

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

Gavin Brown, University of Manchester, UK

Cairo Microsoft Innovation Centre  
...adapted (slightly) from MCS 2010



# What type of document is this?

- ▶ potentially thousands of “features”
- ▶ highly spatially correlated, often noisy

## 1.1. Statistical framework

Throughout the paper,  $\xi_1, \dots, \xi_n \in \Xi$  denote some random variables with common distribution  $P$  (the observations). Except in Section 8.1, the  $\xi_n$  are assumed to be independent. The purpose of statistical inference is to estimate from the data  $\{\xi_i\}_{1 \leq i \leq n}$  some target feature  $s$  of the unknown distribution  $P$ , such as the density of  $P$  w.r.t. some measure  $\mu$ , or the regression function. Let  $\mathbb{S}$  denote the set of possible values for  $s$ . The quality of  $t \in \mathbb{S}$ , as an approximation to  $s$ , is measured by its loss  $\mathcal{L}(t)$ , where  $\mathcal{L}: \mathbb{S} \rightarrow \mathbb{R}$  is called the loss function; the loss is assumed to be minimal for  $t = s$ . Several loss functions can be chosen for a given statistical problem. Many of them are defined by

$$\mathcal{L}(t) = \mathcal{L}_P(t) := \mathbb{E}_{\xi \sim P}[\gamma(t; \xi)] \quad (1)$$

where  $\gamma: \mathbb{S} \times \Xi \rightarrow [0, \infty)$  is called a contrast function. For  $t \in \mathbb{S}$ ,  $\mathbb{E}_{\xi \sim P}[\gamma(t; \xi)]$  measures the average discrepancy between  $t$  and a new observation  $\xi$  with distribution  $P$ . Several frameworks such as transductive learning do not fit definition (1); nevertheless, as detailed in Section 1.2, definition (1) includes most classical statistical frameworks. Given a loss function  $\mathcal{L}_P(\cdot)$ , two useful quantities are the excess loss

$$\mathcal{E}(s, t) := \mathcal{L}_P(t) - \mathcal{L}_P(s) \geq 0$$

and the risk of an estimator  $\bar{s}(\xi_1, \dots, \xi_n)$  of the target  $s$

$$\mathbb{E}_{\xi_1, \dots, \xi_n \sim P}[\mathcal{E}(s, \bar{s}(\xi_1, \dots, \xi_n))] \quad .$$

## 1.2. Statistical problems

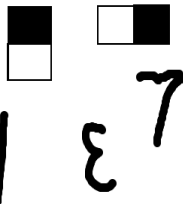
The following examples illustrate how general the framework of Section 1.1 is.

**Density estimation** aims at estimating the density  $s$  of  $P$  with respect to some given measure  $\mu$  on  $\Xi$ . Then,  $\mathbb{S}$  is the set of densities on  $\Xi$  with respect to  $\mu$ . For instance, taking  $\gamma(t; x) = -\ln(\mu(x))$  in (1), the loss is minimal when  $t = s$  and the excess loss

$$\mathcal{E}(s, t) = \mathbb{E}_{\xi \sim P} \left[ \ln \left( \frac{s(\xi)}{t(\xi)} \right) \right] = \int s \ln \left( \frac{s}{t} \right) d\mu$$

is the Kullback-Leibler divergence between distributions  $\mu_s$  and  $\mu_t$ .

**Prediction** aims at predicting a quantity of interest  $Y \in \mathcal{Y}$  given an explanatory variable  $X \in \mathcal{X}$  and a sample  $(X_1, Y_1), \dots, (X_n, Y_n)$ . In other words,  $\Xi = \mathcal{X} \times \mathcal{Y}$ ,  $\mathbb{S}$  is the set of measurable mappings  $\mathcal{X} \rightarrow \mathcal{Y}$  and the contrast  $\gamma(t; (x, y))$  measures the discrepancy between  $y$  and its predicted value  $t(x)$ . Two classical prediction frameworks are regression and classification, which are detailed below.



## Statistics of High-dimensional data

In data mining, lots of features = big problem.

- ▶ overfitting (model only works on in-sample data)
- ▶ computationally expensive
- ▶ low interpretability of final model
- ▶ in certain domains, features cost money!

# Handling Hi-D data?

## Feature **Extraction**

$$S = f(\Omega), \text{ usually } S = \mathbf{w}^T \Omega.$$

With extraction (e.g. PCA), we lose meaning of original features.

## Feature **Selection**

$$S \subseteq \Omega.$$

With selection we identify meaningful subsets of originals.

Combinatorial optimization over space of possible feature subsets.

## 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 relevancy index  $J$  using  $S$ .

    Adapt subset  $S$ .

**30** Return  $S$ .

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

**Con:** task-specific design is open problem

## How 'relevant' is any given feature?

$Y$  is a human-labeled document set.

