Advanced Topics in Ensemble Learning
ECML/PKDD 2012 Tutorial

Daniel Hernández-Lobato[2], Gonzalo Martínez-Muñoz [2], Ioannis Partalas [1]

[1]Equipe Apprentissage: Modèles et Algorithmes
Laboratoire d' Informatique de Grenoble

[2]Computer Science Department,
Universidad Autónoma de Madrid

Outline

Outline

Parallel Ensembles I

General Category of Ensembles Methods:

The ensemble members are built on **independent** realizations of a randomized learning algorithm:

$$h_t(\cdot) \equiv h_t(\cdot|\mathcal{D}, \boldsymbol{\theta}_t) \,.$$

The ensemble output is computed by **majority voting**:

$$H_T(\mathbf{x}) = \arg\max_{c_k} \sum_{t=1}^{T} I(h_t(\mathbf{x}) = c_k), \qquad c_k \in \mathcal{C} = \{c_k\}_{k=1}^{K} \,.$$

Examples: Bagging, Random Forest, Class-switching, Extra-trees, Sub-bagging, Randomizing Outputs, Rotation Forest, Random Subspaces, Randomization, etc.

Parallel Ensembles II

Important **Property**:

When **conditioned to the training data** $\mathcal{D}$, the predictions of two ensemble classifiers for a given test instance **x** are **independent**:

$$\mathcal{P}(h_i(\mathbf{x}) = c', h_j(\mathbf{x}) = c'') = \mathcal{P}(h_i(\mathbf{x}) = c')\mathcal{P}(h_j(\mathbf{x}) = c'') \quad i \neq j,\, c', c'' \in \mathcal{C}\,.$$

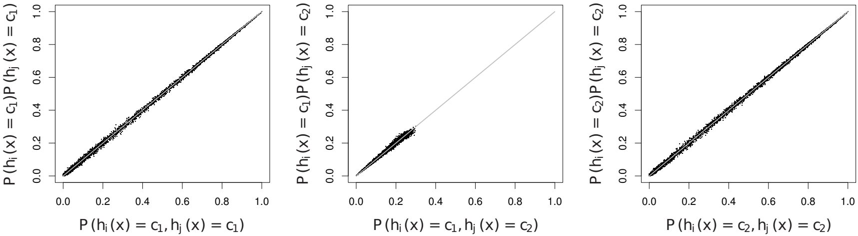
Does not Imply **Independent Prediction Errors** in General:

Both classifiers may **err** in the **same data instances**:

$$\mathbb{E}_{\mathbf{x},y}\left[\mathcal{P}(h_i(\mathbf{x}) \neq y, h_j(\mathbf{x}) \neq y)\right] \neq \mathbb{E}_{\mathbf{x},y}\left[\mathcal{P}(h_i(\mathbf{x}) \neq y)\right]\mathbb{E}_{\mathbf{x},y}\left[\mathcal{P}(h_j(\mathbf{x}) \neq y)\right]\,.$$

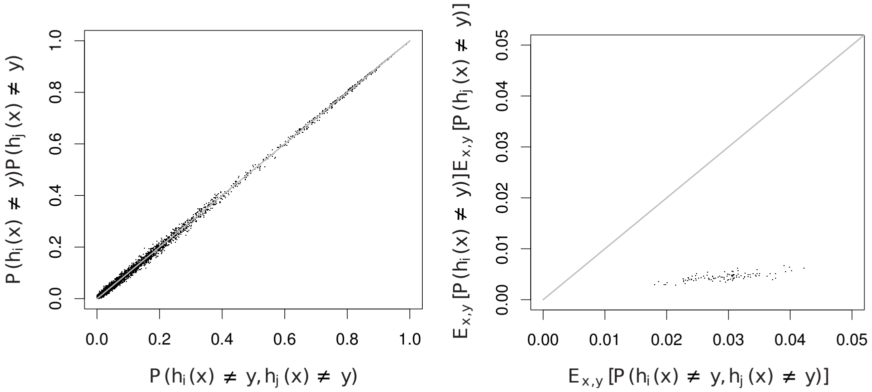
Empirical Validation I

Random Forest: Breast Cancer Dataset.



Empirical Validation II

Random Forest: Breast Cancer Dataset.



Applications

The independence of the predictions for a fixed test instance has **different uses** in parallel ensemble methods:

- Identify instances that are **difficult to classify** by the ensemble.
- Make inference about the prediction of ensembles of **infinite size**.
- Estimate an **adequate size** for the ensemble.

Outline

Ensemble Prediction I

For a fixed instance **x** the predictions of the ensemble members follow a **multinomial** distribution. This distribution is **binomial** when $\mathcal{C} = \{c_1, c_2\}$.

$$\mathcal{P}(\mathbf{T}|\boldsymbol{\pi}(\mathbf{x})) = \frac{T!}{T_1!T_2!}\pi_1(\mathbf{x})^{T_1}\pi_2(\mathbf{x})^{T_2},$$

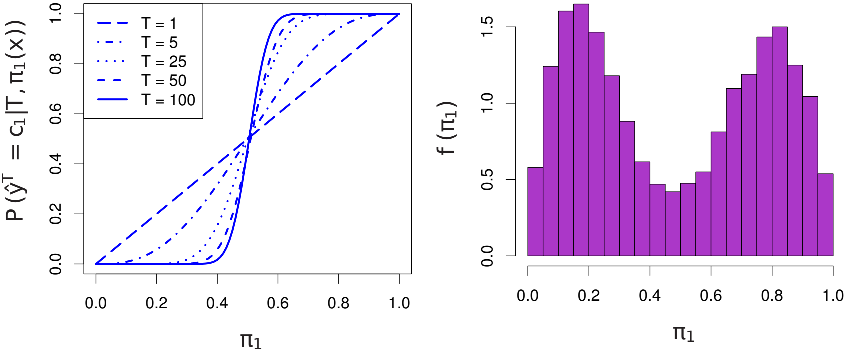
where $\mathbf{T} = (T_1, T_2)$ encodes the predictions for **x** and $\boldsymbol{\pi}(\mathbf{x}) = (\pi_1(\mathbf{x}), \pi_2(\mathbf{x}))$ summarizes the prob. of observing c_1 and c_2, respectively.

