

# Some Thoughts at the Interface of Ensemble Methods & Feature Selection



Gavin Brown  
School of Computer Science  
University of Manchester, UK

## Two Lands

*“England and America are two lands separated by a common language”*

–George Bernard Shaw / Oscar Wilde

Separated, but no common language?

- ▶ The Land of Multiple Classifier Systems
- ▶ The Land of Feature Selection

We're not talking about...

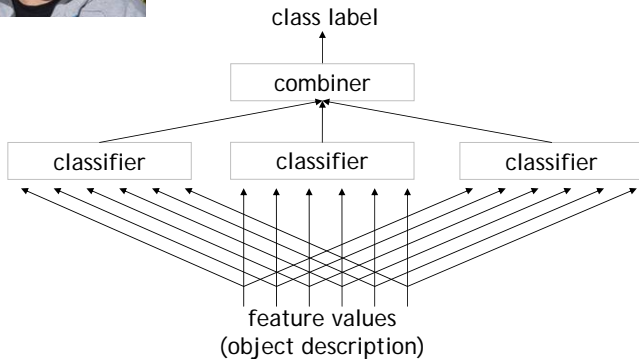
- ▶ Feature selection **for** ensemble members
- ▶ Combining feature sets (e.g. Somol et al, MCS 2009)

So what are we talking about...?

# The “Duality” of MCS and FS



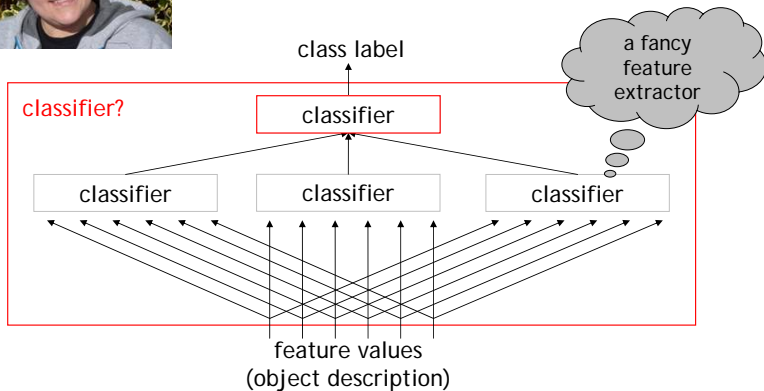
© L.I. Kuncheva, ICPR Plenary 2008



# The “Duality” of MCS and FS



© L.I. Kuncheva, ICPR Plenary 2008



# The “Duality” of MCS and FS

*“Dimensionality reduction and MCS should complement, not compete with each other” ... “aspects of feature selection/extraction procedures may suggest new ideas to MCS designers that should not be ignored.”*

**Sarunas Raudys (Invited talk, MCS 2002)**



*... “[MCS] can therefore be viewed as a multistage classification process, whereby the a posteriori class probabilities generated by the individual classifiers are considered as features for a second stage classification scheme. Most importantly [...] one can view classifier fusion in a unified way. ”*

**Josef Kittler (PAA vol 1(1), pg18–27, 1998)**



## A common language?

Mutual Information : zero *iff*  $X \perp\!\!\!\perp Y$

$$I(X; Y) = \sum_{x \in X} \sum_{y \in Y} p(xy) \log \frac{p(xy)}{p(x)p(y)}$$

Conditional Mutual Information : zero *iff*  $X \perp\!\!\!\perp Y | Z$

$$I(X; Y | Z) = \sum_{x \in X} \sum_{y \in Y} \sum_{z \in Z} p(xyz) \log \frac{p(xy|z)}{p(x|z)p(y|z)}$$

To understand why, we journey to the Land of Feature Selection...

# The Land of Feature Selection - “Wrappers”

## PROCEDURE : WRAPPER

**Input:** large feature set  $\Omega$

**Returns:** useful feature subset  $S \subseteq \Omega$

**10** Identify candidate subset  $S \subseteq \Omega$

**20** While !stop\_criterion()

    Evaluate **error of a classifier** using  $S$ .

    Adapt subset  $S$ .

**30** Return  $S$ .

**Pro:** high accuracy from your classifier

**Con:** computationally expensive!

## The Land of Feature Selection - “Filters”

### PROCEDURE : FILTER

**Input:** large feature set  $\Omega$

**Returns:** useful feature subset  $S \subseteq \Omega$

**10** Identify candidate subset  $S \subseteq \Omega$

**20** While !stop\_criterion()

    Evaluate utility function  $J$  using  $S$ .

