Forêts aléatoires : aspects théoriques, sélection de variables et applications

Robin Genuer

24 Novembre 2010

Soutenance de thèse, sous la direction de Jean-Michel Poggi

Laboratoire de mathématiques d'Orsay

Plan

Forêts aléatoires

- introduites par Breiman (2001)
- famille des méthodes d'ensemble, Dietterich (1999,2000)
- algorithme d'apprentissage statistique très performant, à la fois pour des problèmes de classification et de régression.

Forêts aléatoires

- introduites par Breiman (2001)
- famille des méthodes d'ensemble, Dietterich (1999,2000)
- algorithme d'apprentissage statistique très performant, à la fois pour des problèmes de classification et de régression.

Notations :

$\mathcal{L}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ v.a. i.i.d. de même loi que (X, Y).

$X \in \mathbb{R}^p$ (variables)
$Y \in \mathcal{Y}$ (réponse)

- $\mathcal{Y} = \mathbb{R}$: régression
- $\mathcal{Y} = \{1, \ldots, L\}$: classification

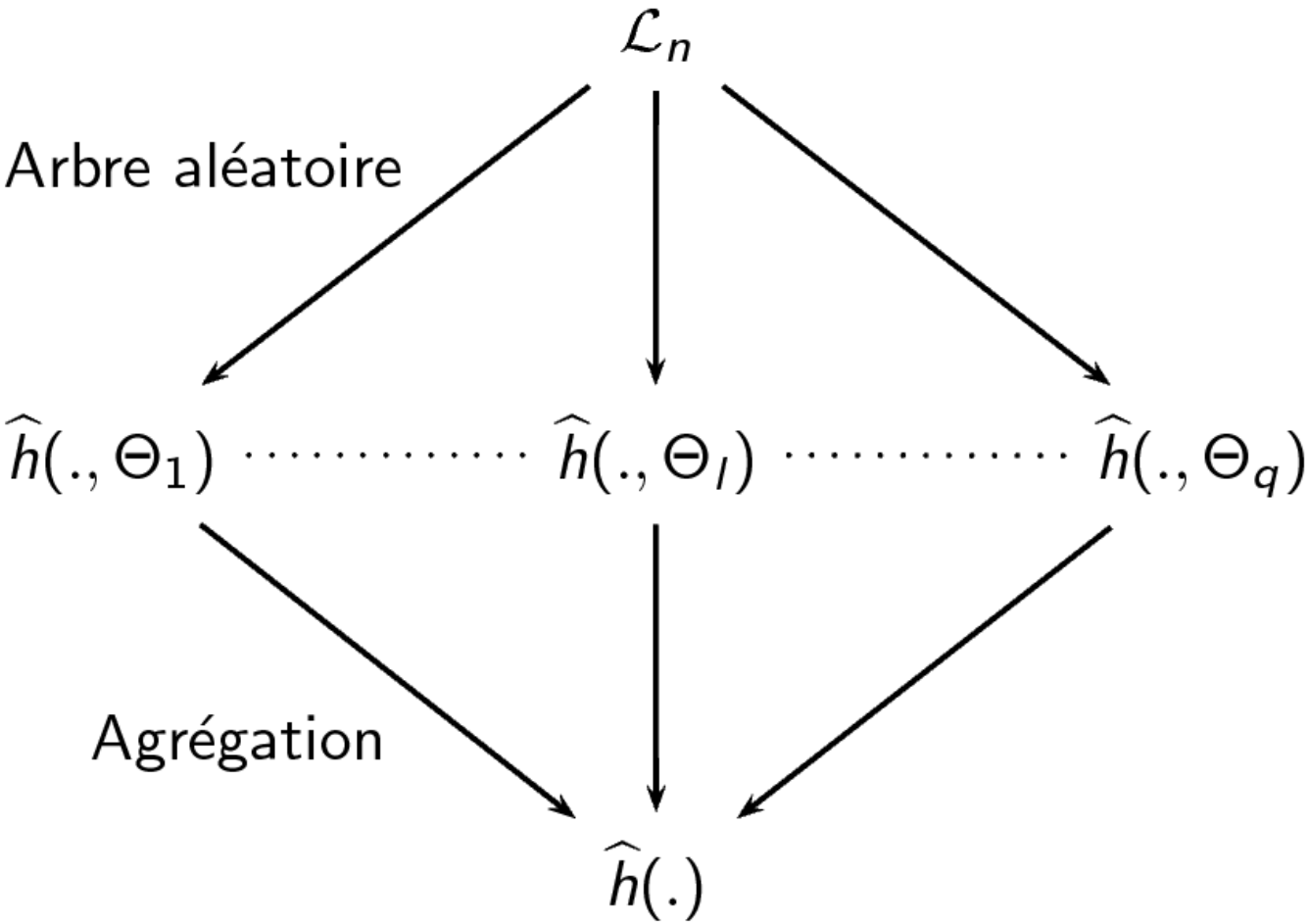
But : construire un prédicteur $\widehat{h} : \mathbb{R}^p \to \mathcal{Y}$

Définition : Forêts aléatoires (Breiman 2001)

$\left\{\widehat{h}(.,\Theta_\ell), 1 \leq \ell \leq q\right\}$ collection de prédicteurs par arbre,
$(\Theta_\ell)_{1\leq\ell\leq q}$ v.a. i.i.d. indépendantes de $\mathcal{L}_n$.
Prédicteur des forêts aléatoires $\widehat{h}$ obtenu en agrégeant la collection d'arbres.

