Arbres de classification pour variables groupées

Audrey Poterie [1], Jean-François Dupuy [1], Valérie Monbet [2] & Laurent Rouvière [3]

[1] *IRMAR, INSA de Rennes, 20 Avenue des Buttes de Coesmes, 35708 Rennes*
apoterie@insa-rennes.fr, jean-francois.dupuy@insa-rennes.fr

[2] *IRMAR, Univerité de Rennes 1, Place du recteur Henri Le Moal, 35043 Rennes*
valerie.monbet@univ-rennes1.fr

[3] *IRMAR, Univerité de Rennes 2, 263 avenue du Général Leclerc, 35042 Rennes*
laurent.rouviere@univ-rennes2.fr

Résumé. Nous nous plaçons dans le cas de la classification supervisée et cherchons à expliquer une variable binaire Y par un vecteur $\mathbf{X}$ à valeurs dans $\mathbb{R}^p$. Nous supposons que le vecteur des variables explicatives est structuré en J groupes connus. L'objectif est de prendre en compte cette structure de groupes pour construire le classifieur. Nous proposons une approche par arbre qui consiste à sélectionner un groupe de variables et à appliquer une analyse discriminante linéaire sur ce groupe pour scinder un nœud. Le procédé est répété jusqu'à ce qu'un critère d'arrêt soit satisfait. L'arbre ainsi défini ayant tendance à sur-apprendre, nous proposons une méthode d'élagage permettant de sélectionner un arbre performant.

Mots-clés. Arbres de classification, variables groupées, analyse discriminante linéaire pénalisée.

Abstract. In the supervised classification setting, we consider the problem of predicting a binary variable Y, based on a vector $\mathbf{X}$ which takes values in $\mathbb{R}^p$. We assume that $\mathbf{X}$ is structured in J known groups. The objective is to take account of this structure to construct the classification rule. We propose a tree-based approach, which consists in selecting a group of inputs and then applying a linear discriminant analysis, based on this group, to split a node. This process is repeated until some stopping criterion is satisfied. The resulting tree is prone to overfitting, thus we propose a pruning strategy that allows to select an optimal tree.

Keywords. Classification tree, groups of variables, linear discriminant analysis, penalized linear discriminant analysis.

1 Introduction

Dans de nombreux problèmes de classification supervisée, les variables explicatives (inputs) ont une structure de groupes connue et/ou clairement identifiable. Par exemple, en biologie, lorsque l'on souhaite étudier la composition chimique d'un sérum à l'aide de la spectrométrie de masse, les variables explicatives, de nature fonctionnelle, peuvent être divisées en groupes représentants différentes parties de la courbe.
Dans ce travail, l'objectif est d'élaborer une règle de classification qui prenne en compte cette structure de groupes. Ce problème a bien entendu déjà été étudié. Par exemple, la régression logistique régularisée par la pénalité "group lasso" permet d'élaborer une règle de classification basée sur la structure groupée des données (Meier, Van De Geer et Bühlman (2008)). A notre connaissance, ce problème n'a pas encore été abordé pour les arbres de classification. En effet, bien qu'un grand nombre d'algorithmes d'arbres de classification ait été développé (Breiman et al. (1984), Quinlan. (1986), Loh et Shih (1997), Wickramarachchi et al. (1997),...), aucun ne permet la prise en compte d'une structure de groupes connue et/ou fixée a priori.

2 Arbres de classification pour variables groupées

On considère un vecteur aléatoire $(\mathbf{X}, Y)$ à valeurs dans $\mathbb{R}^p \times \{0,1\}$. On note $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_{n+m}, Y_{n+m})$ $n+m$ vecteurs aléatoires indépendants et de même loi que $(\mathbf{X}, Y)$. Le vecteur $\mathbf{X}_i$, $i = 1, \ldots, n+m$, représente le vecteurs des variables explicatives et Y_i indique la classe ou label du i-ème individu. Le vecteur $(\mathbf{x}_i, y_i)$ désigne une réalisation du vecteur aléatoire $(\mathbf{X}_i, Y_i)$. L'échantillon est divisé en un ensemble d'apprentissage de taille n et un ensemble de validation de taille m. L'objectif est de construire une règle de classification, c'est-à-dire une fonction mesurable $g : \mathbb{R}^p \times (\mathbb{R}^p \times \{0,1\})^{n+m} \to \{0,1\}$ qui affecte un nouvel individu $\mathbf{x} \in \mathbb{R}^p$ à la classe $g(\mathbf{x}, (\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_{n+m}, Y_{n+m}))$. Elle sera notée $g(\mathbf{x})$ pour simplifier.

