Getting Lost in the Wealth of Classifier Ensembles?

Ludmila Kuncheva

School of Computer Science
Bangor University
mas00a@bangor.ac.uk

Supported by Project RPG-2015-188 Sponsored by the Leverhulme trust
Doing research these days...
Number of publications
08 February 2016

"Classifier ensemble***"

"Big data"
Number of publications
08 February 2016

- Classifier ensembles
- Big data

Graph showing the number of publications over time for "Classifier ensemble" and "Big data".
Number of publications
08 February 2016

"Classifier ensemble**"
Yahoo releases largest ever machine learning dataset to research

Martin Anderson Thu 14 Jan 2016 4.05pm

110 billion events

1.5 terabytes zipped
Ronald Fisher

The FAMOUS Iris Data

150 data points
4 features + labels

150*(4+1)*8 (bytes) = 6Kb

k-i-l-o-b-y-t-e-s, that is

WEB OF SCIENCE™

Search

09 February 2016

"iris data"

568 hits

8 inch floppy
250 Kb

Dre-e-e-e-am
dream dream dream dreamdream...
1.44 Mb
Wow!!!
Let's say, generously, 1.5 Mb
Petabyte of storage is about 666,666,667 floppies

What was the largest dataset you analysed / data mined?

- Over 100 Pb
- 11 to 100 Pb
- 1.1 to 10 Pb
- 101 Tb to 1 Petabyte
- 11 to 100 Tb
- 1.1 to 10 Tb
- 101 Gb to 1 Terabyte
- 11 to 100 Gb
- 1.1 to 10 Gb
- 101 Mb to 1 Gb
- 11 to 100 MB
- 1.1 to 10 Mb
- Less than 1 Mb

459 voters

Pattern recognition algorithms
Training and testing protocols
Clever density approximation
Ingenious models

bye-bye...

Data management and organisation
Efficient storage
Distributed computing
Fast computing
Optimisation / search
Computer clusters

A bit depressing...
Enjoying the ride but kind of... insignificant
All the Giants

You...
What chance do you have with your little...

You...
Classifier Ensembles?
Classifier ensembles

```
Classifier
ensembles

class label

"combiner"

classifier
classifier
classifier

feature values
(object description)
```
VERY well liked classification approach. And this is why:

1. Typically better than the individual ensemble members and other individual classifiers.

2. And the above is enough 😊
Classifier ensembles


Second edition
Classifier ensembles

Netflix Prize
2009 – Winner 1,000,000 USD – an ensemble method

US Patents
Satellite classifier ensemble
US 7769701 B2
Methods for feature selection using classifier ensemble based genetic algorithms
US 8762303 B

Rapid Object Detection using a Boosted Cascade of Simple Features
Paul Viola
viola@merl.com
Mitsubishi Electric Research Labs
201 Broadway, 8th FL
Cambridge, MA 02139

Michael Jones
mjonas@crl.dec.com
Compaq CRL
One Cambridge Center
Cambridge, MA 02142

