -\ell = -\frac{1}{N}\Bigg(&\sum_{i=1}^{N} \log \frac{q(y^i|\mathbf{x}^i, \boldsymbol{\theta}, \tau)}{p(y^i|\mathbf{x}^i)} \\ &+ \sum_{i=1}^{N} \log p(y^i|\mathbf{x}^i) + \log p(\boldsymbol{\theta}, \tau)\Bigg). \end{aligned} \tag{4}$$

This expands the likelihood into 3 terms: the log-likelihood ratio between the true model and our predictive model, the log-likelihood of the true model, and the prior term. This middle term is a finite sample approximation to the conditional entropy $H(Y|X)$

and represents the total amount of uncertainty there is about the class label given the data. The conditional entropy is the log-likelihood of the true model when taking the limit of data points.

We are concerned with separating out the influence of feature selection and classification in our model, and thus introduce an extra ratio $\frac{p(y^i|\mathbf{x}^i,\boldsymbol{\theta})}{p(y^i|\mathbf{x}^i,\boldsymbol{\theta})}$ into the first term. This is the probability of the correct class given the features we have selected with $\boldsymbol{\theta}$. We can then further expand the first logarithm as follows,

$$-\ell = -\frac{1}{N}\Bigg(\sum_{i=1}^{N}\log\frac{q(y^i|\mathbf{x}^i,\boldsymbol{\theta},\tau)}{p(y^i|\mathbf{x}^i,\boldsymbol{\theta})}+\sum_{i=1}^{N}\log\frac{p(y^i|\mathbf{x}^i,\boldsymbol{\theta})}{p(y^i|\mathbf{x}^i)} + \sum_{i=1}^{N}\log p(y^i|\mathbf{x}^i)+\log p(\boldsymbol{\theta},\tau)\Bigg). \tag{5}$$

As before we have the log likelihood of the true model and the prior term. We have now separated out the first log likelihood ratio into two terms. The first term is the ratio between our predictive model and the true distribution of the labels given our selected subset of features. This represents how well our model fits the data given the current set of features. When it is zero our model has the best possible fit given the features selected. The second term is the ratio between the true distribution given the selected features, and the true distribution of the labels given all the data. This measures the quality of the selected feature set $\boldsymbol{\theta}$, based on how close the conditional distribution of y is to the one conditioned on all the data. We can see that this term is a finite sample approximation to the KL-Divergence between $p(y|\mathbf{x})$ and $p(y|\mathbf{x},\boldsymbol{\theta})$. Thus we can write $-\ell$ as the sum of information theoretic quantities plus the prior over $\{\boldsymbol{\theta},\tau\}$ like so

$$-\ell \approx \mathbb{E}_{\mathbf{x},y}\left\{\log\frac{p(y^i|\mathbf{x}^i,\boldsymbol{\theta})}{q(y^i|\mathbf{x}^i,\boldsymbol{\theta},\tau)}\right\}+\mathbb{E}_{\mathbf{x},y}\{p(y|\mathbf{x})||p(y|\mathbf{x},\boldsymbol{\theta})\} + H(Y|X)-\frac{1}{N}\big(\log p(\boldsymbol{\theta},\tau)\big). \tag{6}$$

Assuming for the moment that we have the optimal feature set or a superset thereof (i.e. $X^* \subseteq X_{\boldsymbol{\theta}}$) then $p(y|\mathbf{x},\boldsymbol{\theta}) = p(y|\mathbf{x})$. Then as the expectation in the first term is over $p(y,\mathbf{x})$, the first term can be seen as a finite sample approximation to the expected KL-Divergence over $p(\mathbf{x})$ representing how well the predictive model fits the true distribution, given a superset of the optimal feature set. It is interesting to note that the second term in Eq (6) is precisely that introduced by Koller & Sahami (1996) in their definitions of optimal feature selection. In their work, the term was adopted as a sensible objective to follow—with Eq (6) we show it to be a direct consequence of adopting the discriminative model in Eq (1). As $\mathbf{x} = \{\mathbf{x}_{\boldsymbol{\theta}}, \mathbf{x}_{\neg\boldsymbol{\theta}}\}$, this term can be developed thus:

