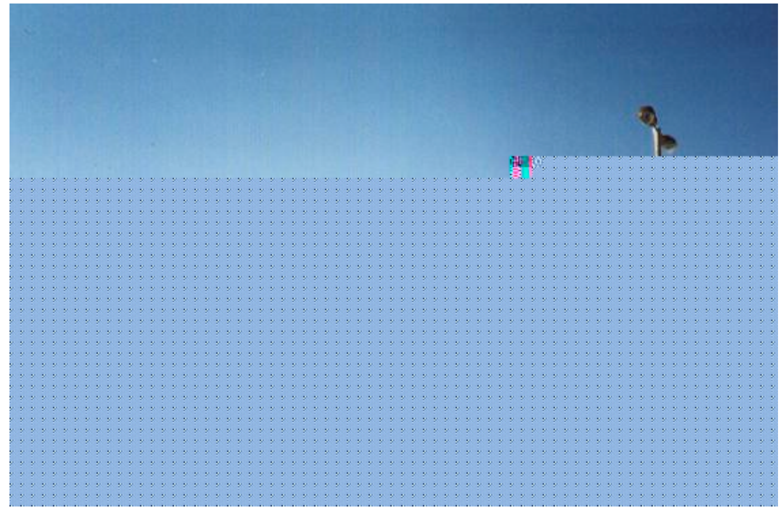
Measures of Diversity in Combining Classifiers

Part 1. General idea of diversity and pairwise measures

Ludmila I. Kuncheva

School of Informatics, University of Wales, Bangor
Bangor, Gwynedd, LL57 1UT
mas00a@bangor.ac.uk



North Wales

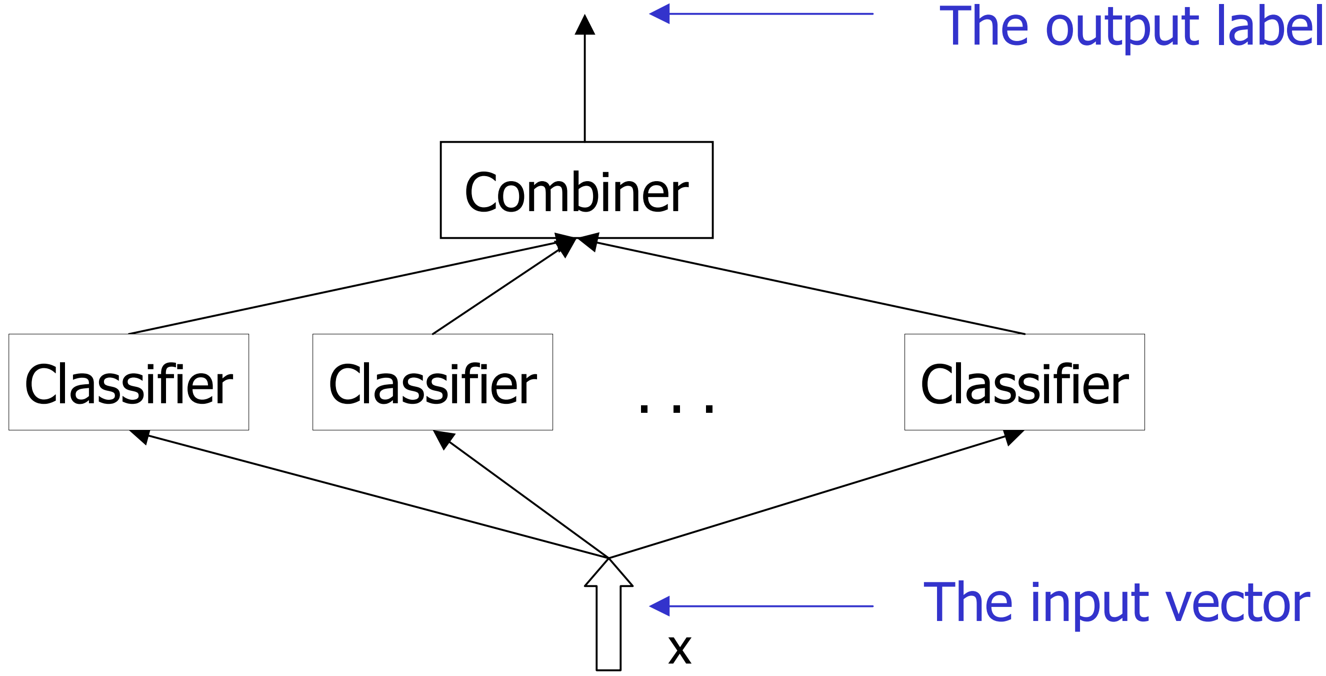
Bangor

Vietri sul Mare, 24 09 02

Part 1: General idea of diversity and pairwise measures

- Combining classifiers. Is independence the best scenario? Pattern of success and pattern of failure.
- An intuitive idea of diversity. “Good” and “bad” diversity.
- Measures of diversity and their various groupings. Pairwise measures.
- Why do the measures disagree? Diversity-accuracy dilemma.
- A synthetic enumerative experiment and the grim reality.
- Why is it difficult to design an experiment?
- What has been done so far?

Combining classifiers:

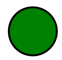


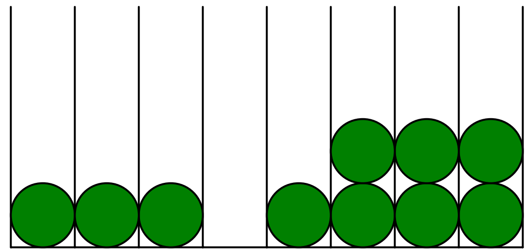
The important question: Should the classifiers agree, disagree or be independent?

The common (inaccurate!) belief: The classifiers should best be INDEPENDENT.

An example:

A possible distribution of votes of L=3 classifiers, p=0.6

10 objects, ; 2 possible votes: correct Π or wrong O



Classifier 1	O Π O O Π Π O Π	… 6 correct
Classifier 2	O O Π O Π O Π Π	… 6 correct
Classifier 3	O O O Π O Π Π Π	… 6 correct
Majority vote	O O O O Π Π Π Π	… 7 correct

(2 or 3 Π)

Suppose the classifiers were independent, each with accuracy p=0.6.