Dans cette étude, nous nous intéressons au cas où le vecteur $\mathbf{X}$ est structuré en J groupes connus. On note p_j, $j = 1, \ldots, J$, le cardinal du j-ème groupe et $G_j = \{G_j(1), \ldots, G_j(p_j)\}$ l'ensemble des indices des composantes de $\mathbf{X}$ inclues dans le j-ème groupe. On a :

$$\bigcup_{j=1}^{J} G_j = \{1, \ldots, p\} \quad \text{et} \quad G_j \cap G_{j'} = \emptyset \quad \text{pour} \quad j \neq j'.$$

Les méthodes par arbre consistent à construire des règles de classification à partir d'un partitionnement récursif et binaire de l'espace $\mathbb{R}^p$ des variables explicatives. Dans un premier temps, l'algorithme considère l'espace de définition des variables explicatives et le divise en deux sous-ensembles (appelés nœuds). Puis, le processus de division est appliqué sur ces deux nœuds. Ce découpage est répété de manière récursive sur chaque nouveau nœud jusqu'à ce qu'un critère d'arrêt soit satisfait.

Nous proposons d'élaborer des arbres de classification qui tiennent compte de la structure de groupes des variables explicatives. La division d'un nœud $\mathcal{N} \subset \mathbb{R}^p$ est définie par le choix conjoint d'un groupe de coupure et d'une division définie par les variables du groupe retenu. Ce processus se compose de deux étapes.

Étape 1 : choix d'une division pour chaque groupe. Une analyse discriminante linéaire est tout d'abord réalisée sur chaque groupe de variables $\mathbf{X}^j = (X_{G_j(1)}, \ldots, X_{G_j(p_j)})$, $j = 1, \ldots, J$. L'analyse discriminante linéaire suppose que conditionnellement à la classe (et en se restreignant au nœud $\mathcal{N}$), les observations $\mathbf{X}^j | Y = k$, $k = 0, 1$, suivent la loi normale d'espérance $\mu^j_{k,\mathcal{N}} \in \mathbb{R}^{p_j}$ et de variance $\Sigma^j_{\mathcal{N}}$ de dimension $p_j \times p_j$. Un nouvel individu $\mathbf{x} \in \mathbb{R}^p$ est alors affecté à la classe 1 si $\delta^j_{1,\mathcal{N}}(\mathbf{x}) \geq \delta^j_{0,\mathcal{N}}(\mathbf{x})$, il est affecté à la classe 0 sinon. Les $\delta^j_{k,\mathcal{N}}, k = 0, 1$, sont les fonctions discriminantes linéaires définies par :

$$\delta^j_{k,\mathcal{N}}(\mathbf{x}) = {\mathbf{x}^j}^\top {\Sigma^j_{\mathcal{N}}}^{-1} \mu^j_{k,\mathcal{N}} - \frac{1}{2} {\mu^j_{k,\mathcal{N}}}^\top {\Sigma^j_{\mathcal{N}}}^{-1} \mu^j_{k,\mathcal{N}} + \log \pi_{k,\mathcal{N}} \tag{1}$$

où $\pi_{k,\mathcal{N}} = \mathbf{P}_{\mathcal{N}}(Y = k)$ représente la probabilité a priori qu'un individu dans le nœud $\mathcal{N}$ appartienne à la classe k ($k = 0, 1$) et $\top$ désigne la transposée. En pratique, les lois conditionnelles de $\mathbf{X}$ sont inconnues et les fonctions discriminantes donnent une approximation de la vraie frontière entre les classes. Les paramètres sont naturellement estimés par :