    Adapt subset  $S$ .

**30** Return  $S$ .

**Pro:** generic feature set, and fast!

**Con:** possibly less accurate, task-specific design is open problem

## A Problem in FS-Land: Design of Filter Criteria

Feature space  $\Omega = \{X_1, \dots, X_M\}$ .

Consider features for inclusion/exclusion one-by-one.

Question: What is the utility of feature  $X_i$ ?

Mutual Information with target  $Y$ .

$$J_{mi}(X_i) = I(X_i; Y)$$

Higher MI means more discriminative power.

- ✓ Encourages relevant features.
- ✗ Ignores possible **redundancy**.
  - *may select two almost identical features*
  - ... *waste of resources!*
  - ... *possible overfitting!*

## A Problem in FS-Land: Design of Filter Criteria

Q. What is the utility of feature  $X_i$ ?

$$J_{mi}(X_i) = I(X_i; Y)$$

**“its own mutual information with the target”**

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

**“as above, but penalised by correlations with features already chosen”**

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

**“as above, but averaged, smoothing out noise”**

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

**“how well it pairs up with other features chosen”**

# The Confusing Literature of Feature Selection Land

<u>Criterion</u>	<u>Full name</u>	<u>Author</u>
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)
JMI-AVG	Averaged Joint Mutual Information	Scanlon et al (2004)
MRMR	Max-Relevance Min-Redundancy	Peng et al (2005)
ICAP	Interaction Capping	Jakulin (2005)
CIFE	Conditional Infomax Feature Extraction	Lin & Tang (2006)
DISR	Double Input Symmetrical Relevance	Meyer (2006)
MINRED	Minimum Redundancy	Duch (2006)
IGFS	Interaction Gain Feature Selection	El-Akadi (2008)
MIGS	Mutual Information Based Gene Selection	Cai et al (2009)

**Why should we trust any of these? How do they relate?**

# The Land of Feature Selection: A Summary

Problem: construct a useful set of features

- ▶ Need features to be relevant and not redundant.

Accepted research practice: invent heuristic measures

- ▶ Encouraging “relevant” features
- ▶ Discouraging correlated features

Sound familiar? For ‘feature’ above, read ‘classifier’...

## What would someone from the Land of MCS do?

MCS inhabitants believe in their (undefined) Diversity God.

But, MCS-Land is just one district in the **Land of Ensemble Methods**.

Other districts are:

- ▶ The Land of Regression Ensembles
- ▶ The Land of Cluster Ensembles
- ▶ The Land of Semi-Supervised Ensembles
- ▶ The Land of Non-Stationary Ensembles

And possibly others, as yet undiscovered...

# The Land of Regression Ensembles

Loss function :  $(\bar{f}(x) - y)^2$

Combiner function :  $\bar{f}(x) = \frac{1}{M} \sum_{i=1}^M f_i(x)$

## Method:

Take objective function, decompose into constituent parts.

$$(\bar{f} - y)^2 = \frac{1}{M} \sum_{i=1}^M (f_i - y)^2 - \frac{1}{M} \sum_{i=1}^M (f_i - \bar{f})^2$$

# An MCS native visits the Land of Feature Selection

Loss function :  $I(F; Y)$

'Combiner' function :  $F = X_{1:M}$  (joint random variable)

## Method:

Take objective function, decompose into constituent parts.

$$\begin{aligned} I(X_{1:M}; Y) = & \sum_{\forall i} I(X_i; Y) \\ & + \sum_{\forall i, j} I(X_i, X_j, Y) \\ & + \sum_{\forall i, j, k} I(X_i, X_j, X_k, Y) \\ & + \sum_{\forall i, j, k, l} I(X_i, X_j, X_k, X_l, Y) \\ & \dots \dots \dots \end{aligned}$$

Multiple "levels" of correlation!

