Correlation et importance des variables dans les forêts aléatoires

Baptiste Gregorutti[12], Bertrand Michel[2] & Philippe Saint-Pierre[2]

[1] *Safety Line*
15 rue Jean-Baptiste Berlier, 75013 Paris, France

[2] *LSTA, Université Pierre et Marie Curie - Paris VI*
Boîte 158, Tour 15-25, 2ème étage
4 place Jussieu, 75252 Paris Cedex 05, France
baptiste.gregorutti@safety-line.fr

Résumé. La sélection de variables dans un contexte de grande dimension est une tache difficile, en particulier lorsque les variables explicatives sont corrélées. L'algorithme des forêts aléatoires est une méthode très compétitive pour traiter de problèmes de classification et de régression. En effet, il présente de bonnes performances prédictives en pratique et peut être utilisé dans un objectif de sélection de variables au moyen de mesures d'importance. Dans ce travail, nous étudions les aspects théoriques de la mesure d'importance par permutation dans le cas d'un modèle de régression additive. Plus particulièrement, nous sommes en mesure de mieux comprendre l'effet de la corrélation sur la mesure d'importance et par suite sur la sélection de variables. Nos résultats motivent l'utilisation de l'algorithme Recursive Feature Elimination (RFE) pour sélectionner les variables dans ce contexte. Cet algorithme élimine récursivement les variables en utilisant la mesure d'importance comme critère de rang. Des simulations numériques confirment d'une part les résultats théoriques et indiquent d'autre part que l'algorithme RFE tend à sélectionner un faible nombre de variables avec une bonne erreur de prédiction.

Mots-clés. Forêts aléatoires, Importance des variables, Sélection de variables

Abstract. In high-dimensional regression or classification frameworks, variable selection is a difficult task, that becomes even more challenging in the presence of highly correlated predictors. The Random Forests algorithm is a very attractive tool for both classification and regression. Indeed, it has good predictive performances in practice and can be used to perform variable selection thanks to importance measures. Firstly we provide a theoretical study of the permutation importance measure for an additive regression model. This allows us to describe how the correlation between predictors impacts the permutation importance. Our results motivate the use of the Recursive Feature Elimination (RFE) algorithm for variable selection in this context. This algorithm recursively eliminates the variables using permutation importance measure as a ranking criterion. Next various simulation experiments illustrate the efficiency of the RFE algorithm for selecting a small number of variables together with a good prediction error.

Keywords. Random Forests, Variable Importance, Variable Selection

1 Introduction

Dans un contexte d'apprentissage en grande dimension, toutes les variables explicatives ne sont pas nécéssairement importantes pour la prédiction de la variable d'intérêt. En effet, les variables non informatives peuvent avoir un effet néfaste sur la précision du modèle. Les techniques de sélection de variables fournissent une réponse naturelle à ce problème en éliminant les covariables qui n'apportent pas assez d'informations prédictives au modèle. La réduction du nombre de variables explicatives présente un double avantage. D'une part, un modèle contenant peu de variables est plus interprétable. D'autre part, l'erreur de prédiction se trouve réduite de fait de la suppression de variables non informatives.

L'algorithme des forêts aléatoires est une methode non paramétrique traitant à la fois de problèmes de classification et de régression. Il présente de bonnes performances prédictives en pratique, même dans un cadre de très grande dimension. En outre, les forêts aléatoires offrent la possibilité de mesurer l'importance des variables d'entrée sur la prédiction de la variable de sortie. Plusieurs mesures existent et peuvent être intégrées dans un algorithme de type "backward" qui élimine itérativement les variables les moins importantes (voir par exemple Díaz-Uriarte et Alvarez de Andrés (2006), Genuer et al. (2010)). Cependant, des études numériques montrent que ces mesures d'importance sont affectées par la corrélation entre les variables. En particulier, Tolo̧si et Lengauer (2011) ont montré que les valeurs d'importance varient en fonction du nombre de variables corrélées et du niveau de corrélation.