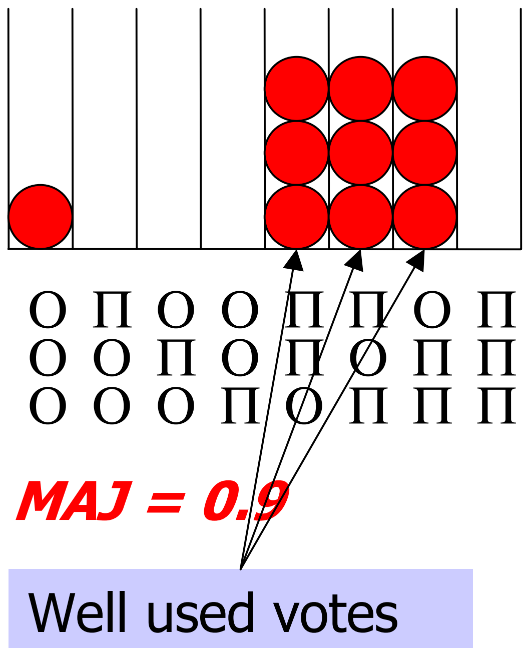
The probability of correct majority vote (at least 2 out of the 3) is

$$P(\text{exactly } 2) + P(\text{exactly } 3) = 3 \times (0.6)^2 \times 0.4 + (0.6)^3 = \mathbf{0.648}$$

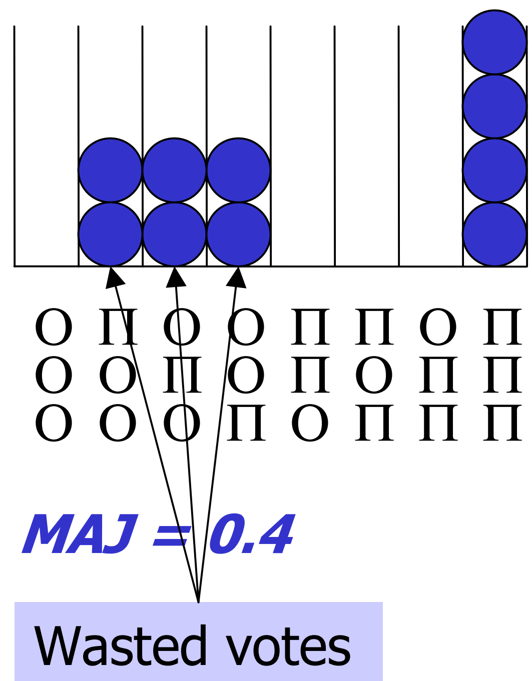
Can we beat that???

The same individual accuracies p=0.6, but a very different majority vote accuracy!

Pattern of Success



Pattern of Failure



The majority vote accuracy for 3 INDEPENDENT classifiers of p=0.6 would be less than 0.7!

The two probability distributions are:

Combination	Pattern of Success	Pattern of Failure
0 0 0	0.1	0.0
0 0 1	0.0	0.2
0 1 0	0.0	0.2
1 0 0	0.0	0.2
0 1 1	0.3	0.0
1 0 1	0.3	0.0
1 1 0	0.3	0.0
1 1 1	0.0	0.4

Try all combinations of votes for 10 patterns and 3 classifiers so that p=0.6. There are 28 possible combinations.

The frequencies of the ensemble accuracy are

- MV=0.4 🏆
- MV=0.5 🏆 🏆 🏆 🏆
- MV=0.6 🏆 🏆 🏆 🏆 🏆 🏆 🏆 🏆 🏆 🏆 🏆

- MV=0.7 🏆 🏆 🏆 🏆 🏆 🏆 🏆
- MV=0.8 🏆 🏆 🏆 🏆
- MV=0.9 🏆

Better than the single best (3/7)

CAUTION: This is not a “real” distribution; real ensembles might not span the whole possible range of accuracies.

GENERALLY and INTUITIVELY, a diverse ensemble is better than a non-diverse ensemble:

■ Correct □ Wrong

■■■■■□□□□□□□□□□
■■■■■□□□□□□□□□□
■■■■■□□□□□□□□□□

Non-diverse: MV accuracy 5/15

■■□■■□■□□□□□□□□
■□■■□■□■□□□□□□□
□■■□■■■□□□□□□□□

Diverse: MV accuracy 7/15
GOOD diversity

■□□□■□□□■□□■□□■
□□□■□■□■□■■□□□□
□■■□□□■□□□□□■■□

Diverse: MV accuracy 0
BAD diversity

How do we measure diversity? (Be it good or bad.)

Measures of Diversity

Pair-wise
Non-pair-wise

For oracle (binary) outputs
For label outputs
For continuous outputs

Ascending
Descending

Symmetric
Non-symmetric

Pair-wise measures of diversity:

Consider 2 classifiers at a time: D1 and D2.

■ Correct   □ Wrong

D1 ■ ■ □ ■ ■ □ ■ □ □ □ □ □ □ □ □

D2 ■ □ ■ ■ □ ■ □ ■ □ □ □ □ □ □ □

	D2 Π	D2 O
D1 Π	2/15	3/15
D1 O	3/15	7/15

Pair-wise measures of diversity:

Calculate the values for all L(L-1)/2 pairs of classifiers and then take the average.

Oracle outputs (correct/wrong)

	D2 Π	D2 O
D1 Π	*a*	*b*
D1 O	*c*	*d*

$a + b + c + d = 1$

Label outputs $(\omega_1, \ldots, \omega_c)$

	ω_1	ω_2	ω_3	ω_4
ω_1	a_{11}	a_{12}	a_{13}	
ω_2	a_{21}	...		
ω_3			...	
ω_4				a_{44}

$\Sigma\, a_{ki} = 1$

Measure	Reference	Notation	Formula
Q statistic	Yule (1900)*	Q	$\frac{ad - bc}{ad + bc}$
Correlation coefficient	Sneath and Sokal (1973)*	ρ	$\sqrt{\frac{ad - bc}{(a+b)(c+d)(a+c)(b+d)}}$
Disagreement measure	Skalak (1996) Ho (1998)	D	b + c
Double fault measure	Giacinto and Roli (2000)	DF	d
Interrater agreement	Margineantu & Dietterich (1997)	k	$\frac{2(ad - bc)}{(a+c)(c+d)+(a+b)(b+d)}$
Mutual information	Masulli and Valentini (2001)	mi	Ugh… Too large to show ☺

`*'means that the reference is not in the context of classifier combination

	D2 Π	D2 O
D1 Π	a	b
D1 O	c	d

Desperate to know that last one? There you go …

$$\text{mi} = a \log \frac{a}{(a+b)(a+c)} + b \log \frac{b}{(a+b)(b+d)} + c \log \frac{c}{(a+c)(c+d)} + d \log \frac{d}{(b+d)(c+d)}$$