AdaBoost
cited 12,403
times by 19/02/2016
(Google Scholar)
combination of multiple classifiers [Lam95, Woods97, Xu92, Kittler98]
classifier fusion [Cho95, Gader96, Grabisch92, Keller94, Bloch96]
mixture of experts [Jacobs91, Jacobs95, Jordan95, Nowlan91]
committees of neural networks [Bishop95, Drucker94]
consensus aggregation [Benediktsson92, Ng92, Benediktsson97]
voting pool of classifiers [Battiti94]
dynamic classifier selection [Woods97]
composite classifier systems [Dasarathy78]
classifier ensembles [Drucker94, Filippi94, Sharkey99]
bagging, boosting, arcing, wagging [Sharkey99]
modular systems [Sharkey99]
collective recognition [Rastrigin81, Barabash83]
stacked generalization [Wolpert92]
divide-and-conquer classifiers [Chiang94]
pandemonium system of reflective agents [Smieja96]
change-glasses approach to classifier selection [KunchevaPRL93]
etc.
combination of multiple classifiers [Lam95, Woods97, Xu92, Kittler98]
classifier fusion [Cho95, Gader96, Grabisch92, Keller94, Bloch96]
mixture of experts [Jacobs91, Jacobs95, Jordan95, Nowlan91]
committees of neural networks [Bishop95, Drucker94]
consensus aggregation [Benediktsson92, Ng92, Benediktsson97]
voting pool of classifiers [Battiti94]
dynamic classifier selection [Woods97]
composite classifier systems [Dasarathy78]
classifier ensembles [Drucker94, Filippi94, Sharkey99]
bagging, boosting, arcing, wagging [Sharkey99]
modular systems [Sharkey99]
collective recognition [Rastrigin81, Barabash83]
stacked generalization [Wolpert92]
divide-and-conquer classifiers [Chiang94]
pandemonium system of reflective agents [Smieja96]
change-glasses approach to classifier selection [Kuncheva93]
etc.

Out of fashion

Subsumed
The big machine learning

The little pattern recognition

instance
classifier
object

attribute
learner

feature

example
The big machine learning

The little pattern recognition

attribute

learner

instance

classifier

object

feature

element

example
The big machine learning

The little pattern recognition

And don’t even start me on what they call things :)

instance
classifier
object

attribute
learner
featuere

example

Australia
International Workshops on Multiple Classifier Systems 2000 – 2014 – continuing
How do we design/build/train classifier ensembles?
Building ensembles

Levels of questions

A Combination level
- selection or fusion?
- voting or another combination method?
- trainable or non-trainable combiner?
- and why not another classifier?

B Classifier level
- same or different classifiers?
- decision trees, neural networks or other?
- how many?

C Feature level
- all features or subsets of features?
- random or selected subsets?

D Data level
- independent/dependent bootstrap samples?
- selected data sets?
Building ensembles

The two strategies: fusion versus selection

Combiner/Selector

class label

Classifier
Classifier
.
.
.
Classifier

Pick one
Use them all

Classifier selection
Classifier fusion

feature values

Building ensembles

The two strategies: fusion versus selection

Combiner/Selector

class label

Classifier
Classifier
.
.
.
Classifier

Pick one
Use them all

Classifier selection
Classifier fusion

feature values
Building ensembles

The two strategies: fusion versus selection

class label

Combiner/Selector

In fact, any classifier can be applied to these “intermediate features”

Classifier Classifier ... Classifier

feature values
Classifier ensembles: the “CLASSICS”

- **Bagging**: Independent bootstrap samples
  - sample sample sample sample

- **Boosting**: Dependent bootstrap samples
  - sample sample sample sample

- **Random Subspace**: Independent feature subsamples
  - sample sample sample sample

- **Random Forest**: Random trees
  - sample sample sample sample
  - Independent bootstrap samples
Classifier ensembles: the not-so-classics

- Rotation Forest
  - Sample
  - Sample
  - Sample
  - Sample
  - Standard or random trees
  - Independent bootstrap samples

- Linear Oracle
  - S1, S2
  - S1, S2
  - S1, S2
  - S1, S2
  - Independently split samples
Building ensembles

Levels of questions

A Combination level
- selection or fusion?
- voting or another combination method?
- trainable or non-trainable combiner?
- and why not another classifier?

B Classifier level
- same or different classifiers?
- decision trees, neural networks or other?
- how many?

C Feature level
- all features or subsets of features?
- random or selected subsets?

D Data level
- independent/dependent bootstrap samples?
- selected data sets?

Random subspace
Rotation Forest
Linear Oracle
Boosting
Bagging
“Anchor” points

1. Combiner
Building ensembles

Levels of questions

A Combination level
- selection or fusion?
- voting or another combination method?
- trainable or non-trainable combiner?
- and why not another classifier?