$X$  is presence/absence of a particular image feature.

$I(X; Y)$  - measure of dependence between feature  $X$  and target  $Y$ .

Zero when  $X$  is independent of  $Y$ .

Increases as  $X$  and  $Y$  become dependent.

Defined as KL-divergence

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

## Filters using Mutual Information

Rank features  $X_i, \forall i$  by their values of  $J = I(X_i; Y)$ .

Retain the highest ranked features, discard the lowest ranked.

$i$	$J(X_i)$
35	0.846
42	0.811
10	0.810
654	0.611
22	0.443
59	0.388
...	...
212	0.09
39	0.05

Cut-off point decided by user, e.g.  $|S| = 5$ , so  
 $S = \{35, 42, 10, 654, 22\}$ .

Problem: F42 and F10 are almost identical!

## A Problem: Design of Filter Criteria

Q. What is the relevance 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

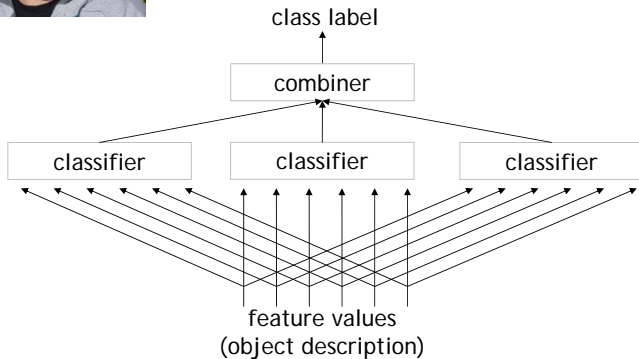
- ▶ Encouraging “relevant” features
- ▶ Discouraging correlated features

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

# The “Duality” of MCS and FS



© L.I. Kuncheva, ICPR Plenary 2008



## 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.

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

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

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

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

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

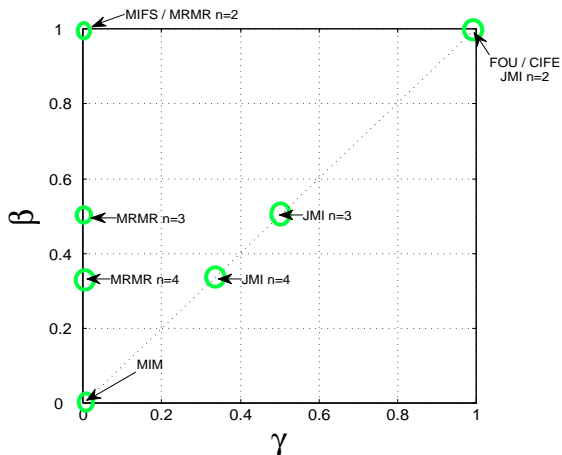
$$J_{cife} = I(X_i; Y) - \sum_{k \in S} I(X_i; X_k) + \sum_{k \in S} I(X_i; X_k | Y)$$

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

---

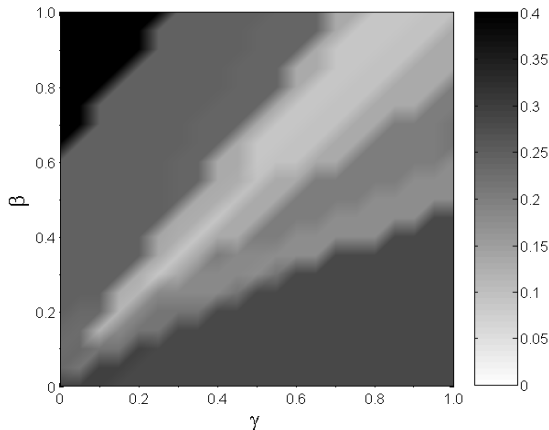
$$J = I(X_n; Y) - \beta \sum_{\forall k \in S} I(X_n; X_k) + \gamma \sum_{\forall k \in S} I(X_n; X_k | Y)$$

# 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

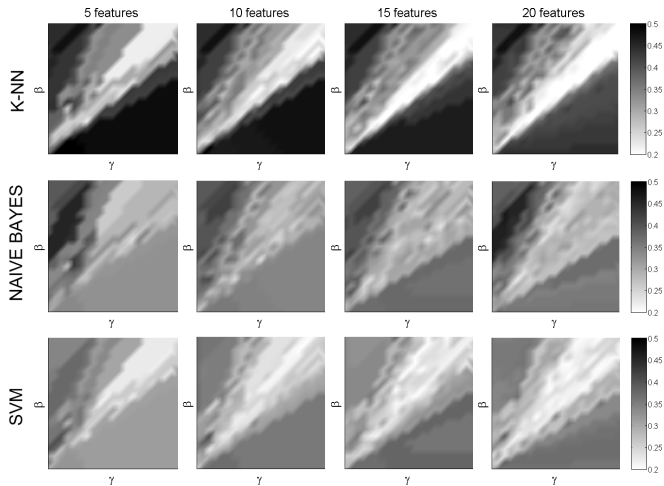


Figure 3: ARCENE data (cancer diagnosis).

