

REVUE DE STATISTIQUE APPLIQUÉE

B. GHATTAS

Agrégation d'arbres de classification

Revue de statistique appliquée, tome 48, n° 2 (2000), p. 85-98

http://www.numdam.org/item?id=RSA_2000__48_2_85_0

© Société française de statistique, 2000, tous droits réservés.

L'accès aux archives de la revue « *Revue de statistique appliquée* » (<http://www.sfds.asso.fr/publicat/rsa.htm>) implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques

<http://www.numdam.org/>

AGRÉGATION D'ARBRES DE CLASSIFICATION

B. GHATTAS

Université de la Méditerranée - GREQAM
ghattas@lumimath.univ-mrs.fr

RÉSUMÉ

Après avoir empiriquement mis en évidence l'instabilité des procédures CART de classification par arbre, nous présentons des méthodes d'agrégation de classificateurs obtenues à l'aide d'un rééchantillonnage de type bootstrap. Enfin, nous mettons en œuvre ces procédures sur un problème de prévision quotidienne de pic de pollution en ozone.

Mots-clés : *Bootstap, CART, classificateur agrégé, prévision, rééchantillonnage adaptatif.*

ABSTRACT

After we describe empirically the instability of CART (classification trees) methods, we present two aggregation algorithms for classifiers, based on bootstrap resampling. We apply then these techniques to the problem of daily maximum ozone prediction.

Keywords : *Adaptive resampling, agregating predictors, bootstrap, CART, prediction, Bagging, Arcing classifiers.*

1. Position du problème

Certaines méthodes de prévision statistiques sont *instables* au sens où de petites modifications effectuées sur l'échantillon d'apprentissage peuvent avoir des effets importants sur le prédicteur construit à partir de cet échantillon. Dans un travail récent Breiman (1996a), met en particulier en évidence l'instabilité des arbres de régression et ceux de classification (méthodes CART).

L'idée de combiner des prédicteurs ou des classificateurs instables est ancienne, Breiman (1996b) la met en œuvre en agrégeant des prédicteurs construits sur différents échantillons obtenus en perturbant l'échantillon \mathcal{E} d'apprentissage de référence. Cette première idée conduit dans le cadre des méthodes CART, à la construction d'arbres agrégés par bootstrap (*Bagging predictors*); nous l'avons mise en œuvre avec succès pour les arbres de régression dans le cadre d'une étude de prévision des pics quotidiens d'ozone sur l'aire métropolitaine Marseillaise (Ghattas 1999). L'amélioration des prédicteurs obtenus est significative, et c'est une procédure d'agrégation par bootstrap

qui constitue l'outil de prévision utilisé actuellement par le réseau de surveillance de la qualité de l'air, AIRMARAIX.

Pendant la même période et dans un environnement informatique Freund *et al.* (1995, 1996) proposent un algorithme dans lequel les échantillons perturbés ne sont plus obtenus par bootstrap mais à l'aide d'une procédure de rééchantillonnage adaptatif avec remise, par laquelle les observations mal classées à l'étape k de l'algorithme ont une probabilité plus grande d'être tirées à l'étape $k + 1$ de l'algorithme. Ces auteurs, de même que Breiman (*Arcing classifiers*) mettent en évidence les effets positifs de cette procédure sur quelques exemples de données artificielles. Nous l'utiliserons nous même dans le cadre de la prévision quotidienne de l'ozone; en effet les jours mal prévus sont en général du type «peu fréquent», pour lesquels le niveau d'ozone est important et qu'on doit arriver à détecter. Le rééchantillonnage adaptatif semble pouvoir accorder un poids plus important à ces niveaux peu fréquents.

Dans cet article nous analysons d'abord l'instabilité des arbres construits sur les données relatives au problème de la prévision des pics quotidiens de l'ozone. Ensuite nous exposons deux algorithmes, l'agrégation par bootstrap et le rééchantillonnage adaptatif. Ceux-ci sont analysés en détail en tant qu'outil de classification. Les performances relatives des différents algorithmes sont comparées à partir de plusieurs essais et dans le cas où la variable dépendante est à deux modalités, ce qui suffit pour fournir une réponse au problème posé de la prévision du franchissement du seuil de $180 \mu\text{g}/\text{m}^3$ d'ozone.

2. L'instabilité : une illustration empirique

Les méthodes CART (Breiman *et al.*, 1984) sont maintenant bien connues et ont été présentées dans cette revue par Gueguen *et al.* (1988) pour la classification et Ghattas (1999) pour la régression. Commençons par mettre en évidence le phénomène d'instabilité à l'aide d'un exemple empirique.

L'échantillon \mathcal{E} de base contient $N = 822$ observations du maximum quotidien d'ozone observé à Vitrolles («*maxo3*») – variable à prévoir, qualitative à deux modalités – et de variables météorologiques et de pollution – variables explicatives qui sont décrites dans le paragraphe 5.1.

