Informative Priors for Markov Blanket Discovery

Adam Pocock, Mikel Luján, Gavin Brown

The University of Manchester, UK

Introduction

- We present an interpretation of the feature selection problem as the maximisation of the joint likelihood of a discriminative model.
- From the joint likelihood we derive information theoretic criteria which maximise the likelihood with respect to the selected features.
- We show that the IAMB algorithm [3] for Markov Blanket discovery also optimises this model, using a sparsity prior.
- Finally we extend IAMB to include a domain knowledge prior, and show how this improves performance.

Model Specification

- We define our discriminative model [2] as follows:

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

- $\mathcal{D}$ is d-dimensional dataset with N samples, $\boldsymbol{\theta}$ is a d-dimensional binary vector denoting the selected features, τ represents other model parameters controlling classification, and λ represents the parameters controlling data generation.
- We work with the scaled negative log-likelihood, which converts our maximisation into a minimisation:

$$-\ell = -\frac{1}{N}\left(\sum_{i=1}^{N}\log q(y^i|\mathbf{x}^i,\boldsymbol{\theta},\tau) + \log p(\boldsymbol{\theta},\tau)\right) \qquad (2)$$

Expanding the likelihood

- We can expand the joint likelihood of our model into a sum of multiple terms:

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

- We can then interpret some of these terms as finite sample approximations to the information theoretic quantities of Entropy (H) and Mutual Information (I).

$$-\ell \approx \mathbb{E}_{\mathbf{x},y}\left\{\log\frac{p(y^i|\mathbf{x}^i,\boldsymbol{\theta})}{q(y^i|\mathbf{x}^i,\boldsymbol{\theta},\tau)}\right\} + I(X_{\neg\boldsymbol{\theta}};Y|X_{\boldsymbol{\theta}}) + H(Y|X) - \frac{1}{N}\log p(\boldsymbol{\theta},\tau). \qquad (4)$$

- Each term measures a different part of the performance:
  - The first term measures the performance of our classifier q compared to the true distribution p.
  - The second term measures the quality of our selected feature set $X_{\boldsymbol{\theta}}$, and is small when we have captured most of the available information.
  - The third term measures the quality of the data for the task of predicting Y, and is large when Y is not constrained by the data.
- We now make the same assumption inherent in all *filter* feature selection algorithms, that our feature selection parameters and model parameters are independent. We do this by specifying $p(\boldsymbol{\theta},\tau) = p(\boldsymbol{\theta})p(\tau)$. Now we turn to the problem of maximising this likelihood.

Iterative maximisation

- We derive greedy iterative updates which maximise our joint likelihood with respect to the selected features.
- The forward update (which includes a feature in the selected set) is:

$$X_k^* = \arg\max_{X_k\in X_{\neg\boldsymbol{\theta}^t}}\left(I(X_k;Y|X_{\boldsymbol{\theta}^t}) + \frac{1}{N}\log\frac{p(\boldsymbol{\theta}^{t+1})}{p(\boldsymbol{\theta}^t)}\right). \qquad (5)$$

- This selects the feature X_k^* in the timestep $t+1$, which is most informative given all the other selected features $\boldsymbol{\theta}^t$.
- We note that with an flat prior the final term vanishes, and we recover the maximum likelihood selection criterion. We investigate the links between this criterion and the information theoretic feature selection literature in [1].

Defining an informative prior

- We assume the probability of selecting each feature is independent, and Bernoulli distributed.
- We define the success probability of each Bernoulli as a logistic function $\beta_i = \frac{1}{1+e^{-\alpha w_i}}$, where α is a scaling factor, and w_i is the weight given to that feature.
  - Positive weights indicate we believe the feature to be in the optimal feature set.
  - Negative weights indicate we believe the feature is not in the optimal set.
  - Zero weights indicate no preference.

Defining an informative prior (cont.)

- This prior can be expressed as $p(\boldsymbol{\theta}) \propto e^{\alpha\mathbf{w}^T\boldsymbol{\theta}}$, where $\mathbf{w}$ is the d-dimensional vector of feature weights.
- We define two classes of informative priors:
  - Sparsity priors $p_s(\boldsymbol{\theta})$: set each $w_i = -1$, and the prior becomes $p_s(\boldsymbol{\theta}) \propto e^{-\alpha|\boldsymbol{\theta}|}$, which penalises large selected feature sets.
  - Knowledge priors $p_d(\boldsymbol{\theta})$: set each $w_i \in [-1,1]$ according to prior knowledge.
- When using these factored priors we can simplify the forward update rule to,

$$X_k^* = \arg\max_{X_k\in X_{\neg\boldsymbol{\theta}^t}}\left(I(X_k;Y|X_{\boldsymbol{\theta}^t}) + \frac{\alpha w_k}{N}\right). \qquad (6)$$

IAMB

- IAMB (Incremental Association Markov Blanket) is an algorithm for finding the Markov Blanket of a target node in a Bayesian Network.
- The Markov Blanket is the set of all parents, children and co-parents of a node.
- The algorithm proceeds by using a conditional independence test to add nodes which are dependent on the target.
- The test is usually the Conditional Mutual Information between a candidate node X_j, the target node Y, conditioned on the selected set $X_{\boldsymbol{\theta}}$,

$$X_k = \arg\max_{X_k\in X_{\neg\boldsymbol{\theta}}} I(X_k;Y|X_{\boldsymbol{\theta}}) > t \qquad (7)$$

- The test is against a threshold value, t, which we interpret as a sparsity prior.
- We can then see that the IAMB update is a special case of our general iterative maximisation rule.