Agrégation :

- $\widehat{h}(x) = \dfrac{1}{q}\sum_{\ell=1}^{q}\widehat{h}(x,\Theta_\ell)$ en régression
- $\widehat{h}(x) = \underset{1\leq c\leq L}{\operatorname{argmax}} \sum_{\ell=1}^{q}\mathbb{1}_{\widehat{h}(x,\Theta_\ell)=c}$ en classification


$\mathcal{L}_n$
Arbre aléatoire
$\widehat{h}(.,\Theta_1)$
$\widehat{h}(.,\Theta_l)$
$\widehat{h}(.,\Theta_q)$
Agrégation
$\widehat{h}(.)$

Arbre : prédicteur constant par morceaux, obtenu par partitionnement récursif dyadique de $\mathbb{R}^p$.

Restriction : coupures parallèles aux axes.

Classiquement, à chaque étape du partitionnement, on cherche à séparer "au mieux" les données de $\mathcal{L}_n$.

Exemple : CART, Breiman et.al. (1984).

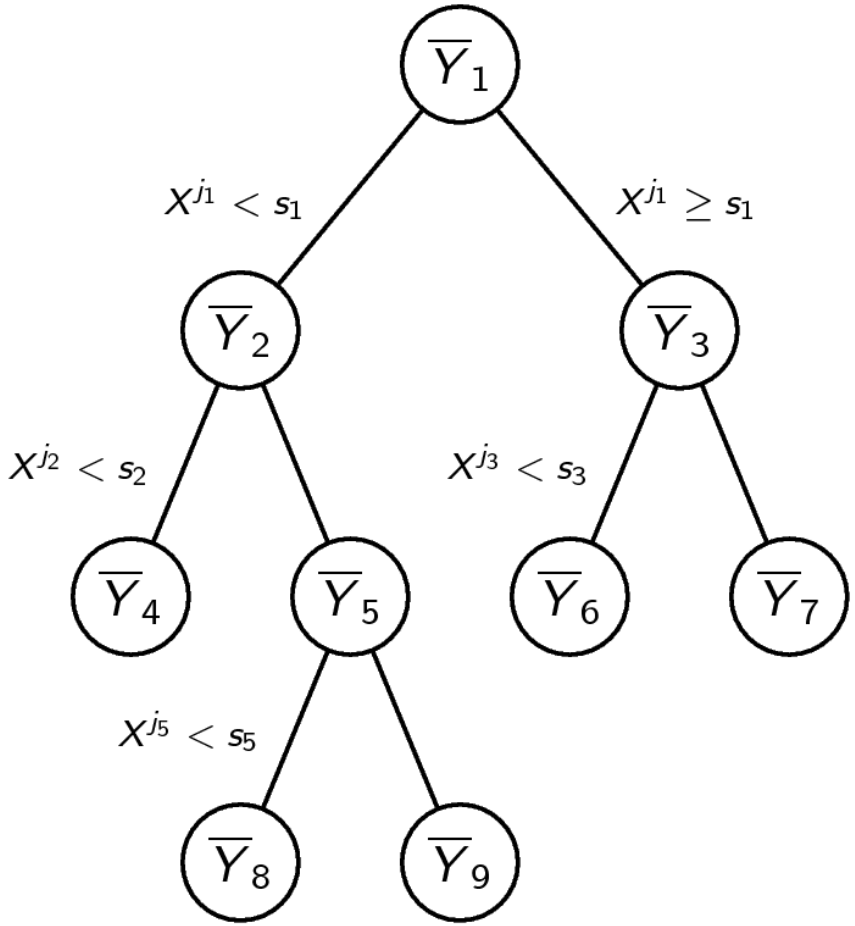


FIGURE: Arbre de régression

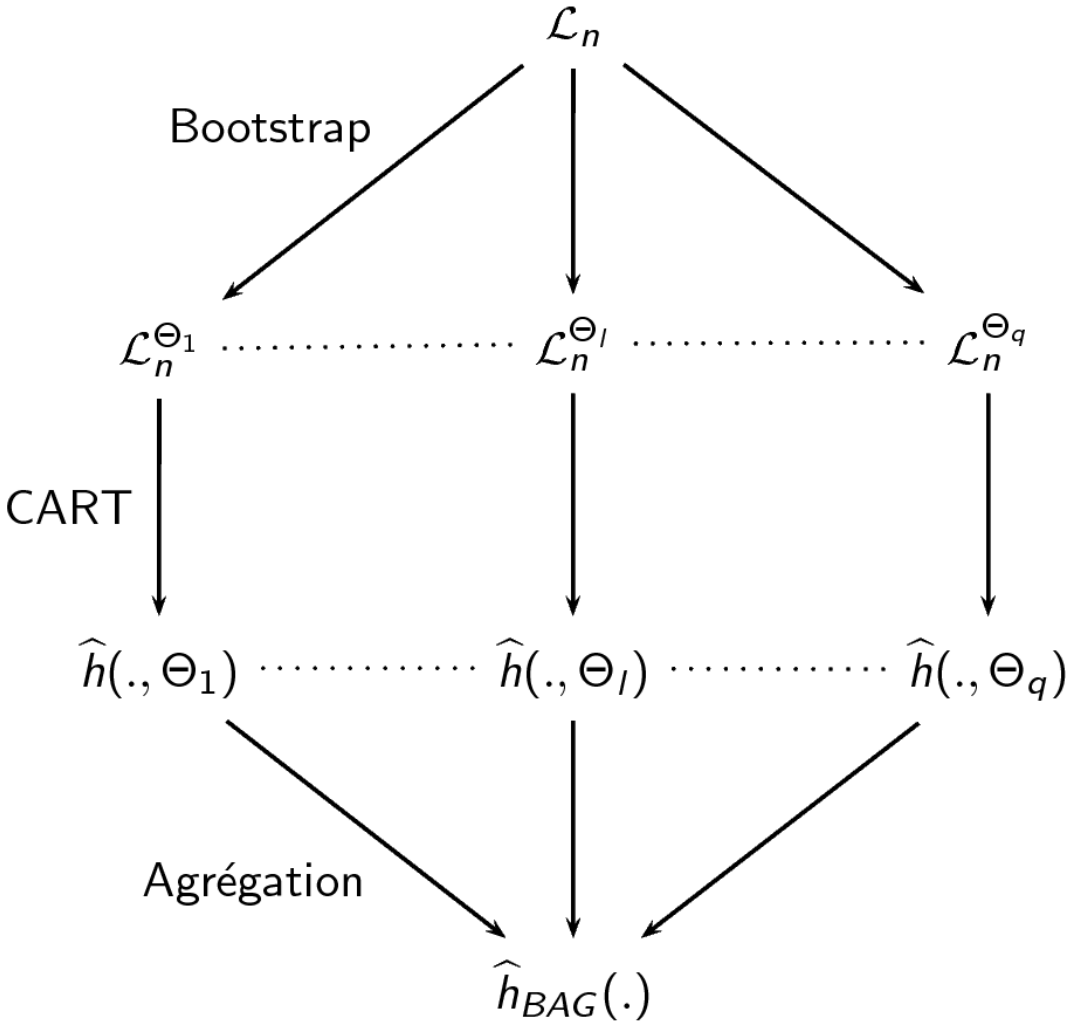
Exemples d'aléas supplémentaires :