Each term is a multi-variate mutual information! (McGill, 1954)

## Linking theory to heuristics....

Take only terms involving  $X_i$  we want to evaluate - exact expression:

$$I(X_i; Y|S) = I(X_i; Y) - \sum_{k \in S} I(X_i; X_k) + \sum_{k \in S} I(X_i; X_k|Y) + \sum_{j, k \in S} I(X_i, X_j, X_k, Y) + \dots +$$

$$J_{mi} = I(X_i; Y)$$

$$J_{mifs} = I(X_i; Y) - \sum_{k \in S} I(X_i; X_k)$$

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

---

and others can be re-written to this form...

$$\begin{aligned} J_{jmi} &= \sum_{k \in S} I(X_i X_k; Y) \\ &= I(X_i; Y) - \frac{1}{|S|} \sum_{k \in S} I(X_i; X_k) + \frac{1}{|S|} \sum_{k \in S} I(X_i; X_k|Y) \end{aligned}$$

## A “Template” Criterion

$$J_{mifs} = I(X_i; Y) - \sum_{k \in S} I(X_i; X_k)$$

$$J_{mrrmr} = I(X_i; Y) - \frac{1}{|S|} \sum_{k \in S} I(X_i; X_k)$$

$$J_{jmi} = I(X_i; Y) - \frac{1}{|S|} \sum_{k \in S} I(X_i; X_k) + \frac{1}{|S|} \sum_{k \in S} I(X_i; X_k | Y)$$

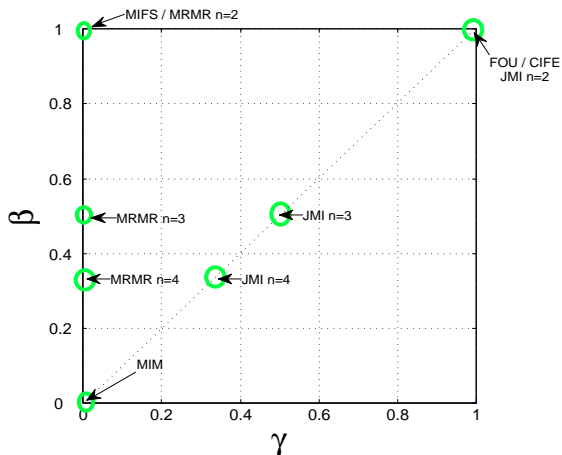
$$J_{cife} = I(X_i; Y) - \sum_{k \in S} I(X_i; X_k) + \sum_{k \in S} I(X_i; X_k | Y)$$

$$J_{cmim} = I(X_i; Y) - \max_k \left\{ I(X_i; X_k) - I(X_i; X_k | Y) \right\}$$

---

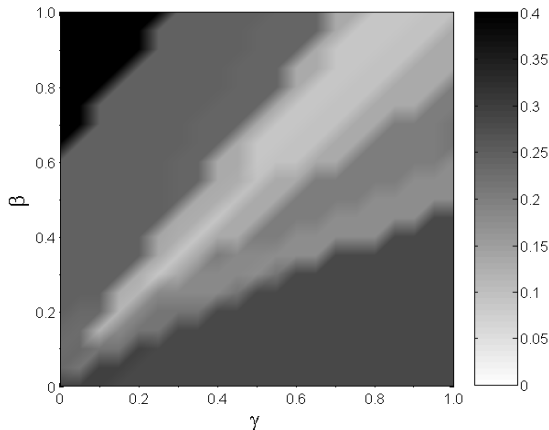
$$J = I(X_n; Y) - \beta \sum_{\forall k \in S} I(X_n; X_k) + \gamma \sum_{\forall k \in S} I(X_n; X_k | Y)$$

# The $\beta/\gamma$ Space of Possible Criteria



$$I(X_{1:M}; Y) \approx \underbrace{I(X_n; Y)}_{\text{relevancy}} - \underbrace{\beta \sum_k I(X_n; X_k)}_{\text{redundancy}} + \underbrace{\gamma \sum_k I(X_n; X_k | Y)}_{\text{conditional redundancy}} .$$

# The $\beta/\gamma$ Space of Possible Criteria



$$I(X_{1:M}; Y) \approx \underbrace{I(X_n; Y)}_{\text{relevancy}} - \underbrace{\beta \sum_k I(X_n; X_k)}_{\text{redundancy}} + \underbrace{\gamma \sum_k I(X_n; X_k | Y)}_{\text{conditional redundancy}} .$$