Kullback-Leibler divergence between 2 probability distributions p(x) and q(x): $\Sigma\ p(x) \log (p(x)/q(x))$

	D2 Π	D2 O
D1 Π	a	b
D1 O	c	d

observed

	D2 Π	D2 O
D1 Π	$(a+c)(a+b)$	$(a+b)(b+d)$
D1 O	$(a+c)(c+d)$	$(b+d)(c+d)$

predicted by independence

	D2 Π	D2 O
D1 Π	*a*	*b*
D1 O	*c*	*d*

Let's think about Q, ρ, and κ:

Why do they all have $(ad - bc)$ in the numerator?

$$(ad - bc) = \boxed{(a + d)} - \boxed{[(a + b)(a + c) + (b + d)(c + d)]}$$

observed agreement

agreement by chance

There are more of them, and just for BINARY outputs! Why having so many of them? Because they behave differently…

	D2 Π	D2 O
D1 Π	98/100	1/100
D1 O	1/100	0/100

Q: Are these D1 and D2 diverse or not? (N=100)

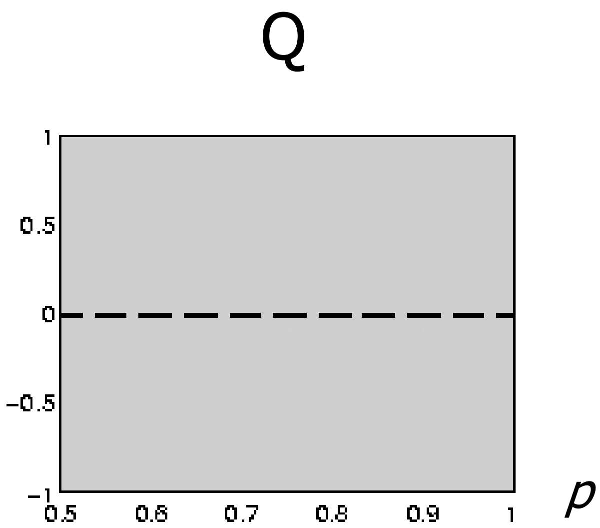
■□■■■■■■■■■ … ■■■■

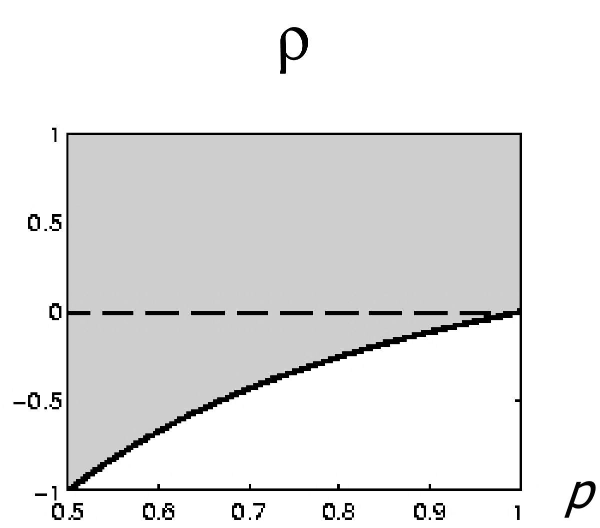
□■■■■■■■■■■ … ■■■■

Q statistic, Q	$\frac{ad - bc}{ad + bc}$	↓	[-1,1]	**-1**
Correlation coefficient, ρ	$\sqrt{\frac{ad - bc}{(a+b)(c+d)(a+c)(b+d)}}$	↓	[-1,1]	**- 0.01**
Disagreement, D	$b + c$	↑	[0,1]	**0.02**
Double fault, DF	d	↓	[0,1]	**0**
Interrater agreement, κ	$\frac{2(ad - bc)}{(a+c)(c+d)+(a+b)(b+d)}$	↓	[-1,1]	**-0.01**
Mutual infromation, mi	☺	?	[0,?]	**0**

The <u>DIVERSITY-ACCURACY dilemma</u>: Very accurate classifiers cannot be very diverse.

Caveat: This depends on which measure you are using!





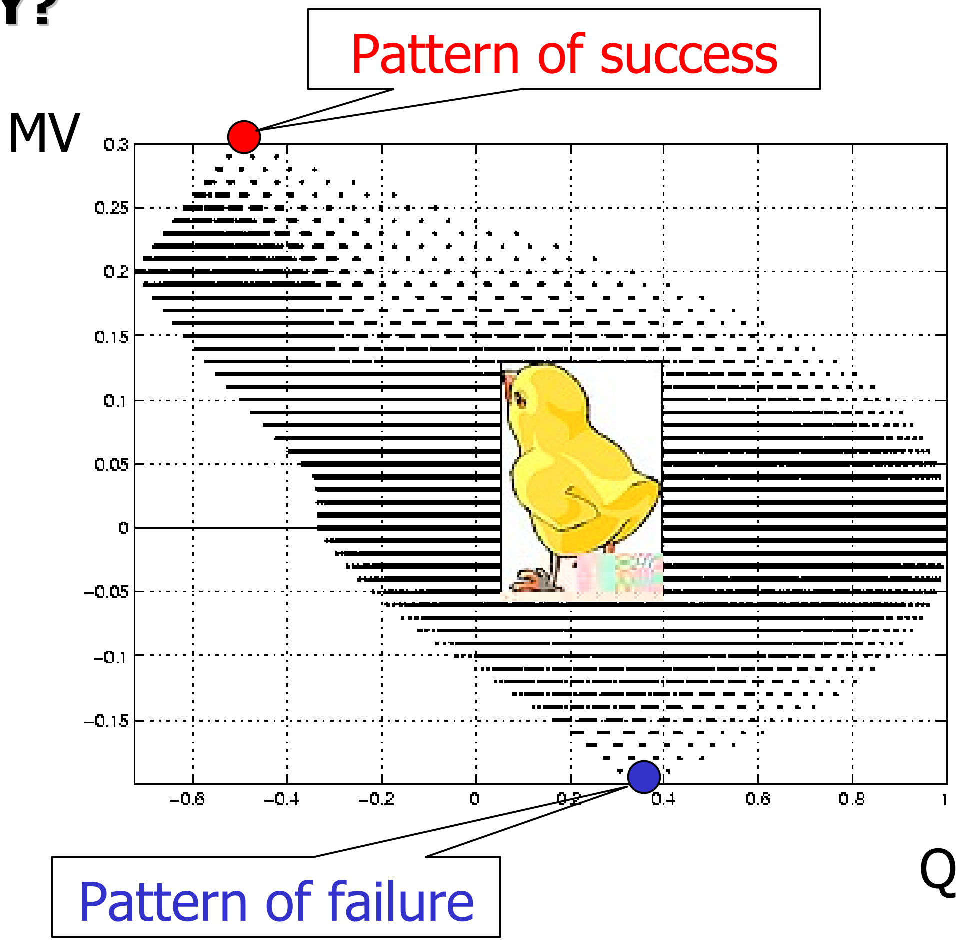
So, whom do we believe? Or do they measure DIFFERENT diversities?