The **probability** that the ensemble **assigns** a particular class label is:

$$\mathcal{P}(\hat{y}^T = c_1|T, \mathbf{x}) = \sum_{T_1 > T_2} \mathcal{P}(\mathbf{T}|\boldsymbol{\pi}(\mathbf{x})) = I_{\pi_1(\mathbf{x})}\left(\lfloor\frac{T}{2}\rfloor + 1, T - \lfloor\frac{T}{2}\rfloor\right),$$

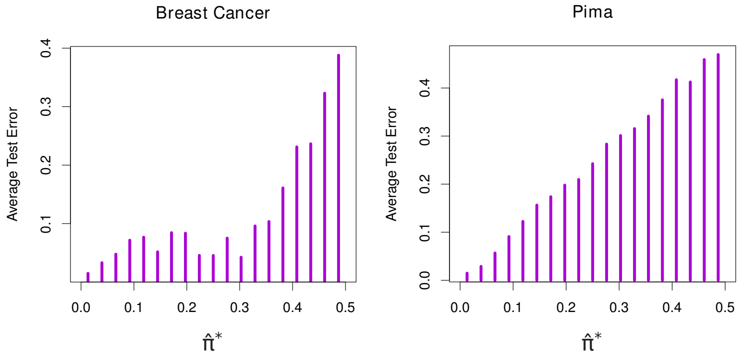
where $I_p(a, b)$ is the regularized incomplete beta function.

Ensemble Prediction II



As the ensemble size increases it is more and more certain the ensemble prediction. The samples of π_1 are obtained using Random Forest and the classification problem is *Twonorm*.

Dependence of the Ensemble Error



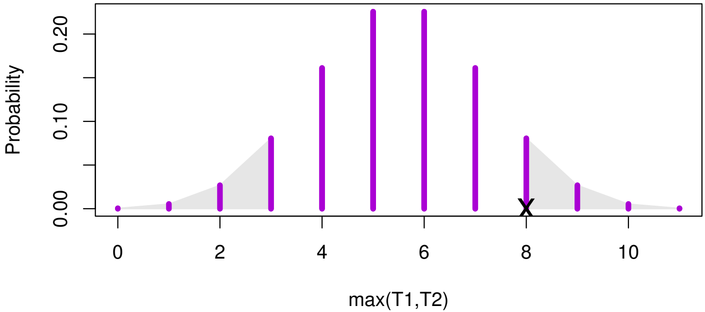
As the estimate $\hat{\pi}^\star = \min(\hat{\pi}_1, \hat{\pi}_2)$ increases, the ensemble error **grows** and approaches $1/2$. The estimates are obtained using Random Forest.

A Statistical Test to Identify Difficult Instances

Motivation:

- For the examples with $\pi_1 = \pi_2 = 1/2$ we know that $\mathcal{P}(\hat{y}^T = c_1 | T, \mathbf{x}) = 1/2$, **independently** of the value of T.
- These instances are **located near the decision boundary** of the problem and are **misclassified** with prob $\approx 50\%$.
- Since **T** follows a binomial distribution, we can use a **binomial test** to evaluate the null hypothesis that $\pi_1 = \pi_2 = 1/2$ and obtain a p-value (Hernández-Lobato *et al.*, 2012).

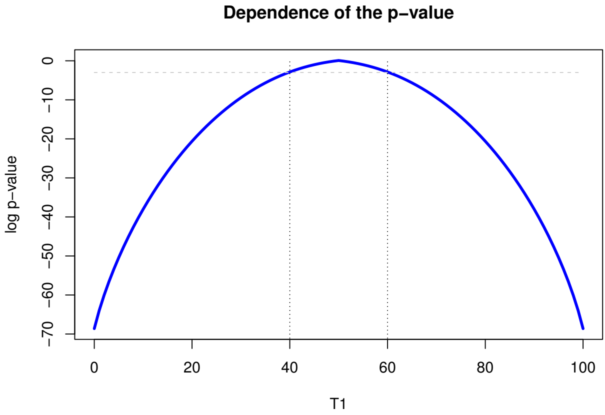
Binomial Test I



The p-value is the prob. of observing under the null-hypothesis a result **at least as unlikely** as the predictions observed $\mathbf{T} = (T_1, T_2)$:

$$\text{p-value} = 2I_{\frac{1}{2}}\left(T - \min(T_1, T_2), 1 + \min(T_1, T_2)\right) .$$

Binomial Test II

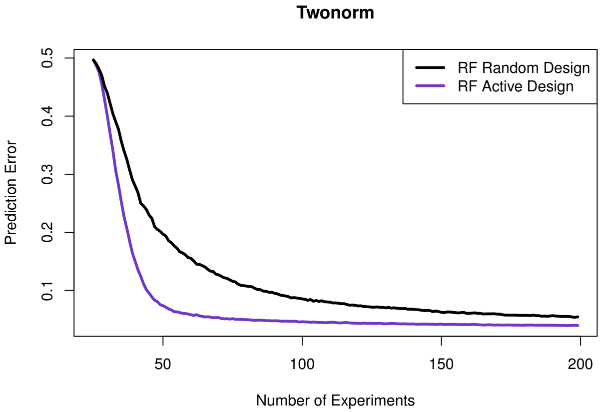


When the p-value is above 5% there is **evidence** that **x** is **difficult** to classify.
For $T = 100$, when $\min(T_1, T_2)$ is between 40 and 50 the p-value exceeds 5%.

