Etude de l'influence des paramètres sur les performances des forêts aléatoires

Simon Bernard, Laurent Heutte, Sébastien Adam

Etude de l'influence des paramètres sur les performances des Forêts Aléatoires

Simon Bernard – Laurent Heutte – Sébastien Adam

Université de Rouen, LITIS EA 4108
BP 12 - 76801 Saint-Etienne du Rouvray, France.

{simon.bernard,laurent.heutte,sebastien.adam}@univ-rouen.fr

Résumé : *Dans cet article nous présentons nos travaux sur la paramétrisation des Forêts Aléatoires (RF pour Random Forests), et plus précisément sur la paramétrisation de l'algorithme d'induction Forest-RI introduit par Breiman en 2001. Nous nous sommes particulièrement intéressés au nombre K de caractéristiques sélectionnées aléatoirement pour le partitionnement de chaque nœud des arbres de décision.*

***Mots Cléfs** : Ensemble de Classifieurs, Combinaison de Classifieurs, Forêts Aléatoires, Arbres de Décision.*

1 Introduction

Les Forêts Aléatoires (RF pour Random Forests) forment une famille de méthodes de classification, basées sur une combinaison d'arbres de décision que l'on note $\{h(x, \Theta_k),\ k = 1, ...L\}$, où $\{\Theta_k\}$ est une famille de vecteurs aléatoires, indépendants et identiquement distribués, et où x représente une donnée à classer. La particularité de cet ensemble est que chacun de ces arbres de décision est construit à partir d'un vecteur aléatoire de paramètres. La définition que Breiman donne dans [BRE 01] est délibérément générique pour ne pas contraindre la nature de ces paramètres. Ainsi une Forêt Aléatoire peut être induite par exemple via un tirage aléatoire des caractéristiques qui définissent l'espace de description des données d'apprentissage, ou encore via un tirage aléatoire des données d'apprentissage utilisées pour chaque classifieur de base.

Depuis qu'elles ont été introduites en 2001, les Forêts Aléatoires ont fait l'objet de plusieurs études prospectives et comparatives [BER 07, BRE 01, GEU 06, ROD 06]. Elles se sont montrées compétitives face au Boosting [BRE 01, ROD 06], réputé pour être un des principes d'apprentissage les plus efficaces [BRE 01, KUN 04]. Pourtant les mécanismes qui expliquent le bon fonctionnement de ce principe de génération d'ensembles de classifieurs basés sur l'aléatoire ne sont à ce jour toujours pas clairement identifiés ; et bien que plusieurs hyperparamètres puissent être utilisés pour modifier le comportement des Forêts Aléatoires, il n'existe pas à notre connaissance dans la littérature d'études pragmatiques qui examinent en détail leur influence sur les performances. Par exemple, s'agissant de l'algorithme de référence appelé Forest-RI, introduit par Breiman dans [BRE 01], un hyperparamètre important a clairement été identifié : le nombre K de caractéristiques sélectionnées aléatoirement pour le partitionnement de chaque nœud au cours du processus d'induction des arbres. Les valeurs de cet hyperparamètre sont toujours arbitrairement ou empiriquement choisies, et parfois sans qu'aucune justification théorique ou expérimentale ne soit fournie [BRE 01, GEU 06].

C'est pourquoi nous proposons d'étudier l'influence de cet hyperparamètre sur les performances en classification, dans le but de distinguer de façon expérimentale des heuristiques de paramétrisation, notamment en fonction des caractéristiques du problème telles que la taille de la base d'apprentissage, le nombre de classes ou encore la dimension l'espace de description. L'idée de ces travaux est à terme de pouvoir construire des forêts aléatoires "optimales", en fonction des caractéristiques intrinséques des données à classer [BER 08].

2 Expérimentations

Le principe de ces expérimentations est d'étudier les performances des RF en fonction de la valeur du paramètre K. Comme nous l'avons mentionné précédemment, nous avons testé l'algorithme Forest-RI sur la base de données de reconnaissance de chiffres manuscrits MNIST au cours d'une première série d'expérimentations, présentée dans [BER 07]. Nous avions ainsi réussi à percevoir le potentiel de cette nouvelle méthode et à avoir une première idée des performances globales en fonction du paramètre K. Un intervalle de valeurs a été trouvé qui ne contient ni $K = 1$ ni $K = M$, et pour lequel on obtient des performances optimales. Nous avions alors conjecturé que les bornes de cet intervalle dépendent de la dimension de l'espace de description. Cependant ces premières conclusions nécessitent d'être confirmées en répétant le même protocole expérimental sur plusieurs bases de données différentes. Nous avons donc ajouté à la base MNIST deux autres bases de données se rapportant à la problèmatique de reconnaissance d'écriture manuscrite, mais présentant des caractéristiques différentes.

Ces caractéristiques sont résumées dans le tableau 1. Les bases Letter et Pendigits sont deux bases de données de l'UCI repository [ASU 07] ; la première concerne un problème de reconnaissance de lettres manuscrites et la deuxième un problème de reconnaissance de chiffres manuscrits. La troisième qui s'ajoute à ces deux premières est donc la base MNIST pour laquelle 84 valeurs moyennes de niveau de gris ont été extraites comme expliqué dans [BER 07].