- rééchantillonnage préalable à la construction de l'arbre,
- choix aléatoire de la variable de coupure à chaque noeud,
- choix aléatoire du point de coupure à chaque noeud.

Deux grandes familles de forêts aléatoires :

- Classiques : partition optimisée sur les données d'apprentissage $\mathcal{L}_n$.
- Purement aléatoires : partition tirée aléatoirement, indépendamment de $\mathcal{L}_n$.

Bagging (Breiman 1996)



Instabilité de CART $\Rightarrow$ amélioration des performances

Random Forests-Random Inputs (Breiman 2001)

Définition : Arbre RI

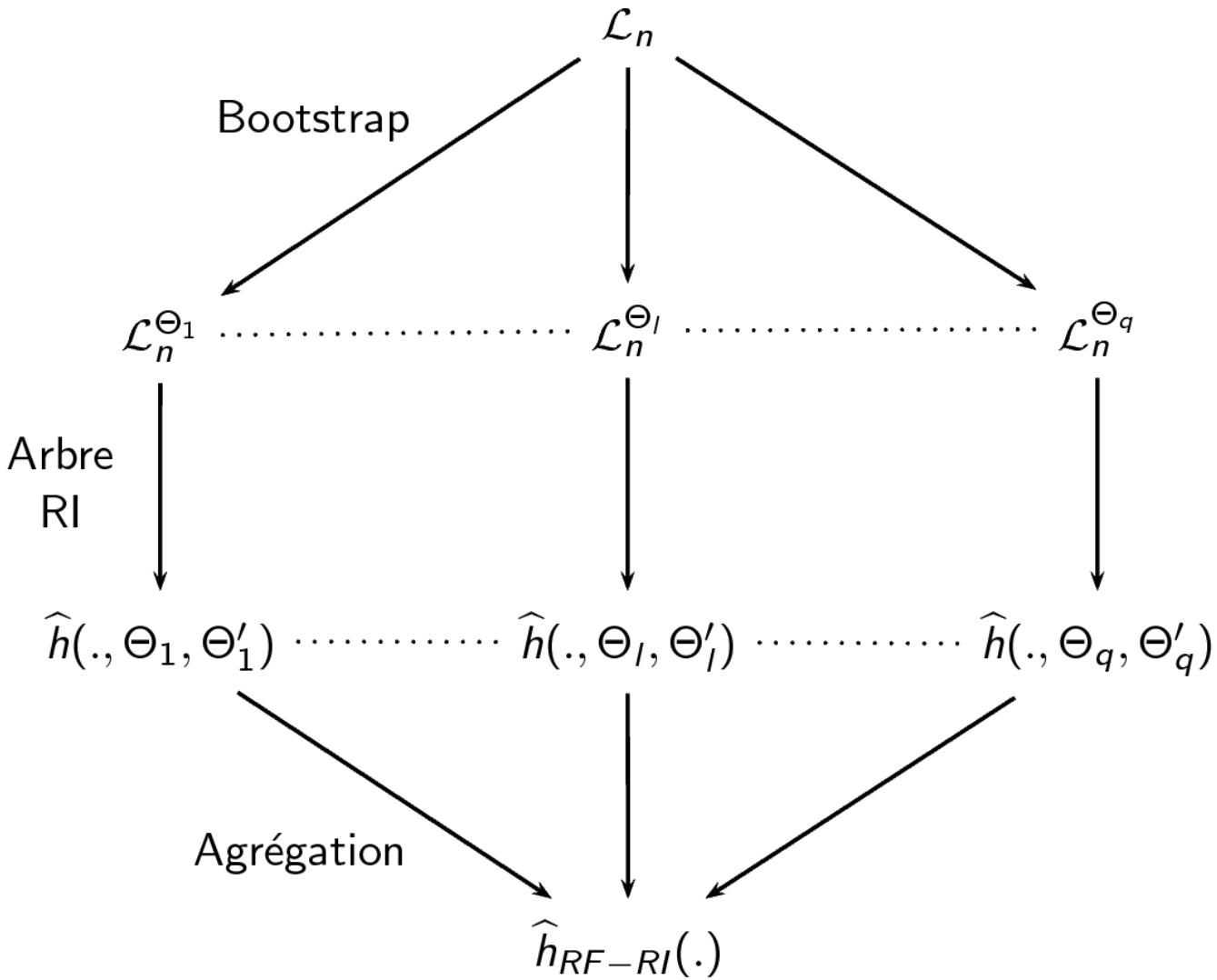
Un arbre RI consiste à tirer aléatoirement, à chaque noeud `mtry` variables, puis à chercher la meilleure coupure uniquement parmi les variables sélectionnées.