# Exploring $\beta/\gamma$ space

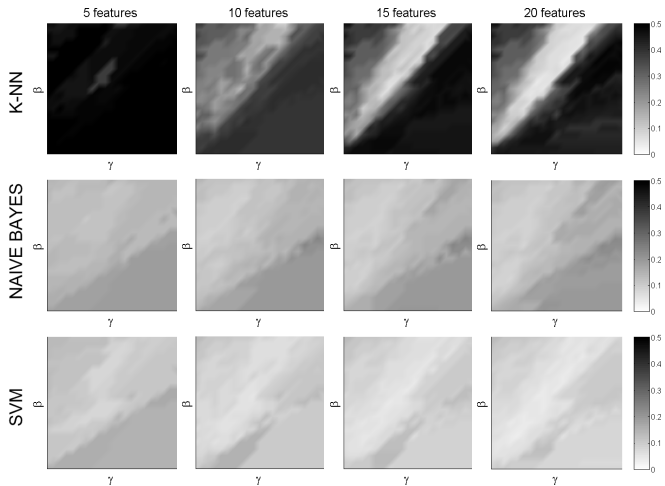
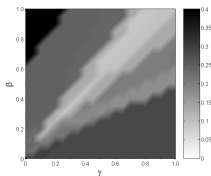


Figure 4: GISETTE data (handwritten digit recognition).

## Exploring $\beta/\gamma$ space

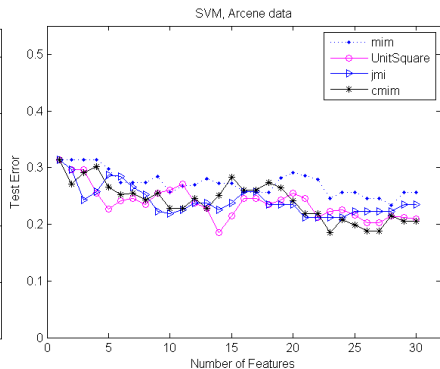
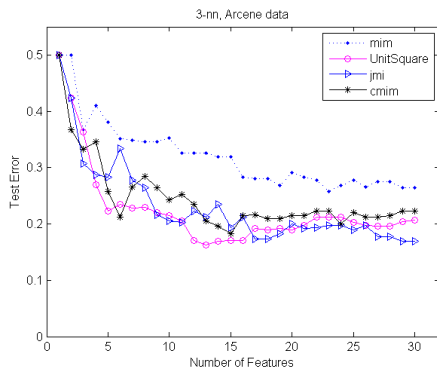


Seems straightforward? Just use the diagonal? Top right corner?  
But... remember these are only low order components.

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

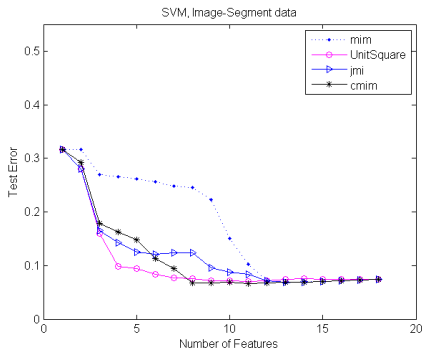
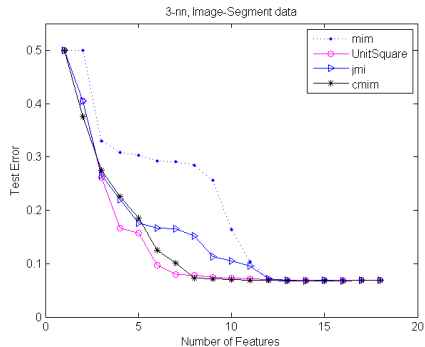
- ▶ e.g. data drawn from a Bayesian net with some nodes exhibiting deterministic behavior. (e.g. parity problem).