Dans un premier temps, chacune des bases de données a été découpée aléatoirement en deux sous-ensembles de don-

TAB. 1 – Description des bases de données

Base	taille	Caract.	Classes
Letter	20000	16	26
Pendigits	10992	16	10
Mnist	60000	84	10

nées d'apprentissage et de test, le premier contenant deux tiers des données de la base et le deuxième le tiers restant. Ce découpage a été répété 50 fois pour les bases Letter et Pendigits, et 20 fois pour la base MNIST. Nous désignons par $T_i = (Tr_i, Ts_i)$ un tel découpage, avec $i \in [1, 50]$ (ou $i \in [1, 20]$ pour MNIST), et où Tr_i et Ts_i représentent respectivement le sous-ensemble d'apprentissage et le sous-ensemble de test. Ensuite pour chacun des T_i, l'algorithme Forest-RI est lancé pour chaque valeur successive de K, de sorte à obtenir un classifieur différent pour chaque entier $K \in [1, M]$ — seulement $\frac{M}{2} + 1$ forêts ont été construites pour la base de données MNIST, c'est à dire pour des valeurs de $K \in \{1, 2, 4, 6, ..., M\}$.

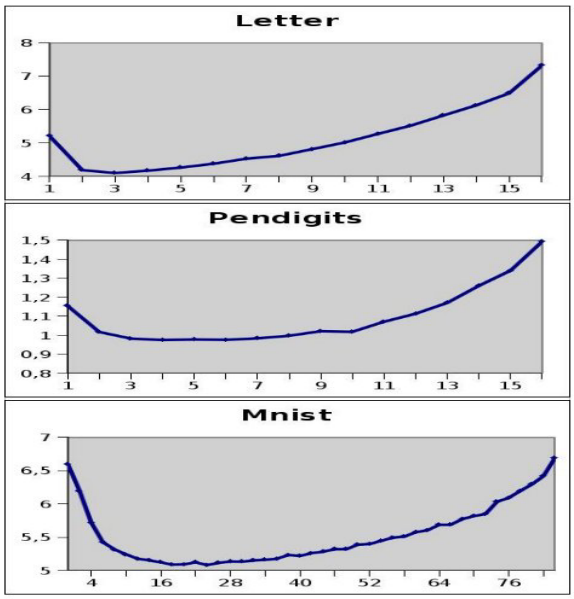
La figure 1 présente les diagrammes des taux d'erreur en fonction de la valeur de K pour les trois bases de données testées. Si on s'intéresse dans un premier temps aux valeurs de K communément utilisées dans la littérature, *i.e.* $K = 1$, $K = \sqrt{M}$ et $K = integer(log_2 M + 1)$, on constate qu'aucune d'elles ne permet d'obtenir le meilleur taux d'erreur en classification. On note cependant que la valeur $K = \sqrt{M}$ semble être un meilleur compromis que la valeur $K = log_2 M + 1$, puisque c'est celle qui se rapproche le plus de la valeur "optimale" obtenu, que l'on note K^*.

Ces courbes permettent également de remarquer que ces taux suivent la même tendance pour un K allant de 1 à M : une première chute rapide des taux d'erreur pour K immédiatement supérieur à 1, puis une stabilisation de cette diminution dans une zone de minima, correspondant à un intervalle borné de valeurs de K ; et enfin une augmentation plus lente de ces taux jusqu'à atteindre la valeur maximale pour $K = M$.

3 Conclusions

Les expérimentations sur la paramétrisation des RF présentées dans cet article ont permis de constater que les valeurs par défaut de l'hyperparamètre K communément utilisées dans la littérature, ne permettent pas d'atteindre les meilleurs performances en généralisation. Il a été possible avec ces expériences d'appréhender le comportement des forêts aléatoires vis à vis de ce paramètre K, et notamment de constater que les valeurs particulières $K = 1$ et $K = M$ n'étaient pas non plus conseillées. Bien qu'aucune régle n'ait pu être établie qui permettrait de déterminer la valeur de K^*, il est possible de percevoir des intervalles de valeurs optimales, dont les bornes semblent dépendre du nombre et de la nature des caractéristiques. Il faut également préciser que bien que ces expérimentations aient été menées sur plusieurs bases de données, elles n'en concernent pas moins exclusivement des problématiques de reconnaissance d'écriture manuscrite, ce qui ne permet donc pas de généraliser ces conclusions à tout type de problémes d'apprentissage automatique. Il serait donc intéressant pour de prochaines expérimentations de sélectionner un ensemble de bases de données, concernant des problématiques d'apprentissage automatique plus variées et présentant des spécificités plus marquées. Cela permettrait d'étendre nos conclusions à d'autres contextes applicatifs et de savoir s'il est possible expérimentalement d'établir une régle générale de paramétrisation permettant de déterminer la valeur K^* en fonction de ces spécificités.

FIG. 1 – Taux d'erreur en fonction des valeurs de K.



Références

[ASU 07] ASUNCION A., NEWMAN D., UCI Machine Learning Repository, 2007.

[BER 07] BERNARD S., HEUTTE L., , ADAM S., Using Random Forests for Handwritten Digit Recognition, *International Conference on Document Analysis and Recognition*, , 2007, pp. 1043–1047.

[BER 08] BERNARD S., HEUTTE L., , ADAM S., Etude de l'influence des paramètres sur les performances des Forêts Aléatoires, *Technical report, Université de Rouen*, , 2008.

[BRE 01] BREIMAN L., Random Forests, *Machine Learning*, vol. 45, n° 1, 2001, pp. 5–32.

[GEU 06] GEURTS P., ERNST D., , WEHENKEL L., Extremely Randomized Trees, *Machine Learning*, vol. 36, n° 1, 2006, pp. 3–42.

[KUN 04] KUNCHEVA L., *Combining Pattern Recognition. Methods and Algorithms*, John Wiley and Sons, 2004.

[ROD 06] RODRIGUEZ J., KUNCHEVA L., , ALONSO C., Rotation Forest : A New Classifier Ensemble Method, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, n° 10, 2006.