Forêts aléatoires

UE IME204 STA307, M2 SITIS et EPI, ISPED

Justine Remiat, support de cours Robin Genuer

30/01/2024

Introduction

- Introduites par Breiman (2001), famille des méthodes d'ensemble Dietterich (1999,2000) : *Bagging*, *Boosting*, *Randomizing Outputs*, *Random Subspace*...

- Algorithme d'apprentissage statistique très performant, à la fois pour des problèmes de classification et de régression. De plus en plus utilisées pour traiter de nombreuses données réelles dans des domaines d'application variés :
  - biopuces Díaz-Uriarte et Alvarez De Andres (2006)
  - l'écologie Prasad et al. (2006)
  - la prévision de la pollution Ghattas (1999)
  - la génomique Goldstein et al. (2010) et Boulesteix et al. (2012)
  - et pour une revue plus large, voir Verikas et al. (2011)

- "Couronnées" dans Fernández-Delgado et al. (2014) (absentes de Wu et al. (2008) qui mentionne CART)

Définition

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

- $\widehat{h}(x) = \dfrac{1}{q} \displaystyle\sum_{\ell=1}^{q} \widehat{h}(x, \Theta_\ell)$ en régression
- $\widehat{h}(x) = \underset{1 \leq c \leq L}{\arg\max} \displaystyle\sum_{\ell=1}^{q} \mathbf{1}_{\widehat{h}(x, \Theta_\ell) = c}$ en classification

Schéma

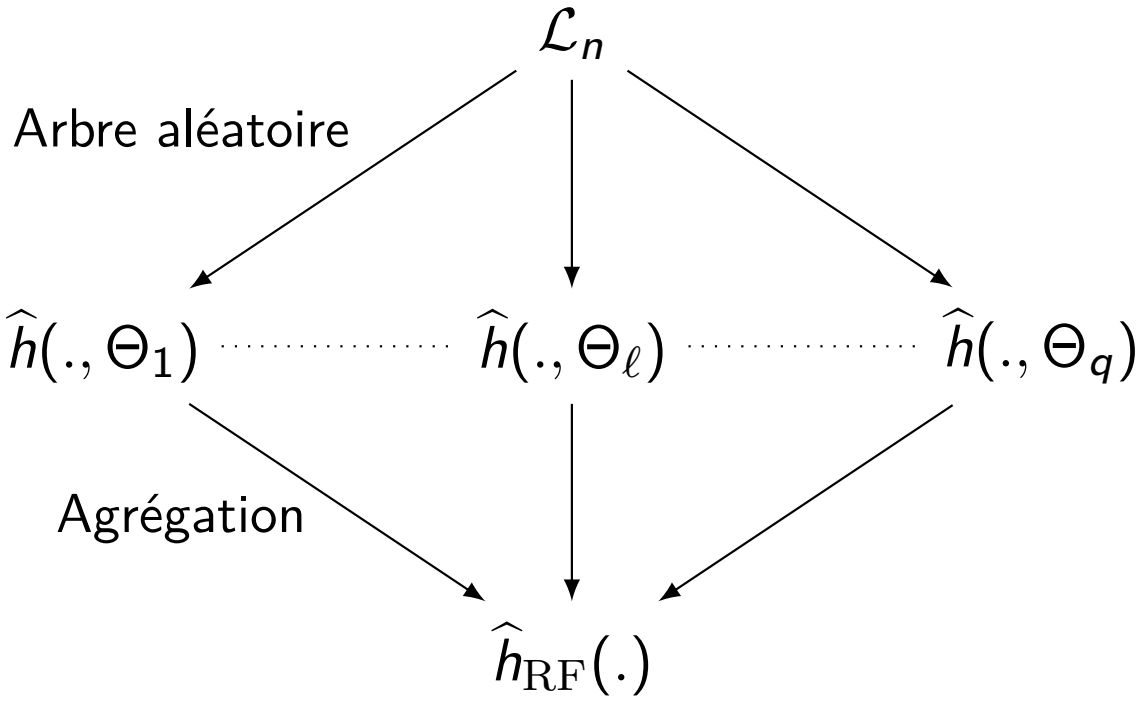


Figure 1: Schéma général RF.

