Combinaison de la classification de variables et de la sélection de variables par forêts aléatoires

Robin Genuer[1], Marie Chavent[2] & Jérôme Saracco[3]

[1] *Inserm U-1219, ISPED, Université de Bordeaux, Inria Bordeaux Sud-Ouest, équipe SISTM, 146 rue Léo Saignat, 33076 Bordeaux*
[2] *Institut de Mathématiques de Bordeaux, UMR CNRS 5251, Université de Bordeaux, Inria Bordeaux Sud-Ouest, équipe CQFD, 351 cours de la libération, 33405 Talence*
[3] *Institut de Mathématiques de Bordeaux, UMR CNRS 5251, Bordeaux INP, Inria Bordeaux Sud-Ouest, équipe CQFD, 351 cours de la libération, 33405 Talence*

Résumé. Les approches standard pour aborder la classification supervisée en grande dimension font souvent intervenir une sélection de variables et/ou une réduction de la dimension. La méthodologie proposée dans ce travail combine la classification de variables et la sélection de variables. La classification hiérarchique des variables permet de construire des groupes de variables corrélées et résume chaque groupe par une variable synthétique. L'originalité est que les groupes de variables sont inconnus a priori. De plus, l'approche de classification traite à la fois des variables numériques et des variables catégorielles. Parmi toutes les partitions possibles, les variables synthétiques les plus pertinentes sont sélectionnées à l'aide d'une procédure utilisant des forêts aléatoires. Les performances numériques sont illustrées sur des ensembles de données simulées et réelles. La sélection de groupes de variables peut permettre d'améliorer les performances en prédiction et facilite l'interprétation des résultats.

Mots-clés. Classification de variables, Forêts aléatoires, Classification supervisée, Sélection de variables

Abstract. Standard approaches to tackle high-dimensional supervised classification often include variable selection and dimension reduction. The proposed methodology combines clustering of variables and feature selection. Hierarchical clustering of variables allows to built groups of correlated variables and summarizes each group by a synthetic variable. Originality is that groups of variables are unknown a priori. Moreover clustering approach deals with both numerical and categorical variables. Among all the possible partitions, the most relevant synthetic variables are selected with a procedure using random forests. Numerical performances are illustrated on simulated and real datasets. Selection of groups of variables can improve prediction performance and provides easier interpretation of results.

Keywords. Clustering of variables, Random forests, Supervised classification, Variable selection

1 Introduction

Cette communication traite des problèmes de réduction de dimension et de sélection des variables. dans le contexte de la classification supervisée. Dans ce cadre, il y a souvent deux objectifs : la prédiction (être capable de prédire les classes associées à de nouvelles observations) et la sélection de caractéristiques (pour extraire les variables les plus intéressantes). Typiquement dans un contexte médical, le premier objectif pourrait être de réussir à prédire si les patients répondront ou ne répondront pas bien à un traitement, par exemple en raison de l'expression de leurs gènes, alors que le deuxième objectif pourrait être de déterminer quel profil d'expression d'une partie du génome est responsable de la bonne ou mauvaise réponse au traitement. Ces objectifs sont en effet particulièrement pertinents lorsqu'il s'agit, par exemple, de la génomique ou de la protéomique, où le nombre p des variables dépasse largement le nombre n d'observations disponibles.

Une façon classique d'aborder ces questions est d'utiliser une méthode de sélection de variables (voir par exemple Guyon and Elisseeff, 2003; Tibshirani, 1996). L'idée est que la méthode doit pouvoir sélectionner les variables les plus intéressantes, tout en préservant les bonnes performances de prédiction. Cependant, les données contiennent souvent de nombreuses variables fortement corrélées, et réussir à sélectionner toutes les variables dans un groupe de variables corrélées peut être très difficile.

Dans cette communication, nous proposons et évaluons une nouvelle méthodologie pour la réduction de dimension et la sélection de variables, qui combine la classification de variables d'une part et la sélection de variables à l'aide des forêts aléatoires (**RF**) d'autre part. La méthode de classification de variables a été introduite par Chavent et al. (2012) et est notée **CoV** ci-après. Notez que cette méthode regroupe les *variables*, ce qui diffère du problème (plus classique) de la classification d'*individus*. La classification de variables regroupe des variables hautement corrélées et fournit pour chaque groupe (cluster) une variable synthétique, qui est une variable numérique résumant les variables du groupe. Ainsi cette approche permet d'éliminer la redondance et de garder toutes les variables ensemble dans un cluster pendant le reste de l'analyse. Ce faisant, elle permet de réduire la dimension des données en remplaçant les p variables originales par K variables synthétiques (où K indique le nombre de clusters sélectionnés). Dans la méthodologie proposée, le nombre K est optimisé en fonction de la performance prédictive d'un classifieur de **RF** construit avec ces variables synthétiques. Soit K^* le nombre optimal de groupes de variables. La réduction de dimension fournit ainsi K^* variables synthétiques qui n'utilisent que les variables originales dans les groupes sélectionnés, contrairement, par exemple, aux composantes principales de l'analyse en composantes principales. Ainsi, dans **CoV**, une variable originale intervient dans la construction d'une variable synthétique unique, ce qui facilite l'interprétation. Une fois la réduction de dimension effectuée, les variables synthétiques les plus importantes sont ensuite sélectionnés à l'aide d'une procédure basée sur les **RF**, introduite par Genuer et al. (2010). Cette méthode de sélection de variables notée **VSURF** ci-après, est appliquée aux K^* variables synthétiques, et conduit à fournir