Les contributions de ce travail sont doubles. Dans un premier temps, nous validons les observations de Tolo̧si et Lengauer (2011) sur les effets de la corrélation sur la mesure d'importance dite par permutation (Breiman (2001)). Plus précisément, nous considérons un modèle de régression additive pour lequel il est possible d'exprimer l'importance d'une variable en fonction de sa corrélation avec les autres covariables. Dans un second temps, nous montrons numériquement que l'algorithme backward Recursive Feature Elimination du à Guyon et al. (2002) appliqué aux forêts aléatoires corrige l'effet du biais de corrélation. Dans Gregorutti et al. (2013), une étude de simulation a également été effectuée afin de montrer que cet algorithme permet de sélectionner des modèles de petite taille avec une bonne précision.

2 Forêts aléatoires et mesure d'importance par permutation

On considère une variable d'intérêt Y à valeur dans $\mathbb{R}$ et un vecteur aléatoire $\mathbf{X} = (X_1, \ldots, X_p)$. La régression vise à estimer la fonction $f(x) = \mathbb{E}[Y|\mathbf{X} = x]$ à partir d'un échantillon d'apprentissage $\mathcal{D}_n = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$ de n vecteurs aléatoires indépendants et de même loi que $(\mathbf{X}, Y)$ où $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})$. L'erreur commise par

un estimateur $\hat{f}$ est alors $R(\hat{f}) = \mathbb{E}\left[(\hat{f}(\mathbf{X}) - Y)^2\right]$. Cette quantité étant inconnue en pratique, nous en considérons un estimateur empirique basé un échantillon de validation $\bar{\mathcal{D}}$:

$$\hat{R}(\hat{f}, \bar{\mathcal{D}}) = \frac{1}{|\bar{\mathcal{D}}|} \sum_{i:(\mathbf{X}_i, Y_i) \in \bar{\mathcal{D}}} (Y_i - \hat{f}(\mathbf{X}_i))^2.$$

Les forêts aléatoires sont une méthode non paramétrique très compétitive pour l'estimation de f. Introduites par Breiman en 2001, elles consistent en l'agrégation d'un grand nombre d'arbres aléatoires basés sur une partition dyadique et récursive de l'espace des observations, ici $\mathbb{R}^p$. Plus précisément, n_{tree} arbres aléatoires sont construits à partir d'échantillons bootstrap $\mathcal{D}_n^1, \ldots, \mathcal{D}_n^{n_{tree}}$ des données d'apprentissage $\mathcal{D}_n$. En conséquence, une collection d'estimateurs $\hat{f}_1, \ldots, \hat{f}_{n_{tree}}$ de f sont considérés.

Une différence majeure dans la construction des arbres par rapport aux algorithmes initiaux est la suivante : le critère de découpe intervenant dans le partitionnement est optimisé sur un faible nombre de variables choisi aléatoirement. L'estimateur final de la forêt est alors défini comme la prédiction moyenne de chaque arbre ainsi randomisé. L'aléa induit par l'échantillonnage bootstrap ainsi que le choix aléatoire des variables à chaque étape du partitionnement permet à l'estimateur agrégé de la forêt d'être meilleur que les arbres pris individuellement.

L'algorithme des forêts aléatoires propose également des critères permettant d'évaluer l'importance des covariables sur la prédiction de Y. Nous considérons ici la mesure d'importance par permutation due à Breiman (2001). Une variable X_j est considérée comme importante pour la prédiction de Y si en brisant le lien entre X_j et Y, l'erreur de prédiction augmente. Pour briser le lien entre X_j et Y, Breiman propose de permuter aléatoirement les valeurs de X_j. Plus formellement, considérons une collection d'ensembles "out-of-bag" (OOB) $\{\bar{\mathcal{D}}_n^t = \mathcal{D}_n \setminus \mathcal{D}_n^t,\, t = 1, \ldots, n_{tree}\}$ contenant les observations non retenues dans les échantillons bootstrap. Ces ensembles seront utilisés pour calculer l'erreur de chaque arbre $\hat{f}_t$. À partir de ces ensembles, définissons les ensembles out-of-bag permutés $\{\bar{\mathcal{D}}_n^{tj},\, t = 1, \ldots, n_{tree}\}$ en permutant les valeurs de la j-ème variable des échantillons out-of-bag. La mesure d'importance par permutation est alors définie par

$$\hat{I}(X_j) = \frac{1}{n_{tree}} \sum_{t=1}^{n_{tree}} \left[\hat{R}(\hat{f}_t, \bar{\mathcal{D}}_n^{tj}) - \hat{R}(\hat{f}_t, \bar{\mathcal{D}}_n^t)\right].$$