Schéma du Bagging (Breiman 1996)

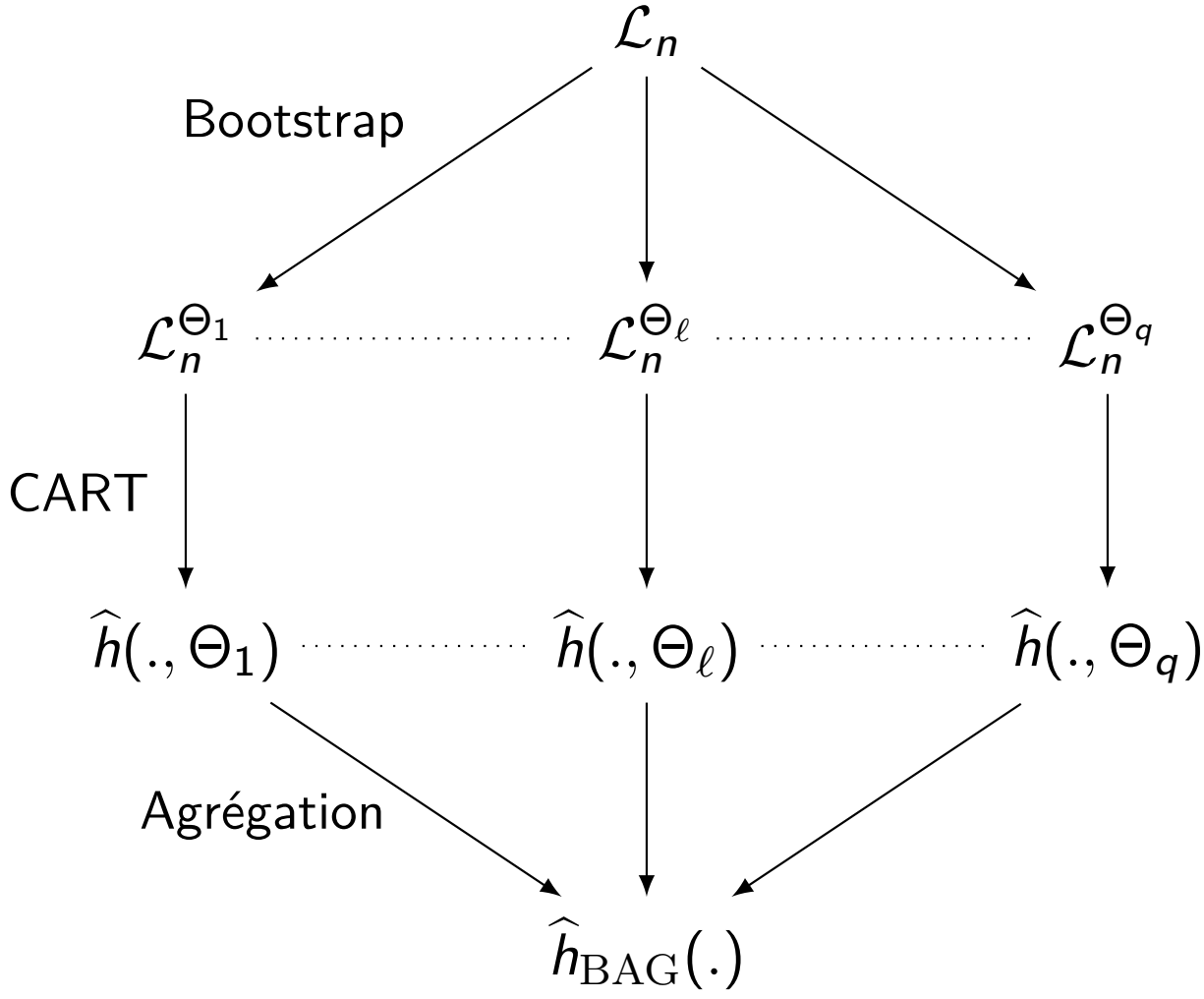


Figure 2: Schéma du Bagging.