Can we relate ANY OF THESE DIVERSITY VALUES with the ENSEMBLE ACCURACY?

Experiment:

L=3; p=0.6;

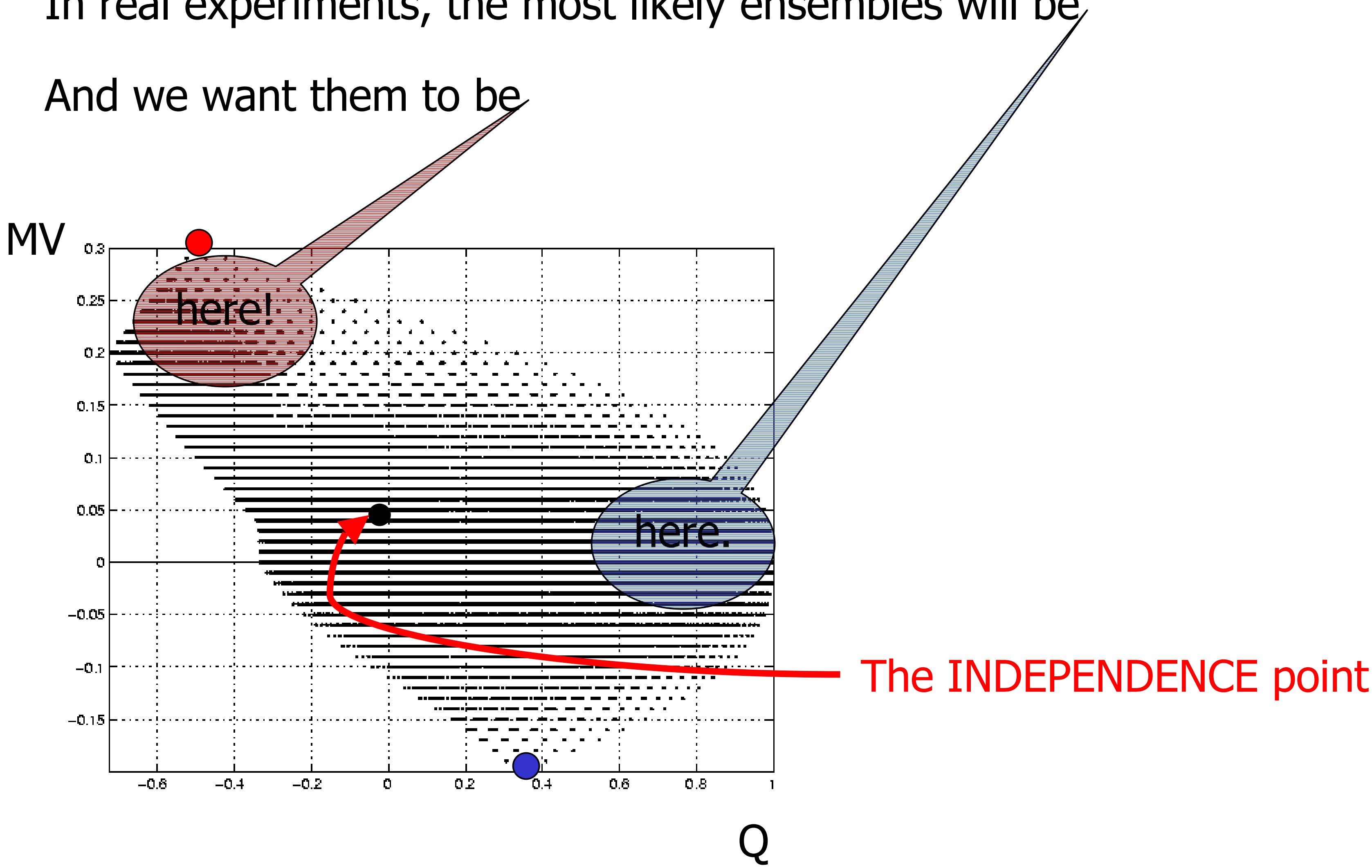
ALL possible vote distributions for N=100 objects