This seems under-researched...

B Classifier level
- same or different classifiers?
- decision trees, neural networks or other?
- how many?

Random Forest

Boosting

Bagging

Linear Oracle

C Feature level
- all features or subsets of features?
- random or selected subsets?

D Data level
- independent/dependent bootstrap samples?
- selected data sets?

Rotation Forest

Random subspace

Combiner

Classifier 1

Classifier 2

…

Classifier L

Features

Data set
Continuous-valued outputs

Decision profile

\[ P_3(\omega_2 | \mathbf{x}) \]
Let's call this data “The Tropical Fish” or just the fish data.

50-by-50 = 2500 objects in 2-d

Bayes error rate = 0%
Example: 2 ensembles

Throw 50 “straws” and label the sides so that the accuracy is greater than 0.5

Train 50 linear classifiers on bootstrap samples
Example: 2 ensembles

Each classifier returns an estimate for class "Fish"

\[ P(F \mid [x, y]^T) \]

And, of course, we have

\[ P(\overline{F} \mid [x, y]^T) = 1 - P(F \mid [x, y]^T) \]

but we will not need this.
Example: 2 ensembles

5% label noise – Majority vote

Combiners
- Majority vote
- Average
- Product
- Trained Linear
- Weighted Avg...
- Ridge Regression
- Tree Combiner
- Decision Tree...
- Naive Bayes
- BKS

Noise

4.98

69.72 %

69.60 %
Example: 2 ensembles

5% label noise – trained linear combiner

Combiners
- Majority vote
- Average
- Product
- Trained Linear
- Weighted Avg...
- Ridge Regres...
- Tree Combiner
- Decision Tem...
- Naive Bayes
- BKS

Noise 4.98

96.92 %

92.40 %
What does the example show?

- The combiner matters (a lot)
- The trained combiner works better

However, nothing is as simple as it looks...
The Combining Classifier: to Train or Not to Train?
The Combining Classifier: to Train or Not to Train?
Train the COMBINER if you have “enough” data!

Otherwise, like with any classifier, we may over-fit the data.

Get this: Almost NOBODY trains the combiner, not in the CLASSIC ensemble methods anyway.

Ha-ha-ha, what is “enough” data?
The Combining Classifier: to Train or Not to Train?

Train the COMBINER if you have “enough” data! Otherwise, like with any classifier, we may overfit the data.

Get this: Almost NOBODY trains the combiner, not in the CLASSIC ensemble methods anyway.

Don’t you worry... BIG DATA is coming

Ha-ha-ha, what is “enough” data?
Train the combiner and live happily ever after!
“Anchor” points

2. Diversity
All ensemble methods we have seen so far strive to keep the individual accuracy high while increasing diversity.

- How can we measure diversity?
- WHAT can we do with the diversity value?
  - Compare ensembles
  - Explain why a certain ensemble heuristic works and others don’t
  - Construct ensemble by overproducing and selecting classifiers with high accuracy and high diversity
Are we still talking about diversity?

classifier + ensemble + diversity: 713 papers, 6543 citations

Published each year (713)

Cited each year (6543)

Search on 22 Feb 2016
Measure diversity for a PAIR of classifiers

**independent outputs ≠ independent errors**

hence, use ORACLE outputs

<table>
<thead>
<tr>
<th></th>
<th>Classifier 1</th>
<th>Classifier 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>correct</td>
<td></td>
<td>correct</td>
</tr>
<tr>
<td></td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>wrong</td>
<td></td>
<td>wrong</td>
</tr>
<tr>
<td></td>
<td>c</td>
<td>d</td>
</tr>
</tbody>
</table>

Number of instances labelled correctly by classifier 1 and mislabelled by classifier 2
Measure diversity for a PAIR of classifiers