Experiments: Results for Random Forest

Dataset	% difficult	Error Difficult	Error Rest	Total Error
Breast Cancer	0.6±0.4	46.6±37.4	2.7±0.7	3.0±0.7
Ionosphere	1.5±1.0	43.4±36.1	6.2±1.5	6.8±1.6
Pima	5.9±1.2	49.8±9.9	22.5±1.7	24.1±1.6
Sonar	9.5±3.0	47.5±16.6	16.9±4.7	19.9±4.5

Sequential Experimental Design using Ensembles



When the design matrix **X** is **sequentially generated** by including the instances that are **most difficult** to classify, RF shows a **steeper decrease** of the generalization error. (Freund *et al.*, 1997) (Abe and Mamitsuka, 1998)

Summary

- The **independence** property of parallel ensembles is used to **analyze** the ensemble prediction.
- A **statistical test** can be used to identify difficult instances.
- On these instances the ensemble error is typically around **50%.**
- The fraction of difficult instances is strongly **problem dependent.**
- Gives a **natural justification** for active learning using ensembles.

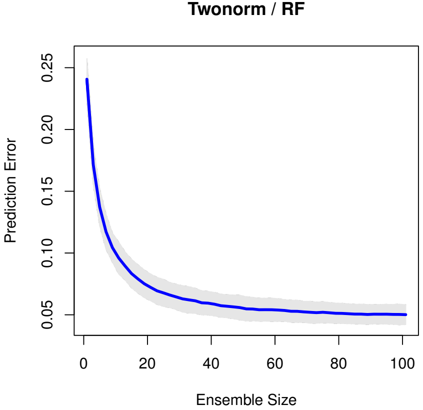
Outline

1. Parallel Ensembles
   - Detection of Instances that are Difficult to Classify
   - Classification in the Infinite Ensemble Limit
   - Optimal Ensemble Size

Classification in the Infinite Ensemble Limit I

Parallel Ensembles:

- The ensemble error **decreases** with the **ensemble size**.
- The **improvements** become progressively **smaller**.
- The **costs** of the ensemble **increase linearly** with the size.



We try to estimate the prediction of an ensemble of **infinite size** based on the predictions of a **finite set of classifiers** (Hernández-Lobato *et al.*, 2011).

Classification in the Infinite Ensemble Limit II

For a fixed instance $\mathbf{x}$ the predictions of the ensemble members follow a **multinomial** distribution:

$$\mathcal{P}(\mathbf{t}|\boldsymbol{\pi}(\mathbf{x})) = \frac{t!}{t_1!t_2!\cdots t_K!}\pi_1(\mathbf{x})^{t_1}\pi_2(\mathbf{x})^{t_2}\cdots\pi_K(\mathbf{x})^{t_K},$$

where $\mathbf{t} = (t_1, t_2, \dots, t_K)$ encodes the predictions for $\mathbf{x}$ and $\boldsymbol{\pi}(\mathbf{x}) = (\pi_1(\mathbf{x}), \pi_2(\mathbf{x}), \cdots, \pi_K(\mathbf{x}))$ summarizes the prob. of observing each class label.

When $t \to \infty$, the **class outputted** by the ensemble is

$$\hat{y}^\infty = c_k \quad \text{with} \quad \arg\max_k \quad \pi_k(\mathbf{x}).$$

Thus, $\boldsymbol{\pi}(\mathbf{x})$ **fully determines** the asymptotic ensemble prediction.

Inference on the Infinite Ensemble Prediction I

After observing $\mathbf{t}$ votes, under the assumption of a uniform prior for $\boldsymbol{\pi}(\mathbf{x})$, Bayes' theorem gives:

$$\mathcal{P}(\boldsymbol{\pi}(\mathbf{x})|\mathbf{t}) = \frac{\Gamma(\sum_{k=1}^{K} t_k + K)}{\prod_{k=1}^{K} \Gamma(t_k + 1)} \pi_1(\mathbf{x})^{t_1} \pi_1(\mathbf{x})^{t_1} \cdots \pi_K(\mathbf{x})^{t_K},$$

i.e. a **Dirichlet distribution** of order K with parameters $t_1 + 1, t_2 + 1, \ldots, t_K + 1$.

We can use this distribution to make inference on the asymptotic ensemble prediction:

$$\mathcal{P}(\hat{y}^{\infty} = c_k|\mathbf{t}) = \mathcal{P}\left(\bigcap_{i \neq k} \pi_k(\mathbf{x}) > \pi_i(\mathbf{x}) \,|\mathbf{t}\right) .$$

Unfortunately, this probability is **difficult to compute** in general.

Inference on the Infinite Ensemble Prediction II

In the binary case, *i.e.* $\mathcal{C} = \{c_1, c_2\}$, this probability is:

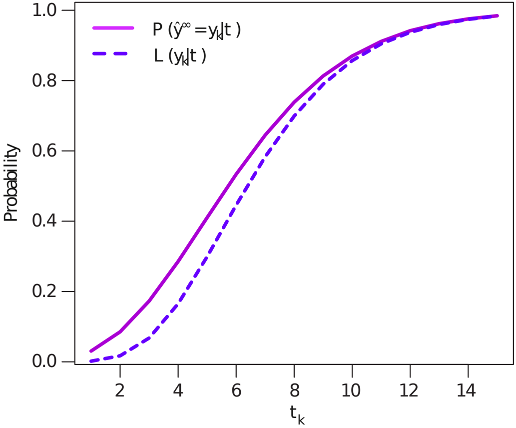
$$\mathcal{P}(\hat{y}^\infty = c_1|\mathbf{t}) = \mathcal{P}(\pi_1(\mathbf{x}) > \pi_2(\mathbf{x})|\mathbf{t}) = I_{\frac{1}{2}}\left(t_2 + 1, t_1 + 1\right) .$$