$$\begin{aligned}\Delta_{KS} &= \mathbb{E}_{\mathbf{x},y}\{\ \ p(y|\mathbf{x}_{\boldsymbol{\theta}},\mathbf{x}_{\neg\boldsymbol{\theta}})\ ||\ p(y|\mathbf{x}_{\boldsymbol{\theta}})\ \ \}\\ &= \sum_{\mathbf{x},y} p(\mathbf{x},y)\log\frac{p(y|\mathbf{x}_{\boldsymbol{\theta}},\mathbf{x}_{\neg\boldsymbol{\theta}})}{p(y|\mathbf{x}_{\boldsymbol{\theta}})}\frac{p(\mathbf{x}_{\neg\boldsymbol{\theta}}|\mathbf{x}_{\boldsymbol{\theta}})}{p(\mathbf{x}_{\neg\boldsymbol{\theta}}|\mathbf{x}_{\boldsymbol{\theta}})}\\ &= \sum_{\mathbf{x},y} p(\mathbf{x},y)\log\frac{p(\mathbf{x}_{\neg\boldsymbol{\theta}},y|\mathbf{x}_{\boldsymbol{\theta}})}{p(\mathbf{x}_{\neg\boldsymbol{\theta}}|\mathbf{x}_{\boldsymbol{\theta}})p(y|\mathbf{x}_{\boldsymbol{\theta}})}\\ &= I(X_{\neg\boldsymbol{\theta}};Y|X_{\boldsymbol{\theta}}).\end{aligned} \tag{7}$$

This is the conditional mutual information between the class label and the remaining features, given the selected features.

Thus we can decompose the negative log-likelihood into three data dependent terms and the prior term,

$$-\ell \approx \mathbb{E}_{\mathbf{x},y}\left\{\log\frac{p(y^i|\mathbf{x}^i,\boldsymbol{\theta})}{q(y^i|\mathbf{x}^i,\boldsymbol{\theta},\tau)}\right\}+I(X_{\neg\boldsymbol{\theta}};Y|X_{\boldsymbol{\theta}}) + H(Y|X)-\frac{1}{N}\log p(\boldsymbol{\theta},\tau). \tag{8}$$

The first term is a measure of the difference between the predictive model q, and the true distribution p. When a superset of the optimal feature set has been found, it becomes the KL-Divergence between p and q. The second term, $I(X_{\neg\boldsymbol{\theta}};Y|X_{\boldsymbol{\theta}})$, depends solely on the choice of features, and is zero when the unselected features $X_{\neg\boldsymbol{\theta}}$ contain no more useful information about Y. Note that due to the chain rule, $I(AB;Y) = I(A;Y)+I(B;Y|A)$, and $X = X_{\boldsymbol{\theta}}\cup X_{\neg\boldsymbol{\theta}}$,

$$I(X;Y) = I(X_{\boldsymbol{\theta}};Y)+I(X_{\neg\boldsymbol{\theta}};Y|X_{\boldsymbol{\theta}}). \tag{9}$$

Since $I(X;Y)$ is constant, minimizing $I(X_{\neg\boldsymbol{\theta}};Y|X_{\boldsymbol{\theta}})$ is equivalent to maximizing $I(X_{\boldsymbol{\theta}};Y)$. The third term in Eq (8) is $H(Y|X)$, the conditional entropy of the labels given *all features*; this is an irreducible constant, independent of all parameters. This term bounds the Bayes error rate (Fano, 1961), and measures the total amount of information available in all of the features.

We now make an assumption made *implicitly* by all filter methods, that model fitting can be separated from the feature selection process. We make this assumption *explicit* by specifiying the prior $p(\boldsymbol{\theta},\tau)$ factorizes into $p(\boldsymbol{\theta})p(\tau)$, thus decoupling model fitting from feature selection. We note that τ is independent of the second term in our expansion, and by factorising the prior we can select features before fitting the model. This assumption is valid if our model q is a consistent estimator of p, as with increasing N it will more closely approximate the true distribution, and the ratio in Eq (8) will approach zero. Then to maximize the likelihood of the feature set, we are only concerned with how $p(y|\mathbf{x},\boldsymbol{\theta})$ approximates $p(y|\mathbf{x},\boldsymbol{\theta}^*)$, and so we specify the optimization problem that defines the feature

selection task for the model in Eq (1) as,

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \left(I(X_{\neg\boldsymbol{\theta}}; Y | X_{\boldsymbol{\theta}}) - \frac{1}{N} \log p(\boldsymbol{\theta}) \right). \quad (10)$$

Until this point we have worked with the true distribution p. In practice we only have access to an estimate of this distribution, $\hat{p}$, where we have estimated the various quantities from our training data; in our experiments we use histogram estimators. The quality of our estimate $\hat{p}$ will depend upon the number of samples, and the dimensionality of the distribution, with a poorer estimate in low sample, high dimensionality spaces. We note that our estimated mutual information $\hat{I}$, which is calculated based on $\hat{p}$, is a Monte-Carlo estimate and converges *almost surely* to the true mutual information I in the limit of infinite samples. For the remainder of this paper, we use notation $I(X;Y)$ 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 $\hat{I}(X;Y)$. For a detailed study of entropy estimation we refer the reader to Paninski (2003).