$$\hat{\pi}_{k,\mathcal{N}} = \frac{n_{k,\mathcal{N}}}{n_{\mathcal{N}}},\ \hat{\mu}^j_{k,\mathcal{N}} = \frac{1}{n_{k,\mathcal{N}}} \sum_{i \in R_k} \mathbf{x}^j_i,\ \widehat{\Sigma}^j_{\mathcal{N}} = \frac{1}{n_{\mathcal{N}} - 2} \sum_{k=0}^{1} \sum_{i \in R_k} (\mathbf{x}^j_i - \hat{\mu}^j_{k,\mathcal{N}})(\mathbf{x}^j_i - \hat{\mu}^j_{k,\mathcal{N}})^\top, \tag{2}$$

où $R_k = \{i \leq n : \mathbf{x}_i \in \mathcal{N} \text{ et } y_i = k\}$ et $n_{k,\mathcal{N}} = \text{Card}(R_k)$ et $n_{\mathcal{N}}$ est le nombre d'observations dans le nœud $\mathcal{N}$. L'analyse discriminante linéaire divise $\mathcal{N}$ en deux nœuds fils :

$$\mathcal{N}_0(j) = \{\mathbf{x} \in \mathcal{N} \mid \hat{\delta}^j_{0,\mathcal{N}}(\mathbf{x}) \geq \hat{\delta}^j_{1,\mathcal{N}}(\mathbf{x})\} \quad \text{et} \quad \mathcal{N}_1(j) = \{\mathbf{x} \in \mathcal{N} \mid \hat{\delta}^j_{0,\mathcal{N}}(\mathbf{x}) < \hat{\delta}^j_{1,\mathcal{N}}(\mathbf{x})\},$$

où $\hat{\delta}^j_{k,\mathcal{N}}(\mathbf{x})$ sont les fonctions discriminantes linéaires (1) dans lesquelles les paramètres inconnus ont été remplacés par leur estimateurs (2).

Étape 2 : choix du groupe de coupure. L'étape précédente permet de définir une division du nœud $\mathcal{N}$ pour chaque groupe de variables. Le critère de Gini est alors utilisé pour sélectionner le groupe servant à découper $\mathcal{N}$. Ce critère est défini par :

$$\mathcal{I}(\mathcal{N}) = \pi_{1,\mathcal{N}}(1 - \pi_{1,\mathcal{N}}).$$

Cette fonction est naturellement estimée par :

$$\widehat{\mathcal{I}}(\mathcal{N}) = \hat{\pi}_{1,\mathcal{N}}(1 - \hat{\pi}_{1,\mathcal{N}}).$$

Le groupe j est sélectionné en maximisant la décroissance d'impureté $\Delta_j(\mathcal{N})$ définie par :

$$\Delta_j(\mathcal{N}) = \widehat{\mathcal{I}}(\mathcal{N}) - \left[\frac{n_{\mathcal{N}_0(j)}}{n_{\mathcal{N}}}\hat{\mathcal{I}}(\mathcal{N}_0(j)) + \frac{n_{\mathcal{N}_1(j)}}{n_{\mathcal{N}}}\hat{\mathcal{I}}(\mathcal{N}_1(j))\right].$$

Ce processus de découpage est tout d'abord appliqué sur l'espace $\mathbb{R}^p$ des variables explicatives. Il en résulte la création de deux nœuds. Le processus est ensuite répété sur ces deux nœuds et sur tous les nœuds créés jusqu'à ce qu'un des critères d'arrêt suivants soit satisfait :

— chaque nœud $\mathcal{N}$ est pur (ou presque), c'est-à-dire qu'il ne contient qu'une seule classe d'observations, soit

$$\hat{\pi}_{1,\mathcal{N}} < \epsilon \quad \text{ou} \quad \hat{\pi}_{1,\mathcal{N}} > 1 - \epsilon$$

pour une valeur de ϵ fixée par l'utilisateur ;