Cette quantité est l'équivalent empirique de la mesure d'importance $I(X_j)$ comme l'ont formulé récemment Zhu et al. (2012) :

$$I(X_j) = \mathbb{E}\left[\left(Y - f(\mathbf{X}_{(j)})\right)^2\right] - \mathbb{E}\left[(Y - f(\mathbf{X}))^2\right], \tag{1}$$

où $\mathbf{X}_{(j)} = (X_1, \ldots, X_j', \ldots, X_p)$ est un vecteur aléatoire tel que X_j' est une réplication indépendante de X_j. La permutation de X_j dans la définition de $\hat{I}(X_j)$ revient donc à remplacer X_j par une variable indépendante et de même loi dans (1).

Dans la suite, nous étudions la mesure d'importance théorique $I(X_j)$ afin de comprendre comment elle varie en fonction des corrélations entre les variables.

3 Effets de la corrélation sur la mesure d'importance par permutation

L'ensemble des études traitant de l'effet de la corrélation sur la mesure d'importance est basé sur des expérimentations numériques. En particulier, Tolo̧si et Lengauer (2011) montrent que ce critère est affecté par la corrélation entre les variables. Plus précisément, il est observé une diminution des valeurs d'importance lorsque le niveau de corrélation et le nombre de variables corrélées augmente. Nous proposons une validation théorique de ces observations.

Pour cela, supposons que la distribution de $(\mathbf{X}, Y)$ satisfait le modèle de régression additive suivant

$$Y = \sum_{j=1}^{p} f_j(X_j) + \varepsilon, \tag{2}$$

où ε est une variable aléatoire centrée conditionnellement à $\mathbf{X}$ et les fonctions f_j sont supposées mesurables. Dans la suite, la variance et la covariance seront notées respectivement $\mathbb{V}$ et $\mathbb{C}$.

Proposition 1. *1. Sous le modèle (2), pour tout $j \in \{1, \ldots, p\}$, la mesure d'importance par permutation satisfait*

$$I(X_j) = 2\mathbb{V}[f_j(X_j)].$$

2. Supposons de plus que pour $j \in \{1, \ldots, p\}$ la variable $f_j(X_j)$ est centrée. Alors

$$I(X_j) = 2\mathbb{C}[Y, f_j(X_j)] - 2\sum_{k \neq j} \mathbb{C}[f_j(X_j), f_k(X_k)].$$

Ce premier résultat montre que la mesure d'importance correspond à la variance de $f_j(X_j)$. Le second point de la Proposition montre clairement un lien entre $I(X_j)$ et la dépendance entre X_j et les autres variables X_k. Ce résultat peut se préciser si l'on suppose que

$$(\mathbf{X}, Y) \sim \mathcal{N}_{p+1}\left(0, \begin{pmatrix} C & \boldsymbol{\tau} \\ \boldsymbol{\tau}^t & \sigma_y^2 \end{pmatrix}\right), \tag{3}$$

où C est la matrice de variance-covariance du vecteur $\mathbf{X}$, $\boldsymbol{\tau}$ est le vecteur des covariances entre les variables $\mathbf{X}$ et Y et σ_y^2 est la variance de Y. Dans ce contexte, le modèle (2) est vérifié et devient

$$Y = \sum_{j=1}^{p} \alpha_j X_j + \varepsilon.$$

Supposons de plus que

$$C = \begin{pmatrix} 1 & c & \cdots & c \\ c & 1 & \cdots & c \\ \vdots & \vdots & \ddots & \vdots \\ c & c & \cdots & 1 \end{pmatrix}$$

et $\boldsymbol{\tau} = (\tau_0, \ldots, \tau_0)^t$. Autrement dit, nous supposons observer p variables corrélées ayant le même niveau de corrélation avec Y. Dans ce cas, nous obtenons une expression analytique de la mesure d'importance théorique en fonction du nombre de variables corrélées et du niveau de corrélation.