In the following section we consider an iterative minimization of Eq (10), and show how this fits with the current filter feature selection literature.

2.3 Iterative minimization

Many filter feature selection techniques adopt a stepwise maximization/minimization of their objective functions (Peng et al., 2005; Brown, 2009). We derive iterative update rules to minimize our objective function, Eq (10), and show how this links to the current literature. We first introduce extra notation, $\boldsymbol{\theta}^t$ and $\boldsymbol{\theta}^{t+1}$, denoting the selected feature set at timesteps t and $t+1$. We use a sequential search, so only one feature is added/removed at each timestep, so there is exactly one bit different between $\boldsymbol{\theta}^t$ and $\boldsymbol{\theta}^{t+1}$. The flipped bit we denote as θ_k.

Theorem 1. *The forward step that optimizes Eq (10) at timestep $t+1$ from timestep t is to add the feature,*

$$X_k^* = \arg\max_{X_k \in X_{\neg\boldsymbol{\theta}^t}} \left(I(X_k; Y | X_{\boldsymbol{\theta}^t}) + \frac{1}{N} \log \frac{p(\boldsymbol{\theta}^{t+1})}{p(\boldsymbol{\theta}^t)} \right). \quad (11)$$

Proof. Denoting the objective at timestep t with J_t, we wish to minimize J_{t+1}. This is equivalent to maximizing the difference $(J_t - J_{t+1})$. The objective at an arbitrary timestep t is:

$$J_t = I(X_{\neg\boldsymbol{\theta}^t}; Y | X_{\boldsymbol{\theta}^t}) - \frac{1}{N} \log p(\boldsymbol{\theta}^t). \quad (12)$$

We wish to add the feature X_k that minimizes J_{t+1}, and thus maximizes the difference $J_t - J_{t+1}$,

$$J_t - J_{t+1} = I(X_{\neg\boldsymbol{\theta}^t}; Y | X_{\boldsymbol{\theta}^t}) - \frac{1}{N} \log p(\boldsymbol{\theta}^t) - I(X_{\neg\boldsymbol{\theta}^{t+1}}; Y | X_{\boldsymbol{\theta}^{t+1}}) + \frac{1}{N} \log p(\boldsymbol{\theta}^{t+1}). \quad (13)$$

After applying the chain rule of mutual information we arrive at:

$$\begin{aligned} X_k^* &= \arg\max_{X_k \in X_{\neg\boldsymbol{\theta}^t}} \left(J_t - J_{t+1} \right) \\ &= \arg\max_{X_k \in X_{\neg\boldsymbol{\theta}^t}} \left(I(X_k; Y | X_{\boldsymbol{\theta}^t}) + \frac{1}{N} \log \frac{p(\boldsymbol{\theta}^{t+1})}{p(\boldsymbol{\theta}^t)} \right). \end{aligned}$$

□

Theorem 2. *The backward step that optimizes (10) at timestep $t+1$ from timestep t is to remove the feature,*

$$X_k^* = \arg\min_{X_k \in X_{\boldsymbol{\theta}^t}} \left(I(X_k; Y | X_{\boldsymbol{\theta}^t} \setminus X_k) + \frac{1}{N} \log \frac{p(\boldsymbol{\theta}^{t+1})}{p(\boldsymbol{\theta}^t)} \right). \quad (14)$$

Proof. Omitted due to space considerations, follows a similar procedure to the forward step. □

We note that with a flat uninformative prior $p(\boldsymbol{\theta}) \propto 1$, the prior term in the update cancels and we recover the maximum likelihood estimate of the optimal feature set, with the forward update becoming

$$X_k^* = \arg\max_{X_k \in X_{\neg\boldsymbol{\theta}^t}} I(X_k; Y | X_{\boldsymbol{\theta}^t}). \quad (15)$$

In Brown et al. (2012) we present a review of filter criteria based on mutual information based upon an earlier version of this expansion. There we maximise the *conditional* likelihood and show it leads to a framework incorporating many common criteria (Yang & Moody, 2000; Fleuret, 2004; Peng et al., 2005). These criteria are coupled with various termination conditions to create feature selection algorithms, so as to avoid the problems inherent in the maximum likelihood solution to the feature selection problem, i.e. selecting all the features. Given our probabilistic perspective we can see that a natural solution to this problem would be to impose a sparsity prior to regularize the feature selection criteria. We note that Eq (15) is very similar to the criteria in the IAMB algorithm, a topic we investigate in Section 4. We will now define sparsity and factored domain knowledge priors for our framework.

