Calibration de la profondeur des forêts aléatoires

Erwan Scornet [1,2]

[1] *LSTA, Université Pierre et Marie Curie – Paris VI*
Boîte 158, Tour 15-25, 2ème étage
4 place Jussieu, 75252 Paris Cedex 05, France
[2] *Institut Curie - Biologie du développement - U900*
26 rue d'ulm
75248 Paris Cedex 05, France
erwan.scornet@upmc.fr

Résumé. Les forêts aléatoires, proposées par L. Breiman (2001), comptent parmi les méthodes les plus utilisées dans les problèmes d'estimation de la régression en grande dimension, particulièrement dans des domaines comme la génomique. Bien que les forêts aléatoires montrent de très bonnes performances en pratique, la compréhension théorique des phénomènes mis en œuvre dans ces algorithmes demeure incomplète. Dans cet exposé, nous prouverons une borne théorique sur la vitesse de convergence de la forêt médiane et comparerons différents choix des paramètres de la forêt (profondeur, sous-échnatillonnage).

Mots-clés. Forêts aléatoires, apprentissage, estimation, arbres de décision.

Abstract. Random forests, designed by Breiman (2001), are among the most powerful methods used in high dimensional regression problems, especially in genomic field. Although random forests are acknowledged to have a good performance in practice, some theoretical mechanisms at work in these algorithms remains difficult to understand. In this talk, we prove a theoretical upper bound on the rate of consistency of median forests and we compare different parameter tuning as the subsampling rate or the depth of each tree.

Keywords. Random forest, machine learning, estimation, decision tree

1 Introduction

Les arbres de décision sont l'un des outils dont dispose le statisticien pour résoudre des problèmes d'estimation de la régression et de classification supervisée. L'algorithme CART, imaginé par Breiman et al. (1984), est l'un des arbres de décision les plus utilisés de nos jours.

Afin de pallier la grande sensibilité de cette méthode aux variations dans les données (rajout de données, données fausses...), Breiman (2001) a proposé d'introduire de l'aléatoire dans le processus de construction de cet arbre. L'agrégation des arbres aléatoires ainsi obtenus est appelée forêt aléatoire. L'estimation est alors effectuée grâce à la forêt tout entière et non plus grâce à un seul arbre. Elle est obtenue en calculant la moyenne des estimations des différents arbres de la forêt. Intuitivement, si les arbres sont assez distincts, les estimateurs seront suffisamment différents pour que la moyenne de ces estimateurs permettent de s'affranchir d'aberrations ponctuelles dans les données.

Depuis 2001, des avancées significatives ont été faites dans la compréhension du comportement de l'algorithme des forêts aléatoires. Ainsi, Biau, Devroye et Lugosi (2008) ont prouvé la convergence de plusieurs modèles de forêts ; Biau (2012) a montré, dans un contexte adapté, que la vitesse de convergence d'une forêt aléatoire particulière ne dépendait pas de la dimension de l'espace ambiant mais seulement du nombre de variables réellement pertinentes. Les forêts ont également fait l'objet de nombreuses publications dans des domaines plus appliqués. Citons par exemple les travaux de Díaz-Uriarte et de Andrés (2006) ou de Genuer (2010) qui ont, tous deux, proposés des méthodes de sélection de variables utilisant les forêts aléatoires.

Cependant, une différence essentielle subsiste entre la théorie et l'application. En effet, dans l'approche originale de Breiman, chaque arbre est développé de sorte que les nœuds terminaux ne contiennent qu'une seule donnée. Ce type de forêt, dite complètement développée, est le plus utilisé en pratique. Cependant, en raison de difficultés théoriques, les auteurs choisissent en général de considérer l'approche duale qui consiste à fixer le nombre de cellules et à faire croître le nombre de points tombant dans chacune de ces dernières.

L'objet de cette présentation est de présenter la vitesse de convergence des forêts aléatoires médianes, qui sont un bon compromis entre la simplicité des forêts indépendantes des données et la complexité des forêts de Breiman. En particulier, montrera comment cette vitesse dépend du taux de sous-échantillonnage et de la profondeur de chaque arbre.

2 Position du problème

Étant donné un échantillon $(\mathbf{X}_1, Y_1), \ldots (\mathbf{X}_n, Y_n)$ de variables aléatoires à valeurs dans $[0,1]^d \times \mathbb{R}$, on cherche dans ce travail à estimer la fonction de régression de Y sur $\mathbf{X}$, définie par

$$m(\mathbf{x}) = \mathbb{E}\left(Y|\mathbf{X} = \mathbf{x}\right), \ \mathbf{x} \in [0,1]^d.$$

Pour ce faire, on construit une partition de $[0,1]^d$ en hyperrectangles. La construction de cette partition est récursive. Chaque cellule est obtenue par découpage d'une cellule pré-existante. La structure naturellement associée à ce type de construction est un arbre.

Une fois la partition construite, on estime, pour tout $\mathbf{x} \in [0,1]^d$, $m(\mathbf{x})$ par la moyenne des valeurs Y_i associées aux points X_i tombant dans la même cellule que $\mathbf{x}$. On note cette estimation $m_n(\mathbf{x})$. Afin d'évaluer la qualité de l'estimation, on s'intéresse à l'erreur en moyenne quadratique, définie par

$$\mathbb{E}\left[(m_n(\mathbf{X}) - m(\mathbf{X}))^2\right].$$

Il existe de nombreuses manières de construire des arbres de décision. Néanmoins, tous ces algorithmes sont caractérisés par les deux éléments suivants :

1. Le critère de coupure, qui permet de déterminer la manière dont les coupes sont effectuées à chaque étape.
2. Le critère d'arrêt, qui détermine quand l'algorithme doit arrêter de couper une cellule (usuellement, un nombre de points minimal par cellule).