<table>
<thead>
<tr>
<th>Objects</th>
<th>C1</th>
<th>C2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Classifier 1</th>
<th>correct</th>
<th>wrong</th>
</tr>
</thead>
<tbody>
<tr>
<td>correct</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>wrong</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Classifier 2

Objects

<table>
<thead>
<tr>
<th>1</th>
<th>1</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
Measure diversity for a PAIR of classifiers

Diversity

Not diverse

Classifier 2

<table>
<thead>
<tr>
<th>Objects</th>
<th>C1</th>
<th>C2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Classifier 1

<table>
<thead>
<tr>
<th></th>
<th>correct</th>
<th>wrong</th>
</tr>
</thead>
<tbody>
<tr>
<td>correct</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>wrong</td>
<td>0</td>
<td>3</td>
</tr>
</tbody>
</table>
Measure diversity for a PAIR of classifiers

<table>
<thead>
<tr>
<th>Classifiers</th>
<th>Correct</th>
<th>Wrong</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classifier 1</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Classifier 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Objects</td>
<td>C1</td>
<td>C2</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
Measure diversity for a PAIR of classifiers

Diversity resides here

<table>
<thead>
<tr>
<th></th>
<th>Classifier 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>correct</td>
<td>a</td>
</tr>
<tr>
<td></td>
<td>b</td>
</tr>
<tr>
<td>wrong</td>
<td>c</td>
</tr>
<tr>
<td></td>
<td>d</td>
</tr>
</tbody>
</table>

Joint error
## Diversity

<table>
<thead>
<tr>
<th>Classifier 2</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>correct</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>wrong</td>
<td>c</td>
<td>d</td>
</tr>
</tbody>
</table>

- $Q$
- kappa
- correlation (rho)
- disagreement
- double fault

### A Survey of Binary Similarity and Distance Measures

Seung-Seok Choi, Sung-Hyuk Cha, Charles C. Tappert  
Department of Computer Science, Pace University  
New York, US

**ABSTRACT**

The binary feature vector is one of the most common representations of patterns and measuring similarity and distance measures play a critical role in many problems such as clustering, classification, etc. Ever since Jaccard ecological 25 fish species [21]. Tubbs summarized seven conventional similarity measures to solve the template matching problem [28], and Zhang et al. compared those seven measures to show the recognition capability in handwriting identification [31]. Willett evaluated 13 similarity measures for binary fingerprint code [30]. Ch
Diversity

Table 2 lists definitions of 76 binary similarity and distance measures used over the last century where S and D are similarity and distance measures, respectively.

Table 2: Definitions of Measures for binary data

<table>
<thead>
<tr>
<th>Measure</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>S_TARANTULA</td>
<td>( \frac{a}{c} + \frac{d}{a} )</td>
</tr>
<tr>
<td>S_SAMPLE</td>
<td>( \frac{a + b}{c + d} )</td>
</tr>
<tr>
<td>S_EXTRA</td>
<td>( \frac{(a + b)(a + c)(b + d)(c + d)}{n^2(na - (a + b)(a + c))} )</td>
</tr>
</tbody>
</table>

The inclusion or exclusion of negative matches, \( d \), in the binary similarity measures have been an ongoing issue. Most methods, such as the Tarantula, Tauc, the Faith, the Ochiai, the Yale, the Cosine, Pearson r, and the Sibson, are included in the negative match inclusive measures. The Accord, the Tarantula, the Dice & Sorensen, and the Kulsinski 1, the Simple matching, and the Jaccard are included in the negative match exclusive measures. The average of the standard measures is used in the calculation of the final similarity measure. The results obtained were used to observe the differences in the results obtained for different measures.
Do we need more “NEW” pairwise diversity measures?

Looks like we don’t...

And the same holds for non-pairwise measures...
Far too many already.
Take just ONE measure – kappa – \( \kappa \) – not because it is “the best” but because one is enough.