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 mais, bien sûr, les variables considérées en chaque noeud pour le choix de la meilleure découpe changent aléatoirement

Définition : Random Forests-RI

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

Schéma des Random Forests-RI

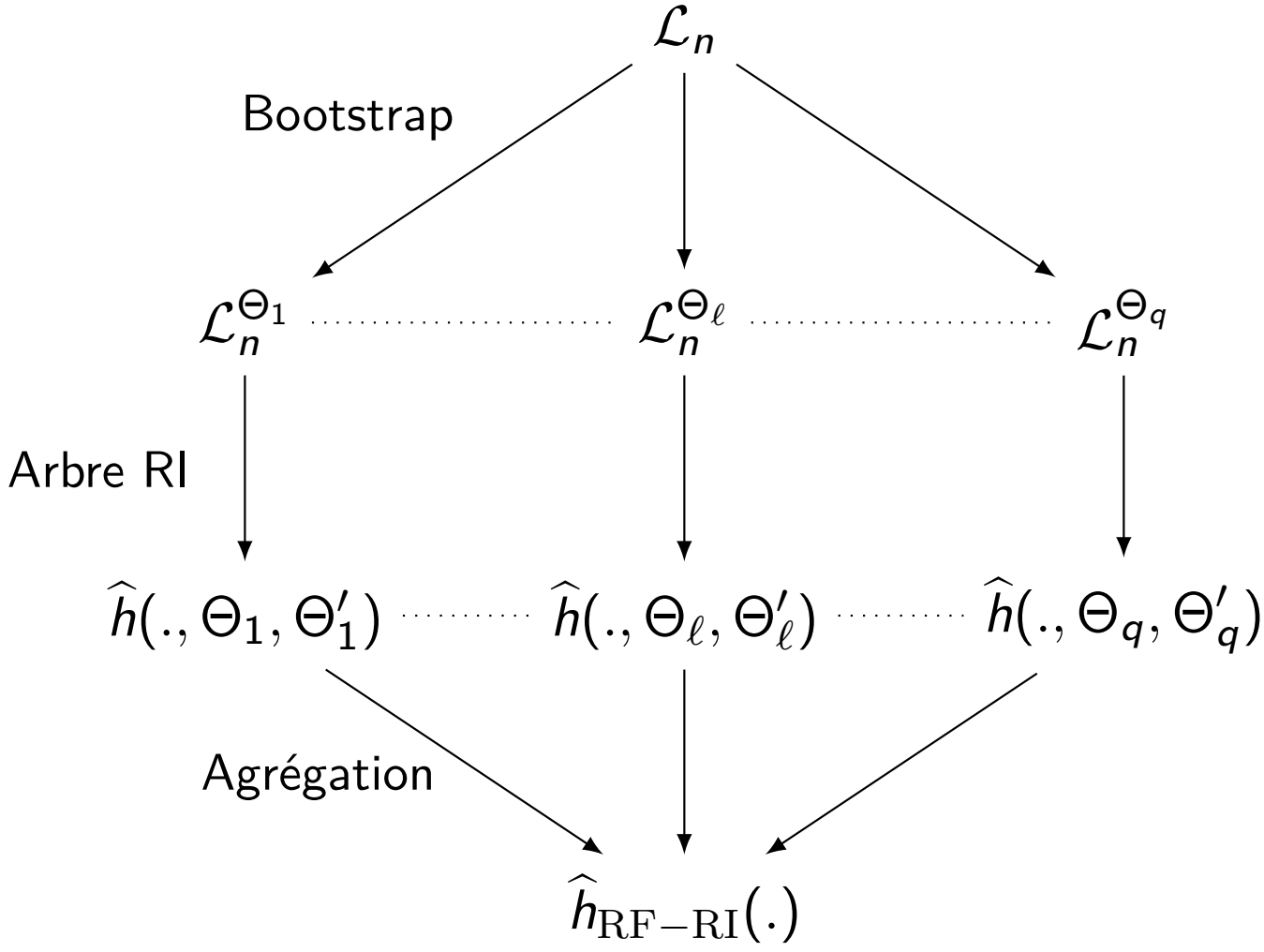


Figure 3: Schéma des RF-RI.