Nous avons construit 10 échantillons bootstrap (c'est-à-dire 10 échantillons de 822 observations obtenus par tirage *avec remise* dans l'échantillon de base) et sur chaque échantillon nous avons examiné les 10 arbres obtenus par validation croisée (donc tous de taille optimale). D'une part la taille de ces arbres est très variable, d'autre part les arbres sont différents à partir de la racine. Le tableau 1 indique pour les 10 arbres (en colonne) le nom des variables apparaissant au niveau des trois premiers noeuds (en ligne) de chaque arbre.

Deux variables apparaissent au niveau de la racine : «*tc10*» (température à 10 heures) pour 5 arbres et «*maxo3v*» (maximum d'ozone de la veille) pour les cinq autres. Le nombre de variables différentes augmente avec la profondeur de l'arbre. L'examen des noeuds suivants accentue la mise en évidence de cette instabilité même si les coupures sont interprétables.

TABLEAU 1

Les variables figurant aux trois premiers noeuds de 10 arbres construits sur des échantillons bootstrap.

	Arbre1	Arbre2	Arbre3	Arbre4	Arbre5
N1	«tc10»	«maxo3v»	«maxo3v»	«tc10»	«tc10»
N2	«neb10»	«tc10»	«tc10»	«maxo3v»	«maxo3v»
N3	«dv13»	«vv13»	«vv10»	«dv16»	«vv10»

	Arbre1	Arbre2	Arbre3	Arbre4	Arbre5
N1	«maxo3v»	«maxo3v»	»tc10»	«maxo3v»	«tc10»
N2	«tc13»	«dv16»	«vv13»	«tc10»	«maxo3v»
N3	«vv13»	«tc7»	«ray»	«dv10»	«vv10»

TABLEAU 2

Discordance exprimée en pourcentage, entre les prévisions données par les 10 arbres construits sur des échantillons bootstrap.

0	24.8	27.3	24.7	22.8	28.1	26.3	31.6	31.4	20.3
	0	21.4	29.4	28.8	31.4	14.6	32.4	23.8	23.7
		0	23.8	24.2	33.9	27.4	33.6	21.4	22.1
			0	6.6	29.7	26.3	36.6	24.8	14.6
				0	27.5	24.7	33.8	27.5	12.8
					0	28.4	15.8	31.4	20.2
						0	31.8	25.9	21.1
							0	32.9	25.6
								0	24.3
									0

Cette instabilité se manifeste aussi au niveau des prévisions données par ces arbres. Pour montrer cela, on effectue la prévision sur l'ensemble des observations de l'échantillon de construction avec les 10 arbres construits. Le tableau 2 donne la discordance¹ entre les prévisions faites par les différents arbres. La discordance entre les prévisions données par deux arbres construits à partir de deux échantillons bootstrap peut atteindre 37 %.

¹ C'est la proportion d'observations pour lesquelles les prévisions sont différentes.

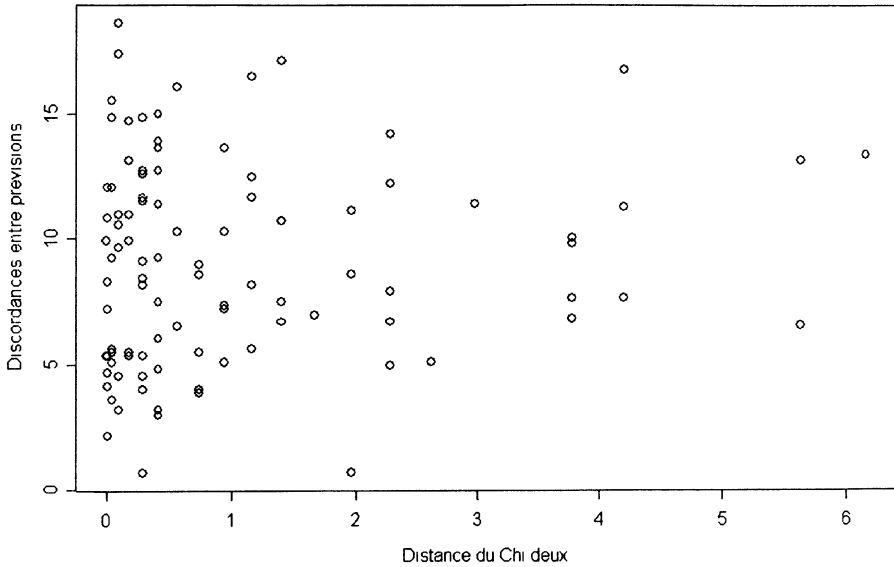


FIGURE 1

En abscisse on présente la distance du Chi deux entre les distributions de la variable de classification dans l'échantillon de base et dans les échantillons bootstrap. En ordonnée, on présente la mesure de discordance entre les prévisions faites à partir des arbres construits sur les échantillons bootstrap et l'arbre construit sur l'échantillon de base.