$m \leq K^*$ variables synthétiques pertinentes. Ainsi, la prédiction de nouvelles observations peut être faite avec un prédicteur basé sur ces m variables synthétiques sélectionnées (i.e. une liste de groupes de variables). Par conséquent, la méthode proposée produit un classifieur et un ensemble de variables sélectionnées avec comme information supplémentaire la structure de groupe de ces variables.

L'approche proposée ne nécessite pas la connaissance a priori des groupes de variables. Par conséquent, elle diffère du group-Lasso (Yuan and Lin, 2006). De plus, notre méthode s'écarte également des techniques sparse-PLS (Chun and Keleş, 2010; Lê Cao et al., 2011), qui effectuent une réduction de dimension et une sélection de variables sans aucune information de structure de groupe.

Enfin, soulignons que l'objectif principal de ce document est de présenter une nouvelle méthodologie qui : i) sélectionne des groupes de variables informatives, ii) peut traiter des données mixtes, iii) fournit de bonnes performances en prédiction, iv) et facilite l'interprétation des résultats.

Mentionnons enfin que la méthode proposée est disponible dans le package R `CoVVSURF`[1] et que ce travail a été publié (Chavent et al., 2019).

2 La procédure CoV/VSURF

Nous décrivons maintenant la méthodologie proposée dans l'algorithme suivant.

Input: un ensemble de n observations $(\mathbf{X}, \mathbf{y})$ et $\mathbf{x}$ le vecteur associé à une nouvelle observation. $\mathbf{X}$ est la matrice des variables explicatives et $\mathbf{y}$ est le vecteur des observations de la variable réponse.

But: sélectionner des groupes de variables informatives et prédire la classe de $\mathbf{x}$.

(a) Sélection de groupes de variables informatives:

1. Appliquer **CoV** sur $\mathbf{X}$ pour obtenir une classification hiérarchique des variables (dendrogramme ou arbre).
2. Pour chaque $K = 2, \ldots, p$, couper l'arbre **CoV** en K groupes, entraîner une **RF** avec les variables synthétiques $\mathbf{f}^1, \ldots, \mathbf{f}^K$ comme prédicteurs et $\mathbf{y}$ comme variable de sortie et calculez son taux d'erreur OOB.
3. Choisir le nombre optimal K^* de clusters, qui conduit au taux d'erreur OOB minimum.

   Coupez l'arbre **CoV** en K^* groupes.
4. Appliquer **VSURF** avec les K^* variables synthétiques $\mathbf{f}^1, \ldots, \mathbf{f}^{K^*}$ comme prédicteurs et $\mathbf{y}$ comme variable de sortie. Notons $m \leq K^*$ le nombre de variables synthétiques informatives sélectionnées (correspondant à l'ensemble d'interprétation de **VSURF**).

[1] `https://github.com/robingenuer/CoVVSURF`

(b) Prédiction d'une nouvelle observation $\mathbf{x}$:

1. Entraîner une **RF**, $\hat{f}$, sur le jeu de données constitué des m variables synthétiques sélectionnées et de $\mathbf{y}$.
2. Calculez les scores de $\mathbf{x}$ sur les m variables synthétiques sélectionnées et prédire la classe en utilisant $\hat{f}$.

Une caractéristique intéressante de la procédure est que, même si l'algorithme ascendant de classification hiérarchique des variables est *non-supervisée* (dans le sens où il n'utilise pas la variable réponse $\mathbf{y}$), la sélection des variables finales est *supervisée* puisque le nombre de clusters est optimisé en terme d'erreur de prédiction du classifieur **RF**.

Cette procédure est illustrée sur des jeux de données simulées et un jeu de données réelles de protéomique. La méthode **CoV/VSURF** a été implémentée dans un package R, qui est disponible en ligne accompagné d'une vignette [2].