Aléa supplémentaire $\Rightarrow$ amélioration des performances

Données spam : taux d'erreurs

	Arbre optimal	Bagging	RF
Erreur test	0.086	0.062	0.052
Erreur empirique	0.062	0.000	0.004

Table 1: Erreurs tests et empiriques de l'arbre optimal, du Bagging et des RF, données spam.

Random Forests-RI en pratique

Package 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 (par défaut 500)
- `mtry` : le nombre de variables tirées aléatoirement à chaque noeud. C'est le paramètre le plus important :
  - par défaut : $\sqrt{p}$ en classification, $p/3$ en régression
  - l'étude empirique Genuer et al. (2008) précise :
    - en régression, hors du temps de calcul, pas d'amélioration drastique par rapport au Bagging non élagué ($mtry = p$)
    - en classification standard, la valeur par défaut est bonne
    - mais pour des problèmes de classification de grande dimension, des valeurs plus grandes pour $mtry$ donnent parfois des résultats bien meilleurs

Erreur OOB et Estimation de l'erreur de prédiction

Erreur OOB, Out Of Bag (≈ "En dehors du Bootstrap")

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

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

Estimation semblable aux estimateurs classiques de l'erreur de généralisation (par échantillon test ou par validation croisée)

Pas de découpage de l'échantillon d'apprentissage, inclus dans la génération des échantillons bootstrap

Mais attention : c'est bien une sous-forêt différente (en général) qui est utilisée pour calculer chaque $\widehat{Y}_i$