Proposition 2. *Sous les conditions précédentes et pour tout $j \in \{1, \ldots, p\}$, on a*

$$I(X_j) = 2\left(\frac{\tau_0}{1 - c + pc}\right)^2.$$

Ce résultat montre que la mesure d'importance théorique décroit dès lors que le nombre de variables corrélées ainsi que le niveau de corrélation augmente. En pratique, une variable fortement prédictive appartenant à un groupe de variables corrélées peut être estimée moins importante qu'une variable indépendante et moins informative. Cela confirme les observations de Tolo̧si et Lengauer (2011).

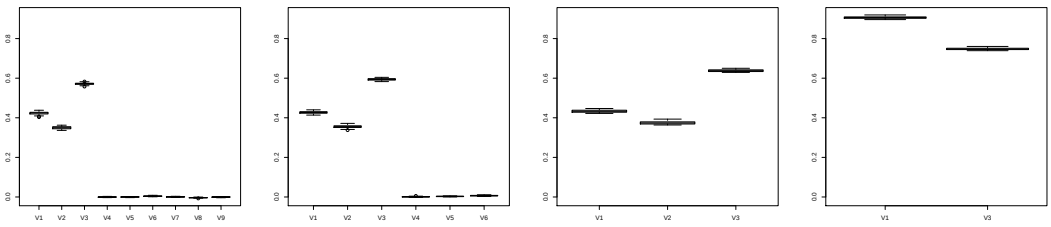
4 Application à la sélection backward

Les résultats de la Section précédente ont montré que le choix des variables prédictives ne peut se faire uniquement grâce à la seule évaluation de la mesure d'importance. Nous l'intégrons donc dans l'algorithme Recursive Feature Elimination (ou RFE, Guyon et al. (2002)) qui élimine récursivement les variables les moins importantes. L'algorithme se décrit comme suit :

1. Construire une forêt aléatoire et calculer l'erreur
2. Calculer la mesure d'importance par permutation
3. Éliminer la variable la moins importante
4. Répéter les étapes 1 à 3 jusqu'à ce que toutes les variables soient éliminées

Cette procédure corrige les effets de la corrélation. En effet, la Figure 1 représente les valeurs d'importance pour neuf variables simulées selon l'hypothèse (3) à chaque étape de l'algorithme RFE : trois variables sont informatives et six variables non informatives. Les deux premières sont simulées corrélées et plus informatives que la troisième qui est indépendante. La Figure 1a montre clairement l'effet de la corrélation sur les deux premières variables qui sont estimées moins importantes que la variable 3 bien qu'elles soient plus informatives pour la prédiction de Y. Cet effet est corrigé du fait du calcul du

critère d'importance à chaque étape de l'algorithme : la variable 2 est éliminée à l'étape 3 (Figure 1c) et la variable 1 est ensuite estimée plus importante que la variable 3 à la dernière étape de l'algorithme (Figure 1d).



(a) Étape 1 (b) Étape 2 (c) Étape 3 (d) Étape 4

FIGURE 1 – Algorithme RFE par étapes avec trois variables informatives et six variables non informatives. Les deux premières sont corrélées et les suivantes sont indépendantes.

Bibliographie

[1] Breiman, L. (2001), *Random Forests*, Machine Learning, Vol. 45, pp 5–32.

[2] Díaz-Uriarte, R. and Alvarez de Andrés, S. (2006), *Gene selection and classification of microarray data using random forest*, BMC Bioinformatics, Vol. 7.

[3] Gregorutti, B., Michel, B. and Saint-Pierre, P. (2013), *Correlation and variable importance in Random Forests*, arXiv preprint.

[4] Genuer, R., Poggi, J.-M. and Tuleau-Malot, C. (2010), *Variable selection using random forests*, Pattern Recognition Letters, 31 :2225–2236.

[5] Guyon, I., Weston, J., Barnhill, S. and Vapnik, V. (2002), *Gene selection for cancer classification using support vector machines*, Machine Learning, Vol. 46, pp. 389–422.

[6] Toloşi, L. and Lengauer, T. (2011), *Classification with correlated features : unreliability of feature ranking and solutions*, Bioinformatics, Vol. 27, pp 1986–1994.

[7] Zhu, R., Zeng, D. and Kosorok, M. R. (2012), *Reinforcement learning trees*, The University of North Carolina at Chapel Hill Department of Biostatistics Technical Report Series.