

'Good' and 'Bad' Diversity in Classifier Ensembles

Gavin Brown and Ludmila Kuncheva

MCS 2010, Cairo

University of Manchester

University of Bangor

Overview

- ▶ The perspective...
- ▶ Good Diversity, Bad Diversity
- ▶ Toy data & Results
- ▶ Relation to Patterns of Success/Failure
- ▶ Questions

Perspective: Diversity should be natural

Diversity should be a **natural consequence** of:

1. The loss function
2. The combiner function

and nothing more.

Perspective: Diversity should be natural

Loss function: **quadratic**

Combiner: **linear**

$$h \in \mathcal{R}$$

$$\bar{h} = \frac{1}{T} \sum_t h_t$$

$$(\bar{h} - y)^2 = \frac{1}{T} \sum_t (h_t - y)^2 - \underbrace{\frac{1}{T} \sum_t (h_t - \bar{h})^2}_{\text{diversity}}.$$

Perspective: Diversity should be natural

Loss function: **KL-divergence**

Combiner: **product rule**

$$h \in \mathcal{R}$$

$$\bar{h} = Z^{-1} \prod_t h_t^{\frac{1}{T}}$$

$$D_{KL}(y||\bar{h}) = \frac{1}{T} \sum_t D_{KL}(y||h_t) - \underbrace{\frac{1}{T} \sum_t D_{KL}(\bar{h}||h_t)}_{\text{diversity}}$$

Perspective: Diversity should be natural

Loss function: **0/1 loss**

Combiner: **majority vote**

$$h_t \in \{-1, +1\}$$

$$\bar{h} = \text{sign}\left(\frac{1}{T} \sum_t h_t\right)$$

$$\delta(\bar{h}, t) = \dots\dots\dots + \dots\dots\dots$$

Decomposing Majority Vote Error

$$y, h_t \in \{-1, +1\}$$

Decomposing Majority Vote Error

$$y, h_t \in \{-1, +1\}$$

$$H(\mathbf{x}) = \text{sign}\left(\frac{1}{T} \sum_t h_t(\mathbf{x})\right)$$

Decomposing Majority Vote Error

$$y, h_t \in \{-1, +1\}$$

$$H(\mathbf{x}) = \text{sign}\left(\frac{1}{T} \sum_t h_t(\mathbf{x})\right)$$

$$\delta(H(\mathbf{x}), y) = \frac{1}{2}(1 - yH(\mathbf{x})) \quad \dots \quad 1 \text{ if } H(\mathbf{x}) \neq y, \text{ otherwise } 0.$$

Decomposing Majority Vote Error

$$y, h_t \in \{-1, +1\}$$

$$H(\mathbf{x}) = \text{sign}\left(\frac{1}{T} \sum_t h_t(\mathbf{x})\right)$$

$$\delta(H(\mathbf{x}), y) = \frac{1}{2}(1 - yH(\mathbf{x})) \quad \dots \quad 1 \text{ if } H(\mathbf{x}) \neq y, \text{ otherwise } 0.$$

$$\Delta = \delta(H(\mathbf{x}), y) - \frac{1}{T} \sum_t \delta(h_t(\mathbf{x}), y)$$

Decomposing Majority Vote Error

$$y, h_t \in \{-1, +1\}$$

$$H(\mathbf{x}) = \text{sign}\left(\frac{1}{T} \sum_t h_t(\mathbf{x})\right)$$

$$\delta(H(\mathbf{x}), y) = \frac{1}{2}(1 - yH(\mathbf{x})) \quad \dots \quad 1 \text{ if } H(\mathbf{x}) \neq y, \text{ otherwise } 0.$$

$$\Delta = \delta(H(\mathbf{x}), y) - \frac{1}{T} \sum_t \delta(h_t(\mathbf{x}), y)$$

— Boring proof goes here —

Cut to the Chase

0/1 loss can be re-written:

$$\delta(H, y) = \frac{1}{T} \sum_t \delta(h_t, y) - yH \frac{1}{T} \sum_t \delta(h_t, H)$$

Cut to the Chase

0/1 loss can be re-written:

$$\delta(H, y) = \frac{1}{T} \sum_t \delta(h_t, y) - y_H \frac{1}{T} \sum_t \delta(h_t, H)$$

Integrate over $p(\mathbf{x})$

$$\begin{aligned} \int_{\mathbf{x}} \delta(H, y) &= \int_{\mathbf{x}} \frac{1}{T} \sum_t \delta(h_t, y) - \int_{\mathbf{x}(+)} \frac{1}{T} \sum_t \delta(h_t, H) \\ &\quad + \int_{\mathbf{x}(-)} \frac{1}{T} \sum_t \delta(h_t, H) \end{aligned}$$

$\mathbf{x}(+)$... data subspace where ensemble is correct.