3 Simulations

Un échantillon d'apprentissage de $n = 600$ observations, $p = 120$ variables explicatives et avec une variable réponse Y binaire, est généré. 30 variables explicatives sont des variables de bruit (indépendantes entre elles et avec Y). Les autres variables explicatives sont générées pour obtenir 9 groupes de variables très corrélées entre elles, dont 6 sont informatifs (reliés à Y) et les autres ne le sont pas (indépendants de Y). La méthode **CoV** est appliquée à la matrice de 600×120 des variables explicatives. Le dendrogramme de la hiérarchie des 120 variables (aussi appelé arbre **CoV**) est calculé (non montré ici). Ce dendrogramme suggère une partition en 9 groupes. Cependant, dans notre méthodologie, le nombre de clusters n'est pas choisi en fonction de la forme du dendrogramme mais en fonction de la prédiction de la variable de réponse binaire Y. Pour chaque valeur de K entre 2 et 120, nous coupons l'arbre **CoV**, construisons une **RF** sur les K variables synthétiques de la partition en K clusters correspondante et calculons le taux d'erreur OOB. La valeur optimale que nous obtenons pour cet ensemble de données d'apprentissage est $K^* = 9$ (Fig. 1). Cette partition en $K^* = 9$ clusters récupère presque toute la structure des données. Nous récupérons 8 des 9 groupes de variables corrélées, alors que toutes les variables de bruit sont ensemble dans un grand groupe avec le dernier groupe de variables corrélées. Cependant, la variable synthétique de ce cluster ne prend pas vraiment en compte les variables de bruit, dont les coefficients dans la combinaison linéaire sont très faibles.

VSURF est alors appliqué sur les $K^* = 9$ variables synthétiques de la classification de variables précédemment choisie. **VSURF** sélectionne 6 variables synthétiques, correspondant aux 6 groupes de variables informatives. Du point de vue de l'interprétation,

[2] `https://robingenuer.github.io/CoVVSURF/`

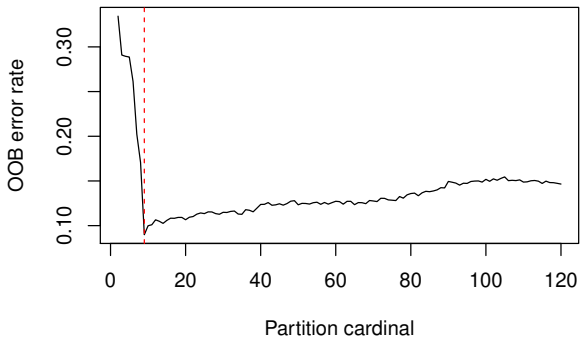

Figure 1: Taux d'erreur OOB des forêts aléatoires en fonction du nombre de grappes de partitions obtenues en coupant l'arbre **CoV** pour l'ensemble de données d'apprentissage simulé de $n = 600$ observations. La ligne verticale en pointillés rouges correspond à $K^* = 9$ clusters.

nous avons réussi à sélectionner toutes les variables informatives, avec en plus la structure de clustering. Pour comparaison, **VSURF** appliqué directement sur les 120 variables initiales sélectionne 39 variables parmi les 54 variables informatives, avec au moins une par groupe informatif.

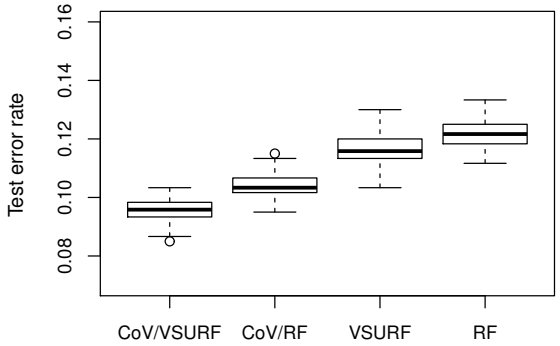

Figure 2: Comparaison des taux d'erreur test (moyens sur 100 **RF**) de 4 procédures. **CoV/VSURF** et **CoV/RF** correspondent à **VSURF** et **RF** appliqués sur les K^* variables synthétiques données par **CoV**. **VSURF** et **RF** font référence aux méthodes appliquées sur les 120 variables originales.

Les taux d'erreur (moyennés sur 100 **RF**) calculés sur un échantillon test contenant 600 nouvelles observations sont illustrés à la Figure 2. Ainsi, l'utilisation de **CoV** dans une étape de réduction de dimension conduit —au moins pour cet exemple— à un gain en prédiction tant par rapport à **RF** qu'à **VSURF**. De plus, l'application de **VSURF** dans

la procédure **CoV/VSURF**permet de sélectionner des groupes informatifs de variables liées sans perte de prédiction.

On trouvera dans Chavent et al. (2019) des schémas de simulations supplémentaires ainsi qu'une application sur données réelles portant sur des mesures de protéomiques dans le cadre d'un essai clinique.

Bibliographie

Chavent, M., Genuer, R., and Saracco, J. (2019). Combining clustering of variables and feature selection using random forests. *Communications in Statistics - Simulation and Computation*, 0(0):1–20.

Chavent, M., Kuentz-Simonet, V., Liquet, B., and Saracco, J. (2012). ClustOfVar: An R package for the clustering of variables. *Journal of Statistical Software*, 50(13):1–16.

Chun, H. and Keleş, S. (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(1):3–25.

Genuer, R., Poggi, J.-M., and Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters*, 31(14):2225–2236.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182.

Lê Cao, K.-A., Boitard, S., and Besse, P. (2011). Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. *BMC Bioinformatics*, 12(1):1.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, pages 267–288.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67.