`mtry` est le même pour tous les noeuds de tous les arbres de la forêt.

Définition : Random Forests-RI

Une forêt Random Forests-RI est obtenue en effectuant du Bagging avec des arbres RI.

Random Forests-RI



Aléa supplémentaire $\Rightarrow$ amélioration des performances

Random Forests-RI

Paquet R `randomForest` :

- basé sur le code de Breiman, Cutler (2000)
- décrit dans Liaw, Wiener (2002)

Principaux paramètres de l'algorithme `randomForest` :

- `ntree` : nombre d'arbres dans la forêt
- `mtry` : le nombre de variables tirées aléatoirement à chaque noeud

Estimation de l'erreur de prédiction

OOB = Out Of Bag ($\approx$ "En dehors du Bootstrap")

Erreur OOB

Pour prédire X_i, on agrège uniquement les prédicteurs $\widehat{h}(.,\Theta_\ell)$ construits sur des échantillons bootstrap ne contenant pas (X_i, Y_i).

$\Rightarrow$ Erreur OOB :

- $\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2$ en régression
- $\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{Y_i \neq \widehat{Y}_i}$ en classification

Importance des variables

Breiman (2001), Strobl *et al.* (2007,2008), Ishwaran (2007), Archer *et al.* (2008).

Importance des variables

Soit $j \in \{1, \ldots, p\}$. Pour chaque échantillon OOB, on permute aléatoirement les valeurs de la j-ième variable des données.

Importance de la j-ième variable = augmentation moyenne de l'erreur d'un arbre après permutation.

Plus l'augmentation d'erreur est forte,
plus la variable est importante.

Genuer, Poggi, Tuleau (2010)

Deux objectifs différents de sélection de variables :

1. sélectionner toutes les variables importantes, même si elles sont redondantes, dans un but d'interprétation
2. trouver un ensemble parcimonieux de variables importantes suffisant pour la prédiction

Notre but est de proposer une procédure automatique qui atteint ces deux objectifs.

Travaux antérieurs :

- Díaz-Uriarte, Alvarez de Andrés (2006)
- Ben Ishak, Ghattas (2008)

Deux applications :

- Toussile, Genuer, Morlais (2010) (soumis)
- Genuer, Michel, Eger, Thirion (2010)

Deux applications :

- Toussile, Genuer, Morlais (2010) (soumis)
- Genuer, Michel, Eger, Thirion (2010)

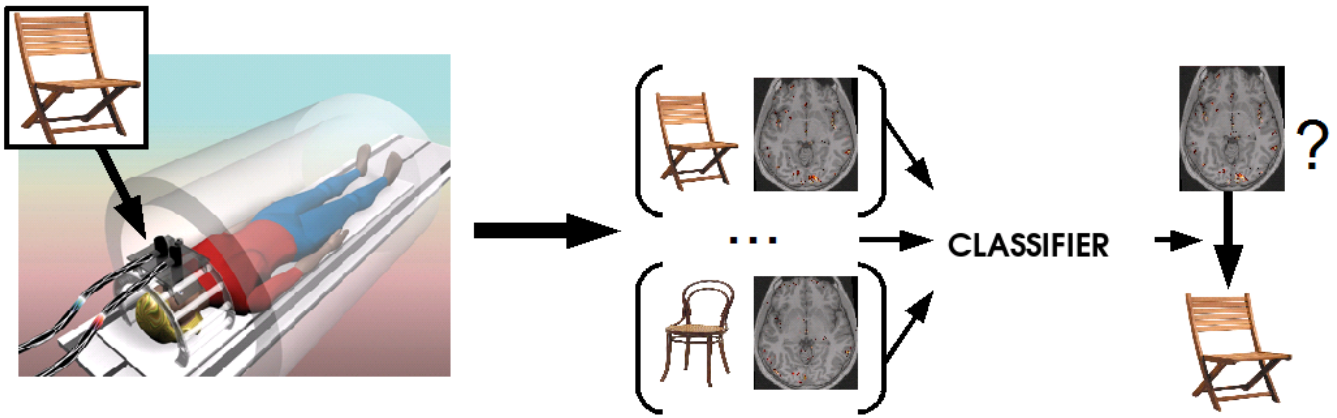


FIGURE: Expérience

12 sujets : 100 000 voxels, 72 observations.
Etape préliminaire : réduction à 1000 parcelles par un algorithme de Ward.