# Exploring $\beta/\gamma$ space

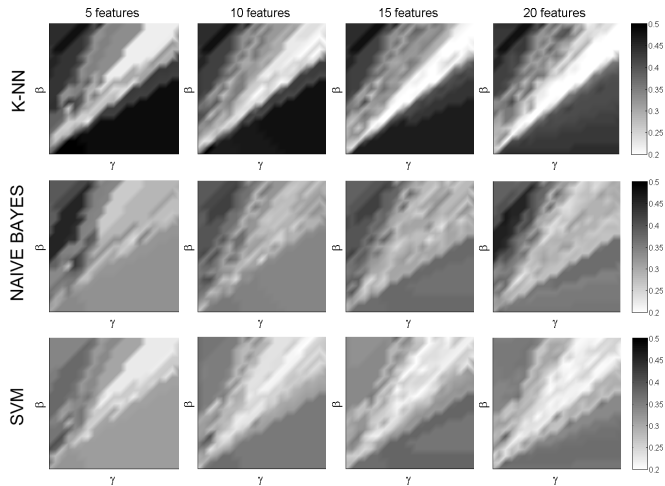


Figure 3: ARCENE data (cancer diagnosis).

# Exploring $\beta/\gamma$ space

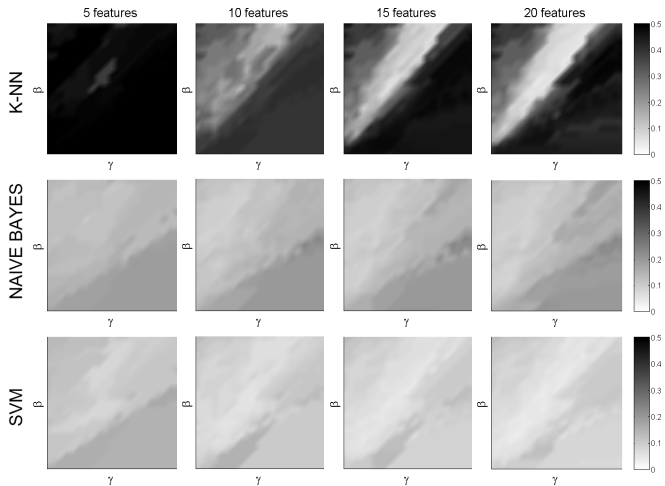
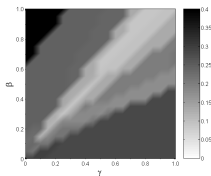


Figure 4: GISETTE data (handwritten digit recognition).

## Exploring $\beta/\gamma$ space

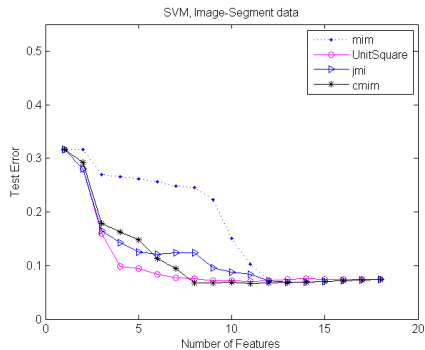
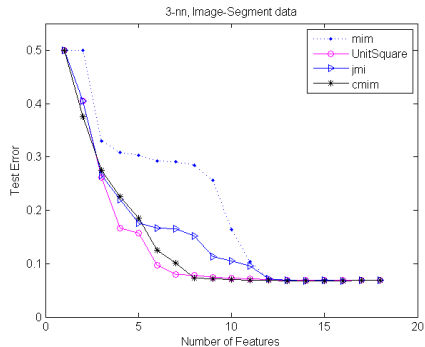


Seems straightforward? Just use the diagonal? Top right corner?  
But... remember these are only low order components.

Easy to construct problems that have ZERO in low orders, and positive terms in high orders.

- ▶ e.g. data drawn from a Bayesian net with some nodes exhibiting deterministic behavior. (e.g. parity problem).

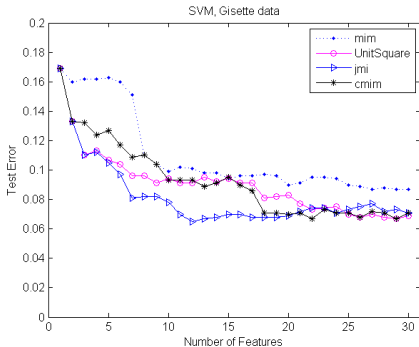
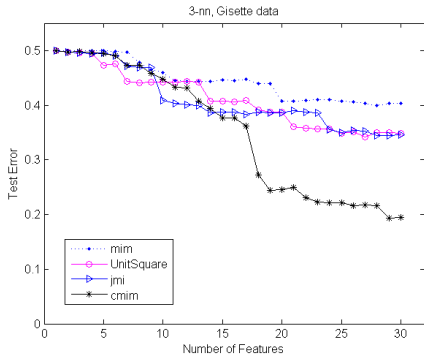
# Image Segment data



3-nn classifier (left), and SVM (right).