A partir de \mathcal{E} , nous avons construit cette fois $K = 50$ échantillons bootstrap $(\mathcal{E}_k^b)_{1 \leq k \leq K}$, tous de même taille que \mathcal{E} , obtenus avec remise dans \mathcal{E} . Les échantillons bootstrap sont considérés comme des échantillons obtenus à partir de petites perturbations sur \mathcal{E} , la perturbation étant effectuée sur le tirage et non sur les valeurs.

Sur \mathcal{E} comme sur chacun des \mathcal{E}_k^b , nous avons construit un arbre de classification à 10 feuilles. Nous avons alors calculé d'une part la distance du χ^2 entre les distributions de la variable expliquée dans chacun des \mathcal{E}_k^b et sa distribution dans \mathcal{E} et d'autre part une mesure de la discordance entre les classificateurs construits sur (\mathcal{E}_k^b) et celui construit sur \mathcal{E} . L'examen de la figure 1 montre que la discordance entre les classificateurs n'est pas une fonction croissante de la distance entre les distributions des échantillons bootstrap. A des petits écarts entre les distributions des échantillons perturbés peuvent correspondre des discordances importantes entre les classificateurs.

3. Combinaisons d'arbres de classification

3.1. Définitions

Notons $\mathcal{E} = (\mathbf{x}_n, j_n)_{1 \leq n \leq N}$ l'échantillon d'apprentissage où les j_n sont les observations d'une variable aléatoire Y à valeurs dans $\{1, 2, \dots, J\}$, un ensemble

de classes, et les \mathbf{x}_n les observations du vecteur aléatoire \mathbf{X} à valeurs dans \mathcal{X} . Construisons à l'aide de \mathcal{E} un classificateur (prédicteur) pour Y , par exemple un arbre de classification, noté $Cl(\cdot, \mathcal{E})$. Si \mathbf{x} est observé, Y est prédit par $Cl(\mathbf{x}, \mathcal{E})$. Nous renvoyons le lecteur à différents auteurs concernant la construction des arbres de classification (Breiman *et al.* (1984) par exemple).

Étant donné l'échantillon d'apprentissage \mathcal{E} , notons $P_{\mathbf{X}, Y}$ (resp. $P_{\mathbf{X}}$) la loi de (\mathbf{X}, Y) (resp. la loi marginale de \mathbf{X}) et $P_{\mathbf{x}, Y}$ la loi de Y lorsque $\mathbf{X}=\mathbf{x}$ et posons :

$$Q(j|\mathbf{x}) = P_{\mathcal{E}} [Cl(\mathbf{x}, \mathcal{E}) = j]$$

et

$$P(j|\mathbf{x}) = P[Y = j | X = \mathbf{x}]$$

Définissons le classificateur agrégé $Cl_A(\cdot)$ par :

$$Cl_A(\mathbf{x}) = \arg \max_j Q(j|\mathbf{x})$$

et son taux de mauvais classement par :

$$\tau_{Cl_A}(\mathbf{x}) = 1 - \max_j Q(j|\mathbf{x})$$

Définissons la probabilité de mauvais classement pour \mathbf{x} par $Cl(\cdot, \mathcal{E})$ par :

$$P_{\mathbf{x}, Y} [Cl(\mathbf{x}, \mathcal{E}) \neq Y]$$

et le taux de mauvais classement de \mathbf{x} par $Cl(\cdot, \mathcal{E})$ par l'espérance par rapport à \mathcal{E} de cette probabilité, soit :

$$\begin{aligned} \tau_{Cl}(\mathbf{x}) &= E_Y [P_{\mathbf{x}, Y} [Cl(\mathbf{x}, \mathcal{E}) \neq Y]] \\ &= 1 - \sum_j Q(j|\mathbf{x}) P(j|\mathbf{x}) \end{aligned}$$

Le taux de mauvais classement de $Cl(\cdot, \mathcal{E})$ est alors :

$$\tau_{Cl} = 1 - \int_{\mathcal{X}} \sum_j Q(j|\mathbf{x}) P(j|\mathbf{x}) P_{\mathbf{X}}(d\mathbf{x})$$

3.2. Essais de justification de la procédure d'agrégation

On dispose de trois classificateurs : $Cl(\cdot, \mathcal{E})$, $Cl_A(\cdot)$ et le classificateur de Bayes noté $Cl^*(\cdot)$, défini par :

$$Cl^*(\mathbf{x}) = \arg \max_j P(j|\mathbf{x})$$

Le taux de mauvais classement de Cl^* pour \mathbf{x} est

$$\tau^*(\mathbf{x}) = 1 - \max_j P(j|\mathbf{x})$$

et le taux de mauvais classement de Cl^* est :

$$\tau^* = 1 - \int_{\mathcal{X}} \max_j P(j|\mathbf{x}) P_{\mathbf{X}}(d\mathbf{x})$$