In the multi-class setting, there is no closed-form expression. We use a **lower bound** which guarantees a conservative estimation:

$$\mathcal{P}(\hat{y}^\infty = c_k|\mathbf{t}) \geq \mathcal{L}(c_k|\mathbf{t}) = \prod_{i \neq k} \mathcal{P}(\pi_k(\mathbf{x}) > \pi_i(\mathbf{x})|\mathbf{t}) = \prod_{i \neq k} I_{\frac{1}{2}}\left(t_i + 1, t_k + 1\right) ,$$

where we have used that $\mathcal{P}(\pi_k(\mathbf{x}) > \pi_i(\mathbf{x})|\pi_k(\mathbf{x}) > \pi_j(\mathbf{x})) \geq \mathcal{P}(\pi_k(\mathbf{x}) > \pi_i(\mathbf{x}))$.
In binary classification problems $\mathcal{L}(c_k|\mathbf{t})$ gives the **exact** result.

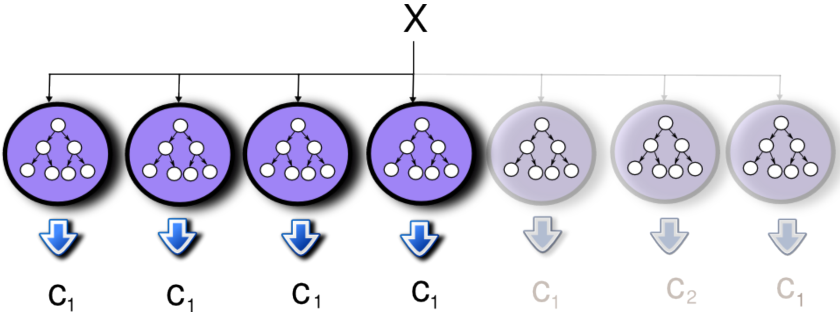
Lower Bound vs The Exact Post. Probability



When $\mathcal{P}(\hat{y}^{\infty} = c_1|\mathbf{t})$ is **large** the lower bound becomes more and more **accurate**. $\{t_j : j \neq i\} = \{5, 3, 2, 1\}$.

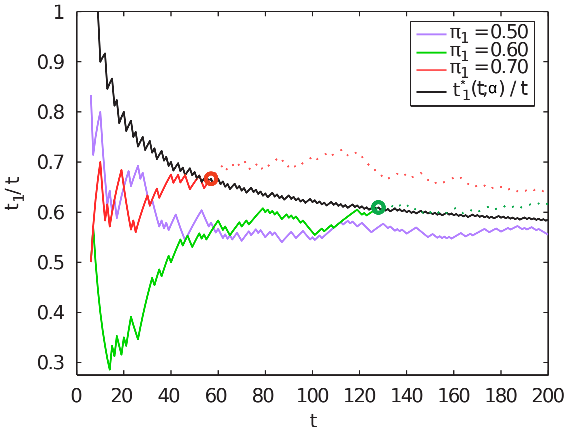
Dynamic Pruning Criterion I

When $\mathcal{L}(c_k|\mathbf{t}) > \alpha$ we stop the querying process for instance **x**.



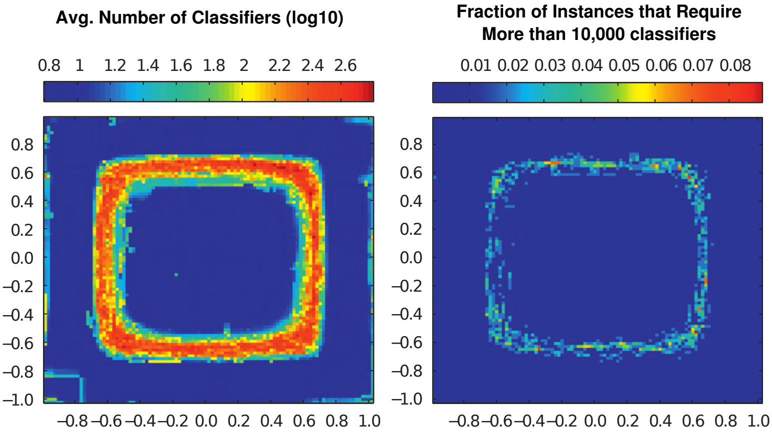
The ensemble prediction should **coincide** with the asymptotic prediction with at least prob. α. The differences in prediction error should be **below** $1 - \alpha$. The values of $\mathcal{L}(c_k|\mathbf{t})$ can be **precomputed**.

Dynamic Pruning Criterion II



We consider a **binary problem** with $\alpha = 99\%$. $t_1^\star(t;\alpha)$ represents the **minimum** number of prediction for class c_1 to stop, for a fixed t.

Dynamic Pruning Criterion III



RF ensembles. Most instances require querying a **small number** of classifiers. Others a potentially **infinite number** of classifiers.