Données spam : réglage de mtry

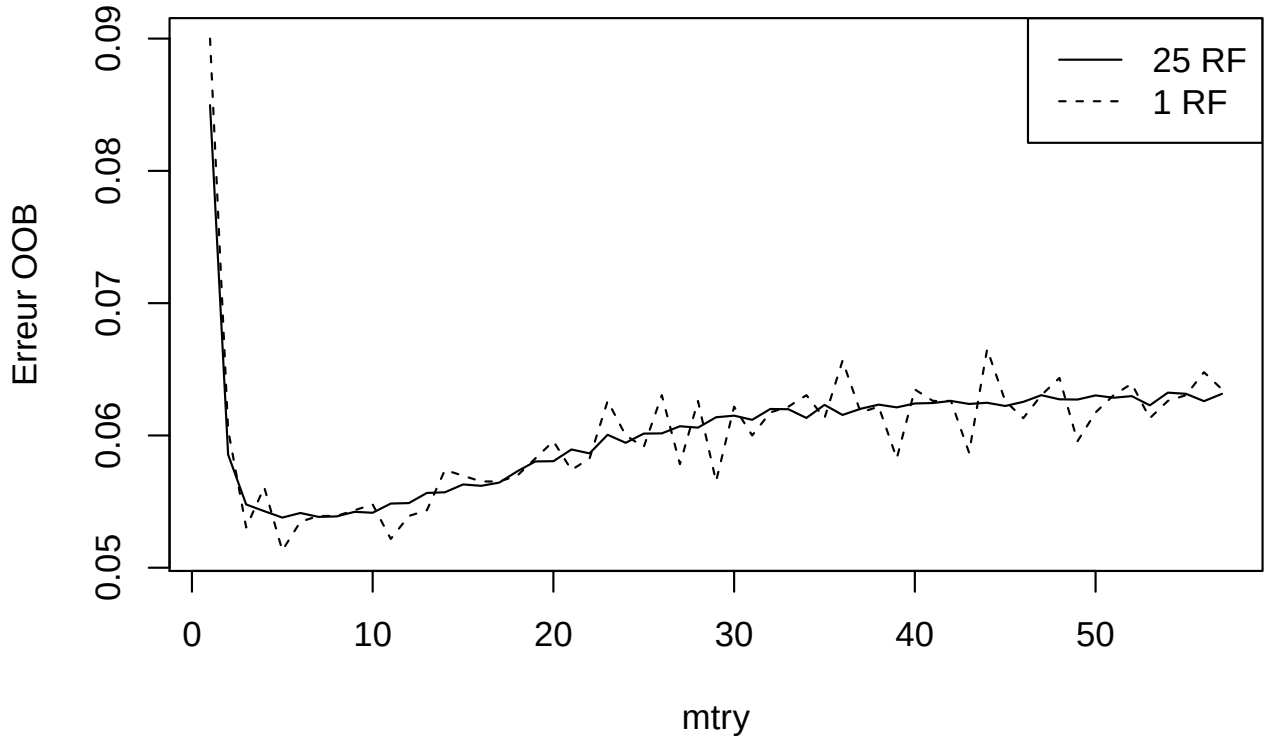


Figure 4: Évolution de l'erreur OOB en fonction de la valeur de mtry, données spam

Importance des variables

- Au delà des performances et du caractère automatique des RF, l'un des aspects les plus importants sur le plan appliqué est la quantification de l'importance des variables
- Azen et Budescu (2003) : discussion générale sur cette notion
- Notion relativement peu examinée par les statisticiens et principalement dans le cadre des modèles linéaires, Grömping (2015) ou la thèse de Wallard (2015)

- Les RF offrent un cadre idéal alliant
