MANCHESTER
1824
The University of Manchester

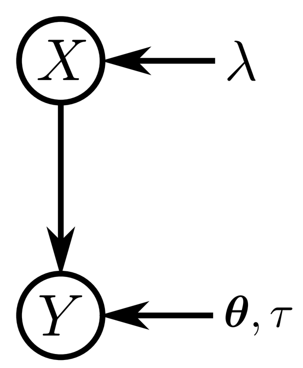
A Unifying Framework for Information Theoretic Feature Selection

Gavin Brown, Adam Pocock, Ming-Jie Zhao, Mikel Luján

Machine Learning and Optimization Group, University of Manchester, UK

The Short Story

- Feature selection using mutual information is very popular.
- Accepted research practice is to *hand-design* filter criteria to maximise **"relevancy"** and minimise **"redundancy"**.
- In contrast, here we *derive* a criterion, which naturally includes these concepts. This criterion provably maximises the joint likelihood of the discriminative model on the left.
- This enables us to retrofit numerous heuristics — we find that 20 years of heuristics can be understood within a single theoretical interpretation.



The Relevancy/Redundancy mystery...

Many successful criteria try to maximise relevancy / minimise redundancy:

- MIM - $J_{MIM}(X_i) = I(X_i; Y)$
- mRMR - $J_{mRMR}(X_i) = I(X_i; Y) - \frac{1}{|S|}\sum_{X_j \in S} I(X_i; X_j)$
- JMI - $J_{JMI}(X_i) = I(X_i; Y) - \frac{1}{|S|}\sum_{X_j \in S} I(X_i; X_j) + \frac{1}{|S|}\sum_{X_j \in S} I(X_i; X_j|Y)$

There are numerous suggested criteria 1994-2012... (incomplete list!)

Criterion	*Full name*	*Author*
MI	Mutual Information Maximisation	Various (1970s -)
MIFS	Mutual Information Feature Selection	Battiti (1994)
JMI	Joint Mutual Information	Yang & Moody (1999)
MIFS-U	MIFS-'Uniform'	Kwak & Choi (2002)
IF	Informative Fragments	Vidal-Naquet (2003)
FCBF	Fast Correlation Based Filter	Yu et al (2004)
CMIM	Conditional Mutual Info Maximisation	Fleuret (2004)
mRMR	min-Redundancy Max-Relevance	Peng et al (2005)
ICAP	Interaction Capping	Jakulin (2005)
CIFE	Conditional Infomax Feature Extraction	Lin & Tang (2006)
DISR	Double Input Symmetrical Relevance	Meyer (2006)
IGFS	Interaction Gain Feature Selection	El-Akadi (2008)
MIGS	Mutual Information Based Gene Selection	Cai et al (2009)
mIMR	min-Interaction Max-Relevance	Bontempi & Meyer (2010)
CMIFS	Conditional MIFS	Cheng (2011)

But... each is motivated from a different direction! Which can we trust?

Defining a Model

- We define our discriminative model [1] as follows:

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

- $\mathcal{D}$ 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 data generation parameters.
- We use scaled negative log-likelihood, and so we minimise:

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

Expanding the likelihood

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

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

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

$$-\ell \approx \underbrace{\mathbb{E}_{\mathbf{x}}\left(D_{KL}\{p_{\boldsymbol{\theta}}||q_{\boldsymbol{\theta}}\}\right)}_{\text{Classification error}} + \underbrace{I(X_{\neg\boldsymbol{\theta}}; Y|X_{\boldsymbol{\theta}})}_{\text{Feature Selection}} + \underbrace{H(Y|X)}_{\text{Data Quality}} - \frac{1}{N}\underbrace{\log p(\boldsymbol{\theta}, \tau)}_{\text{Prior}}. \quad (4)$$

- Minimising each of these terms maximises the likelihood.
- 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 $p(\boldsymbol{\theta}, \tau) = p(\boldsymbol{\theta})p(\tau)$.
- Then the iterative forward update which maximises the likelihood is (assuming an uninformative prior):

$$X_k^* = \arg\max_{X_k \in X_{\neg\boldsymbol{\theta}^t}} I(X_k; Y|X_{\boldsymbol{\theta}^t}) \quad (5)$$

- We considered the case of informative priors in [2].

Investigating the assumptions of the literature

Most of the criteria can be written in a common functional form, as the relevancy minus the redundancy plus the complementarity.

$$J(X_i) = I(X_i; Y) - \beta\sum_{X_j \in S} I(X_i; X_j) + \gamma\sum_{X_j \in S} I(X_i; X_j|Y) \quad (6)$$

But how does this relate to the optimal criterion derived above?

- Each combination of terms (or value of β and γ) makes an *assumption*.
- This factorises the likelihood, resulting in an *approximate* update rule.
  - MIM assumes complete independence, *i.e.* $\forall x_i, x_j\ p(x_i, x_j) = p(x_i)p(x_j)$.
  - mRMR and JMI assume the selected features are independent given the one under consideration, *i.e.* $p(x_{\boldsymbol{\theta}}|x_i) = \prod_{j \in S} p(x_j|x_i)$ and $p(x_{\boldsymbol{\theta}}|x_i, y) = \prod_{j \in S} p(x_j|x_i, y)$.
  - mRMR makes one further assumption, that all the features are pairwise class-conditionally independent (similar to the Naïve Bayes assumption), *i.e.* $\forall x_i, x_j\ p(x_i, x_j|y) = p(x_i|y)p(x_j|y)$.