# ARCENE data



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

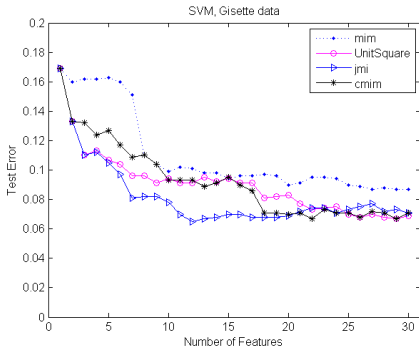
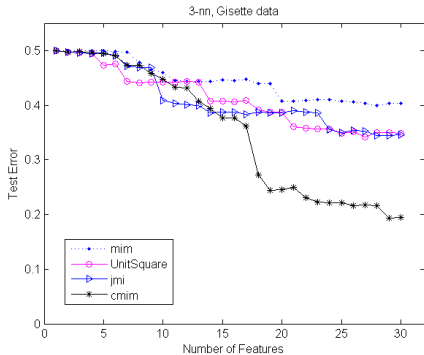
# Image Segment data



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...

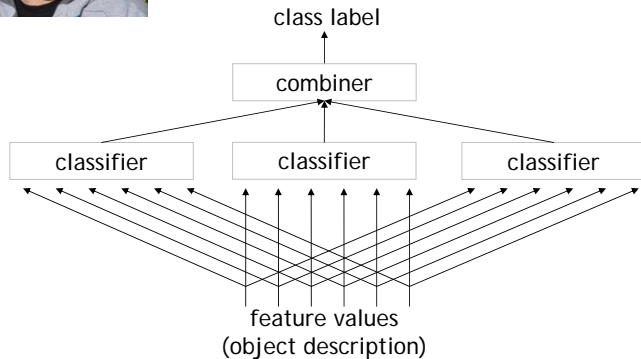
...solved an open problem in the Land of FS.

But could the MCS natives also learn from this?

# Committee members should be 'diverse'



© L.I. Kuncheva, ICPR Plenary 2008



# 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~~

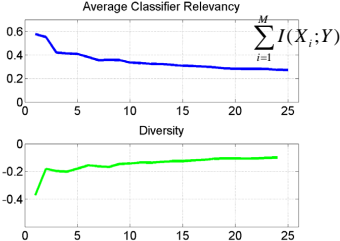
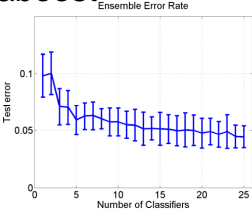
~~+ M-way diversity~~

---

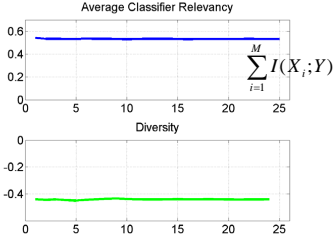
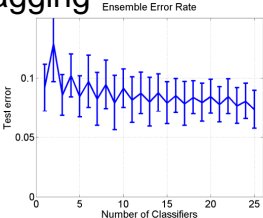
*"An Information Theoretic Perspective on Multiple Classifier Systems", MCS 2009.*

# Exports & Imports : Understanding Ensemble Diversity

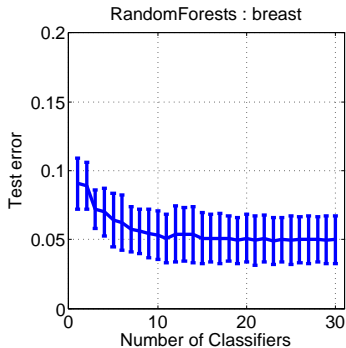
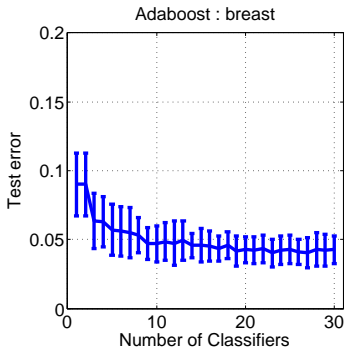
## Adaboost



## Bagging



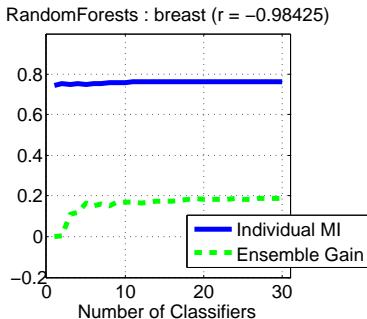
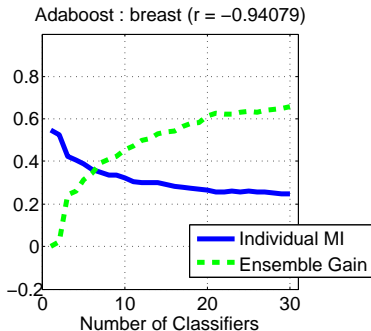
# Exports & Imports : Understanding Ensemble Diversity



---

(ongoing work with Zhi-Hua Zhou)

# Exports & Imports : Understanding Ensemble Diversity



---

(ongoing work with Zhi-Hua Zhou)

# 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
  
- ▶ Unified framework for feature selection
- ▶ Deeper understanding of committee learning methods