3 Constructing a prior

We have derived general update rules for feature selection incorporating priors, and now show a specific case of an independent Bernoulli prior over each feature. We show how this prior can be used to impose sparsity or to include prior knowledge.

3.1 A factored prior

We will treat each feature independently, and assume each $p(\theta_i)$ is a Bernoulli random variable. Therefore

$$p(\boldsymbol{\theta}) = \prod_i^d p(\theta_i) = \prod_i^d \beta_i^{\theta_i}(1-\beta_i)^{1-\theta_i}. \quad (16)$$

We then define the success probability, β_i of the Bernoulli as a logistic function,

$$\beta_i = \frac{e^{\alpha w_i}}{1+e^{\alpha w_i}} = \frac{1}{1+e^{-\alpha w_i}}. \quad (17)$$

We define $\alpha > 0$ as a scaling factor and w_i as a per-feature weight with $w_i = 0$ denoting no preference, $w_i < 0$ indicating we believe $X_i \notin X^*$, and $w_i > 0$ indicating we believe $X_i \in X^*$. We then define $\mathbf{w}$ as the vector of w_i elements. This is equivalent to specifying $p(\boldsymbol{\theta})$ as

$$p(\boldsymbol{\theta}) \propto e^{\alpha \mathbf{w}^T \boldsymbol{\theta}}. \quad (18)$$

We note that this formulation is of a similar exponential form to the priors given in Mukherjee & Speed (2008), and we could extend our framework to incorporate many of their graph structure priors.

3.2 Update rules

When using the factored prior above, we can further simplify the update rules in Equations (11) and (14), as there is a single bit difference between each step, the ratio of $p(\boldsymbol{\theta}^{t+1})$ to $p(\boldsymbol{\theta}^t)$ in the forward case is

$$\frac{p(\boldsymbol{\theta}^{t+1})}{p(\boldsymbol{\theta}^t)} = e^{\alpha w_k} \quad (19)$$

where w_k denotes the weight of the candidate feature. The ratio in the backwards step is the same but with a negated exponent. This gives the factored prior forward update as:

$$X_k^* = \underset{X_k \in X_{-\boldsymbol{\theta}^t}}{\arg\max} \left(I(X_k; Y | X_{\boldsymbol{\theta}^t}) + \frac{\alpha w_k}{N} \right) \quad (20)$$

with a similar update for the backward step.

3.3 Encoding sparsity or prior knowledge

Using the prior formulation in Eq (18) we can specify priors for sparsity or domain knowledge. We can encode sparsity by setting all $w_i = -1$, and using the α parameter to decide how much sparsity we wish to impose. Increasing α in this case lowers the success probability of the Bernoulli, and it is this probability that encodes how sparse a solution we impose. We will denote sparsity priors using the notation $p_s(\boldsymbol{\theta})$ and α_s leading to a sparsity prior where

$$p_s(\boldsymbol{\theta}) \propto e^{-\alpha_s |\boldsymbol{\theta}|}. \quad (21)$$

We use $|\boldsymbol{\theta}|$ to represent the number of selected features in $\boldsymbol{\theta}$. If we allow the w_i values to range freely we can encode varying levels of information into the prior, as these again change the success probability of the Bernoulli, thus encoding how useful *a priori* we think a given feature is. We will denote such knowledge priors with $p_d(\boldsymbol{\theta})$ and α_d leading to an knowledge prior where

$$p_d(\boldsymbol{\theta}) \propto e^{\alpha_d \mathbf{w}^T \boldsymbol{\theta}}. \quad (22)$$

We have now described two kinds of priors which we can integrate into any criterion derived from our discriminative model assumption. We now demonstrate the usefulness of this theoretical understanding by incorporating priors into the the IAMB (Tsamardinos & Aliferis, 2003) algorithm, as it uses a direct implementation of Eq (15) as the selection criteria. We note this process could be applied to many of the mutual information based filters found in Brown et al. (2012).

4 Incorporating a prior into IAMB

IAMB (Tsamardinos & Aliferis, 2003), shown in Algorithm 1, is a Markov Blanket (MB) discovery algorithm that uses a conditional independence test to decide variable inclusion. The test, $f(X; Y|\text{CMB})$, measures the association of a candidate feature X to the target Y, in the context of the currently estimated Markov Blanket. Tsamardinos & Aliferis recommend that instead of a test against zero, a threshold value is used—when the measured association is above this, the variables are considered dependent. IAMB has two phases, a greedy forward search of the feature space until all remaining features are independent of the class given CMB, and a backward search to remove false positives. Equating the notation in Algorithm 1 with our own, we have $\Omega = X$, CMB $= X_{\boldsymbol{\theta}}$, and the independence test $f(X; Y|\text{CMB}) = I(X_k; Y|X_{\boldsymbol{\theta}})$.