Classification $n = 72$ $p = 1000$ $L = 4$

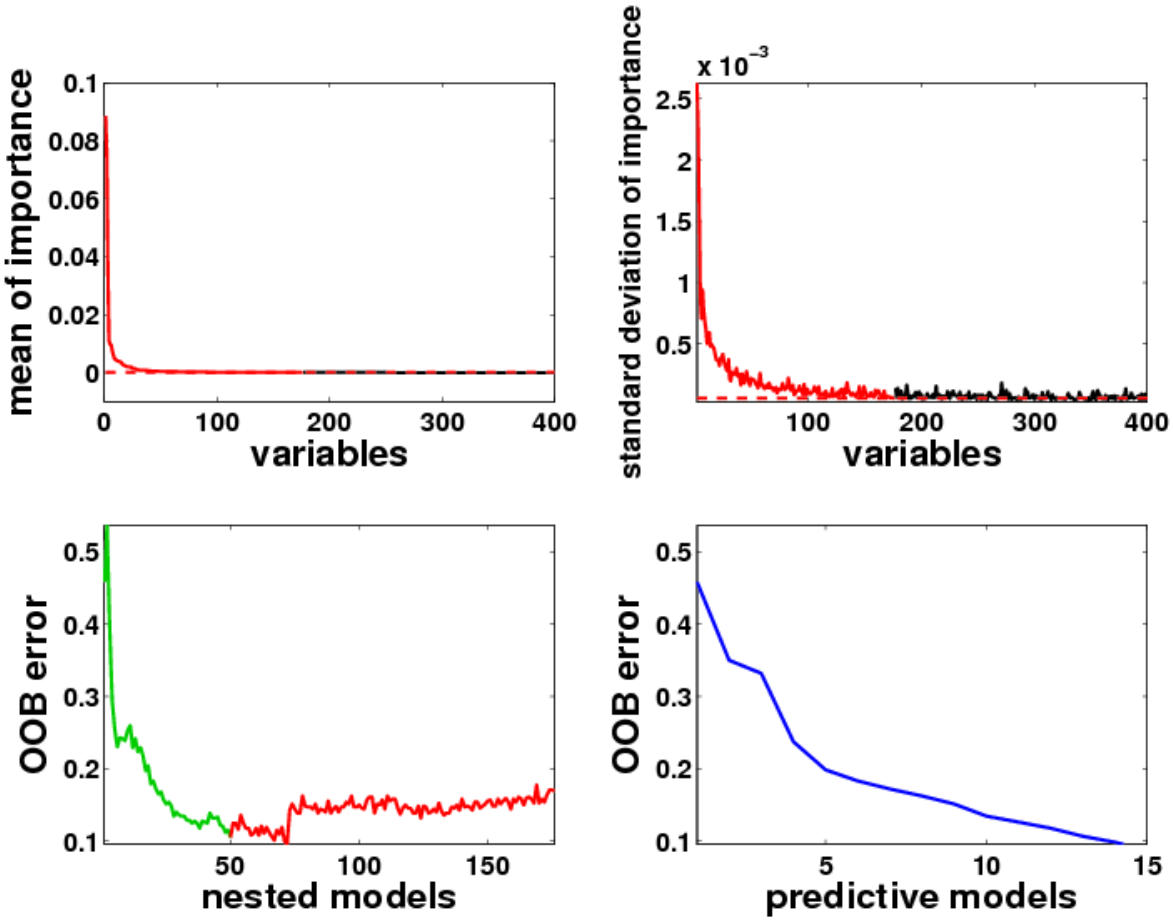


FIGURE: Procédure de sélection de variables pour un sujet ($ntree = 2000$, $mtry = p/3$)
Elimination : 176 variables, Interprétation : 50 variables, Prédiction : 15 variables