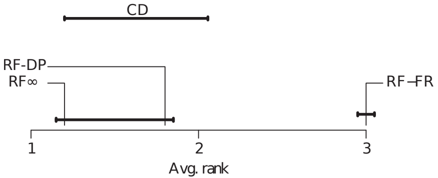
Experiments: Random Forest (size 101 $\alpha = 99\%$) I

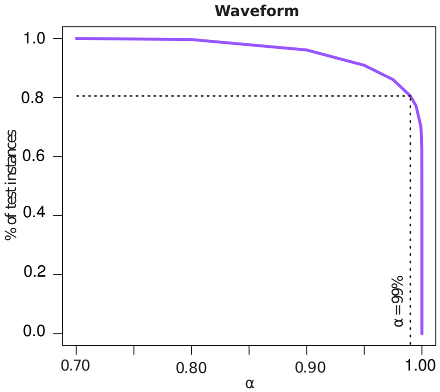
Problem	% of test instances	# Trees RF-DP	Classification Error in % RF-FS	RF-DP	RF∞
breast	98.0±0.8	8.1±0.5	3.1±1.0	2.8±0.9	2.7±0.9
glass*	78.9±4.8	22.9±2.6	18.1±5.5	17.6±5.4	17.6±5.4
heart	86.1±3.1	18.6±2.5	14.4±3.6	13.7±3.6	13.6±3.5
led*	82.7±4.9	23.9±2.9	22.5±2.1	22.2±2.0	22.2±2.0
liver	75.1±4.2	26.7±2.7	24.5±4.2	23.8±4.2	23.4±4.1
new-thyroid*	96.1±2.5	10.6±1.2	3.3±2.2	2.9±2.1	2.9±2.1
pima	83.6±2.3	20.0±1.5	20.6±2.2	20.3±2.4	20.2±2.4
ringnorm	88.3±1.3	20.1±1.3	4.3±1.0	3.3±0.9	3.2±0.9
spam	96.5±0.4	10.3±0.3	4.2±0.5	3.7±0.5	3.7±0.5
threenorm	73.8±1.8	27.6±1.2	11.4±1.5	10.6±1.4	10.3±1.4
twonorm	90.2±1.0	18.7±0.7	2.9±0.5	1.9±0.5	1.7±0.5
vehicle*	77.5±2.7	22.0±1.3	16.4±2.6	16.2±2.6	16.2±2.6
vowel*	86.2±2.1	25.8±1.1	2.7±1.0	2.0±1.0	2.0±1.0
waveform*	80.5±1.7	23.8±1.1	11.9±1.3	11.4±1.2	11.3±1.3
wine*	97.1±2.0	12.2±1.5	1.9±1.7	1.3±1.4	1.3±1.4

Experiments: Random Forest (size 101 $\alpha = 99\%$) II

Problem	**% of disagreement**	
	RF-FS	**RF-DP**
breast	0.7±0.5	0.1±0.2
glass*	1.2±1.5	0.3±0.6
heart	2.1±1.6	0.6±0.9
led*	1.0±1.3	0.2±0.6
liver	2.6±1.8	1.2±1.4
new-thyroid*	1.0±1.3	0.1±0.3
pima	2.1±1.0	0.7±0.5
ringnorm	1.8±0.5	0.5±0.3
spam	1.0±0.3	0.1±0.1
threenorm	2.4±0.6	1.3±0.5
twonorm	1.5±0.4	0.4±0.2
vehicle*	1.8±0.9	0.5±0.5
vowel*	0.9±0.6	0.1±0.2
waveform*	1.8±0.5	0.7±0.3
wine*	0.9±1.3	0.1±0.3

Experiments: Random Forest (size 101 $\alpha = 99\%$) III





The differences with respect to RF-FS are statistically **significant** (Demšar, 2006). Only **small benefits** are obtained by allowing a **lower** confidence level α on the estimates.

Summary

- We have shown how to make Bayesian **inference** about the infinite ensemble prediction.
- We have derived expressions for the probability that the current majority class **coincides** with the asymptotic ensemble prediction.
- Computing this probability is **costly** for multi-class problems and we use an approximation based on a **lower bound**.
- A large fraction of the instances require on average a **small** number of classifiers to get enough evidence on the asymptotic ensemble prediction.
- For some instances it is not possible to get **enough evidence** even after querying a very **large number** of classifiers.

Outline

Motivation

- The error of the ensemble **asymptotically decreases** with its size T.
- How to choose the value of T?
  - If T is too **large** we waste computational resources.
  - If T is too **small** we loose prediction accuracy.

We consider a practical solution:

Stop including classifiers in the ensemble when it is **unlikely** that adding extra classifiers will **change** the ensemble prediction (Hernández-Lobato, 2009).

- The dynamic pruning methods **relay** on receiving an adequate ensemble.
- They **cannot** be used to estimate an adequate ensemble size.

An Adequate Ensemble Size for a Fixed Instance I

If $\mathcal{C} = \{c_1, c_2\}$, given **x** we can compute the **probability** that an ensemble of size T gives the **asymptotic** ensemble prediction:

$$\mathcal{P}(\hat{y}^T = \hat{y}^\infty | T, \mathbf{x}) = I_{\max(\pi_1(\mathbf{x}), 1-\pi_1(\mathbf{x}))}\left(\lfloor\frac{T}{2}\rfloor + 1, T - \lfloor\frac{T}{2}\rfloor\right),$$

We can define $T^\star(\alpha, \mathbf{x})$ as the **minimum ensemble size** whose prediction for **x coincides** with $\hat{y}^\infty$ with **at least** probability α.

$$\alpha \leq I_{\max(\pi_1(\mathbf{x}), 1-\pi_1(\mathbf{x}))}\left(\lfloor\frac{T}{2}\rfloor + 1, T - \lfloor\frac{T}{2}\rfloor\right)$$

Unfortunately there is no **closed form expression** for $T^\star(\alpha, \mathbf{x})$.

An Adequate Ensemble Size for a Fixed Instance II

For large T we can compute an **accurate** Gaussian approximation:

$$\mathcal{P}(\hat{y}^T = \hat{y}^\infty | T, \mathbf{x}) \approx \Phi\left(\frac{T \max\{\pi_1(\mathbf{x}), 1-\pi_1(\mathbf{x})\}}{\sqrt{T\pi_1(\mathbf{x})(1-\pi_1(\mathbf{x}))}}\right),$$

where $\Phi(\cdot)$ is the c.p.f. of a standard Gaussian distribution.