  - une méthode non-paramétrique, ne prescrivant pas de forme particulière à la relation entre Y et les composantes de X
  - le ré-échantillonnage bootstrap

  pour disposer d'une définition à la fois efficace et commode de tels indices

Importance des variables par permutation, principe

Breiman (2001), Strobl *et al.* (2007, 2008), Ishwaran (2007), Archer *et al.* (2008), Genuer et al. (2010), Gregorutti et al. (2013, 2015), Louppe et al. (2013)

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

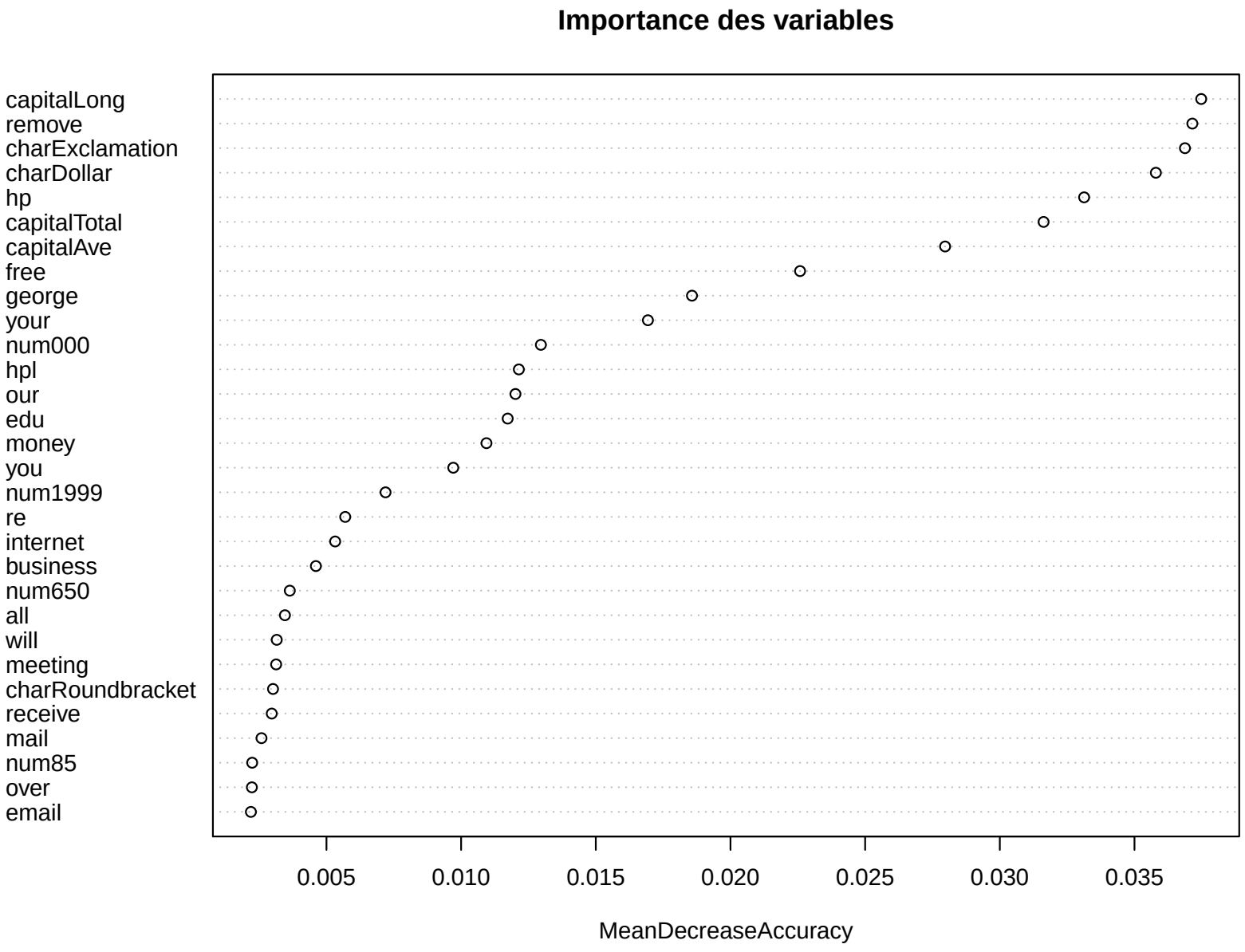
Importance des variables par permutation, définition