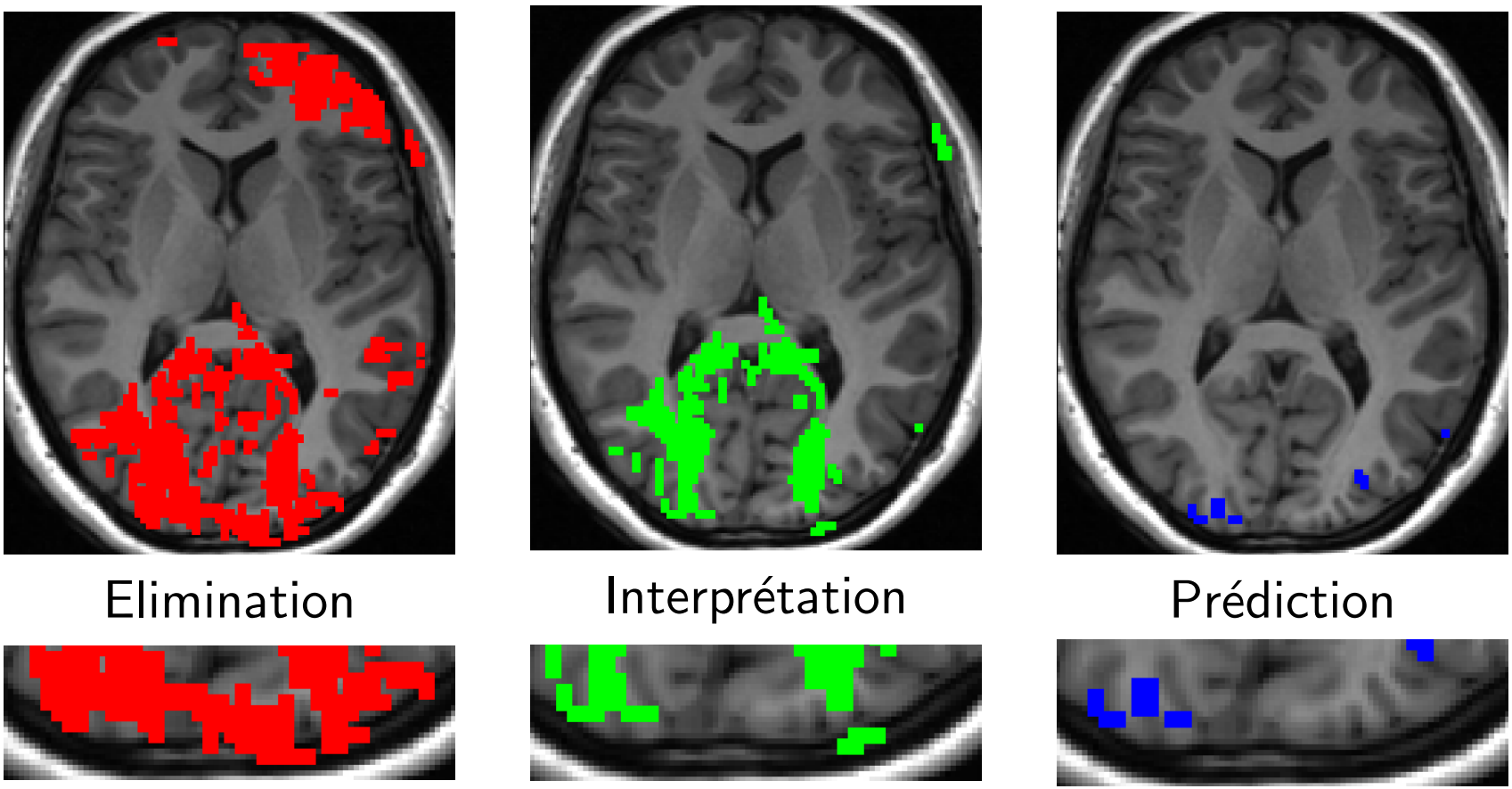


FIGURE: Variables sélectionnnées aux différentes étapes de la procédure

	Initiale	Elim.	Interp.	Préd.	Référence
Erreur	34 %	29 %	27 %	30 %	31 %
Nombre var.	1000	146	23	8	350

FIGURE: Résultats sur les 12 sujets de l'étude

- Méthode de référence : SVM linéaire (F-test + validation croisée)
- Taux d'erreurs comparables
- Beaucoup moins de variables

Biau, Devroye, Lugosi (2008) : résultats de consistance pour une variante de forêts purement aléatoires dans $\mathbb{R}^p$.

Biau, Devroye, Lugosi (2008) : résultats de consistance pour une variante de forêts purement aléatoires dans $\mathbb{R}^p$.

Genuer (2010) (soumis)

- $X \in [0,1]$ de densité marginale μ ;
- $Y \in \mathbb{R}$;
- $Y_i = s(X_i) + \varepsilon_i$ où $(\varepsilon_1, \ldots, \varepsilon_n)$ v.a. i.i.d. $\sim \varepsilon$, indépendantes de $(X_1, \ldots, X_n)$, et :
  - $\mathbb{E}[\varepsilon] = 0$,
  - $\mathrm{Var}(\varepsilon) = \sigma^2$

Arbre

Soit $\mathbb{U} = (U_1, \ldots, U_k)$ v.a. i.i.d. de loi $\mathcal{U}([0,1])$.

Arbre, associé à $\mathbb{U}$, $x \in [0,1]$:

$$\hat{s}_{\mathbb{U}}(x) = \sum_{j=0}^{k} \hat{\beta}_j \mathbb{1}_{U_{(j)} < x \leq U_{(j+1)}}$$

où

$$\hat{\beta}_j = \frac{1}{\sharp\{i : \ U_{(j)} < X_i \leq U_{(j+1)}\}} \sum_{i:\, U_{(j)} < X_i \leq U_{(j+1)}} Y_i$$

$(U_{(1)}, \ldots, U_{(k)})$ statistique d'ordre de $(U_1, \ldots, U_k)$,
$U_{(0)} = 0$, $U_{(k+1)} = 1$.

Arbre

Arbre idéal, $x \in [0,1]$:

$$\tilde{s}_{\mathbb{U}}(x) = \sum_{j=0}^{k} \beta_j \mathbb{1}_{U_{(j)} < x \leq U_{(j+1)}}$$

où

$$\beta_j = \mathbb{E}[Y | U_{(j)} < X \leq U_{(j+1)}] \, .$$

Décomposition :

$$\begin{array}{ccccc} \mathbb{E}[(\hat{s}_{\mathbb{U}}(X) - s(X))^2] & = & \mathbb{E}[(\hat{s}_{\mathbb{U}}(X) - \tilde{s}_{\mathbb{U}}(X))^2] & + & \mathbb{E}[(\tilde{s}_{\mathbb{U}}(X) - s(X))^2] \\ \text{risque} & = & \text{terme de variance} & + & \text{terme de biais} \end{array}$$

Variance

Proposition (Arlot 2008)

$$\mathbb{E}[(\hat{s}_{\mathbb{U}}(X) - \tilde{s}_{\mathbb{U}}(X))^2 | \mathbb{U}] = \frac{1}{n} \sum_{j=0}^{k} (1 + \delta_{n,p_j})[\sigma^2 + \sigma_j^2]$$

où

- $p_j = \mathrm{P}(U_{(j)} < X \leq U_{(j+1)})$,
- $\delta_{n,p} \xrightarrow[np \to +\infty]{} 0$,
- $\sigma_j^2 = \mathbb{E}[(s(X) - \tilde{s}_{\mathbb{U}}(X))^2 | U_{(j)} < X \leq U_{(j+1)}]$.