Par définition du classificateur de Bayes, quels que soient \mathbf{x} et Cl :

$$\tau^*(\mathbf{x}) \leq \tau_{Cl}(\mathbf{x})$$

Notons $Sb = \{\mathbf{x} \in \mathcal{X}; Cl_A(\mathbf{x}) = Cl^*(\mathbf{x})\}$ et Sb^c son complémentaire dans \mathcal{X} . C'est donc l'ensemble des \mathbf{x} pour lesquels Cl_A est optimal. Cela dit, même si $\mathbf{x} \in Sb$, la somme $\sum_j Q(j|\mathbf{x}) P(j|\mathbf{x})$ peut être beaucoup plus petite que $\max_j P(j|\mathbf{x})$. Par exemple si $J = 2$, $P(1|\mathbf{x}) = 0.92$, $P(2|\mathbf{x}) = 0.08$, $Q(1|\mathbf{x}) = 0.57$ et $Q(2|\mathbf{x}) = 0.43$ alors $1 - \tau^*(\mathbf{x}) = 0.92$ et $1 - \tau_{Cl}(\mathbf{x}) = 0.56$.

Si $P_{\mathbf{X}}(Sb) \simeq 1$, la procédure d'agrégation conduit à un bon prédicteur. Par contre si $P_{\mathbf{X}}(Sb)$ est petit, l'agrégation peut conduire à la construction d'un mauvais prédicteur.

Afin de formaliser ces remarques, définissons

$$\begin{aligned} \text{Biais}(Cl) &= P_{\mathbf{X},Y} [Cl^*(\mathbf{X}) = Y, \mathbf{X} \in Sb^c] - E_{\mathcal{E}} [P_{\mathbf{X},Y} [Cl(\mathbf{X}, \mathcal{E}) = Y, \mathbf{X} \in Sb^c]] \\ &= \int_{Sb^c} \max_j P(j|\mathbf{x}) P_{\mathbf{X}}(d\mathbf{x}) - \int_{Sb^c} \sum_j Q(j|\mathbf{x}) P(j|\mathbf{x}) P_{\mathbf{X}}(d\mathbf{x}) \end{aligned} \quad (1)$$

$$\begin{aligned} \text{Var}(Cl) &= P_{\mathbf{X},Y} [Cl^*(\mathbf{X}) = Y, \mathbf{X} \in Sb] - E_{\mathcal{E}} [P_{\mathbf{X},Y} [Cl(\mathbf{X}, \mathcal{E}) = Y, \mathbf{X} \in Sb]] \\ &= \int_{Sb} \max_j P(j|\mathbf{x}) P_{\mathbf{X}}(d\mathbf{x}) - \int_{Sb} \sum_j Q(j|\mathbf{x}) P(j|\mathbf{x}) P_{\mathbf{X}}(d\mathbf{x}) \end{aligned} \quad (2)$$

d'où

$$\tau_{Cl} = \tau_{Cl^*} + \text{Biais}(Cl) + \text{Var}(Cl)$$

Sachant que $\sum_j Q(j|\mathbf{x}) P(j|\mathbf{x}) \leq \max_j P(j|\mathbf{x})$ avec égalité seulement si :

$$Q(j|\mathbf{x}) = 1_{\left(P(j|\mathbf{x}) = \max_i P(i|\mathbf{x}) \right)}$$

où $1_{(\cdot)}$ est la fonction indicatrice d'un ensemble, on en déduit que $\text{Biais}(Cl)$ et $\text{Var}(Cl)$ sont tous deux non négatifs; de plus le biais de Cl^* (comme du reste sa variance) est nul.

Enfin par la définition de l'ensemble Sb , $\text{Var}(Cl_A) = 0$, donc la variance est la composante de l'erreur de classification qui est éliminée par agrégation. Cela dit le biais du classificateur agrégé peut être plus grand que celui du classificateur initial. Ces indicateurs sont définis et étudiés dans ce contexte de classification dans Breiman (1998) et Kohavi *et al.* (1996).

En pratique on dispose de l'échantillon \mathcal{E} , et on cherche à obtenir K échantillons de même loi que \mathcal{E} afin d'estimer $Q(j|\mathbf{x})$ et construire le prédicteur agrégé.

Une première idée, due à Breiman (1996b), consiste à construire $(\mathcal{E}_k^b)_{1 \leq k \leq K}$, c'est-à-dire K échantillons bootstrap de taille N obtenus par tirage avec remise dans \mathcal{E} ; chaque observation de \mathcal{E} est donc tirée avec une probabilité $p_k = \frac{1}{N}$.

Une seconde idée proposée par Freund *et al.* (1995, 1996) et reprise par Breiman (1998) consiste à construire des échantillons bootstrap \mathcal{E}_k^b de taille N à l'aide d'une procédure adaptative où les probabilités p_k , sur les observations de \mathcal{E} , ne sont plus égales, mais changent à chaque itération k de l'algorithme. Les probabilités des observations mal classées à l'étape $k - 1$ sont augmentées à l'étape k .

La présentation des deux algorithmes fait l'objet du paragraphe 4.