Given our probabilistic perspective we can interpret the threshold t in the IAMB algorithm as a sparsity prior, p_s, by rearranging the independence test in Algorithm 1,

$$I(X_k; Y|X_{\boldsymbol{\theta}}) + \frac{1}{N}\log\frac{p_s(\boldsymbol{\theta}^{t+1})}{p_s(\boldsymbol{\theta}^t)} \implies -t = \frac{1}{N}\log\frac{p_s(\boldsymbol{\theta}^{t+1})}{p_s(\boldsymbol{\theta}^t)}. \quad (23)$$

We can then see that the threshold t is a special case of the sparsity prior in Eq (21) with $\alpha_s = tN$, where the strength of the prior is dependent on the number of samples N, and a parameter t.

Algorithm 1 IAMB Tsamardinos & Aliferis (2003).

Phase 1 (forward)
CMB = $\emptyset$
while CMB has changed **do**
  Find $X \in \Omega \setminus \text{CMB}$ to maximise $f(X; Y|\text{CMB})$
  if $f(X; Y|\text{CMB}) > t$ **then**
    Add X to CMB
  end if
end while
Phase 2 (backward)
while CMB has changed **do**
  Find $X \in \text{CMB}$ to minimise $f(X; Y|\text{CMB} \setminus X)$
  if $f(X; Y|\text{CMB} \setminus X) < t$ **then**
    Remove X from CMB
  end if
end while

Theorem 3. *Tsamardinos and Aliferis proved that IAMB returns the true Markov Blanket under the condition of a perfect independence test $f(X; Y|CMB)$. Given this condition is satisfied, then IAMB is an iterative maximization of the discriminative model in Eq (1), under a specific sparsity prior.*

Proof. A perfect independence test comes as a result of sufficient data to estimate all necessary distributions. In this situation, the first KL term in Eq (8) will be zero. In the previous section we derived iterative update steps for our model, in Equations (11) and (14) — if we use a sparsity prior of the form in Eq (21), these coincide exactly with the steps employed by IAMB, therefore it is an iterative maximization of the discriminative model specified in Eq (1). □

We can now extend IAMB by introducing informative priors into the Markov Blanket discovery process. First we define $p(\boldsymbol{\theta}) \propto p_s(\boldsymbol{\theta})p_d(\boldsymbol{\theta})$ where $p_s(\boldsymbol{\theta})$ is the sparsity prior (or threshold), and $p_d(\boldsymbol{\theta})$ is our knowledge prior specified in Eq (22). We can ignore the normalisation constant as we only consider the ratio of the prior terms. We then use

$$I(X_k; Y|X_{\boldsymbol{\theta}}) + \frac{1}{N}\log\frac{p_s(\boldsymbol{\theta}^{t+1})}{p_s(\boldsymbol{\theta}^t)} + \frac{1}{N}\log\frac{p_d(\boldsymbol{\theta}^{t+1})}{p_d(\boldsymbol{\theta}^t)} > 0 \tag{24}$$

as the independence test having expanded out the prior $p(\boldsymbol{\theta})$. Incorporating $p_d(\boldsymbol{\theta})$ into IAMB lowers the "threshold" for features we believe are in the Markov Blanket and increases it for those we believe are not. We call this modified version *IAMB-IP* (IAMB-Informative Prior).

In some cases the knowledge prior, p_d, may be larger than the sparsity prior, p_s, causing the algorithm to unconditionally include feature X_k without reference to the data. In general we wish to blend prior domain knowledge with statistical evidence from the data, so this is undesirable and we recommend a bound on the strength of the knowledge prior by fixing $\alpha_d \leq \alpha_s$. This bounds the knowledge prior from above and below to ensure the prior is not strong enough to blindly include a feature without *some* evidence from the data.

5 Empirical Evaluation

We compare our novel IAMB-IP against the original IAMB algorithm using a selection of problems on MB discovery in artificial Bayesian Networks; these provide a ground truth feature set to compare the selected feature sets against. The networks used are standard benchmarks for MB discovery: Alarm (37 nodes, average MB size is 4) (Beinlich et al., 1989), Barley (48 nodes, average MB size is 5.25) (Kristensen & Rasmussen, 2002), Hailfinder (56 nodes, average MB size is 4.3) (Abramson et al., 1996) and Insurance (27 nodes, average MB size is 5.52) (Binder et al., 1997), downloaded from (Elidan, 1998).