Proposition (Variance d'un arbre)

Si $k \xrightarrow[n\to+\infty]{} +\infty$, $\frac{k}{n} \xrightarrow[n\to+\infty]{} 0$, $\mu > 0$ et s C-Lipschitzienne :

$$\mathbb{E}[(\hat{s}_{\mathbb{U}}(X) - \tilde{s}_{\mathbb{U}}(X))^2] = \frac{\sigma^2(k+1)}{n} + \underset{n\to+\infty}{\mathrm{o}}\left(\frac{k}{n}\right)$$

Proposition (Variance d'un arbre)

Si $k \xrightarrow[n\to+\infty]{} +\infty$, $\frac{k}{n} \xrightarrow[n\to+\infty]{} 0$, $\mu > 0$ et s C-Lipschitzienne :

$$\mathbb{E}[(\hat{s}_{\mathbb{U}}(X) - \tilde{s}_{\mathbb{U}}(X))^2] = \frac{\sigma^2(k+1)}{n} + \underset{n\to+\infty}{\mathrm{o}}\left(\frac{k}{n}\right)$$

Proposition (Biais d'un arbre)

Si $\mu \leq M$ et s C-Lipschitzienne :

$$\mathbb{E}[(\tilde{s}_{\mathbb{U}}(X) - s(X))^2] \leq \frac{6MC^2}{(k+1)^2} \cdot$$

Corollaire (Borne de risque pour un arbre)

$$\mathbb{E}[(\hat{s}_{\mathbb{U}}(X) - s(X))^2] \leq \frac{\sigma^2(k+1)}{n} + \frac{6MC^2}{(k+1)^2} + \underset{n\to+\infty}{\mathrm{o}}\left(\frac{k}{n}\right) .$$

$(k+1) = n^{1/3}$ donne une majoration en :

$$K n^{-2/3} + \underset{n\to+\infty}{\mathrm{o}}(n^{-2/3})$$

vitesse minimax

Forêt

Soit $\mathbb{V} = (\mathbb{U}^1, \ldots, \mathbb{U}^q)$ v.a i.i.d. $\sim \mathbb{U}$.

Forêt, associée à $\mathbb{V}$, $x \in [0,1]$:

$$\hat{s}(x) = \frac{1}{q} \sum_{\ell=1}^{q} \hat{s}_{\mathbb{U}^\ell}(x) \, .$$

Forêt idéale, $x \in [0,1]$:

$$\tilde{s}(x) = \frac{1}{q} \sum_{\ell=1}^{q} \tilde{s}_{\mathbb{U}^\ell}(x) \, .$$

Variance

$$\mathbb{E}[(\hat{s}(X) - \tilde{s}(X))^2] = \mathbb{E}\left[\left(\frac{1}{q}\sum_{\ell=1}^{q}(\hat{s}_{\mathbb{U}^\ell}(X) - \tilde{s}_{\mathbb{U}^\ell}(X))\right)^2\right]$$

$$= \frac{1}{q^2}\sum_{\ell=1}^{q}\mathbb{E}[(\hat{s}_{\mathbb{U}^\ell}(X) - \tilde{s}_{\mathbb{U}^\ell}(X))^2]$$

$$+ \frac{1}{q^2}\sum_{1\leq \ell,m\leq q\,:\,\ell\neq q}\mathbb{E}[(\hat{s}_{\mathbb{U}^\ell}(X) - \tilde{s}_{\mathbb{U}^\ell}(X))(\hat{s}_{\mathbb{U}^m}(X) - \tilde{s}_{\mathbb{U}^m}(X))]$$

Si $q \xrightarrow[n\to+\infty]{} +\infty$:

$$\mathbb{E}[(\hat{s}(X)-\tilde{s}(X))^2] = \mathbb{E}[(\hat{s}_{\mathbb{U}^1}(X) - \tilde{s}_{\mathbb{U}^1}(X))(\hat{s}_{\mathbb{U}^2}(X) - \tilde{s}_{\mathbb{U}^2}(X))]+\mathrm{o}(1)$$

Théorème (Variance d'une forêt)

Si $k \xrightarrow[n\to+\infty]{} +\infty$, $\frac{k}{n} \xrightarrow[n\to+\infty]{} 0$, $\mu > 0$, s C-Lipschitzienne et $q \xrightarrow[n\to+\infty]{} +\infty$:

$$\mathbb{E}[(\hat{s}(X) - \tilde{s}(X))^2] \leq \frac{3}{4}\frac{\sigma^2(k+1)}{n} + \underset{n\to+\infty}{\mathrm{o}}\left(\frac{k}{n}\right)$$

Théorème (Variance d'une forêt)

Si $k \xrightarrow[n\to+\infty]{} +\infty$, $\frac{k}{n} \xrightarrow[n\to+\infty]{} 0$, $\mu > 0$, s C-Lipschitzienne et $q \xrightarrow[n\to+\infty]{} +\infty$:

$$\mathbb{E}[(\hat{s}(X) - \tilde{s}(X))^2] \leq \frac{3}{4}\frac{\sigma^2(k+1)}{n} + \underset{n\to+\infty}{\mathrm{o}}\left(\frac{k}{n}\right)$$

Propostion (Biais d'une forêt)

$$\mathbb{E}[(\tilde{s}(X) - s(X))^2] \leq \mathbb{E}[(\tilde{s}_{\mathbb{U}^1}(X) - s(X))^2] \,.$$

Corollaire (Borne de risque pour une forêt)

$$\mathbb{E}[(\hat{s}(X) - s(X))^2] \leq \frac{3}{4}\frac{\sigma^2(k+1)}{n} + \frac{6MC^2}{(k+1)^2} + \underset{n\to+\infty}{\mathrm{o}}\left(\frac{k}{n}\right) .$$

$(k+1) = n^{1/3}$ donne une majoration en :

$$K n^{-2/3} + \underset{n\to+\infty}{\mathrm{o}}(n^{-2/3})$$

vitesse minimax

Conclusion et perspectives

- Peut-on faire mieux que Random-Forests RI ?
  En injectant encore plus d'aléa (Geurts et.al. 2006) ?
  En utilisant d'autres prédicteurs de base ?
- Application de la procédure de sélection de variables des données d'ultra grande dimension ?
  Problème du temps de calcul ?
- Bornes de risques en dimension supérieure ?
  Réduction du biais ?