This is not a REAL distribution of the classifier ensembles…

In real experiments, the most likely ensembles will be

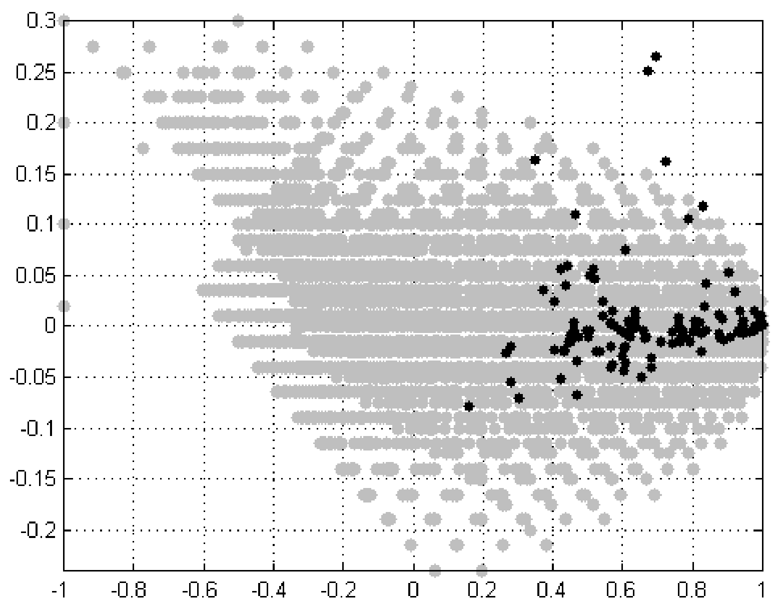
And we want them to be



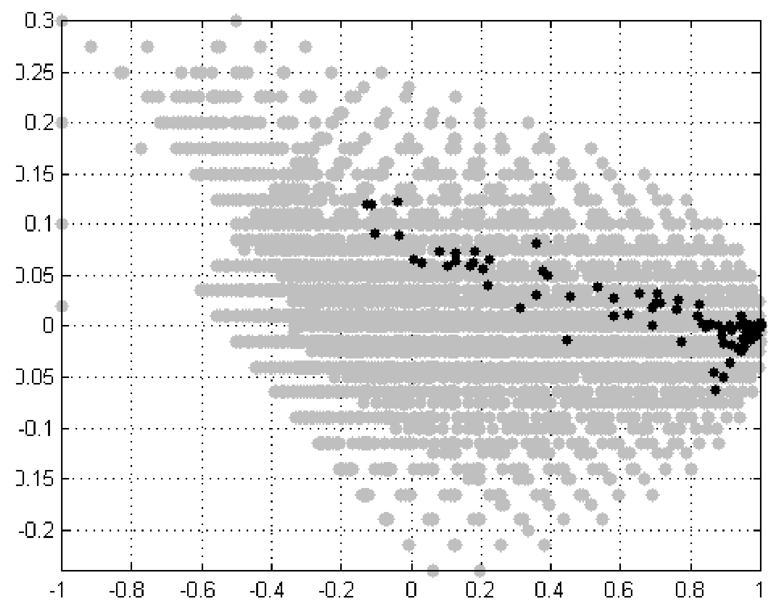
...The grim reality...

Improvement on the single classifier

Bagging



Boosting

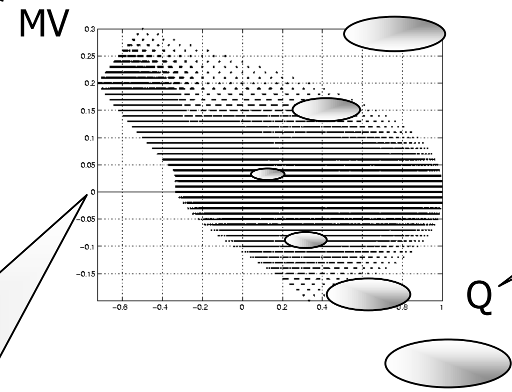


Q

Recall the “chicken” picture. What is wrong with it?

Why just majority vote?

Why just L = 3 classifiers?



Why just Q?

Why just individual accuracy p = 0.6 ?

Why should ALL classifiers have the same p ?

The main difficulty in designing an experiment about the relationship DIVERSITY-ACCURACY:

How do we generate the ensembles?

- Bagging, boosting, arcing, etc.? But there are so many variants! Besides, we cannot control the span of the diversity across the ensembles. So if there appeared to be no relationship between diversity and accuracy, this does not mean the two are not related *in general*.
- Enumeration, exhaustive search? Exhaustive on what? Only small, highly restrictive studies are possible (e.g., the chicken picture). And these do not tell us about the real ensembles.
- Synthetic experiments with pre-set *p*'s and *Q*'s just to see what happens for unequal *p*'s, bigger ensembles and different pairwise *Q*'s. Again, how is this related to real ensembles?

Easy way out : pick an Ensemble Generating Methodology and postulate that the relationship DIVERSITY-ACCURACY is specific for that methodology.

Specify all the details: How are the training sets generated? What feature subsets are used and how? What classifier models are used? (homogeneous/heterogeneous ensemble) What training protocol is adopted? What combination method is used?, etc.

Run the experiments: Vary the specified parameters within the context of the methodology (e.g., the combination formula or training size) and asses the potential of diversity measures.

Use diversity to improve the chosen methodology: If we are lucky we may find a way to encourage "good" diversity and suppress "bad" diversity while constructing the ensemble or selecting its parameters or training protocol.

Broaden the horizon a little

Consider label outputs (and again 2 classifiers at a time)

Label outputs $(\omega_1, \ldots, \omega_c)$ $\quad (\Sigma\, a_{ki} = 1)$

	ω_1	ω_2	ω_3	ω_4
ω_1	a_{11}	a_{12}	a_{13}	
ω_2	a_{21}	…		
ω_3			…	
ω_4				a_{44}

coincidence (confusion) matrix

Q: N/A

ρ: N/A

D: $1 - \Sigma\, a_{ii}$

DF: N/A

κ: $\dfrac{\Sigma a_{ii} - ABC}{1 - ABC}$

Agreement by chance

$ABC = \Sigma_i\, (\, \Sigma_k\, a_{ik}\,)\,(\, \Sigma_k\, a_{ki}\,)$

mi:

$\Sigma_i\, \Sigma_k\;\; a_{ik}\; \log[a_{ik}/(\, \Sigma_s\, a_{sk}\,)\,(\, \Sigma_s\, a_{is}\,)]$

What do we do with pairwise measures of diversity?

<u>1. Look at them.</u> How do we visualize the relationships between the classifiers in the ensemble?

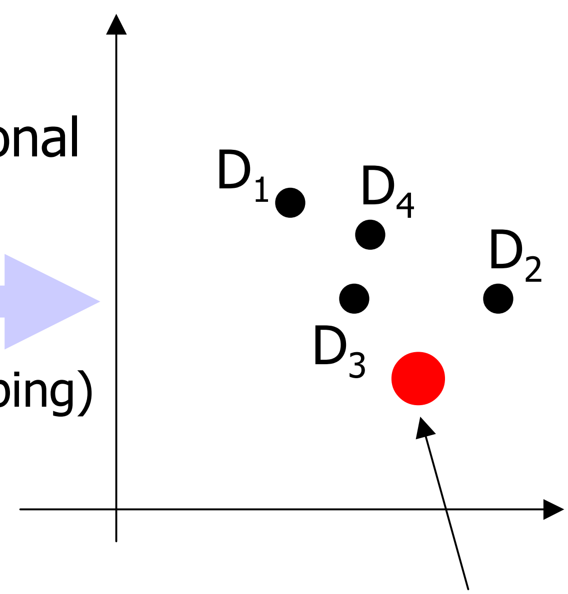
(a) [Pekalska et al., 2002] Use the *Classifier Projection Space*

	D_1	D_2	D_3	D_4
D_1		a	b	c
D_2			d	e
D_3				f
D_4				

Treated as Euclidean distances

Multidimensional scaling

(Sammon mapping)

