Stability, Bagging & Decision Trees

Y. Grandvalet

CNRS, France & IDIAP, Switzerland

NIPS Workshop on Multi-level Inference

Motivations

- Bousquet & Elisseeff (2002): stability relates generalization error to
  - the apparent error (training error)
  - and the leave-one-out error
- Poggio *et al.* (2004): stability characterizes learnability, subsuming Vapnik's empirical risk minimization principle
- Breiman (1994): bagging gains accuracy for unstable methods

Overview

- We have
  - a training sample $\mathcal{T} = \{\mathbf{x}_i, y_i\}_{i=1}^n$
  - a learning algorithm $\mathcal{A} : \mathcal{T} \longrightarrow \widehat{f}$
- Stability ensures non-asymptotic bounds on generalization error

$$R(\widehat{f}) = \mathbb{P}_{XY}[\widehat{f}(X) \neq Y] = \mathbb{E}_{XY}[\mathbb{I}_{\{\widehat{f}(X) \neq Y\}}]$$

- based on the apparent error (empirical risk)

$$R_{\mathrm{emp}}(\widehat{f}) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{\{\widehat{f}(\mathbf{x}_i) \neq y_i\}}$$

- and the leave-one-out error (leave-one-out cross-validation risk)

$$\mathcal{T}^{-i} = \{\mathbf{x}_j, y_j\}_{j \neq i} \ , \quad \widehat{f}^{-i} = \mathcal{A}(\mathcal{T}^{-i}) \ , \quad R_{\mathrm{loo}}(\widehat{f}) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{\{\widehat{f}^{-i}(\mathbf{x}_i) \neq y_i\}}$$

Definition (1)

Pointwise stability

Algorithm $\mathcal{A}$ has pointwise stability β_n w.r.t. the $\{0,1\}$-loss iff

$$\forall i \in \{1,\ldots,n\}, \mathbb{E}_{\mathcal{T}}\left[\left|\mathbb{I}_{\{\widehat{f}(\mathbf{x}_i)\neq y_i\}} - \mathbb{I}_{\{\widehat{f}^{-i}(\mathbf{x}_i)\neq y_i\}}\right|\right] \leq \beta_n$$

Theorem (Bousquet & Elisseeff, 2002)

If algorithm $\mathcal{A}$ has hypothesis stability β_n w.r.t. the $\{0,1\}$-loss, then

$$R(\widehat{f}) \leq R_{\mathrm{emp}}(\widehat{f}) + \sqrt{\frac{1+12n\beta_n}{2n\delta}} \ ,$$

with probability $1-\delta$ over the random draw of the training set $\mathcal{T}$

CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

Definition (2)

Hypothesis stability

Algorithm $\mathcal{A}$ has hypothesis stability β_n w.r.t. the $\{0,1\}$-loss iff

$$\forall i \in \{1,\ldots,n\}, \mathbb{E}_{\mathcal{T},X,Y}\left[\left|\mathbb{I}_{\{\widehat{f}(X)\neq Y\}} - \mathbb{I}_{\{\widehat{f}^{-i}(X)\neq Y\}}\right|\right] \leq \beta_n$$

Theorem (Bousquet & Elisseeff, 2002)

If algorithm $\mathcal{A}$ has hypothesis stability β_n w.r.t. the $\{0,1\}$-loss, then

$$R(\widehat{f}) \leq R_{\mathrm{loo}}(\widehat{f}) + \sqrt{\frac{1+6n\beta_n}{2n\delta}} \enspace ,$$

with probability $1-\delta$ over the random draw of the training set $\mathcal{T}$

CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

IDIAP

Stability is related to robustness

Stability measures the robustness of $\mathcal{A}$ w.r.t. sampling randomness
$\mathcal{A}$ is stable iff there are no leverage examples
In linear regression, the usual leverage statistic is

$$h_i = \frac{\partial \widehat{f}(\mathbf{x}_i)}{\partial y_i} = \frac{\widehat{f}(\mathbf{x}_i) - \widehat{f}^{-i}(\mathbf{x}_i)}{y_i - \widehat{f}^{-i}(\mathbf{x}_i)}$$

In classification, $\widehat{f}$ is a decision function, h_i can transposed as

$$h_i = \mathbb{I}_{\{\widehat{f}(\mathbf{x}_i) \neq \widehat{f}^{-i}(\mathbf{x}_i)\}} \quad ,$$

and a global influence measure is

$$H_i = \mathbb{P}_X[\widehat{f}(X) \neq \widehat{f}^{-i}(X)]$$

For binary classification $H_i = \mathbb{E}_{X,Y}\left[\left|\mathbb{I}_{\{\widehat{f}(X)\neq Y\}} - \mathbb{I}_{\{\widehat{f}^{-i}(X)\neq Y\}}\right|\right]$

Stability is expected influence

What is Bagging?