These different assumptions form an important theoretical difference between criteria, changing what they expect from the data distribution.

- One further difference is the scaling of the redundancy/complementarity terms.
  - Popular criteria such as mRMR and JMI scale β and γ as $|S|$ increases.
  - This balances the size of the redundancy term so it does not dominate the relevancy term.

  Together these properties explain much of the empirical performance of the various criteria.
- Theoretically the JMI criterion makes the fewest assumptions, whilst balancing the terms and ensuring the informations involved are estimable.

Experiment: Similarity

- 50 bootstraps, measure intersection of selected features with a correction for chance. Using Kuncheva's similarity measure (Kuncheva 2007).
- We visualise the results using multi-dimensional scaling.
- Proximity of dots indicates similar selected feature sets, across many datasets.
- **Conclusion:** Methods which balance relevancy/redundancy are clustered – the outliers are different from this cluster and each other.

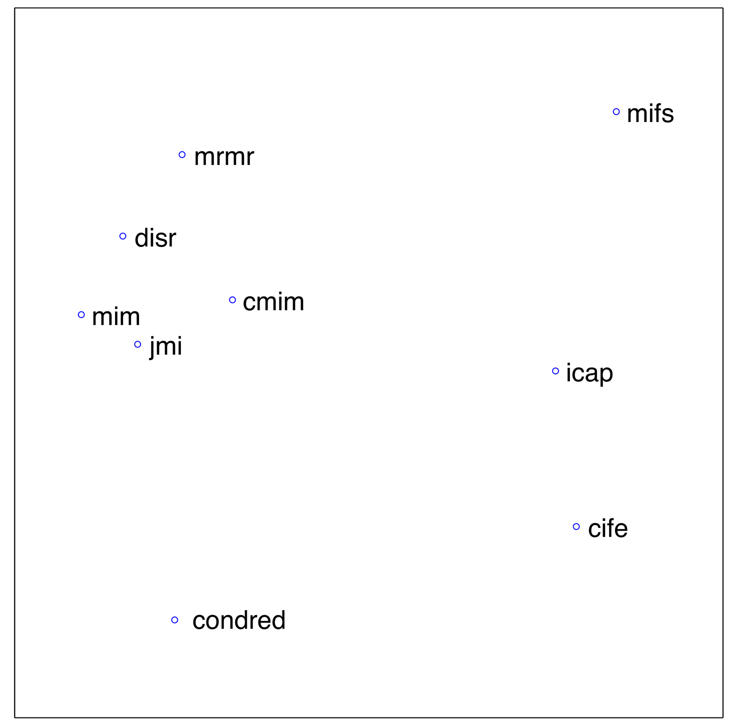
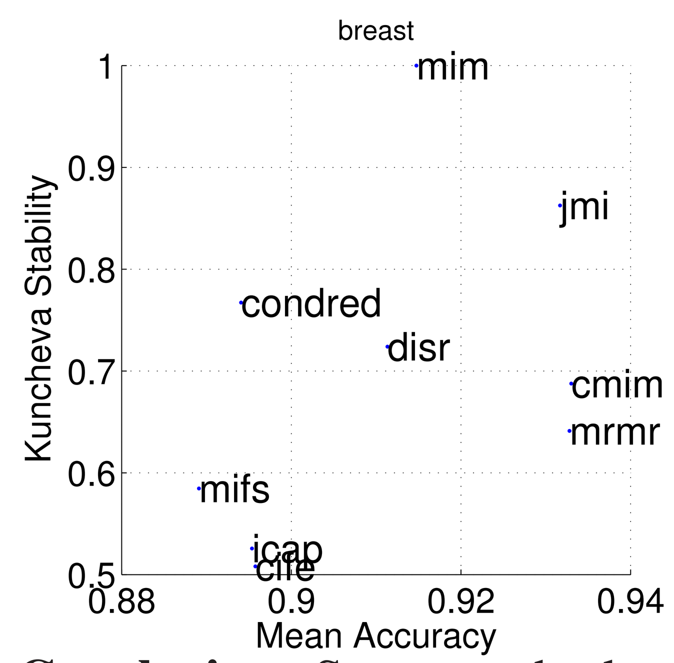


Figure: Similarity results across 9 criteria using Kuncheva's measure.

Experiment: Accuracy and Stability



Average pareto-optimal, non-dominated rank:

Accuracy/Stability	Accuracy
JMI (1.5)	JMI (2.6)
DISR (2.2)	MRMR (3.6)
MIM (2.3)	DISR (3.7)
MRMR (2.5)	CMIM (4.5)
CMIM (3.4)	ICAP (5.3)
ICAP (4.3)	MIM (5.4)
CIFE (4.8)	CIFE (5.9)
MIFS (4.9)	MIFS (6.5)

Conclusion: Some methods are *extremely* unstable with respect to small changes in training data. On average over 15 datasets, we find the JMI criterion (Yang & Moody, NIPS 1999) to have the most favourable properties.

Conclusions

- Unifying framework for over 20 years of heuristics – all are approximate maximisers of the conditional likelihood, with differing probabilistic independence assumptions.
- We have natural definitions of relevancy, redundancy, and complementarity.
- Clear probabilistic framework to devise new methods...

References

[1] 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.

[2] A. Pocock, M. Luján, and G. Brown.
Informative priors for markov blanket discovery.
In *Artificial Intelligence and Statistics (AISTATS 2012)*, volume 22, pages 905–913, 2012.

 email: gbrown@cs.man.ac.uk