$\mathbf{x}(-)$... data subspace where ensemble is wrong.

Good Diversity, Bad Diversity

$$\dots - \underbrace{\int_{\mathbf{x}(+)} \frac{1}{T} \sum_t \delta(h_t, H)}_{\text{'good' diversity}} + \underbrace{\int_{\mathbf{x}(-)} \frac{1}{T} \sum_t \delta(h_t, H)}_{\text{'bad' diversity}}$$

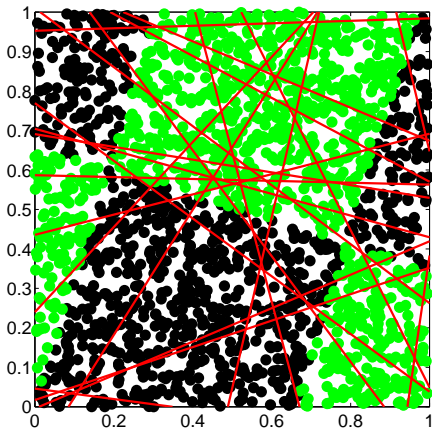
Good diversity subtracts error...

- any disagreement on these \mathbf{x} increases gain of the ensemble relative to the average individual error.

Bad diversity adds error...

- any disagreement on these \mathbf{x} reduces gain of the ensemble relative to the average individual error.

A Toy Dataset



Experimental Protocol

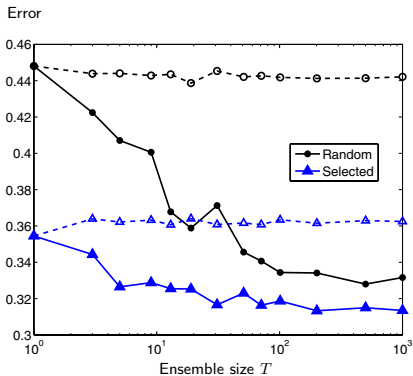
1. Generate pool Φ : 2000 random linear classifiers.
2. The ensemble size T was varied as

$$T \in \{1, 3, 5, 9, 13, 19, 31, 51, 71, 101, 201, 501, 1001\}.$$

3. Generate/test 50 ensembles for each value of T :
 - 3.1 Sample T classifiers without replacement from Φ .
 - 3.2 Generate test set of 1000 2-d points.
 - 3.3 Statistics on ensemble and individual performance stored.
4. The stored values were averaged across the 50 runs.

Results

Majority vote error (solid lines). Average individual error (dashed).

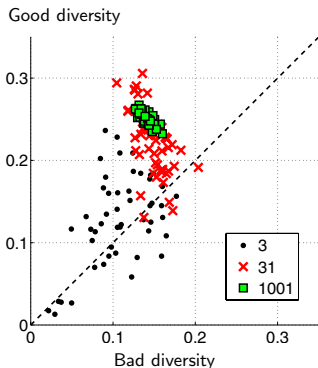


Random : 2000 random classifiers.

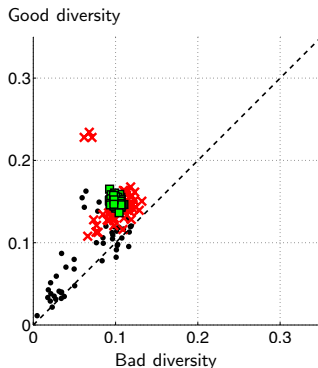
Selected : 2000 best classifiers from 8000 random ones.

Plotting Good vs Bad

(a) RANDOM



(b) SELECTED



Random classifiers (left)

Classifiers selected for good accuracy (right)

Patterns of Success/Failure

Assume T classifiers each with accuracy p .

L. Kuncheva, C. Whitaker, C. Shipp, R. Duin: *Limits on the majority vote accuracy in classifier fusion*. Pattern Analysis & Applications 6(1), 22-31 (2003)

Patterns of Success/Failure

Assume T classifiers each with accuracy p .

Pattern of Success: $\forall \mathbf{x}$, votes cast such that exactly $\frac{T+1}{2}$ correct.

(best gain possible relative to p , no wasted votes!)