En toute généralité, les forêts aléatoires procèdent de la manière suivante. À partir d'un même jeu de données, on construit un ensemble d'arbres de décision grâce à l'introduction d'un paramètre aléatoire influençant la construction de chacun des arbres. L'aléatoire peut prendre de multiple formes : il peut, par exemple, induire un sous-échantillonnage des données ou des coupures effectuées selon certaines lois de probabilités. Dans l'algorithme de Breiman, l'aléatoire est introduit dans le processus de construction de CART mais d'autres arbres de décision peuvent être utilisés en remplacement de CART : tous ces algorithmes sont appelés forêts aléatoires. L'estimation n'est alors plus donnée par un seul arbre mais par la moyenne des estimations de chaque arbre. Autrement dit, si l'on veut construire une forêt contenant M arbres, il faut :

1. Disposer de M variables aléatoires $\Theta_1, ..., \Theta_M$ indépendantes, de même loi Θ, correspondant à l'aléatoire introduit dans l'arbre.
2. Construire les M arbres correspondant et les estimations correspondantes, notées $m_n(\mathbf{x}, \Theta_1), ..., m_n(\mathbf{x}, \Theta_M)$ pour tout $\mathbf{x} \in [0,1]^d$.
3. Construire l'estimateur associé à la forêt, noté $\hat{m}_{M,n}(\mathbf{x})$, en effectuant la moyenne des estimations précédentes.

En pratique, on s'intéresse à l'estimateur $\hat{m}_{M,n}(\mathbf{x})$ car on ne peut construire qu'un nombre fini d'arbres. Cependant, d'après la loi des grands nombres, on sait que, pour tout $\mathbf{x}$,

$$\hat{m}_{M,n}(\mathbf{x}) \underset{M\to\infty}{\to} \mathbb{E}_\Theta\left[m_n(\mathbf{x}, \Theta)\right].$$

L'estimateur limite $\hat{m}_{\infty,n} = \mathbb{E}_\Theta\left[m_n(\mathbf{x}, \Theta)\right]$ est l'estimateur de la forêt aléatoire infinie. Contrairement à $\hat{m}_{M,n}$ qui dépend des M arbres qui composent la forêt et est donc aléatoire, $\hat{m}_{\infty,n}$ ne dépend pas de l'aléatoire Θ. Il est donc plus facile à étudier.

3 Forêts aléatoires médianes

Les arbres formant les forêts aléatoires médianes sont construits de la façon suivante :

1. Avant la construction de chaque arbre de la forêt, on sous-échantillonne le jeu de donnée, c'est-à-dire qu'on tire a_n observations sans remise parmi les n points du jeu de donnée. L'arbre ne dépend que de ces observations.
2. Tant que chaque cellule contient au moins un point, on sélectionne une coordonnée uniformément parmi $\{1, \ldots, d\}$. On coupe alors la cellule selon la coordonnée sélectionnée, à la médiane empirique des observations contenues dans cette cellule.

Les forêts médianes (et plus généralement les forêts quantiles qui coupent à un quantile pré-spécifié de la distribution) constituent un bon compromis entre la simplicité des forêts centrées ou uniformes (dont les coupures ne dépendent pas du jeu de données) et la complexité des forêts de Breiman (dont les coupures dépendent à la fois des X_i et des Y_i). Puisqu'elles contiennent exactement un point par cellule, elles sont mieux à même de modéliser le comportement des forêts de Breiman (qui ont des cellules contenant un faible nombre de points).

On note $m_{\infty,n}$ l'estimateur des forêts aléatoires de Breiman. Nous montrerons une borne supérieure sur la vitesse de convergence de la forêt médiane, c'est-à-dire une borne sur la quantité

$$\mathbb{E}\left[(m_{\infty,n}(\mathbf{X}) - m(\mathbf{X}))^2\right].$$

Bibliographie

[1] Breiman, L. (2001), Random Forest, disponible à l'adresse :
http ://oz.berkeley.edu/users/breiman/randomforest2001.pdf
[2] Breiman, L., Friedman, J., Olshen, R.A., Stone, C.J. (1984), Classification and regression tree, Wadsworth Advanced Book Program, Belmont, California.
[3] Biau, G., Devroye, L. and Lugosi, G. (2008). Consistency of random forests and other averaging classifiers, Journal of Machine Learning Research, Vol. 9, pp. 2015-2033.
[4] Biau, G. (2012). Analysis of a random forests model, Journal of Machine Learning Research, Vol. 13, pp. 1063-1095.
[5] Díaz-Uriarte, R., Alvarez de Andrés, S. (2006), Gene selection and classification of microarray data using random forest, BMC Bioinformatics , 7 :3.
[6] Genuer, R. , Poggi, J.-M. , Tuleau-Malot, C. (2010), Variable selection using Random Forests, Pattern Recognition Letters, 31 :2225-2236.
[7] Györfi, L., Kohler, M., Krzyzak, A., Walk, H.A. (2002), A Distribution-Free Theory of Nonparametric Regression, Springer-Verlag, New York.