4. Algorithmes

4.1. Agrégation par bootstrap («Bagging»)

L'algorithme est le suivant :

On tire aléatoirement, avec remise dans \mathcal{E} (l'échantillon d'apprentissage) K échantillons bootstrap \mathcal{E}_k^b de même taille N (pour le tirage, toutes les observations de \mathcal{E} sont équiprobables).

On construit alors un arbre de classification Cl_k sur chaque échantillon bootstrap \mathcal{E}_k^b . L'échantillon \mathcal{E} est utilisé comme échantillon témoin pour déterminer une taille optimale pour l'arbre Cl_k .

On définit le classificateur agrégé par booststrap à partir des K arbres Cl_k de la manière suivante :

$$Cl^b(\mathbf{x}) = \text{Arg max}_j \text{Card} \{k; Cl_k(\mathbf{x}, \mathcal{E}_k^b) = j\}$$

L'estimateur bootstrap de $Q(j|\mathbf{x})$ est donné par :

$$\hat{Q}^b(j|\mathbf{x}) = \frac{1}{K} \text{Card} \{k; Cl_k(\mathbf{x}, \mathcal{E}_k^b) = j\}$$

4.2. Agrégation avec rééchantillonnage adaptatif («Arcing»)

L'algorithme présenté ici est adapté au cas où $J = 2$; Freund *et al.* (1996) proposent des extensions aux cas où $J > 2$ et nous même avons abordé ce sujet. C'est une procédure basée sur un rééchantillonnage adaptatif permettant la réalisation de K échantillons notés $(\mathcal{E}_k^a)_{1 \leq k \leq K}$ à partir de \mathcal{E} de même taille que \mathcal{E} . A l'étape k de l'algorithme, l'échantillon \mathcal{E}_k^a est tiré aléatoirement dans \mathcal{E} suivant une loi $p_k(\cdot)$. Présentons l'algorithme de construction des $(p_k(\cdot))_{1 \leq k \leq K}$:

- Construction de \mathcal{E}_1^a : la loi *initiale* de tirage dans \mathcal{E} est uniforme :

$$p_1(i) = \frac{1}{N}, \quad 1 \leq i \leq N.$$

- Construction de \mathcal{E}_{k+1}^a : la loi de tirage $p_{k+1}(\cdot)$ dans \mathcal{E} à l'étape $k + 1$ de l'algorithme est définie par :

$$p_{k+1}(i) = \frac{p_k(i)\beta_k^{d_k(i)}}{\sum_i p_k(i)\beta_k^{d_k(i)}} \quad (3)$$

où :

$$d_k(i) = 1_{\{Cl(\mathbf{x}_i, \mathcal{E}_k^a) \neq j_i\}}$$

j_i étant rappelons le, la valeur de y pour l'observation i .

($d_k(i) = 1$ si l'observation d'indice i est mal classée par $Cl(\cdot, \mathcal{E}_k^a)$, et vaut 0 sinon)

$$\epsilon_k = \sum_{i=1}^N p_k(i)d_k(i) \quad (4)$$

est la somme des poids des observations de \mathcal{E} mal classées par $Cl(\cdot, \mathcal{E}_k^a)$

et

$$\beta_k = \frac{1 - \epsilon_k}{\epsilon_k}$$

est le rapport «bien classées / mal classées».

Si $\epsilon_k \geq 1/2$ (ϵ_k étant toujours ≤ 1) on remet $p_k(i)_{1 \leq i \leq N}$ à égalité, car un trop grand nombre d'observations de \mathcal{E} , plus de 50 %, sont mal classées.

Si $\epsilon_k = 0$ on remet $p_k(i)_{1 \leq i \leq N}$ à égalité, car β_k est alors non défini.

Après K itérations, on dispose de K échantillons sur chacun desquels on construit un arbre de classification noté $Cl(\cdot, \mathcal{E}_k^a)$. Le classificateur agrégé par procédure adaptative est ici défini par :

$$Cl^a(\mathbf{x}) = \operatorname{Argmax}_{j=1,2} \sum_{k: Cl(\mathbf{x}, \mathcal{E}_k^a)=j} \log(\beta_k) \quad (5)$$

Dans l'équation (3) les poids des observations mal classées sont modifiés : ils sont multipliés par β_k , qui est supérieur à 1, tant que ϵ_k est inférieur à $\frac{1}{2}$.

Remarque 1 Le choix du classificateur agrégé proposé en (5) se justifie par les propriétés de la règle de Bayes en classification. En effet l'inégalité $\sum_{k: Cl(\mathbf{x}, \mathcal{E}_k^a)=2} \log(\beta_k) > \sum_{k: Cl(\mathbf{x}, \mathcal{E}_k^a)=1} \log(\beta_k)$ est équivalente à $P[Y = 2] > P[Y = 1]$.