**Kappa-error diagrams**

- proposed by Margineantu and Dietterich in 1997
- visualise individual accuracy and diversity in a 2-dimensional plot
- have been used to decide which ensemble members can be pruned without much harm to the overall performance
### Kappa-error diagrams

<table>
<thead>
<tr>
<th></th>
<th>C1</th>
<th></th>
<th>C2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>correct</td>
<td>wrong</td>
<td></td>
</tr>
<tr>
<td>C1</td>
<td>a</td>
<td>b</td>
<td></td>
</tr>
<tr>
<td>correct</td>
<td>c</td>
<td>d</td>
<td></td>
</tr>
<tr>
<td>wrong</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**error**

\[
e_1 = \frac{c + d}{a + b + c + d}; \quad e_2 = \frac{b + d}{a + b + c + d}
\]

\[
e = \frac{b + c + 2d}{2(a + b + c + d)}
\]

**kappa** = \( \frac{\text{observed} - \text{chance}}{1 - \text{chance}} \)

\[
\kappa = \frac{2(ad - bc)}{(a + b)(b + d) + (a + c)(c + d)}
\]
Example

sonar data (UCI): 260 instances, 60 features, 2 classes, ensemble size $L = 11$ classifiers, base model – tree C4.5

- Adaboost: 75.0%
- Bagging: 77.0%
- Random subspace: 80.9%
- Random oracle: 83.3%
- Rotation Forest: 84.7%
Kappa-error diagrams

error \[ e = \frac{b + c + 2d}{2(a + b + c + d)} \]

kappa \[ \kappa = \frac{2 \ (ad - bc)}{(a+b)(b+d) + (a+c)(c+d)} \]

bound (tight) \[ \kappa_{\text{min}} = \begin{cases} 1 & 0 < e \leq 0.5 \\ \frac{1}{1-e} & 0.5 < e \leq 1 \end{cases} \]

Kappa-error diagrams – bounds
Kappa-error diagrams – simulated ensembles $L = 3$

- min = 6%
- median = 50%
- max = 91%
Kappa-error diagrams – simulated ensembles $L = 3$
Kappa-error diagrams – How much SPACE do we have to the bound?

In theory – none. I can design ensembles with accuracy 100%, all on the bottom branch of the bound.
Kappa-error diagrams – How much **SPACE** do we have to the bound?

In practice – this is a different story. We must “engineer” diversity to get better ensembles, but this is not easy...
Kappa-error diagrams – How much SPACE do we have to the bound?

5 real data sets
Is there space for new classifier ensembles?

Looks like yes...

But we need revolutionary ideas about embedding diversity into the ensemble
Why is diversity so baffling?

The problem is that diversity is NOT monotonically related to the ensemble accuracy.

In other words, diverse ensembles may be good or may be bad...
Good and Bad diversity

MAJORITY VOTE

3 classifiers: A, B, C
15 objects, □ wrong vote, ■ correct vote
individual accuracy = 10/15 = 0.667
P = ensemble accuracy

- independent classifiers
  P = 11/15 = 0.733

- identical classifiers
  P = 10/15 = 0.667

- dependent classifiers 1
  P = 7/15 = 0.467

- dependent classifiers 2
  P = 15/15 = 1.000
Good and Bad diversity

MAJORITY VOTE

3 classifiers: A, B, C

15 objects,

- wrong vote,
- correct vote

individual accuracy = 10/15 = 0.667

P = ensemble accuracy

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

independent classifiers

P = 11/15 = 0.733

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

identical classifiers

P = 10/15 = 0.667

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

dependent classifiers 1

P = 7/15 = 0.467

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Dependent classifiers 2

P = 15/15 = 1.000

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Good diversity

Bad diversity
Good and Bad diversity

\( l_i \) number of classifiers with correct output for \( z_i \)

\( L - l_i \) number of classifiers with wrong output for \( z_i \)

\( \bar{p} \) mean individual accuracy

\( N \) number of data points

\[ E_{maj} = (1 - \bar{p}) - \frac{1}{NL} \sum_{maj} (L - l_i) + \frac{1}{NL} \sum l_i \]