As our datasets are Bayesian Networks from fields with which we have no experience, we simulate the process of prior elicitation by selecting certain features at random. Features can be either *upweighted*, i.e. we believe them to be in the MB, or *downweighted*, i.e. we believe they are not in the MB. Upweighting feature X_i corresponds to $w_i = 1$, while downweighting sets $w_i = -1$. With this process, we emulate two types of *correct* prior knowledge: A *true positive* (TP) — a feature $X_j \in$ MB that we *upweight*. A *true negative* (TN) — a feature $X_j \notin$ MB that we *downweight*. Real prior knowledge is unlikely to be completely correct, hence we must also test the resilience of IAMB-IP when presented with false information. A *false positive* (FP) — a feature $X_j \notin$ MB that we *upweight*. A *false negative* (FN) — a feature $X_j \in$ MB that we *downweight*. We will use the term *correct priors* to denote priors which only contain True Positives and True Negatives (e.g. 2 TP, TPTN). We will use the term *misspecified priors* to denote priors which contain a mixture of true and false information (e.g. TPFN, TPFP). We expect that these misspecified priors more accurately reflect the state of domain knowledge. In all experiments we only consider nodes with a Markov Blanket containing two or more features and we assess performance using the F-Measure (harmonic mean of precision & recall), comparing against the ground truth.

We use the protocol in Algorithm 2 to test the relative performance for two groups of sample sizes: 10 to 100 samples in steps of 10 (small sample), and 200 to 1000 samples in steps of 100 (large sample). For the large sample we perform 10 trials over independent data

Algorithm 2 Experimental Protocol

```
for each valid feature do
  for dataRepeats times do
    data ← selectFold()
    MB-I = IAMB(data,feature)
    Calculate MB-I F-measure
    for 40 repeats do
      Generate random prior
      MB-IP = IAMB-IP(data,feature,prior)
      Calculate MB-IP F-measure
    end for
    Calculate mean and std. err. for IAMB-IP
    Determine win/draw/loss
  end for
end for
Average wins/draws/losses over the features
```

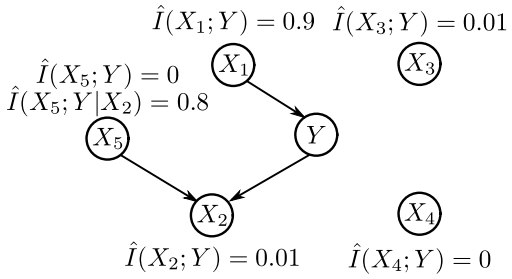

Figure 1: Toy problem, 5 feature nodes ($X_1 \dots X_5$) and their estimated mutual information with the target node Y on a particular data sample. X_1, X_2, X_5 form the Markov Blanket of Y.

samples, and for the smaller sizes we expect a greater variance and thus use 30 trials. The wins/draws/losses were assessed using a 95% confidence interval over the IAMB-IP results, compared to the IAMB result. The variance in IAMB-IP is due to the random selection of features which are included in the prior, which was repeated 40 times. We set $\alpha_d = \log 99$, except when this was above the bound $\alpha_d \leq \alpha_s$ where we set $\alpha_d = \alpha_s$. This is equivalent to setting individual priors $p(\theta_i = 1) = 0.99$ for upweighted features and $p(\theta_i = 1) = 0.01$ for downweighted features. We set α_s so $t = 0.02$ for both IAMB and IAMB-IP. We average these wins/draws/losses over *all valid features in a dataset*, where a valid feature is one with a Markov Blanket containing two or more features.

In Figure 1 we show a toy problem to illustrate the different effects prior knowledge can have on the Markov Blanket discovery process. Features X_1, X_2, X_5 are in the Markov Blanket of Y and features X_3 and X_4 are not. IAMB (with the default threshold, $t = 0.02$) would select only X_1 as the MB, based upon the estimated mutual informations given. The performance of

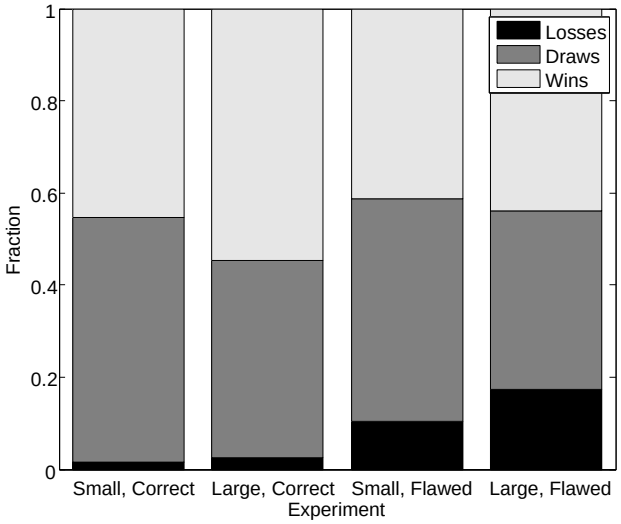

Figure 2: Average results: varying sample size (Small, Large) and prior (Correct, Flawed/Misspecified).