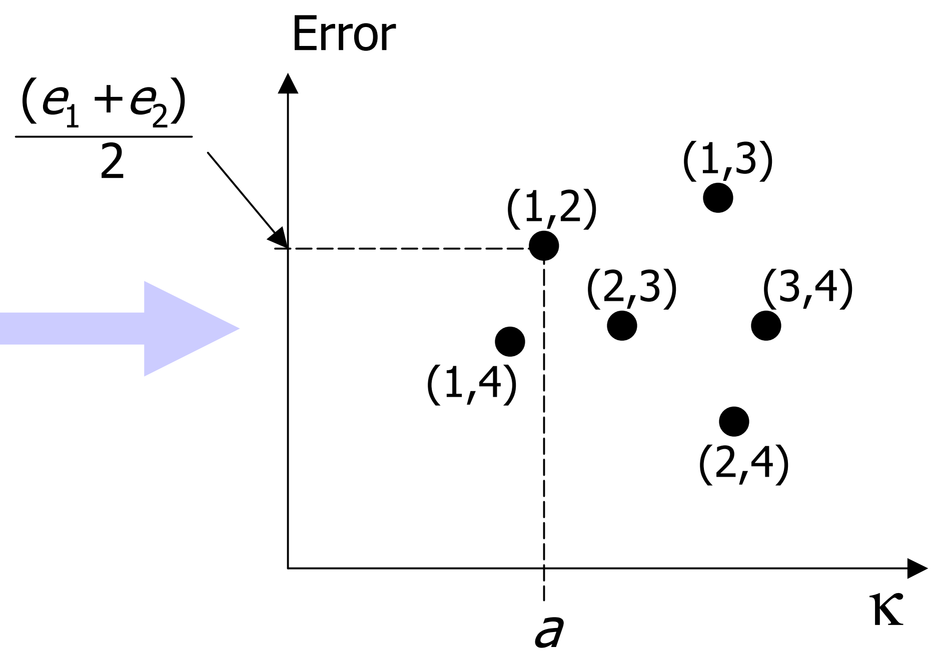


majority vote

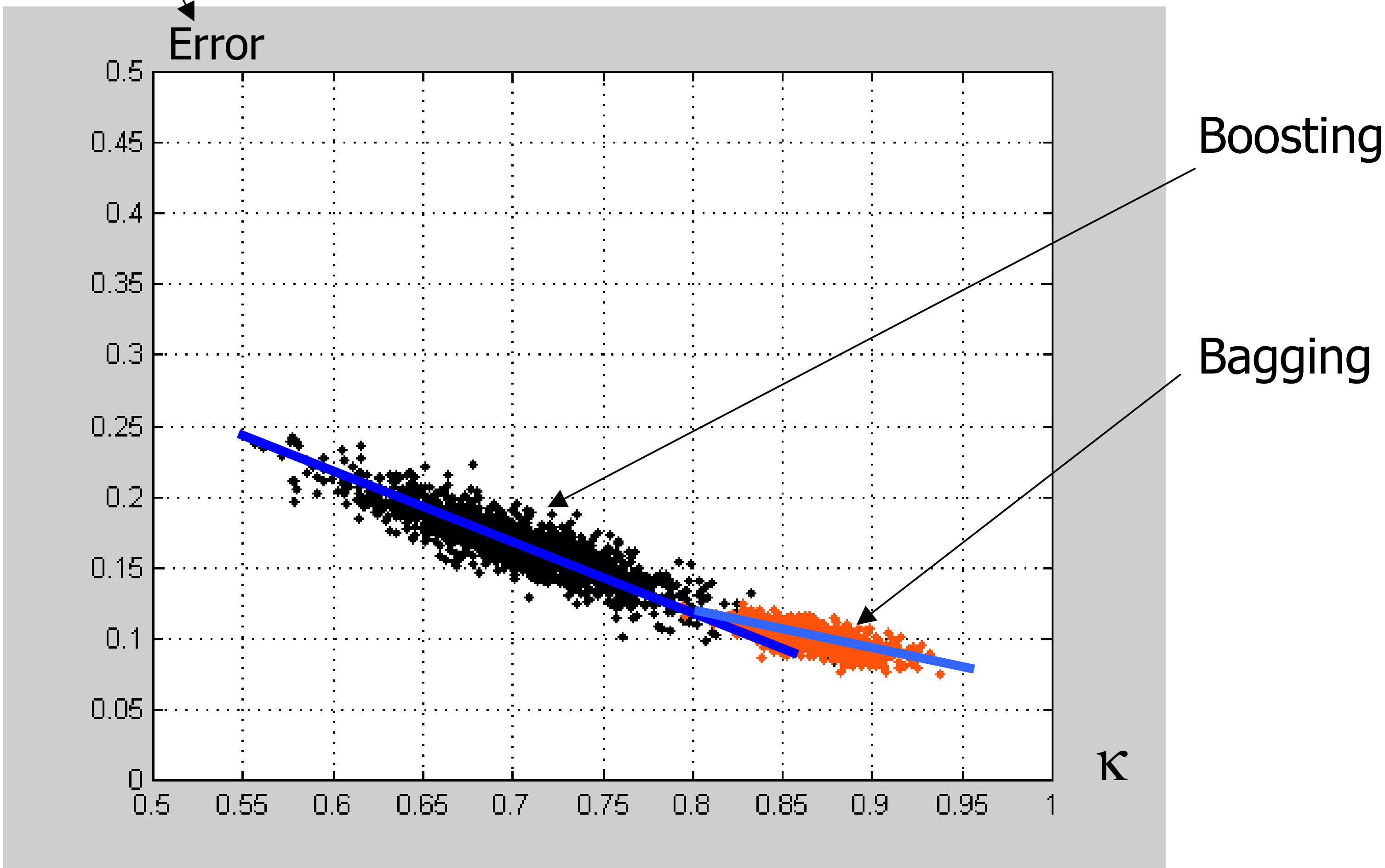
(and any other combination method)

(b) [Margineantu and Dietterich, 1997] Use *kappa-error plots*.