4.3. Combien de rééchantillonnages ?

Pour justifier le nombre d'échantillons bootstrap utilisé dans l'étude empirique décrite ici, nous avons réalisé l'expérience suivante : construction de 1000 échantillons bootstrap sur un échantillon d'apprentissage \mathcal{E} et des 1000 prédicteurs construits par agrégations successives. Un échantillon test \mathcal{T} stratifié a été utilisé pour évaluer les performances des classificateurs à chaque étape. Deux critères de performance sont présentés : le taux de mauvais classement et le threat score (cf. équation 6). Les performances sont d'autant meilleures que le taux de mauvais classement est bas et le threat score est élevé.

Les résultats numériques obtenus et visualisés sur la figure 2 pour l'agrégation par bootstrap, et sur la figure 3 pour le rééchantillonnage adaptatif, semblent indiquer que l'agrégation d'environ 400 arbres permet d'obtenir une amélioration constante de la qualité de prévision. Dans notre application, la qualité des résultats obtenus avec cinquante arbres nous a paru satisfaisante.

Nous avons observé l'évolution de la performance des agrégations successives aussi sur l'échantillon de construction \mathcal{E} . Comme prévu, les résultats sont meilleurs quand le nombre d'échantillons agrégés augmente. De plus, pour le rééchantillonnage adaptatif, le taux de mauvais classement de l'échantillon de construction converge vers zéro au bout de sept itérations. De même, le threat score converge vers 100 % aussi rapidement.

C'est une des caractéristiques particulières à cet algorithme. La convergence vers des performances maximales (sur l'échantillon d'apprentissage) avec un faible nombre d'itérations.

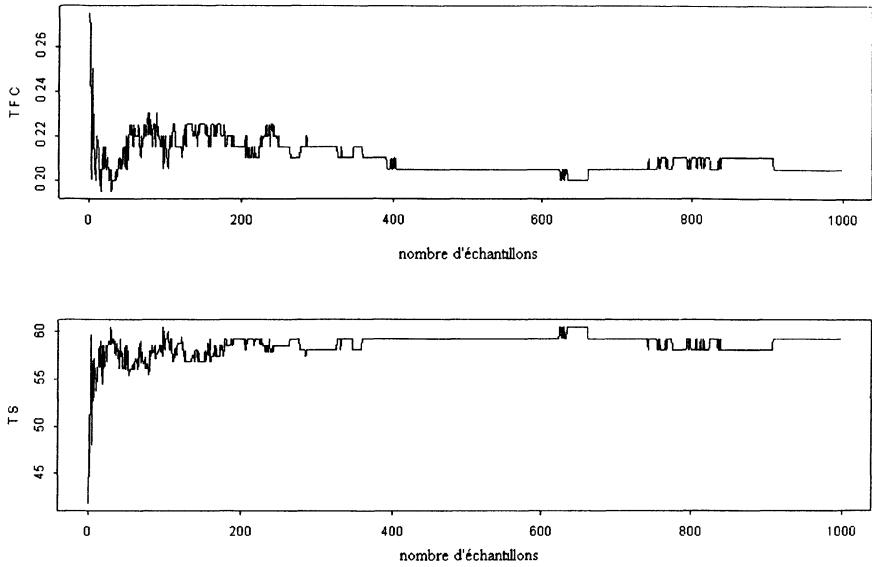


FIGURE 2

Qualité d'un classificateur agrégé en fonction du nombre d'échantillons utilisés (Agrégation par bootstrap). T.F.C. = taux de mauvais classement, T.S.=Threat Score.

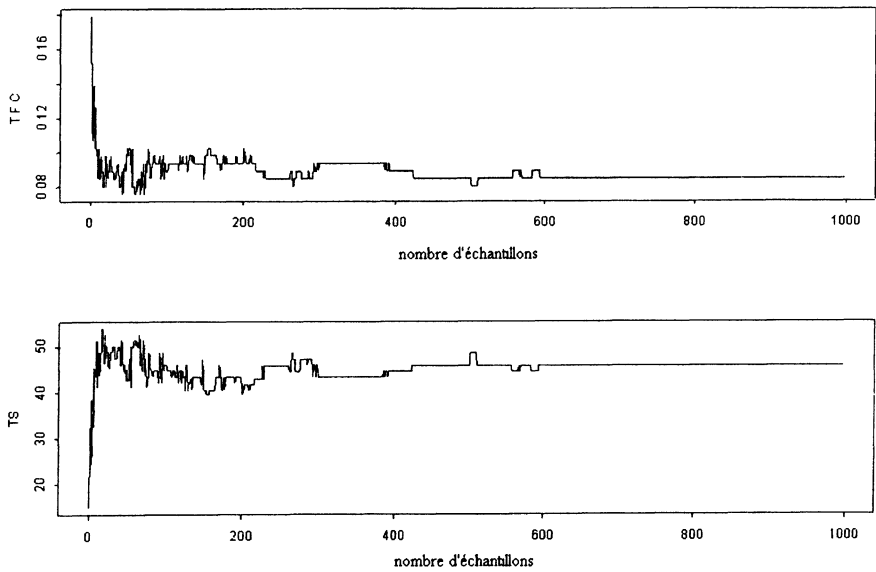


FIGURE 3

Qualité d'un classificateur agrégé en fonction du nombre d'échantillons utilisés (Agrégation avec rééchantillonnage adaptatif). T.F.C. = taux de mauvais classement, T.S.=Threat Score

5. Quelques Résultats

5.1. Le problème étudié et les données utilisées

La présente étude porte sur la prévision quotidienne (le matin) du maximum quotidien des concentrations horaires de l'ozone. Pour cela nous utilisons deux types de variables explicatives.