IAMB-IP will depend upon what knowledge is put into the prior. If we upweight X_1 it is a true positive, as it actually lies in the MB, similarly if we downweight X_3 it is a true negative. If we upweight X_4 it is a false positive, as it does not lie in the MB of Y, and similarly downweighting X_2 is a false negative as it does lie in the MB of Y. If we upweighted only X_1 IAMB-IP would perform similarly to IAMB, as X_1 has a strong measured association with Y, however upweighting X_2 would include that variable and then X_5, as X_2 only has a weak measured association with Y and so the prior will increase it. If X_4 is upweighted, (introducing a false positive into the prior) then it is unlikely to be included, as it has no measured association with Y, however X_3 would be included if it was upweighted. If we downweight X_2, (introducing a false negative) we can see this would remove both X_2 and X_5, as X_5 only becomes relevant when X_2 is included. We can see that false negatives in the prior are more problematic for IAMB-IP, as they can cause multiple variables to be incorrectly removed from the candidate MB.

We first investigate the performance of IAMB-IP when using a correct prior. We tested priors that included 2 true positives, and 1 true positive and 1 true negative. The average results over the 4 datasets are in the first two columns of Figure 2. When we incorporate correct priors IAMB-IP performs better than IAMB or equivalently to it in the vast majority of cases. The draws between IAMB and IAMB-IP are due to the overlap between the statistical information in the data and the information in the prior. When the prior upweights a feature with a strong signal from the data, then the behavior of IAMB-IP is the same as IAMB. It is when the prior upweights a feature with a weak signal that

the behavior of the two algorithms diverges, and similarly for features that are downweighted.

We now investigate the more interesting case of misspecified priors, where the prior contains some incorrect information. We tested priors using 1 true positive & 1 false negative, and 1 true positive & 1 false positive. These are presented in the last two columns of Figure 2. We can see that IAMB-IP performs equivalently or better than IAMB in four-fifths of the repeats, on average. We present full results for Hailfinder in Table 1 results for the other datasets are in the supplementary material. We can see that the algorithm is more sensitive to false negatives than false positives especially when there are small amounts of data, as the prior knowledge is more important in those situations, hence any flaws impact performance more. This is because false negatives may remove children from the MB, which in turn means no spouse nodes (the other parents of the common child node) will be included, which magnifies the effect of the false information.

In summary we can see that the addition of informative priors into IAMB to give IAMB-IP improves performance in many cases, even when half the prior knowledge given is incorrect. This improvement can be seen in extremely small sample environments with as few as 10 datapoints and 56 features, and still provides a performance benefit with 1000 datapoints.

We have focused on adding true positives to the prior, and how they interact with the false information. In our datasets true positives are rarer than true negatives and thus more important, because the Markov Blankets are much smaller than the number of features. Therefore when we construct the prior at random, we are more likely to select true positives where the prior information is useful (i.e. there is not enough statistical information in the data to include the true positive) as there are fewer true positives to select from. When including true negatives the prior only improves performance if the true negative appears to be statistically dependent on the target (and then penalised by the prior and not included), if it does not appear dependent, then the prior information has no effect on its inclusion. Therefore when only including true negatives IAMB-IP performs similarly to IAMB.

Table 1: Win/Draw/Loss results on Hailfinder.

Size	**2 TP**	**TPTN**	**TPFN**	**TPFP**
10	16/14/0	15/15/0	15/12/3	15/14/1
20	13/17/0	13/17/0	12/15/3	12/17/1
30	12/18/0	12/18/0	11/15/4	11/18/1
40	10/19/0	11/19/0	10/16/3	10/19/1
50	14/16/0	14/16/0	12/14/4	12/16/1
60	12/18/0	12/18/0	11/14/5	11/17/1
70	10/20/0	10/20/0	9/16/5	9/20/1
80	11/19/0	10/20/0	10/16/4	10/19/1
90	12/17/0	12/18/0	11/15/4	12/17/1
100	15/15/0	15/14/1	13/12/5	14/14/1
Mean	**12/17/0**	**12/18/0**	**11/14/4**	**12/17/1**

Size	**2 TP**	**TPTN**	**TPFN**	**TPFP**
200	6/4/0	6/4/0	5/4/2	5/4/1
300	6/4/0	6/4/0	5/4/2	6/4/1
400	6/4/0	6/4/0	5/4/2	5/4/1
500	6/4/0	5/4/0	4/4/2	5/4/1
600	6/4/0	5/5/0	4/4/2	4/4/1
700	5/5/0	5/5/0	4/4/2	4/5/1
800	4/6/0	4/6/0	4/4/2	4/5/1
900	4/6/0	3/6/0	3/5/2	4/6/0
1000	4/6/0	3/6/0	3/6/2	3/6/0
Mean	**5/5/0**	**5/5/0**	**4/4/2**	**4/5/1**