Soit $L_{\mathrm{OOB},k}$ le k-ème échantillon OOB et $\widetilde{L}^j_{\mathrm{OOB},k}$ cet échantillon obtenu par permutation aléatoire des valeurs de la j-ème variable

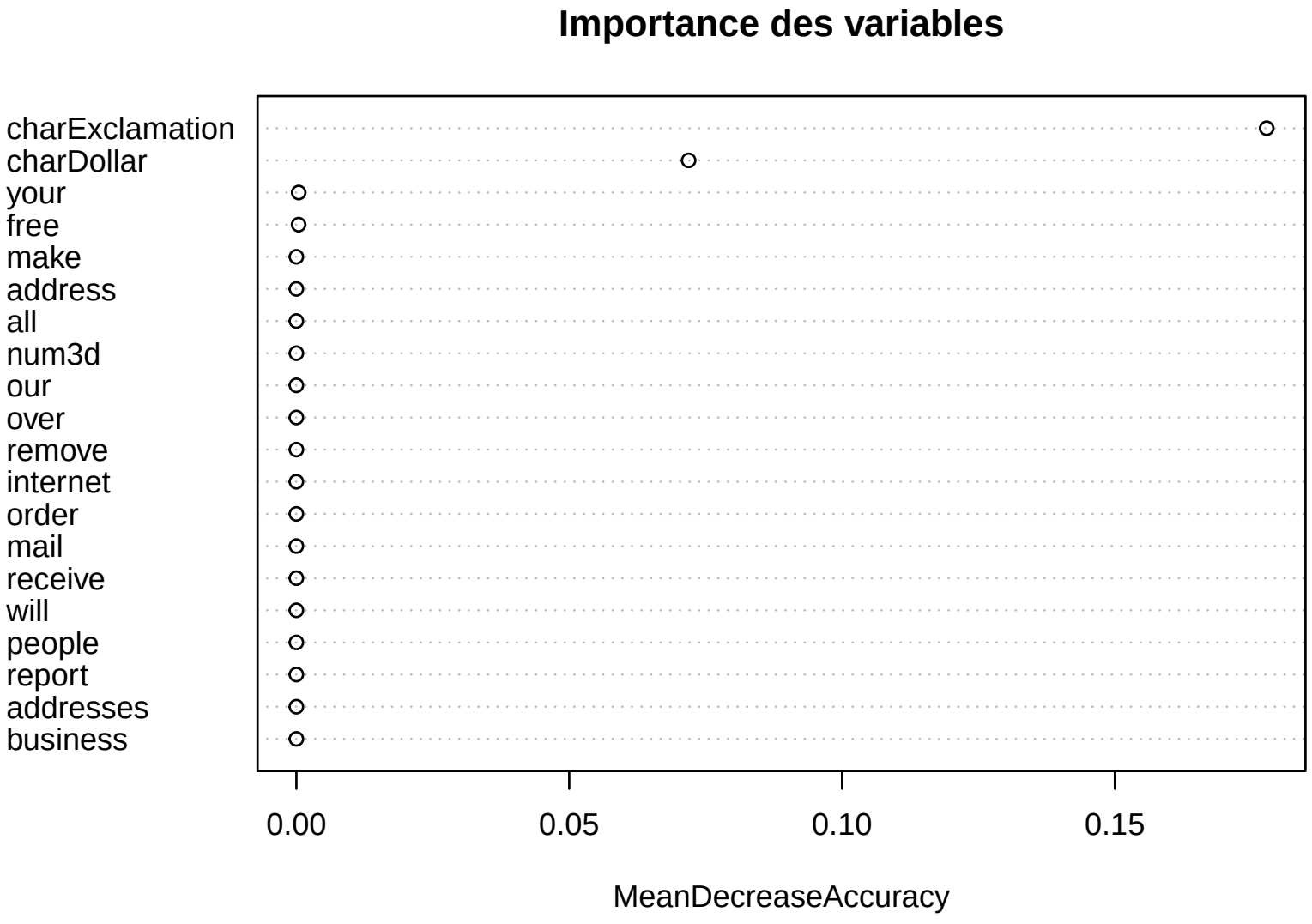
$$\mathrm{VI}(X^j) = \frac{1}{q}\sum_{k=1}^{q}\left[\mathrm{Err}(\widehat{h}_k, \widetilde{L}^j_{\mathrm{OOB},k}) - \mathrm{Err}(\widehat{h}_k, L_{\mathrm{OOB},k})\right]$$

