Les forêts aléatoires avec R

de Robin GENUER et Jean-Michel POGGI (2019)

Chloé FRIGUET[1]
Université de Bretagne-Sud, IRISA

Livre (112 pages)
Auteurs : Robin GENUER et Jean-Michel POGGI
Édition : Presses Universitaires de Rennes (Collection : Pratique de la Statistique) – 2019
ISBN : 978-2-7535-7710-7
http://lesforetsaleatoiresavecr.robin.genuer.fr/

Robin Genuer, Maître de Conférences en Statistique à l'Université de Bordeaux, et Jean-Michel Poggi, Professeur en Statistique à l'Université Paris-Descartes, proposent à travers cet ouvrage de 112 pages de s'approprier une méthode d'apprentissage statistique essentielle pour tout-e praticien-ne des données : les forêts aléatoires (*random forests*).

Comme la plupart des ouvrages de la collection *Pratique de la Statistique* des Presses Universitaires de Rennes dans laquelle celui-ci est paru en ce début 2019, la présentation des concepts se fait par le point de vue des applications, en particulier à travers un exemple « fil rouge » de données publiques traitant de la détection de pourriels dans la messagerie électronique de George, un employé de l'entreprise américaine HP. Cette approche permet de dérouler les différentes étapes de la méthode, pas à pas. Pour favoriser l'assimilation d'un concept, rien ne vaut la pratique : le logiciel R, outil libre et *open-source*, offre un terrain tout trouvé pour reproduire – et s'approprier – la méthodologie des forêts aléatoires sur cet exemple, et d'autres. Ainsi, les données et les codes R (formats : scripts et fichiers Rmarkdown) associés à ces exemples sont disponibles en ligne.

1. chloe.friguet@univ-ubs.fr

Introduction

En apprentissage statistique dit *supervisé*, on cherche à comprendre, à partir d'un échantillon de données observées, le lien supposé entre les informations d'entrée (variable-s explicative-s) et de sortie (variable d'intérêt). On distingue alors les problèmes de régression, où la variable de sortie est quantitative, des problèmes de classification, où la variable de sortie est qualitative. Les forêts aléatoires ont le mérite de permettre de traiter ces deux problèmes. De plus, elles permettent de traiter des informations d'entrée de nature qualitative et/ou quantitative, y compris en grande dimension, lorsque le nombre de ces variables augmente drastiquement. On voit ici les raisons de son succès dans le contexte de données de plus en plus volumineuses auxquelles nous sommes confrontés dans de nombreuses applications.

Commençons par le début

Ce livre débute au **chapitre 1** par une petite rétrospective de l'introduction de cette méthode dans la communauté statistique : des premiers arbres de décision de Leo Breiman dans les années 80 aux forêts aléatoires au début du siècle suivant, en soulignant la présence d'exemples applicatifs variés qui utilisent cette méthode dans la littérature scientifique. Après l'introduction des notations utilisées dans l'ouvrage avec le minimum de formalisme nécessaire, les trois paquets (*packages*) conseillés par les auteurs, et utilisés dans ce livre, sont introduits, respectivement pour la création d'arbres de décision (`rpart`), la mise en œuvre des forêts aléatoires (`randomForest`) et la sélection de variables à l'aide de forêts aléatoires (`VSURF`). Enfin, les cas applicatifs illustrant les concepts tout au long de l'ouvrage, en particulier l'exemple « fil rouge », sont détaillés : données de pourriels (`spam`), mesures environnementales de pollution de l'air et données génomiques, entre autres.

Le **chapitre 2** aborde ensuite le concept d'arbre de décision, et plus particulièrement l'algorithme CART. En quelques pages, le principe et l'objectif de cet outil d'aide à la décision est présenté. L'utilisation du paquet `rpart` qui implémente cet algorithme dans et fait partie des fonctionnalités de base de ce logiciel est détaillée. L'exemple des données `spam` permet ensuite de façon détaillée de construire un arbre de classification pas à pas (cf. figure 1). Deux autres exemples sont ensuite traités, illustrant plus brièvement des situations différentes mais usuelles en pratique (arbre de régression, données manquantes, grande dimension).

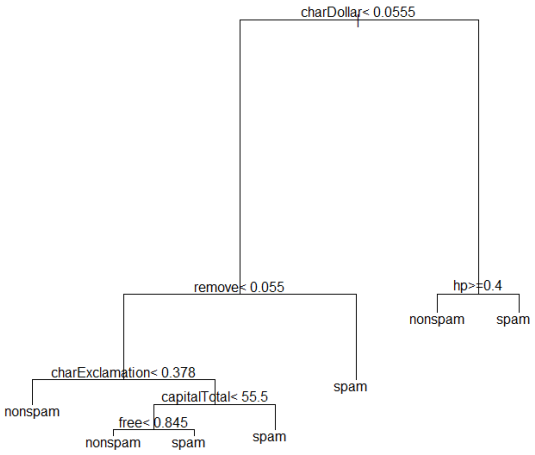

Figure 1 – *Arbre de classification – données* `spam` *– paquet* `rpart`

Une forêt, c'est des arbres

Le **chapitre 3** entre dans le vif du sujet : les forêts aléatoires, dont le principe consiste à agréger plusieurs arbres de décision. Le chapitre s'organise comme le précédent en décrivant d'abord en quelques pages le principe et les étapes de la mise en œuvre de la méthode, puis en illustrant pas à pas son utilisation concrète avec le paquet `randomForest` sur les données `spam`. La méthode est ensuite appliquée sur les trois autres exemples introduits en début d'ouvrage.

Importance des variables

La notion d'*importance* constitue le cœur du sujet du **chapitre 4** : il s'agit d'ordonner les variables explicatives selon leur lien avec la variable d'intérêt. L'algorithme utilisé dans le paquet `randomForest` est explicitement décrit, et illustré sur l'exemple « fil rouge ». Ses caractéristiques, en particulier en présence de données de grande dimension et/ou corrélées, et l'impact des choix des hyper-paramètres sont précisés et illustrés sur des données simulées et les données `spam`. Les autres exemples sont de nouveau utilisés pour illustrer les points clés de ce chapitre.

Sélection de variables

Enfin, le **chapitre 5** aborde la question du choix des variables à inclure dans les arbres. C'est une problématique classique en Statistique pour données de grande dimension notamment, qui se pose lorsqu'on cherche à *prévoir* une information pour de nouvelles données, ou bien pour aider à l'*interprétation* de celles-ci. La procédure est détaillée pour ces deux objectifs pratiques, et illustrée à l'aide des fonctions du paquet `VSURF` sur données simulées et l'exemple « fil rouge ».

Conclusion

Les concepts et méthodes présentés dans cet ouvrage sont clairs. Ils sont décrits avec un niveau de détails techniques suffisant et illustrés sur données simulées et réelles, rendant l'ouvrage accessible pour tout étudiant ou praticien de la Statistique ayant un bagage scientifique de niveau Licence et des connaissances de base en programmation statistique avec  . On peut alors bien comprendre l'implémentation des procédures de forêts aléatoires dans  , mais également avoir des références pour des aspects très pratiques de leur mise en œuvre comme le choix des hyper-paramètres, étape souvent délicate quand on travaille avec des données réelles. Au-delà des paquets  choisis par les auteurs dans cet ouvrage, d'autres sont également évoqués, permettant à l'utilisateur de se faire une idée des autres implémentations disponibles.