	D_1	D_2	D_3	D_4
D_1		a	b	c
D_2			d	e
D_3				f
D_4				

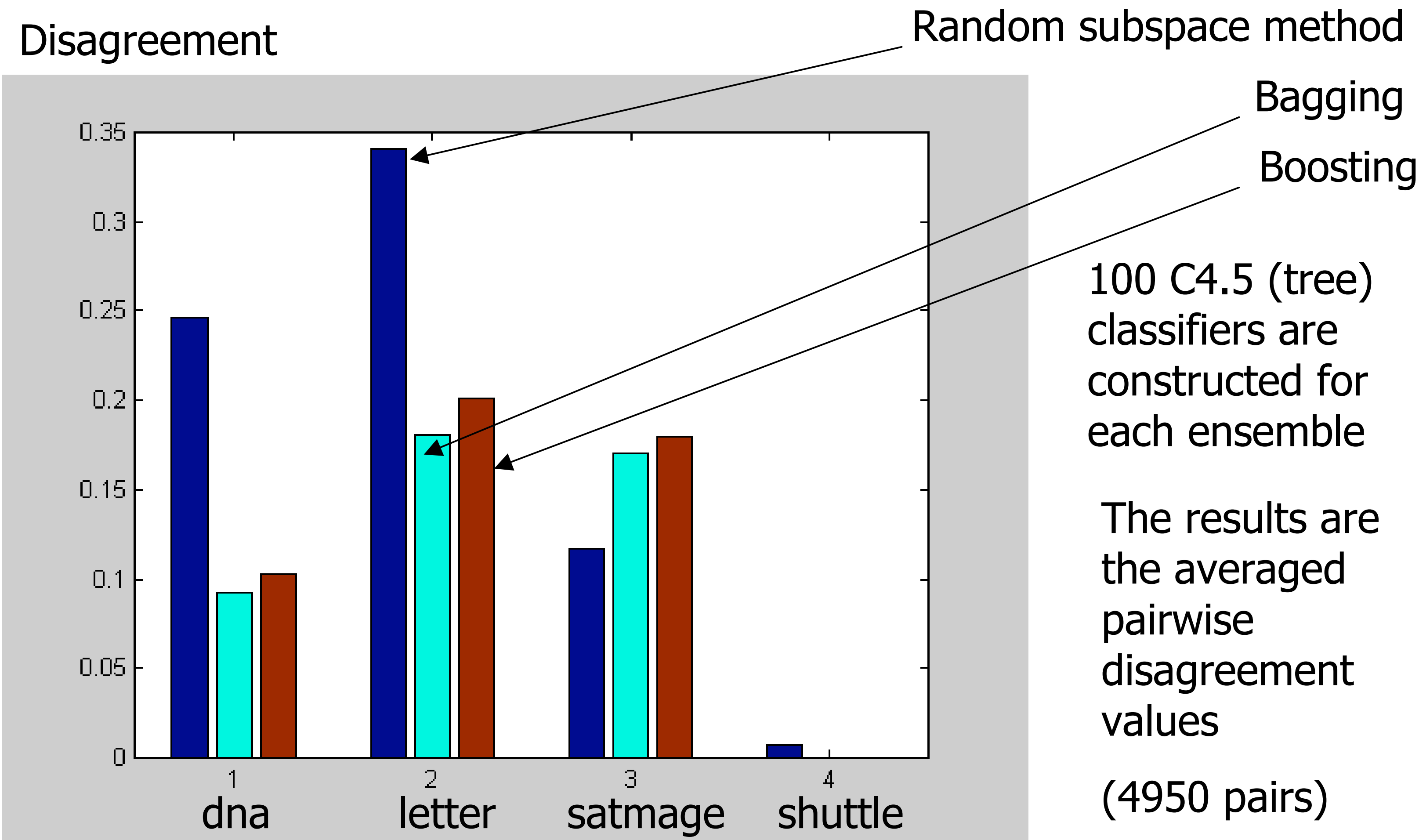


NB: The error is the individual average (of the pair), <u>not the ensemble error</u>!



Diversity-accuracy trade-off

(c) [T.K. Ho, 1998] Use *disagreement measure for label outputs*.



100 C4.5 (tree) classifiers are constructed for each ensemble

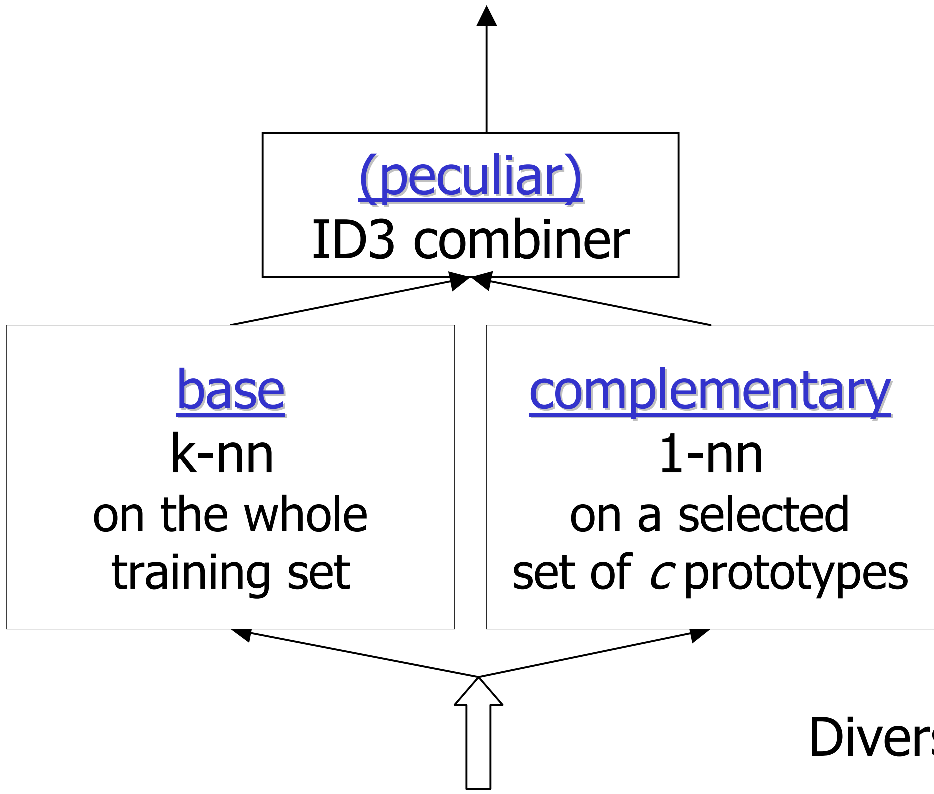
The results are the averaged pairwise disagreement values

(4950 pairs)

"Ideally, one should look for the best individual trees with lowest similarity. But exactly how this dual optimization can be done with an algorithm remains unclear."

(d) [D. Skalak, 1996] Use *disagreement measure for oracle outputs*.