X^1	...	X^j	...	X^p	Y
x_1^1		$x^j_{\pi_j(1)}$		x_1^p	y_1
$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_i^1		$x^j_{\pi_j(i)}$		x_i^p	y_i
$\vdots$		$\vdots$		$\vdots$	$\vdots$
$x_{n_k}^1$		$x^j_{\pi_j(n_k)}$		$x_{n_k}^p$	y_{n_k}

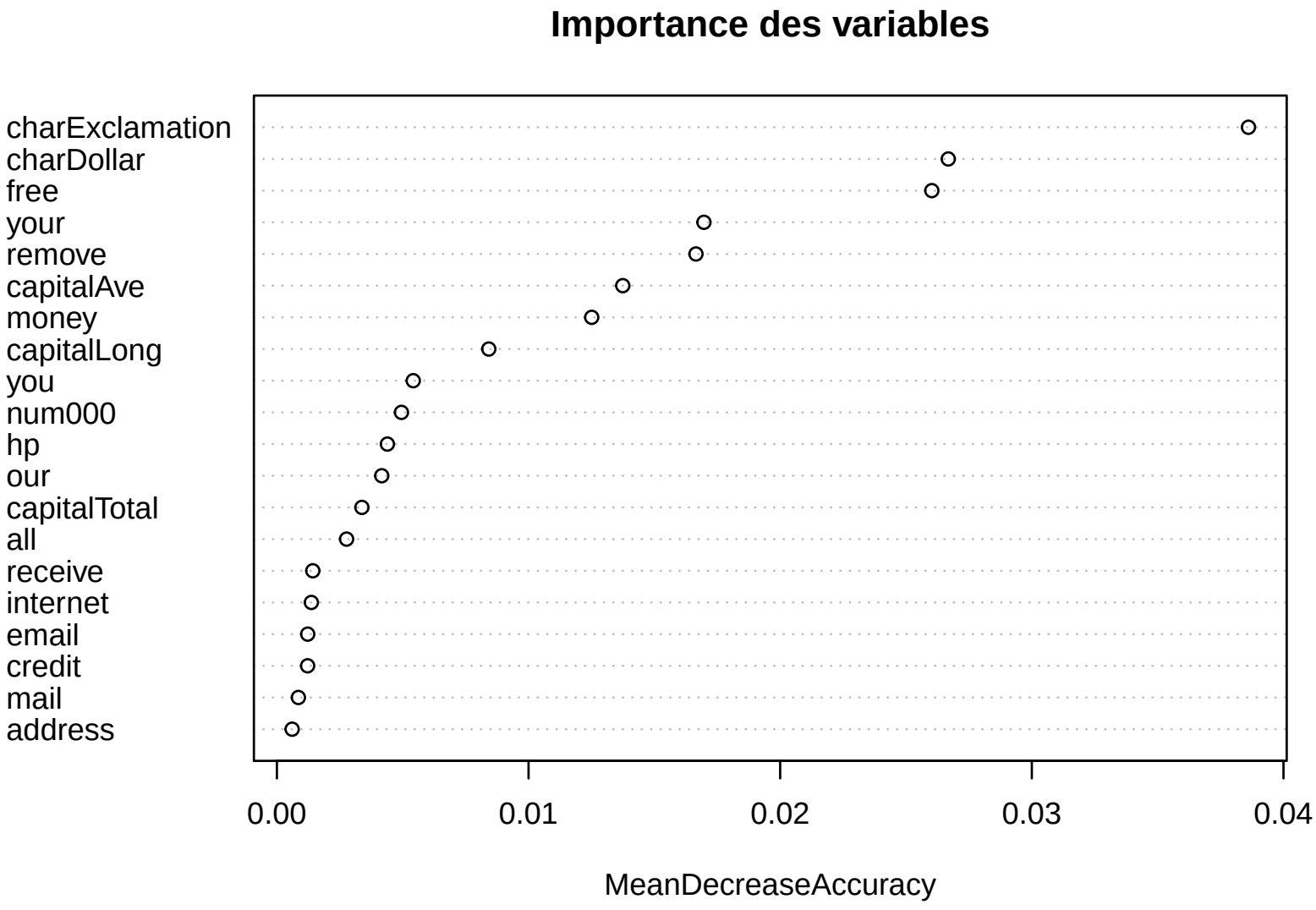
Données `spam` : importance des variables



Bagging d'arbres à 2 feuilles



Forêt d'arbres à 2 feuilles



Extensions et variantes

- Extensions pour des objectifs variés :
  - Problèmes de classement Clemençon et al. (2013)
  - Analyse des données de survie Hothorn et al. (2006) et Ishwaran et al. (2008)
  - Régression quantile Meinshausen (2006)
  - Cluster forests, Yan et al. (2013), Afanador et al. (2016)
  - Régression pour données fonctionnelles Capitaine et al. (2020)
- Variantes :
  - LOFB-DRF vise l'amélioration de la diversité des arbres d'une RF, Fawagreh et al. (2015) utilisent Local Outlier Factor (LOF) pour identifier les arbres divers et sélectionner ceux dont les scores de LOF sont les plus élevés
  - Pondérer *a posteriori* les arbres pour améliorer la performance prédictive, Winham et al. (2013)
  - Random Forests-RC (RC pour "random combination"), coupures non parallèles aux axes, déjà dans Breiman (2001), plus récemment Blaser, Frizlewicz (2015), Menze et al. (2011)
  - Une dernière variante neuronale des RF due à Biau et al. 2016

Quelques références

- Breiman, L. *Bagging*. Machine Learning (1996)
- Breiman, L. *Random Forests*. Machine Learning (2001)
- Capitaine L., Bigot J., Thiébaut R. & Genuer R. *Fréchet random forests for metric space valued regression with non euclidean predictors*. arXiv:1906.01741 (2020)
- Díaz-Uriarte R. & Alvarez de Andrés S. *Gene Selection and classification of microarray data using random forest*. BMC Bioinformatics (2006)
- Genuer R., Poggi J.-M. & Tuleau-Malot C. *VSURF: An R Package for Variable Selection Using Random Forests*. The R Journal (2015)

Deux références choisies non-aléatoirement