Les variables de *pollution* : le maximum d'ozone de la veille et les maximums et les minimums des concentrations de polluants (O_3 , NO_2 et SO_2) observées la nuit.

Les variables *météorologiques* (tri-horaires) observées par Météo France à Marignane : vitesse et direction du vent, température, gradient thermique vertical, humidité et nébulosité. Pour plus de détails on peut se reporter à Ghattas (1999).

5.2. Comparaisons des méthodes

Afin d'évaluer et de comparer les performances des différents classificateurs, nous définissons un estimateur du taux de mauvais classement, à l'aide d'un échantillon témoin \mathcal{T} , pour un classificateur quelconque Cl (qui peut être simple ou agrégé) par la formule suivante :

$$\hat{\tau} = \frac{1}{\text{Card}(\mathcal{T})} \sum_{(\mathbf{x}_n, j_n) \in \mathcal{T}} 1_{(Cl(\mathbf{x}_n) \neq j_n)}$$

Pour effectuer ces comparaisons nous avons mis en œuvre la procédure suivante :

1. L'ensemble des données est divisé aléatoirement en deux parties inégales
 - $N_1\%$ de cet ensemble constitue l'échantillon d'apprentissage $\mathcal{E} = (\mathbf{x}_n, j_n)$
 - $N_2\%$ de cet ensemble constitue l'échantillon témoin $\mathcal{T} = (\mathbf{x}_n, j_n)$
 - $N_1 + N_2 = 100$
2. Un arbre A est construit par validation croisée sur \mathcal{E} . Le taux de mauvais classement τ de ce classificateur $Cl(\cdot, \mathcal{E})$ est estimé à l'aide de l'échantillon témoin \mathcal{T} .
3. Pour la procédure d'agrégation par bootstrap Cl^b , on construit les arbres de classifications associés aux 50 échantillons générés à partir de \mathcal{E} suivant l'algorithme décrit en 4.1. La taille de chaque arbre est optimale vis-à-vis du taux de mauvais classement des éléments de \mathcal{E} . Le taux de mauvais classement du classificateur agrégé Cl^b est noté $\hat{\tau}_b$.
4. Pour la procédure d'agrégation adaptative Cl^a , on construit les arbres de classifications associés aux 50 échantillons bootstrap générés à partir de \mathcal{E} . La taille de chaque arbre est optimale vis-à-vis du taux de mauvais classements des éléments d'un échantillon de même loi que \mathcal{E}_k^a . Le taux de mauvais classement du classificateur agrégé Cl^a est noté $\hat{\tau}_a$.
5. Les échantillons \mathcal{E} et \mathcal{T} sont construits $H = 100$ fois comme en 1. Pour chacun de ces couples \mathcal{E}_h et \mathcal{T}_h , on calcule les taux de mauvais classement :

$\hat{\tau}_h, \hat{\tau}_{b_h}, \hat{\tau}_{a_h} \quad 1 \leq h \leq H$. On calcule aussi les moyennes et les écarts types empiriques de ces trois distributions de taux de mauvais classement.

Ces comparaisons ont été réalisées sur les données issues de deux sites de mesure de l'ozone «VTRL» (Vitrolles) et «RBRT» (Rognac).

Outre le taux de mauvais classement nous utilisons le critère du «threat score» noté TS , afin d'évaluer la performance des méthodes utilisées. Son calcul est basé sur le tableau de croisement entre les observations et les prévisions. Si on note a_{11} (resp. a_{22}) le nombre de bonnes prévisions pour le premier (resp. deuxième) niveau, a_{12} le nombre de fausses alertes et a_{21} le nombre d'alertes manquées, alors :

$$TS = \frac{100 * a_{22}}{a_{22} + a_{12} + a_{21}} \quad (6)$$

C'est un critère très employé en météorologie (Bel *et al.*, 1999), plus il est grand meilleure est estimée la qualité de la classification considérée.

Les résultats sont donnés dans le tableau 3.

TABLEAU 3

Comparaisons des trois méthodes : arbre obtenu par validation croisée, arbres agrégés par bootstrap et rééchantillonnage adaptatif. La comparaison est faite sur deux sites Vitrolles (VTRL) et Rognac (RBRT), en terme des deux indicateurs : taux de mauvais classement (TFC) et threat score (TS).