Given α, we can now **find** $T^\star(\alpha, \mathbf{x})$:

$$T^\star(\alpha, \mathbf{x}) \approx \frac{\Phi^{-1}(\alpha)^2 \pi_1(\mathbf{x})(1-\pi_1(\mathbf{x}))}{(\pi_1(\mathbf{x}) - 1/2)^2}$$

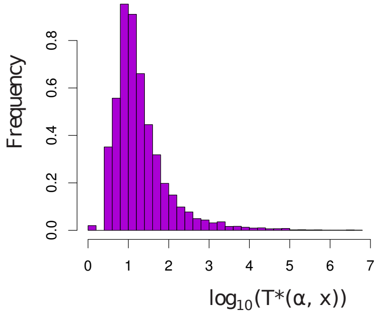
For any $\alpha > 50\%$, if $\pi_1(\mathbf{x}) \to 1/2$, then $T^\star(\alpha, \mathbf{x}) \to \infty$.

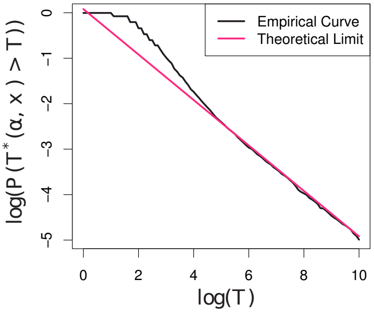
An Adequate Ensemble Size for a Fixed Instance III

If we consider $\pi_1(\mathbf{x})$ as a **random variable**, $T^\star(\alpha,\mathbf{x})$ is also random:

$$\mathcal{P}(T^\star(\alpha,\mathbf{x}) > T) \approx \frac{f(\pi_1(\mathbf{x}) = 1/2)\Phi^{-1}(\alpha)}{\sqrt{T}}, \quad \text{for large } T.$$

$\mathcal{P}(T^\star(\alpha,\mathbf{x}) > T)$ has **universal** heavy-tailed behavior. Only **depends** on the classification problem by $f(\pi_1(\mathbf{x}) = 1/2)$.





An Adequate Ensemble Size in General

We estimate the prob. that $\hat{y}^T$ and $\hat{y}^\infty$ **agree** in general:

$$\mathcal{P}(\hat{y}^T = \hat{y}^\infty) \approx 1 - \frac{f(\pi_1(\mathbf{x}) = 1/2) \int_{-\infty}^{0} \Phi(z) dz}{\sqrt{T}} \quad \text{with } T \to \infty \, .$$

Solving for T we find the **size** $T^\star(\alpha)$ **of the ensemble** that agrees with the infinite ensemble with probability $\alpha = \mathcal{P}(\hat{y}^T = \hat{y}^\infty)$ close to one:

$$T^\star(\alpha) \approx \left(\frac{f(\pi_1(\mathbf{x}) = 1/2) \int_{-\infty}^{0} \Phi(z) dz}{1 - \alpha} \right)^2 .$$

Only **depends** on the classification problem by $f(\pi_1(\mathbf{x}) = 1/2)$.

When $\alpha \to 1$, $T^\star(\alpha) \to \infty$, as expected.

Practical Implementation

Given α, $T^\star(\alpha)$ is obtained as the minimum T such that:

$$\alpha \leq \frac{1}{N}\sum_{i=1}^{N} I_{\max(\hat{\pi}_1^{(i)}(\mathbf{x}),1-\hat{\pi}_1^{(i)}(\mathbf{x}))}\left(\lfloor\frac{T}{2}\rfloor + 1, T - \lfloor\frac{T}{2}\rfloor\right),$$

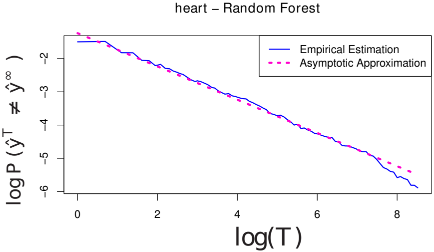
where $\{(\hat{\pi}_1^{(i)}(\mathbf{x})\}_{i=1}^{N}$ are **estimated** using OOB, validation or un-labeled test data using an **initial ensemble** of $T' = 100$ classifiers.

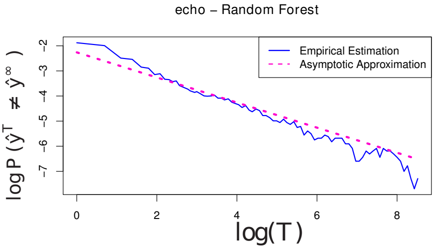
If $T^\star(\alpha) > T'$, we set $T' = \min(T^\star(\alpha), 2T')$ and **repeat**.

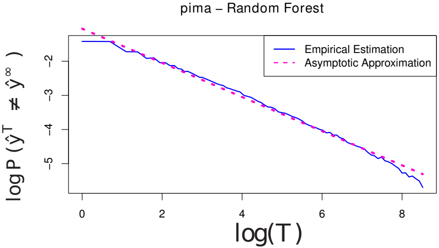
When $T^\star(\alpha) \leq T'$ we stop and return an ensemble of size $T^\star(\alpha)$.

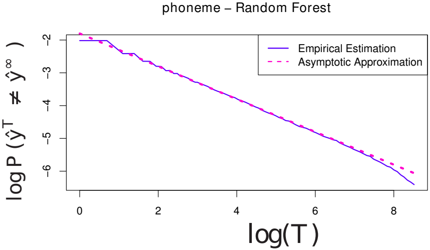
Empirical evaluation: 25 problems from the UCI repository. The infinite ensemble is **approx.** by an ensemble of size $10,000$. We use **RF** and **bagging** and compare results with the method suggested in (Banfield *et al.*, 2007).