Breiman (1994): *Bootstrap aggregating*

Ingredients

- a training sample $\mathcal{T} = \{\mathbf{x}_i, y_i\}_{i=1}^n$
- a learning algorithm $\mathcal{A} : \mathcal{T} \longrightarrow \widehat{f}$

Recipe

- draw B bootstrap samples $\{\mathcal{T}^b\}_{b=1}^B$
- apply algorithm $\mathcal{A}$ on each bootstrap sample $\mathcal{T}^b \longrightarrow \widehat{f}^b$
- output $\widehat{f}^{\text{bag}} : \widehat{f}^{\text{bag}}(\mathbf{x}) = \frac{1}{B} \sum_{b=1}^{B} \widehat{f}^b(\mathbf{x}) \underset{B\to\infty}{\longrightarrow} \mathbb{E}_{\widehat{P}}[\widehat{F}^b(\mathbf{x})]$

CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

Rationale: Bias/Variance Decomposition

Let $f^*(\mathbf{x}) = \mathbb{E}[Y|X = \mathbf{x}]$

$$\underbrace{\mathbb{E}_P[(\widehat{F}(\mathbf{x}) - y)^2]}_{\text{Expected error}} = \underbrace{\mathbb{E}_P[(\widehat{F}(\mathbf{x}) - f^*(\mathbf{x}))^2]}_{\text{Reducable error}} + \underbrace{\mathbb{E}_P[(f^*(\mathbf{x}) - y)^2]}_{\text{Intrinsic variability}}$$

$$\underbrace{\mathbb{E}_P[(\widehat{F}(\mathbf{x}) - f^*(\mathbf{x}))^2]}_{\text{Reducable error}} = \underbrace{(\mathbb{E}_P[\widehat{F}(\mathbf{x})] - f^*(\mathbf{x}))^2}_{\text{Bias}} + \underbrace{\mathbb{E}_P[(\widehat{F}(\mathbf{x}) - \mathbb{E}_P[\widehat{F}(\mathbf{x})])^2]}_{\text{Variance}}$$

"backward plug-in principle" not fully motivated!

$$\widehat{f}^{\text{bag}}(\mathbf{x}) = \mathbb{E}_{\widehat{P}}[\widehat{F}^b(\mathbf{x})] \quad \Rightarrow \quad \text{Bias}_P[\widehat{F}^{\text{bag}}(\mathbf{x})] \simeq \text{Bias}_P[\widehat{F}(\mathbf{x})]$$

$$\Rightarrow \quad \text{Var}_P[\widehat{F}^{\text{bag}}(\mathbf{x})] \simeq 0$$

CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

Bagging *vs.* Bias Reduction

Compared to $\widehat{f}$, $\widehat{f}^{\text{bag}}$ is an over-biased estimate:

$$\widehat{f}^{\text{bag}}(\mathbf{x}) = \frac{1}{B}\sum_{b=1}^{B}\widehat{f}^{b}(\mathbf{x}) = \widehat{f}(\mathbf{x}) + \left(\frac{1}{B}\sum_{b=1}^{B}\widehat{f}^{b}(\mathbf{x}) - \widehat{f}(\mathbf{x})\right)$$

$$\widehat{f}^{\text{bag}}(\mathbf{x}) = \widehat{f}(\mathbf{x}) + \widehat{\text{bias}}$$

The down-biased estimate is $2\widehat{f}(\mathbf{x}) - \widehat{f}^{\text{bag}}(\mathbf{x})$

- $\Rightarrow$ $\widehat{f}(\mathbf{x})$ should have little bias, so that the potential decrease in variance is not counterbalanced by an increase in bias
- $\Rightarrow$ In practice, one uses overcomplex predictors that overfit data

CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

Bagging for Classification

Ingredients

- a learning sample $\mathcal{T} = \{\mathbf{x}_i, y_i\}_{i=1}^{n}$,
- an algorithm $\mathcal{A} : \mathcal{T} \longrightarrow \widehat{f}$,

Recipe

- draw B bootstrap samples $\{\mathcal{T}^b\}_{b=1}^{B}$
- apply algorithm $\mathcal{A}$ on each bootstrap sample $\mathcal{T}^b \longrightarrow \widehat{f}^b$
- majority vote $\widehat{f}^{\mathrm{bag}}(\mathbf{x}) = \underset{y \in \Omega}{\mathrm{Argmax}} \sum_{b=1}^{B} \mathbb{I}_{\{\widehat{f}^b(\mathbf{x}) = y\}}$

Many generalization of the bias/variance decomposition

Does Bagging Work?

Shown to be effective for

- Neural networks
- Naive Bayes classifiers
- Stumps
- Decision trees
- SVMs,...

May not rank as the #1 ensemble method, but...
no failure of bagging has ever been reported in classification...