Pink line ('UnitSquare') is top right corner of  $\beta/\gamma$  space.

# GISETTE data



Low order components insufficient ...  
....heuristics can triumph over theory!

## Exports & Imports

Exported a perspective from the Land of MCS...

...solved an open problem in the Land of FS.

But could the MCS natives also learn from this?

# Exports & Imports : Understanding Ensemble Diversity

**(Step 1)** Take an objective function...

- log-likelihood: ensemble combiner  $g$ , with  $M$  members...

$$\mathcal{L} = \mathcal{E}_{\mathbf{x}y} \left\{ \log g(y|\phi_{1:M}) \right\}$$

**(Step 2)** ...decompose into constituent parts.

$$\mathcal{L} = \text{const} + \underbrace{I(\phi_{1:M}; Y)}_{\text{ensemble members}} - \underbrace{KL( p(y|\mathbf{x}) || g(y|\phi_{1:M}) )}_{\text{combiner}}$$

---

*"Information Theoretic Views of Ensemble Learning".*

G.Brown, Manchester MLO Tech Report, Feb 2010

# Exports & Imports : Understanding Ensemble Diversity

$$I(X_{1:M}; Y) \approx \underbrace{\sum_{i=1}^M I(X_i; Y)}_{\text{"relevancy"}} - \underbrace{\sum_{j=1}^M \sum_{k=j+1}^M I(X_i; X_j)}_{\text{"diversity"}} + \sum_{j=1}^M \sum_{k=j+1}^M I(X_i; X_j | Y)$$

$$I(X_{1:M}; Y) = \text{Individual Mutual Info} + \text{2-way diversity (pairwise)}$$

~~+ 3-way diversity~~

~~+ ... way diversity~~

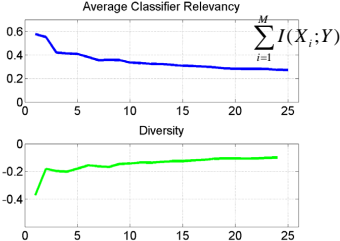
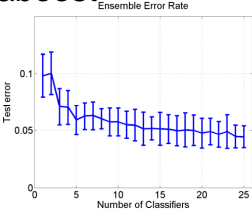
~~+ M-way diversity~~

---

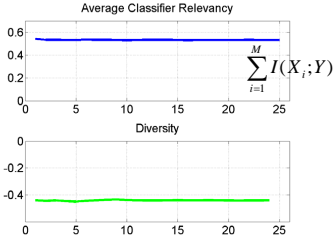
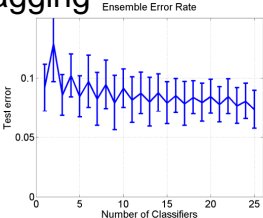
*"An Information Theoretic Perspective on Multiple Classifier Systems", MCS 2009.*

# Exports & Imports : Understanding Ensemble Diversity

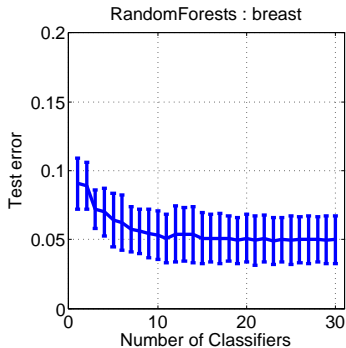
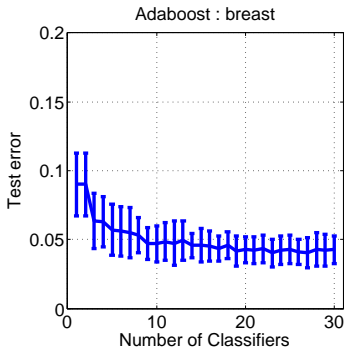
## Adaboost



## Bagging



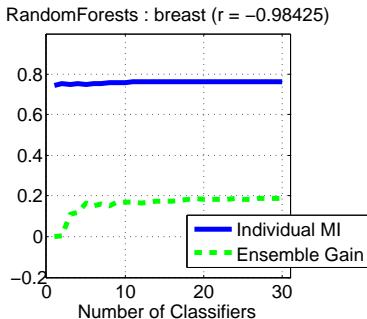
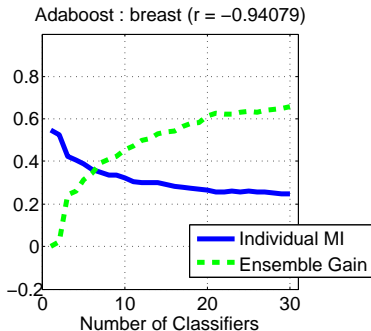
# Exports & Imports : Understanding Ensemble Diversity