Références


Breiman, L. *Random Forests*. Machine Learning (2001)


Biau G., Devroye L., and Lugosi G. *Consistency of random forests and other averaging classifiers*. Journal of Machine Learning Research (2008)


Biau, G. *Analysis of a Random Forests Model*. Soumis


Ben Ishak A. and Ghattas B. *Sélection de variables pour la classification binaire en grande dimension : comparaisons et application aux données de biopuces*. Journal de la SFdS (2008)


Díaz-Uriarte R., Alvarez de Andrés S. *Gene Selection and classification of microarray data using random forest*. BMC Bioinformatics (2006)


Lê Cao, K.-A., Gonçalves, O., Besse, P., Gadat, S. *Selection of Biologically Relevant Genes with a Wrapper Stochastic Algorithm*. Statistical Applications in Genetics and Molecular Biology (2007).

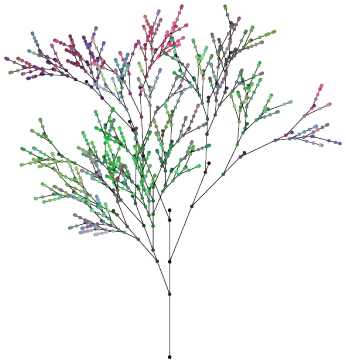
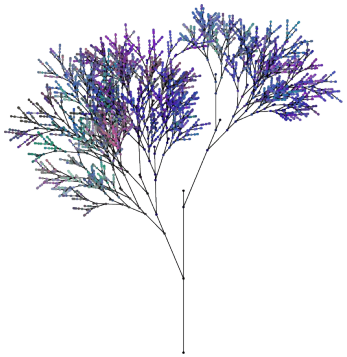
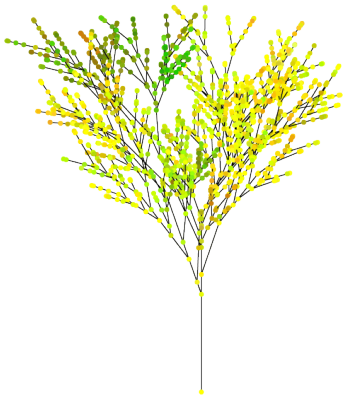
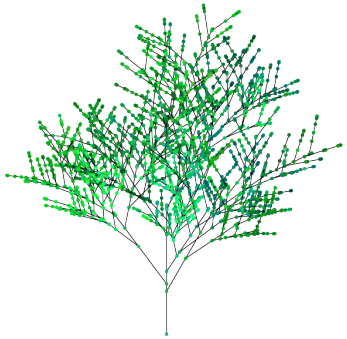

Geurts, P., Ernst, D., Wehenkel, L. *Extremely randomized trees*. Machine Learning (2006)


Genuer R., Poggi J.-M. and Tuleau C. *Variable selection using random forests*. Pattern Recognition Letters (2010)

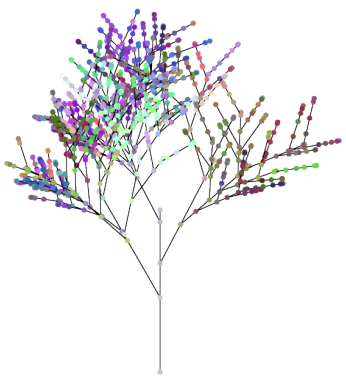
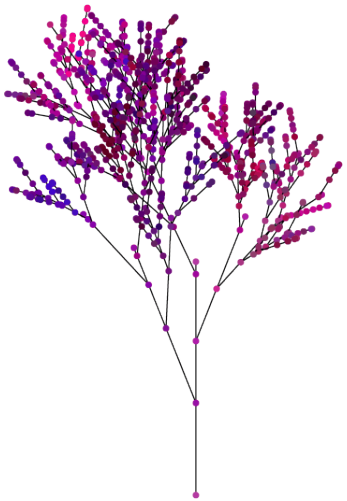
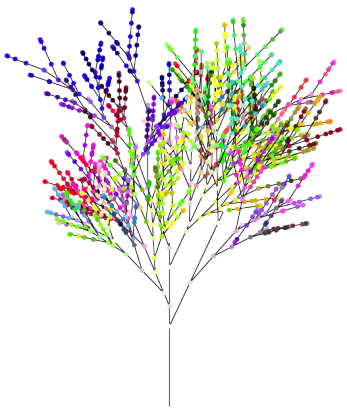
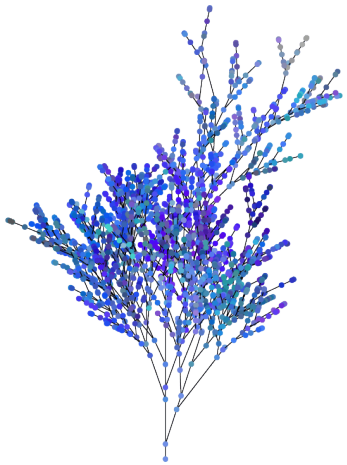

Genuer R., Michel V., Eger E. and Thirion B. *Random Forests based feature selection for decoding fMRI data*. Proceedings of Compstat (2010)


Toussile W., Genuer R., Morlais I. *Gametocytes infectiousness to mosquitoes : variable selection using random forests, and zero inflated models*. Soumis


Genuer R. *Risk bounds for purely uniformly random forests*. Soumis

Merci de votre attention !

Crédits : http ://cg.scs.carleton.ca/∼luc/BRUCE/brucepics.html