		Validation Croisée	Agrégation par bootstrap	Rééchantillonnage adaptatif
		$\hat{\tau}$	$\hat{\tau}_b$	$\hat{\tau}_a$
VTRL	moyenne	0.104	0.079	0.081
	écart-type	0.022	0.019	0.018
RBRT	moyenne	0.072	0.063	0.062
	écart-type	0.017	0.017	0.016
		\widehat{TS}	\widehat{TS}_b	\widehat{TS}_a
VTRL	moyenne	30.978	39.950	41.702
	écart-type	9.778	9.045	8.594
RBRT	moyenne	14.034	22.251	26.710
	écart-type	12.951	9.961	10.983

On remarque tout d'abord que les valeurs moyennes du taux de mauvais classement des classificateurs agrégés sont plus basses que celles des arbres simples obtenus par validation croisée. D'autre part, les valeurs du threat score pour les arbres agrégés sont plus élevées que celles des arbres simples. Ceci est valable pour les deux stations de mesure, et montre que la procédure d'agrégation permet d'avoir un outil de classification plus performant.

De plus les écart types des moyennes obtenues pour les deux indicateurs (threat score et taux de mauvais classement) sont généralement plus faibles que ceux observés pour les arbres simples. Ceci montre en plus, une plus faible variabilité des critères de performance, donc plus de stabilité dans les résultats de classification par agrégation d'arbres.

Les écart types sont dans l'ensemble plus faibles pour le rééchantillonnage adaptatif que pour l'agrégation par bootstrap, sauf pour la deuxième station, pour le threat score. Les taux de mauvais classement sont assez voisins pour les deux techniques d'agrégation, alors que le threat score est toujours plus élevé avec le rééchantillonnage adaptatif. Ceci est dû au fait que le threat score est basé sur les bonnes prévisions des niveaux élevés, qui sont rares et difficiles à classer en général. Le rééchantillonnage adaptatif semble améliorer la qualité de prévision de ces observations.

6. Conclusion et perspectives

L'agrégation d'arbres de classification présente un double avantage par rapport à l'utilisation d'un seul arbre de classification : une meilleure performance sur des échantillons témoins et une stabilité dans les résultats de prévision. Le rééchantillonnage adaptatif semble être particulièrement mieux adapté aux problèmes où l'on dispose d'observations rares et donc difficiles à prévoir. Cependant, la structure de l'arbre de classification n'est plus aussi simple à exploiter (les arbres sont en trop grand nombre) et donc on perd l'aspect visuel de la technique CART.

D'autre part nous nous sommes intéressés ici seulement au cas où la variable de classification est à deux modalités. L'algorithme utilisé dans ce papier est en cours d'amélioration et d'adaptation au cas où le nombre de modalités est supérieur à deux.

Les programmes permettant de réaliser cette étude ont été écrits avec le logiciel S+.

Remerciements

Je tiens à remercier les deux referees pour leurs critiques et suggestions constructives, ainsi que Claude Deniau pour ses conseils.

Références

- [1] BEL L., BELLANGER L., BONNEAU V., CIUPERCA G., DACUNHA-CASTELLE D., DENIAU C., GHATTAS B., MISITI M., MISITI Y., OPPENHEIM G., POGGI J.-M., TOMASSONE R. (1999) Eléments de comparaison de prévisions statistiques des pics d'ozone. À paraître dans la *Revue de Statistique Appliquée*.
- [2] BREIMAN L., FRIEDMAN J.H., OLSHEN R., STONE C.J. (1984) *Classification And Regression Trees*, Wadsworth, Belmont CA.
- [3] BREIMAN L., Heuristic of instability and stabilization in model selection, (1996a) *The Annals of Statistics*, Vol 24, N°6, pp. 2350–2383
- [4] BREIMAN L. (1996b) Bagging Predictors, *Machine Learning*, 24, pp. 123–140
- [5] BREIMAN L. (1998) Arcing Classifiers. *The Annals of Statistics*, Vol 26, N°3, pp. 801–849
- [6] FREUND, Y. and SCHAPIRE R. (1995) A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences* 55 (1) : 119–139.
- [7] FREUND, Y. and SCHAPIRE R. (1996) Experiments with a new boosting algorithm, *Machine Learning : Proceedings of the Thirteenth International Conference*, pp. 148–156
- [8] GHATTAS B. (1999) Prévisions des pics d'ozone par arbres de régression simples et agrégés par bootstrap. *Revue de Statistique Appliquée* 47, 2, 61–80.
- [9] GUEGUEN A., J.-P. NAKACHE (1988), Méthode de discrimination basée sur la construction d'un arbre binaire, *Revue de Statistique Appliquée* 36, 1, 19–37.
- [10] KOHAVIR. and WOLPERT D.H. (1996) Bias Plus Variance Decomposition for Zero-One Loss Functions, *Machine Learning : Proceedings of the Thirteenth International Conference*, pp. 275–283.