IAMB-IP

- We can replace the test used in IAMB with a new test, incorporating both sparsity and domain knowledge priors.

$$I(X_k;Y|X_{\boldsymbol{\theta}}) + \frac{1}{N}\log\frac{p_s(\boldsymbol{\theta}^{t+1})}{p_s(\boldsymbol{\theta}^t)} + \frac{1}{N}\log\frac{p_d(\boldsymbol{\theta}^{t+1})}{p_d(\boldsymbol{\theta}^t)} > 0 \qquad (8)$$

- We call this algorithm IAMB-IP, as it integrates *informative priors* into IAMB.

Experiments

- We tested IAMB and IAMB-IP across 4 artificial Bayesian Networks, finding the MB for each node in turn.
- The results are split into two categories: where the prior knowledge was correct, and where half the knowledge was incorrect. We selected 2 nodes at random to include in the prior, repeated 40 times.
- We also tested using 20 different sample sizes, small samples using $\{10,\ldots,100\}$ datapoints, and large samples using $\{200,\ldots,1000\}$.

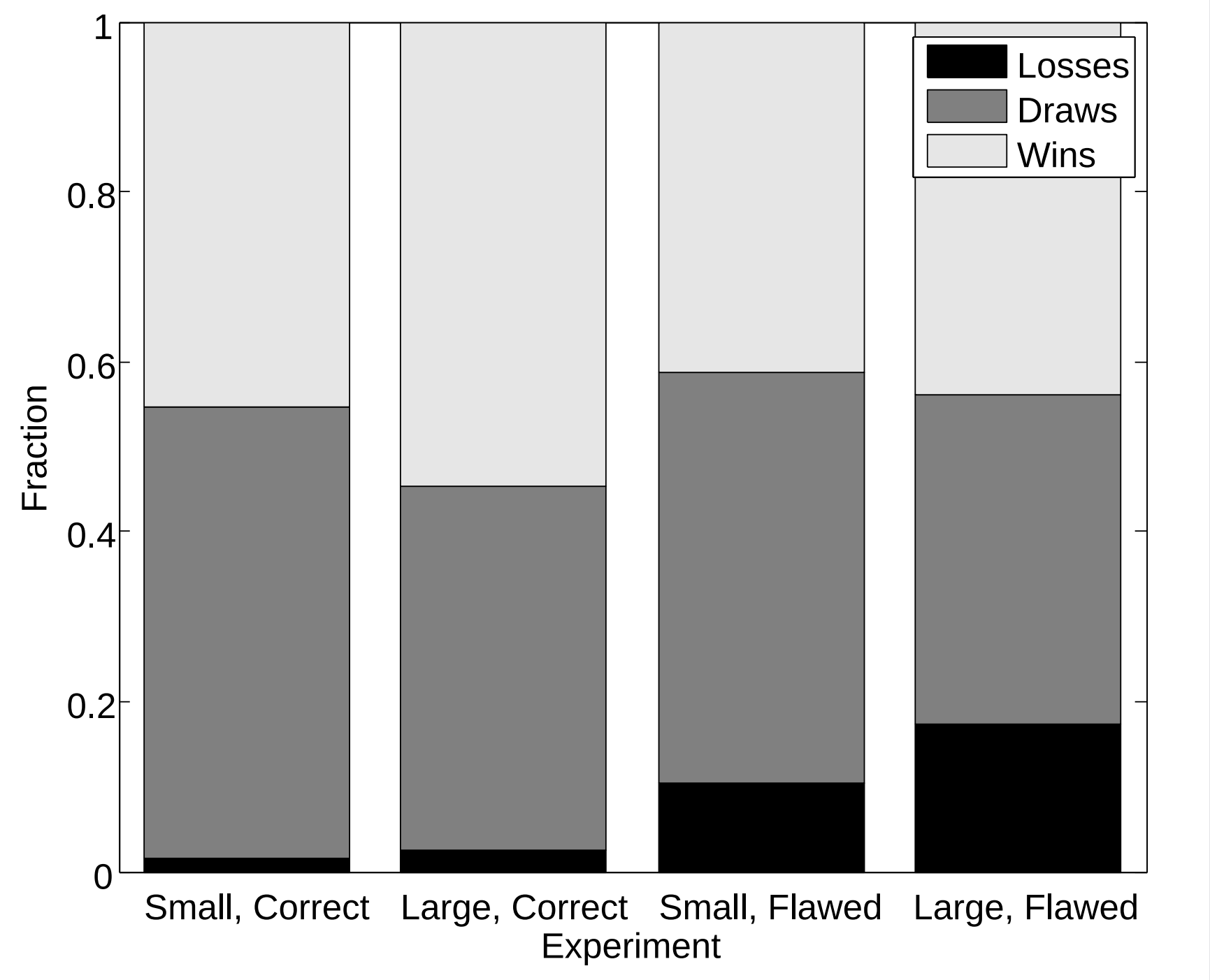

Figure: Average results: varying sample size (Small, Large) and prior (Correct, Flawed).

Conclusions

- We have shown that information theoretic feature selection optimises the joint likelihood of a discriminative model.
- This interpretation allows the incorporation of informative priors into information theoretic algorithms for feature selection.
- The derivation shows that the IAMB algorithm for Markov Blanket discovery also optimises the joint likelihood, using a sparsity prior.
- We thus extend IAMB into IAMB-IP which includes a domain knowledge prior in addition to the sparsity prior, improving Markov Blanket discovery on several benchmark Bayesian Networks.

References

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

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

[3] 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.

http://www.cs.man.ac.uk/~pococka4 email: apocock@cs.manchester.ac.uk