L. Kuncheva, C. Whitaker, C. Shipp, R. Duin: *Limits on the majority vote accuracy in classifier fusion*. Pattern Analysis & Applications 6(1), 22-31 (2003)

Patterns of Success/Failure

Assume T classifiers each with accuracy p .

Pattern of Success: $\forall \mathbf{x}$, votes cast such that exactly $\frac{T+1}{2}$ correct.

(best gain possible relative to p , no wasted votes!)

Pattern of Failure: $\forall \mathbf{x}$, votes cast such that exactly $\frac{T-1}{2}$ correct.

(worst gain possible relative to p , just missed it!)

L. Kuncheva, C. Whitaker, C. Shipp, R. Duin: *Limits on the majority vote accuracy in classifier fusion*. Pattern Analysis & Applications 6(1), 22-31 (2003)

Patterns of Success/Failure

Assume T classifiers each with accuracy p .

Pattern of Success: $\forall \mathbf{x}$, votes cast such that exactly $\frac{T+1}{2}$ correct.

(best gain possible relative to p , no wasted votes!)

Pattern of Failure: $\forall \mathbf{x}$, votes cast such that exactly $\frac{T-1}{2}$ correct.

(worst gain possible relative to p , just missed it!)

$$\max \left\{ 0, E_{\text{ind}} - \frac{T-1}{T+1}(1 - E_{\text{ind}}) \right\} \leq \mathbf{E}_{\text{maj}} \leq p \left(1 + \frac{T-1}{T+1} \right)$$

L. Kuncheva, C. Whitaker, C. Shipp, R. Duin: *Limits on the majority vote accuracy in classifier fusion*. Pattern Analysis & Applications 6(1), 22-31 (2003)

Patterns of Success/Failure

$$E_{maj} = E_{ind} - \int_{\mathbf{x}(+)} \frac{1}{T} \sum_t \delta(h_t, H) + \int_{\mathbf{x}(-)} \frac{1}{T} \sum_t \delta(h_t, H)$$

For each point in $\mathbf{x}(+)$, imagine $\frac{T+1}{2}$ classifiers correct (**PoS**)

Patterns of Success/Failure

$$E_{maj} = E_{ind} - \int_{\mathbf{x}(+)} \frac{1}{T} \sum_t \delta(h_t, H) + \int_{\mathbf{x}(-)} \frac{1}{T} \sum_t \delta(h_t, H)$$

For each point in $\mathbf{x}(+)$, imagine $\frac{T+1}{2}$ classifiers correct (**PoS**)

$$E_{maj} = \max \left\{ 0, E_{ind} - \frac{T-1}{T+1} (1 - E_{ind}) \right\}.$$

... recovers *exactly* the lower bound :-)

Patterns of Success/Failure

$$E_{maj} = E_{ind} - \int_{\mathbf{x}(+)} \frac{1}{T} \sum_t \delta(h_t, H) + \int_{\mathbf{x}(-)} \frac{1}{T} \sum_t \delta(h_t, H)$$

For each point in $\mathbf{x}(+)$, imagine $\frac{T+1}{2}$ classifiers correct (**PoS**)

$$E_{maj} = \max \left\{ 0, E_{ind} - \frac{T-1}{T+1} (1 - E_{ind}) \right\}.$$

... recovers *exactly* the lower bound :-)

For each point $\mathbf{x}(-)$, imagine $\frac{T-1}{2}$ classifiers correct (**PoF**)

Patterns of Success/Failure

$$E_{maj} = E_{ind} - \int_{\mathbf{x}(+)} \frac{1}{T} \sum_t \delta(h_t, H) + \int_{\mathbf{x}(-)} \frac{1}{T} \sum_t \delta(h_t, H)$$

For each point in $\mathbf{x}(+)$, imagine $\frac{T+1}{2}$ classifiers correct (**PoS**)

$$E_{maj} = \max \left\{ 0, E_{ind} - \frac{T-1}{T+1} (1 - E_{ind}) \right\}.$$

... recovers *exactly* the lower bound :-)

For each point $\mathbf{x}(-)$, imagine $\frac{T-1}{2}$ classifiers correct (**PoF**)

$$E_{maj} = p \left(1 + \frac{T-1}{T+1} \right).$$

... recovers *exactly* the upper bound :-)

Conclusion

- ▶ voting error can be decomposed.
- ▶ good diversity helps, bad diversity hinders.
- ▶ strong relation to PoS/PoF.
- ▶ large variance when T small, stabilize with large T
- ▶ not a magic solution - how can we make use of it!?

Questions?