— dans chaque nœud $\mathcal{N}$, aucun découpage ne permet de réduire l'impureté, soit

$$\Delta_j(\mathcal{N}) \leq 0, \quad \forall j = 1, \ldots, J.$$

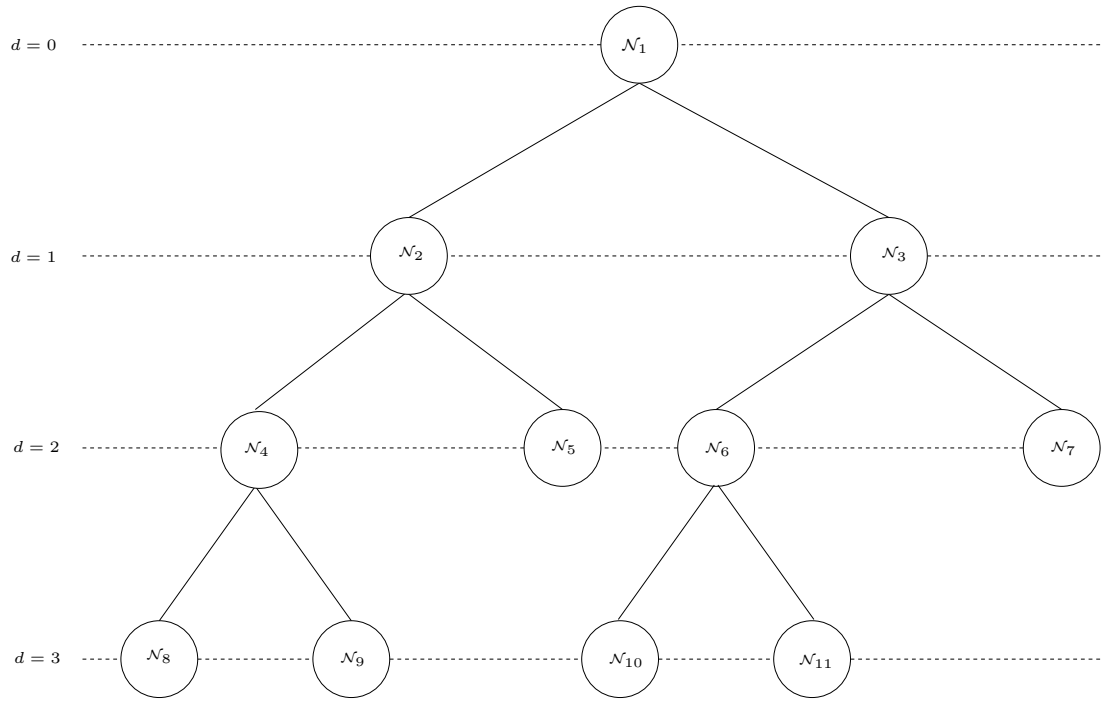

FIGURE 1 – Représentation d'un arbre.

La démarche décrite précédemment conduit à l'élaboration d'un arbre maximal T_{max}. Cet arbre souvent trop complexe n'est généralement pas optimal au sens d'un critère de

performance choisi (l'erreur de classification ou l'aire sous la courbe ROC par exemple). Un nombre excessif de coupures conduit à un arbre qui a tendance à sur-ajuster. Nous proposons donc d'élaguer l'arbre de la manière suivante.

Soit T un sous-arbre de T_{max}, on désigne par $\mathcal{N}_1(T), \ldots, \mathcal{N}_{|T|}(T)$ les nœuds terminaux de T et par $d(\mathcal{N}_\ell(T))$ la profondeur du nœud $\mathcal{N}_\ell(T)$. La profondeur $d(\mathcal{N}_\ell(T))$ de $\mathcal{N}_\ell(T)$ correspond au nombre de conditions qui doivent être satisfaites entre la racine et le nœud $\mathcal{N}_\ell(T)$. La profondeur $D(T)$ d'un arbre T se définit alors comme le maximum des profondeurs des nœuds terminaux de T (voir Figure 1) :

$$D(T) = \max_{\ell=1,\ldots,|T|} d(\mathcal{N}_\ell(T)).$$

La stratégie d'élagage que nous proposons consiste à construire la suite emboîtée de sous-arbres $T_0 \subset T_1 \subset \ldots \subset T_{D(T_{max})} = T_{max}$ de l'arbre maximal T_{max} telle que T_k ($k = 1, \ldots, D(T_{max})$) soit le sous-arbre de T_{max} qui maximise, parmi tous les sous-arbres de T_{max}, la quantité

$$\sum_{\ell=1,\ldots,|T|} d(\mathcal{N}_\ell(T)) \quad \text{sous la contrainte} \quad d(\mathcal{N}_\ell(T)) \leq k.$$

De manière équivalente, T_k est le plus profond des sous-arbres de T_{max} de profondeur k. Le tableau 1 donne la liste des nœuds terminaux associés à chaque arbre de la suite emboîtée définie à partir de l'arbre représenté sur la Figure 1.