Decomposition of the Majority Vote Error

Good and Bad diversity

This object will contribute $L - l_i = (7 - 4) = 3$ to good diversity

This object will contribute $l_i = 3$ to bad diversity

Note that diversity quantity is 3 in both cases
Ensemble Margin

The voting margin for object $z_i$ is the proportion of $\text{correct minus wrong votes}$

$$m_i = \frac{l_i - (L - l_i)}{L}$$

For $z_1$,

$$m_1 = \frac{4 - (7 - 4)}{7} = \frac{1}{7}$$

For $z_2$,

$$m_2 = \frac{3 - (7 - 3)}{7} = -\frac{1}{7}$$

POSITIVE

NEGATIVE
Ensemble Margin

Average margin

\[ \bar{m} = \frac{1}{N} \sum_{i=1}^{N} m_i = \frac{1}{N} \sum_{i=1}^{N} \frac{l_i - (L - l_i)}{L} \]

Large \( \bar{m} \) corresponds to BETTER ensembles...

However, nearly all diversity measures are functions of

**Average absolute margin**

\[ |m| = \frac{1}{N} \sum_{i=1}^{N} |m_i| \]

or

**Average square margin**

\[ \bar{m}^2 = \frac{1}{N} \sum_{i=1}^{N} m_i^2 \]

Margin has no sign...
The bottom line is:
Diversity is not MONOTONICALLY related to ensemble accuracy

So, stop looking for what is not there...
Where next in classifier ensembles?
20 years from now, what will stay in the textbooks on classifier ensembles?
We will branch out like every other walk of science

Leonardo da Vinci 1452 - 1519

Isaac Newton 1643 - 1727

Leo Breiman 1928 - 2005

RZ21 4X@#3 2216 - 2354

A polymath. Invention, painting, sculpting, architecture, science, music, and mathematics

English physicist and mathematician

Statistician

Classifier ensemblist (Concept drift, imbalanced classes)

Ah, and Big Datist too.

A polymath is a person whose expertise spans a significant number of different subject areas.
Instead of conclusions :)
For the winner

by my favourite illustrator Marcello Barenghi

Well, I’ll give you a less crinkled one :)
A guessing game: CITATIONS on Web of Science 23/02/2016

"Deep learning" 1,490
"Big data" 8,151

Sort 2-9 from highest to lowest

1. "Classifier Ensemble*” 788
2. (1) + "concept drift"
3. (1) + "rotation forest"
4. (1) + adaboost
5. (1) + (imbalanced or unbalanced)
6. (1) + diversity
7. (1) + combiner
8. (1) + "big data"
9. (1) + "deep learning"
<table>
<thead>
<tr>
<th>Citation counts</th>
<th>1. “Classifier Ensemble*”</th>
<th>788</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2. (1) + “concept drift”</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>3. (1) + “rotation forest”</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>4. (1) + adaboost</td>
<td>78</td>
</tr>
<tr>
<td></td>
<td>5. (1) + (imbalanced or unbalanced)</td>
<td>37</td>
</tr>
<tr>
<td></td>
<td>6. (1) + diversity</td>
<td>205</td>
</tr>
<tr>
<td></td>
<td>7. (1) + combiner</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>8. (1) + “big data”</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>9. (1) + “deep learning”</td>
<td>0</td>
</tr>
</tbody>
</table>

“Deep learning” 1,490

“Big data” 8,151
Solution

“Classifier Ensemble*”
788

“Deep learning”
1,490

“Big data”
8,151

6. (1) + diversity

4. (1) + adaboost

5. (1) + (imbalanced or unbalanced)

3. (1) + “rotation forest”

2. (1) + “concept drift”

7. (1) + combiner

8. (1) + “big data”

9. (1) + “deep learning”
And thank you for listening to me!