Universal Behavior Verification









Average Disagreement Rates

Problem	RF-Test	RF-OOB	RF-BAN	Bag-Test	Bag-OOB	Bag-BAN
abalone	1.0±0.2	1.0±0.3	2.2±0.7	1.1±0.2	1.0±0.3	2.1±0.5
australian	1.0±0.6	1.2±0.7	2.3±1.1	1.0±0.6	1.1±0.7	2.3±1.3
breast	0.9±0.6	1.0±0.7	0.6±0.5	0.9±0.5	0.9±0.7	0.8±0.6
echo	1.0±1.5	1.1±1.8	2.2±2.4	1.2±1.5	1.1±2.0	2.0±2.6
german	1.1±0.5	1.2±0.6	5.1±1.5	1.1±0.6	1.2±0.6	5.7±2.1
heart	1.2±1.1	1.3±1.2	4.7±3.1	1.3±1.0	1.2±1.1	4.9±3.4
hepatitis	1.5±1.4	1.5±1.8	4.7±3.4	1.3±1.5	1.2±1.8	5.2±3.6
horse	1.2±1.0	1.1±1.1	2.4±1.7	1.1±0.8	1.2±1.1	2.6±2.0
ionosphere	0.9±0.8	1.0±0.8	1.5±1.2	0.9±0.8	1.1±1.0	1.8±1.5
labor	1.8±2.8	1.9±2.9	3.5±4.9	1.4±2.6	1.7±3.7	3.2±4.2
liver	1.5±1.1	1.5±1.2	8.5±3.5	1.3±1.0	1.2±0.9	7.6±4.0
magic	1.0±0.1	1.0±0.1	1.4±0.3	1.0±0.1	1.0±0.1	1.4±0.3
musk	0.9±0.2	0.8±0.2	0.4±0.1	1.0±0.2	0.9±0.2	0.4±0.2
phoneme	1.0±0.2	1.0±0.2	1.7±0.5	1.0±0.2	1.0±0.3	1.6±0.4
pima	1.1±0.7	1.0±0.7	5.2±2.1	1.3±0.6	1.2±0.7	5.2±2.2
ringnorm	1.1±0.3	1.2±0.5	2.8±0.7	1.1±0.3	1.2±0.4	3.3±1.2
spam	1.0±0.3	0.9±0.3	0.8±0.3	1.0±0.3	1.0±0.3	0.8±0.3
sonar	1.4±1.2	1.9±1.7	8.1±3.9	1.3±1.4	1.4±1.6	7.0±4.1
tic-tac-toe	0.9±0.5	0.8±0.5	1.3±0.8	0.9±0.5	0.8±0.6	0.6±0.5
votes	0.8±0.8	0.8±0.9	0.7±0.9	1.1±0.8	1.0±1.0	1.0±0.8
whitewine	1.0±0.3	1.0±0.3	2.6±0.7	1.1±0.2	1.0±0.3	2.6±0.6

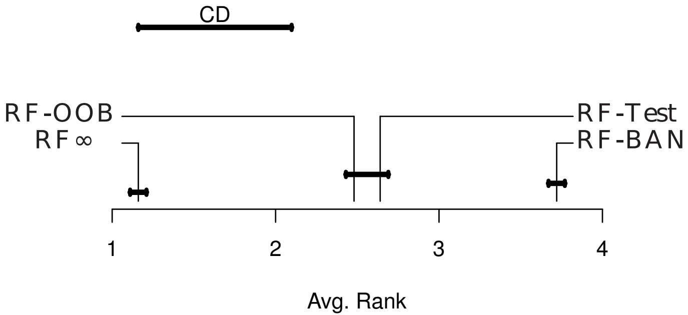
Median of the Ensemble Size

Problem	# Tree RF-Test	# Tree RF-OOB	# Tree RF-BAN
abalone	391 (318, 474)	397 (363, 442)	92 (66, 120)
australian	257 (192, 427)	238 (189, 318)	58 (43, 78)
breast	19 (15, 34)	23 (17, 28)	57 (36, 76)
echo	57 (24, 131)	88 (62, 117)	35 (18, 46)
german	1570 (1216, 2280)	1616 (1422, 2130)	78 (54, 102)
heart	529 (320, 1079)	618 (404, 1088)	47 (32, 74)
hepatitis	313 (178, 767)	532 (288, 768)	30 (20, 61)
horse	191 (126, 350)	241 (164, 368)	73 (49, 110)
ionosphere	66 (39, 100)	71 (53, 96)	41 (29, 61)
labor	64 (37, 117)	78 (53, 175)	21 (14, 37)
liver	2224 (1312, 4062)	2440 (1526, 3631)	54 (33, 81)
magic	247 (226, 276)	257 (243, 270)	144 (109, 175)
musk	17 (15, 19)	17 (17, 19)	84 (66, 107)
phoneme	246 (206, 287)	267 (233, 297)	96 (76, 122)
pima	1194 (798, 1904)	1258 (1000, 1598)	56 (36, 89)
ringnorm	563 (429, 703)	443 (346, 638)	83 (64, 111)
sonar	1975 (954, 3877)	2070 (1198, 3146)	58 (37, 85)
spam	63 (53, 72)	64 (58, 73)	90 (70, 114)
tic-tac-toe	143 (97, 195)	185 (148, 216)	116 (86, 141)
votes	20 (13, 36)	29 (19, 41)	44 (30, 61)
whitewine	714 (570, 842)	716 (644, 788)	100 (78, 127)