Tree	Terminal Nodes
T_0	$\mathcal{N}_0$
T_1	$\mathcal{N}_2, \mathcal{N}_3$
T_2	$\mathcal{N}_4, \mathcal{N}_5, \mathcal{N}_6, \mathcal{N}_7$
T_3	$\mathcal{N}_8, \mathcal{N}_9, \mathcal{N}_5, \mathcal{N}_{10}, \mathcal{N}_{11}, \mathcal{N}_7$

Tableau 1 – Nœuds terminaux des sous-arbres issus de l'arbre représenté sur la Figure 1.

A partir de chaque élément de la séquence $(T_k)_k$ on peut définir une règle de classification

$$g_k(\mathbf{x}) = \begin{cases} 1 & \text{si } \text{card}(\{i : y_i = 1 \text{ et } \mathbf{x}_i \in \mathcal{N}(\mathbf{x}, T_k)\}) \geq \text{card}(\{i : y_i = 0 \text{ et } \mathbf{x}_i \in \mathcal{N}(\mathbf{x}, T_k)\}) \\ 0 & \text{sinon,} \end{cases}$$

où $\mathcal{N}(\mathbf{x}, T_k)$ désigne le nœud terminal de T_k qui contient $\mathbf{x}$. Nous proposons de sélectionner, dans la suite $(T_k)_k$, l'arbre associé à la règle de classification qui minimise l'erreur de classement $\mathbf{P}(g_k(\mathbf{X}) \neq Y)$. Cette erreur est estimée sur l'ensemble de validation $(\mathbf{x}_{n+1}, y_{n+1}), \ldots, (\mathbf{x}_{n+m}, y_{n+m})$, *i.e.* l'algorithme sélectionne la profondeur $\widehat{K}$ vérifiant

$$\widehat{K} \in \operatorname*{argmin}_k \frac{1}{m} \sum_{i=n+1}^{n+m} \mathbf{1}_{g_k(\mathbf{x}_i) \neq y_i}.$$

L'arbre final est alors le sous-arbre $T_{\widehat{K}}$.

Remarque 2.1 *Le découpage successif de l'espace $\mathbb{R}^p$ des variables explicatives entraine la création de nœuds de plus en plus petits. Ainsi, dans certains nœuds, le nombre d'observations peut devenir relativement faible par rapport au nombre de variables contenues dans les groupes. Dans ce contexte, l'analyse discriminante linéaire classique n'est pas performante. Afin de palier à ce "fléau de la dimension", nous proposons de remplacer l'analyse discriminante classique par l'approche dite d'analyse discriminante régularisée proposée par Witten et Tibshirani (2011).*

3 Conclusion

Les performances de l'algorithme seront illustrées et comparées à d'autres méthodes sur plusieurs scénarios de simulations. Des illustrations sur données réelles seront également présentées.

Bibliographie

[1] Meier, L., Van De Geer, S. et Bühlmann, P. (2008), The group lasso for logistic regression, *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 70, 53–71.
[2] Breiman, L., Friedman, J., Stone, C J. et Olshen, R.A. (1984), Classification and regression trees, 1984.
[3] Quinlan, J.R. (1986), Induction of decision trees, *Machine learning*, 1, 81–106.
[4] Loh, W. et Shih, Y. (1997), Split selection methods for classification trees, *Statistica Sinica*, 815–840.
[5] Wickramarachchi, D.C., Robertson, B.L., Reale, M., Price, C.J. et Brown, J (1997), HHCART : An oblique decision tree, *Computational Statistics & Data Analysis*, 96, 12–23.
[6] Witten, D. M. et Tibshirani, R. (2011), Penalized classification using Fisher's linear discriminant, *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 73, 753–772.