---

(ongoing work with Zhi-Hua Zhou)

# Exports & Imports : Understanding Ensemble Diversity



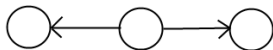
---

(ongoing work with Zhi-Hua Zhou)

# Exports & Imports : Model Selection



COLLIDER STRUCTURE  
(positive McGill information)



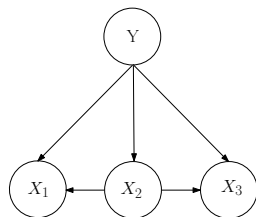
FORK STRUCTURE  
(negative McGill information)



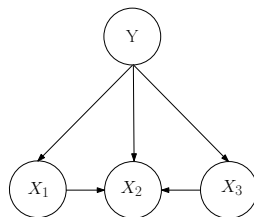
CHAIN STRUCTURE  
(negative McGill information)

Multi-variate mutual information can be positive or negative!

## Exports & Imports : Model Selection



*Fork ANB*



*Collider ANB*

# Exports & Imports : Model Selection

---

**Algorithm 1** Ensemble Method **rsADE**

---

1 Split D into TR, TS

**for**  $i = 1 : T$  **do**

    Randomly pick 3 features  $S_i = \{X_1, X_2, X_3\}$

    Build all possible collider models from TR( $S_i$ )

    Build all possible fork models from TR( $S_i$ )

    Choose the most accurate model class  $\mathcal{C}$  on TR( $S_i$ )

    Build ADE from this model class

**end for**

---

---

**Algorithm 2** Ensemble Method **irsADE**

---

1 Split D into TR, TS

**for**  $i = 1 : T$  **do**

    Randomly pick 3 features  $S = \{X_1, X_2, X_3\}$

**if**  $I(S) > 0$  **then**

        Build all possible collider models from TR( $S_i$ )

**else**

        Build all possible fork models from TR( $S_i$ )

**end if**

    Build ADE

**end for**

---

# Exports & Imports : Model Selection

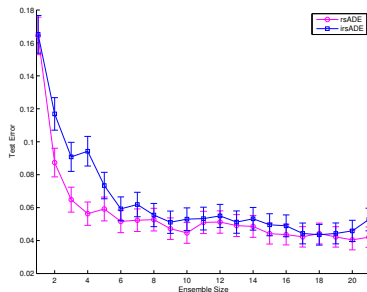


Figure 3.5: Mushroom dataset: rsmADE vs irsmADE - Test Error mean and 95% confidence interval

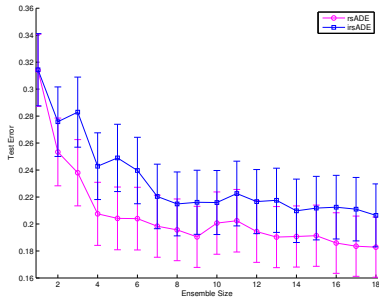


Figure 3.6: idaimage dataset: rsmADE vs irsmADE

Mushroom (left) and Image Segment (right).  
Performance almost as good... but **much** faster to train...

Usually 5 – 10 $\times$  faster, sometimes up to 90 $\times$ .  
Speedup proportional to **arity** of features.

## Current work: other multivariate mutual informations...

The *multi-variate information* used here is not the only one...

- ▶ “Interaction Information” (McGill, 1954) - this work
- ▶ “Multi-Information” (Watanabe, 1960) - **Zhou & Li: Thu 9.45am**
- ▶ “Difference Entropy” (Han, 1980) - similar to McGill
- ▶ “ $\mathcal{I}$ -measure” (Yeung, 1991) - pure set theoretic framework

---

G. Brown,

“Some Properties of Multi-variate Mutual Information”, (in preparation)

# Conclusions

It's getting really hard to contribute meaningful research to MCS.  
*... and to ML/PR in general!*

- ▶ I'm starting to look at importing ideas from other fields
- ▶ Information Theory seems natural
- ▶ Knowledge can flow both ways