Worth trying to understand why

Past explanations

Widely accepted	
Breiman (1994)	variance reduction by averaging
Margin	
Schapire *et al.* (1997)	margin maximization
Breiman (1997)	bounds too far from observed, even misleading
Bayesian	
Rao & Tibshirani (1997)	approximates draws with an uninformative *prior*
Asymptotics	
Friedman & Hall (2000)	reduce the variance of non-linear components
Buja and Stuetzle (2000)	smoothing effect on variance terms in n^{-2}
Bühlman & Yu (2000)	smoothing at discontinuities

Still no definitive argument

Experimental setup

Goal #1: provide experimental evidence of stabilization

1. Define distributions on (X, Y)
2. Choose “unstable” setups: training trees with “small” samples
3. Draw many training samples to estimate $\mathbb{E}_{\mathcal{T}}$ and stability

Four distributions tested

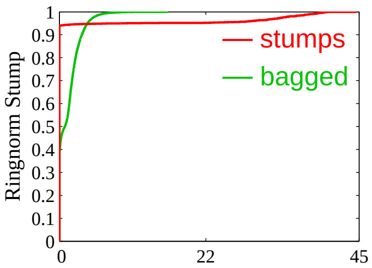
Goal #2: separate stability and variance effects
Two predictors

- Stumps (one node tree): little variance, large bias
- Overcomplex trees (no pruning): large variance, small bias

Bagging’s effects should differ regarding variance, but we hope to observe systematic stabilization

CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

Stabilization – Stumps

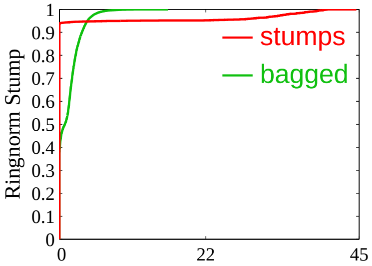


$$100\mathbb{P}_X(\widehat{f}(X) \neq \widehat{f}^{-i}(X)])$$

- More examples have a small influence ⇒ smoothing
- Highest influences reduced ⇒ stability

CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

Stabilization – Stumps

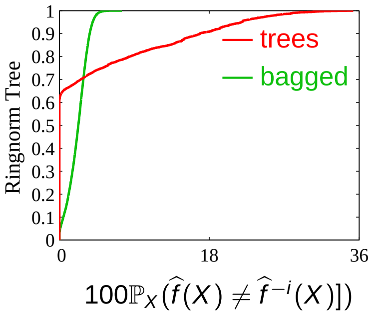


$$100\mathbb{P}_X(\widehat{f}(X) \neq \widehat{f}^{-i}(X)])$$

Variance viewpoint

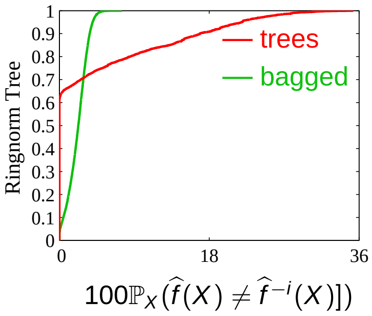
- No effect on $\mathbb{E}_{\mathcal{T}}[R(\widehat{f})]$
- Variability of $R(\widehat{f})$ w.r.t $\mathcal{T}$ increased by bagging

Stabilization – Overfitting trees



- More examples have a small influence $\Rightarrow$ smoothing
- Highest influences reduced $\Rightarrow$ stability

Stabilization – Overfitting trees



Variance viewpoint

- $\mathbb{E}_{\mathcal{T}}[R(\widehat{f})]$ reduced
- Variability of $R(\widehat{f})$ w.r.t $\mathcal{T}$ reduced by bagging

CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

Generalization

		$\widehat{\mathbb{E}}_{\mathcal{T}}[R(\widehat{f})]$	$\widehat{\mathbb{E}}_{\mathcal{T}}[R_{\mathrm{emp}}(\widehat{f})]$	Opt.	$\widehat{\beta}_n$	$1-\delta_0$
Ring.	stump	40.4	34.8	5.6	1.7	0.950
	bagged stump	40.5	35.3	5.2	1.3	0.961
	tree	22.4	5.1	17.4	4.1	0.876
	bagged tree	12.8	0.9	11.8	2.2	0.934
2norm	stump	33.0	26.5	6.5	3.3	0.902
	bagged stump	21.0	14.4	6.6	2.4	0.929
	tree	22.6	4.2	18.4	4.5	0.865
	bagged tree	8.9	0.4	8.6	1.9	0.942
3norm	stump	41.7	35.0	6.7	3.6	0.892
	bagged stump	34.4	25.8	8.6	3.4	0.899
	tree	32.6	6.1	26.5	6.8	0.796
	bagged tree	21.0	0.8	20.2	4.2	0.873
Satim.	tree	14.3	5.3	9.0	0.1	0.998
	bagged tree	11.5	3.6	7.9	0.1	0.997

CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

Conclusion

- Stability characterizes very general learning schemes
- Stability bounds are tight
  - confirms the results of Andonova et al. for subbagging
  - may be used for quantitative prediction
  - needs a practical means to estimate β_n
- Bagging is never detrimental in classification
- Bagging always down-weights the most influential examples
- Stabilization is universally good in classification: no good leverage