6 Conclusion & Future Work

We have developed a novel interpretation of information theoretic feature selection as an optimization of a discriminative model. This approach gives a theoretical grounding to the use of information theoretic criteria, revealing the underlying function which they optimise, namely the joint likelihood of a discriminative model under a flat prior. We show that in light of our formulation, the well-known IAMB algorithm can be seen as an iterative maximization of this discriminative model, using a particular sparsity prior. We then developed IAMB-IP, which incorporates informative priors into Markov Blanket discovery. We empirically tested IAMB-IP against IAMB, and found it to improve performance even when a misspecified prior was used, in which half the supplied "knowledge" was incorrect. This improvement was greatest when using more complex datasets, and small amounts of data.

Our future efforts will be directed towards integrating our framework with other feature selection filters, deriving new algorithms that can incorporate informative priors. This could be achieved either by investigating using the network structure priors presented in Mukherjee & Speed (2008) or more complex MB discovery algorithms such as the one presented by Margaritis (2009) or the GLL framework of Aliferis et al. (2010) to incorporate informative priors. These algorithms are not as closely linked to likelihood optimisation as IAMB, as they do not strictly maximize Eq (11) at each step, but they can use the mutual information as a conditional independence test and thus approximate an iterative maximisation of the model likelihood. Nevertheless we believe the construction of practical algorithms is facilitated by a sound theoretical framework, such as that presented in this paper.

References

Abramson, B., Brown, J., Edwards, W., Murphy, A., and Winkler, R.L. Hailfinder: A Bayesian system for forecasting severe weather. *Int. Journal of Forecasting*, 12(1):57–71, 1996.

Aliferis, C.F., Statnikov, A., Tsamardinos, I., Mani, S., and Koutsoukos, X.D. Local causal and markov blanket induction for causal discovery and feature selection for classification part i: Algorithms and empirical evaluation. *Journal of Machine Learning Research*, 11:171–234, 2010.

Beinlich, I.A., Suermondt, M., Chavez, R.M., and Cooper, G.F. A Case Study with two Probabilistic Inference Techniques for Belief Networks. In *2nd European Conference on Artificial Intelligence in Medicine*, pp. 247, 1989.

Binder, J., Koller, D., Russell, S., and Kanazawa, K. Adaptive probabilistic networks with hidden variables. *Machine Learning*, 29, 1997.

Brown, G. A New Perspective for Information Theoretic Feature Selection. In *12th International Conference on Artificial Intelligence and Statistics*, volume 5, pp. 49–56, 2009.

Brown, G., Pocock, A., Zhao, M.-J., and Luján, M. Conditional likelihood maximisation: A unifying framework for information theoretic feature selection. *Journal of Machine Learning Research*, 13: 27–66, 2012.

Elidan, G. Bayesian network repository. `www.cs.huji.ac.il/~galel/Repository/`, 1998.

Fano, R.M. Transmission of information. *Physics Today*, 14:56, 1961.

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

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

Helleputte, T. and Dupont, P. Partially supervised feature selection with regularized linear models. In *Int. Conf. on Machine Learning*, pp. 409–416, 2009.

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

Kristensen, K. and Rasmussen, I.A. The use of a Bayesian network in the design of a decision support system for growing malting barley without use of pesticides. *Computers and Electronics in Agriculture*, 33(3):197–217, 2002.

Krupka, E., Navot, A., and Tishby, N. Learning to select features using their properties. *Journal of Machine Learning Research*, 9:2349–2376, 2008.

Lasserre, J.A., Bishop, C.M., and Minka, T.P. Principled hybrids of generative and discriminative models. In *Computer Vision and Pattern Recognition*, pp. 87–94, 2006.

Margaritis, D. Toward provably correct feature selection in arbitrary domains. In *Neural Information Processing Systems*, volume 22, pp. 1240–1248, 2009.

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

Mukherjee, S. and Speed, T.P. Network inference using informative priors. *Proc. of the National Academy of Sciences*, 105(38):14313–14318, 2008.

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

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

Tsamardinos, I. and Aliferis, C.F. Towards principled feature selection: Relevancy, filters and wrappers. In *Artificial Intelligence and Statistics*, 2003.

Yang, H. and Moody, J. Data visualization and Feature selection. *Neural Information Processing Systems*, 12:687–702, 2000.