Average Test Error

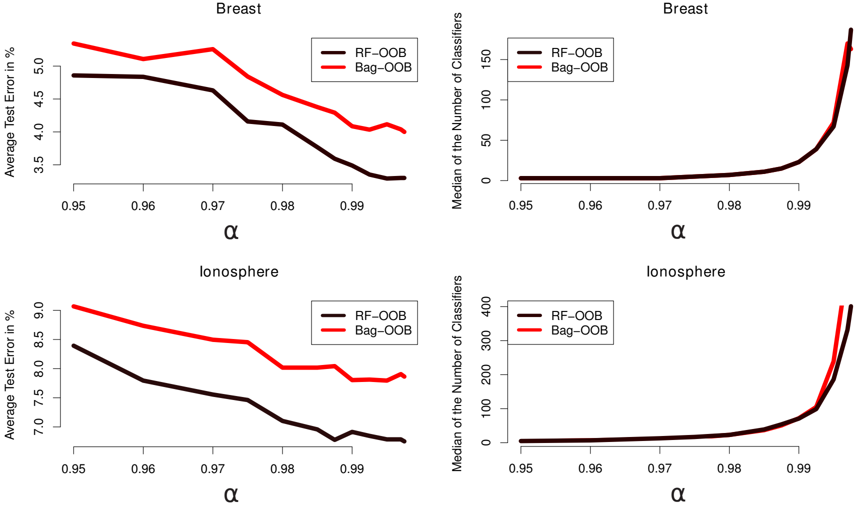
Problem	RF∞	RF-Test	RF-OOB	RF-BAN
abalone	16.67±0.68	16.72±0.69	**16.73±0.71**	**16.88±0.73**
australian	13.13±1.90	13.08±2.02	13.20±2.06	13.24±1.89
breast	3.20±0.89	**3.55±1.00**	**3.57±1.02**	**3.40±0.94**
echo	9.16±3.41	**9.59±3.50**	9.20±3.53	9.52±3.50
german	24.16±1.77	24.21±1.65	24.19±1.74	**24.45±1.92**
heart	17.20±3.42	17.10±3.35	17.22±3.40	**17.90±3.63**
hepatitis	15.44±4.68	15.63±4.53	15.27±4.56	15.73±5.07
horse	14.07±2.83	14.26±2.90	14.22±2.90	**14.67±2.99**
ionosphere	6.72±1.97	6.78±1.93	**6.95±2.03**	**7.26±2.16**
labor	8.42±5.39	**9.53±5.43**	8.74±5.90	**9.89±7.42**
liver	28.16±4.05	28.17±3.86	28.37±3.98	**29.37±4.23**
magic	12.07±0.35	**12.14±0.34**	**12.13±0.33**	**12.18±0.36**
musk	2.46±0.32	**2.78±0.36**	**2.72±0.34**	**2.51±0.31**
phoneme	9.60±0.72	9.63±0.70	9.63±0.69	**9.77±0.66**
pima	24.05±2.10	24.07±2.06	24.05±2.00	**24.41±2.28**
ringnorm	6.17±1.14	**6.29±1.09**	**6.26±1.17**	**6.86±1.15**
sonar	18.30±5.16	18.36±5.28	18.41±5.44	**19.38±5.05**
spam	5.00±0.56	**5.08±0.61**	5.03±0.53	**5.09±0.53**
tic-tac-toe	2.01±0.85	**2.37±0.88**	**2.23±0.93**	**2.49±0.98**
votes	3.82±1.52	**4.01±1.52**	**4.04±1.52**	3.93±1.55
whitewine	16.93±0.87	**17.01±0.88**	16.97±0.91	**17.12±0.86**

Average Ranks



Similar results are obtained in **bagging**. However, in this case the differences between BG-Test and BG-BAN are not significant (Demšar, 2006).

Dependence on the Confidence Level α



Summary

- Determining an adequate size for the ensemble requires balancing **accuracy** and **efficiency**.
- We estimate the **ensemble size** by requiring that the finite and the infinite ensemble predictions **coincide** with probability α.
- The ensemble size is strongly **problem dependent.**
- The **fraction** of instances whose predicted class-label differs from the asymptotic prediction is **proportional** to $T^{-1/2}$.
- The ensemble size is **fully determined** by $f(\pi_1(\mathbf{x}) = 1/2)$.
- The method is **general** and valid for **any** classification problem and **any** parallel ensemble.

References

- Hernández-Lobato, D.; Martínez-Muñoz, G. & Suárez, A. On the Independence of the Individual Predictions in Parallel Randomized Ensembles, ESANN 2012.
- Hernández-Lobato, D.; Martńez-Muñoz, G. & Suárez, A. Inference on the prediction of ensembles of infinite size, Pattern Recognition, 2011, 44, 1426-1434.
- Hernández-Lobato, D. Prediction Based on Averages over Automatically Induced Learners: Ensemble Methods and Bayesian Techniques Computer Science Department, Universidad Autónoma de Madrid, PhD Thesis, 2009.
- Demšar, J. Statistical Comparisons of Classifiers over Multiple Data Sets, Journal of Machine Learning Research, MIT Press, 2006, 7, 1-30.
- Freund, Y.; Seung, H. S.; Shamir, E. & Tishby, N. Selective Sampling Using the Query by Committee Algorithm, Machine Learning, Kluwer Academic Publishers, 1997, 28, 133-168.
- Abe, N. & Mamitsuka, H. Query Learning Strategies Using Boosting and Bagging, ICML 1998.
- Banfield, R. E.; Hall, L. O.; Bowyer, K. W. & Kegelmeyer, W. P., A Comparison of Decision Tree Ensemble Creation Techniques, IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE Computer Society, 2007, 29, 173-180.