2 cute algorithms: (1) Coarse reclassification; (2) Deliberate misclassification



(1) Coarse reclassification =
"radical destabilization of the nn algorithm by choosing a very small number of prototypes" = random editing

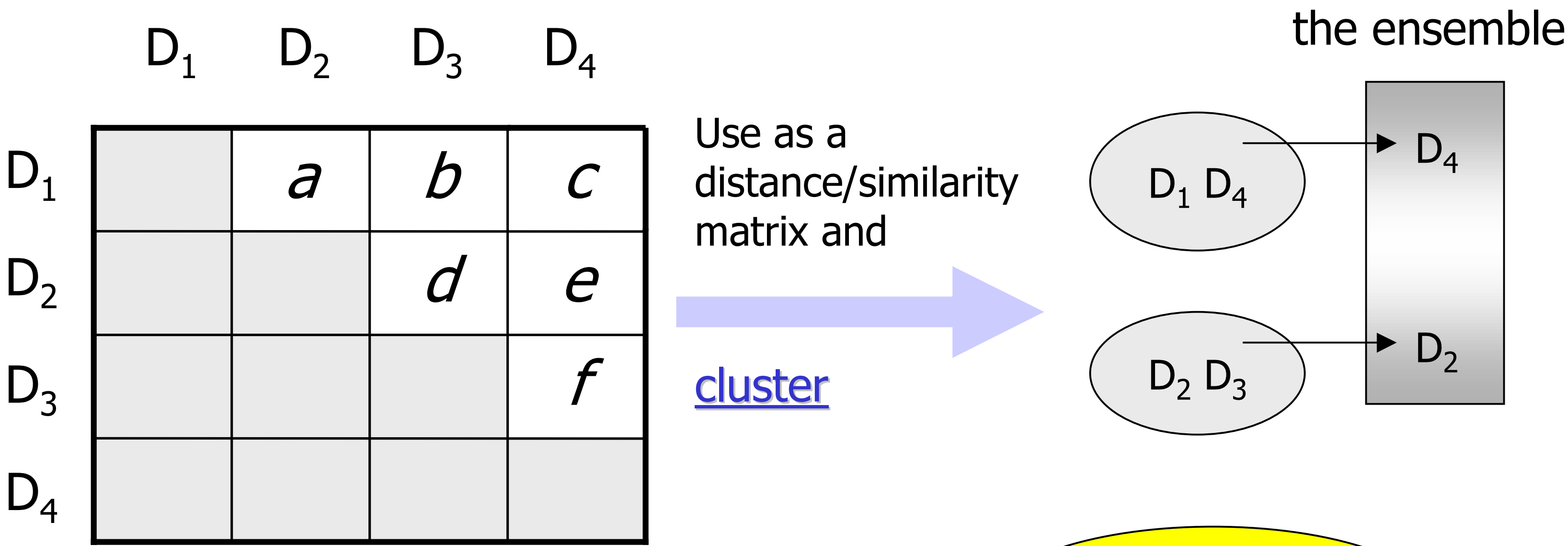
(2) Deliberate misclassification
Form the training set by switching the labels in the vicinity of objects misclassified by the base classifier

Diversity achieved: from 0.03 to 0.54

"Our study provides evidence that it may be useful to investigate families of boosting algorithms that incorporate varying level of accuracy and diversity so as to achieve an **appropriate mix** for a given task and domain."

2. Select the members of the ensemble

(a) [Roli et al., 2001, Giacinto & Roli, 2001] "Overproduce and select"



Diversity measures used: Q, DF, GD

We'll see later what that is.

" …Although these design methods [overproduce and select] exhibited some interesting features they do not guarantee to design the optimal multiple classifier system for the classification task at hand. Accordingly, the main conclusion of this paper [MCS'01] is that the problem of the optimal MCS design still remains open."

(b) [Margineantu and Dietterich, 1997] Use *kappa* and *kappa-error plots* to prune ensembles created by boosting.

κ	D_1	D_2	D_3	D_4
D_1		0.9	0.8	0.6
D_2			0.3	0.1
D_3				0.9
D_4				

Take the pair with the lowest kappa first, keep adding classifiers until the desired number is reached

ensemble of 3

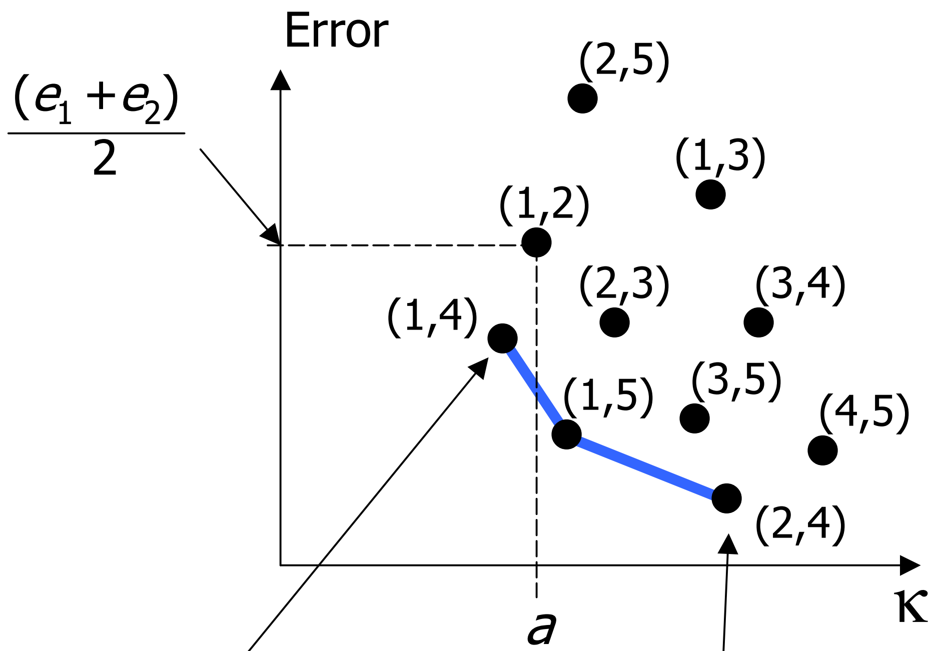
D_2 D_4

(D_2) D_3

This is a greedy algorithm, hence non-optimal. In the process of selection some pairs with high kappa will appear in the ensemble (D_3 and D_4 here).

Suppose we have obtained this kappa-error plot

(don't try to match it with the matrix on the previous slide!)



The ensemble

$D_1\,D_2\,D_4\,D_5$

Construct the *convex hull* or the *Pareto-optimal* set of pairs, i.e., the non-dominated ones. This will include the most accurate (on average) pair and the most diverse one too.

To summarize:

- Our intuition says that diversity is important in combining classifiers
- We don't have a consensus definition of diversity so far
- There are many measures (we looked at some pairwise measures) which might disagree with one another on the same data
- There is no clear-cut relationship between diversity and the ensemble accuracy
- Diversity-accuracy trade-off is measure-related
- Although there are some heuristic ideas about using diversity during building the ensemble we are still far from a consistent guideline, let alone a theoretical one.

And this… is the only piece of food left after **I**’ve been there …

This is Chris Whitaker on this year’s graduation lunch.