Arbres de décision k-means

Célia Châtel, François Brucker, Pascal Préa

Aix-Marseille Université - LIF - École Centrale Marseille
{celia.chatel, francois.brucker, pascal.prea}@lif.univ-mrs.fr

Mots-clés : *classification, arbre de décision, k-means, système totalement équilibré*

1 Introduction

Les arbres de décision (Breiman [2]) sont un modèle très utilisé en apprentissage automatique supervisé. Dans ce travail, nous appliquons des idées de classification non supervisée (clustering) à la construction d'arbres semblables à des arbres de décision. Les arbres sont ainsi construits en utilisant uniquement des critères de distance et non de pureté des classes produites. Nous présentons les arbres de décision obtenus et un critère de performance.

2 Méthode

L'Algorithme (1) détaille la construction d'un arbre de k-means (KMDT). Le système de classes obtenu est une hiérarchie et reprend le principe des k-means hiérarchiques (Böcker [1]). L'utilisation du modèle ainsi construit pour une tâche de prédiction consiste à faire descendre les exemples dans la structure à la manière d'un arbre de décision : l'exemple descend dans le noeud fils dont il est le plus proche du centre de gravité.

On calcule au maximum n k-means, mais en pratique beaucoup moins. Un arbre de décision classique contient moins de noeuds mais doit à chaque noeud optimiser le découpage selon toutes les variables et peut tester de nombreux découpages en cas de variables réelles.

```
Créer un noeud racine contenant tous les exemples
Créer une file F contenant la racine
tant que la file F est non vide faire
    Dépiler de F un noeud
    si tous les exemples que le noeud contient ne sont pas de la même classe alors
        Calculer un k-means avec k = 2
        Créer un noeud fils du noeud courant pour chaque cluster obtenu
        Empiler les noeuds fils dans F
    fin
fin
```

Algorithme 1 : Algorithme de construction d'un arbre de k-means

Les modèles les plus efficaces basés sur des arbres de décision classiques sont les forêts d'arbres aléatoires (Breiman [3]) construits à partir de l'idée de sous-espaces aléatoires (Ho [5]). Les KMDT peuvent être rendus aléatoires grâce à la même idée en appliquant la division de chaque noeud par le k-means à partir d'un nombre limité de dimensions. La prédiction se fait ensuite en faisant voter tous les arbres obtenus.

3 Résultats

Les résultats moyens obtenus par dix répétitions de validation croisée 10-folds apparaissent dans la Table 1 comparés à des arbres classiques, des méthodes d'ensembles d'arbres et à un k-

	n	d	DT	KMDT	RF	ET	KMRF	KN
breast-cancer	683	9	0.9198	**0.9357**	0.9541	0.9537	**0.9546**	0.9601
digits	1797	64	0.8679	**0.9495**	0.9607	0.9682	**0.9789**	0.987
iris	150	4	0.9443	**0.9522**	0.945	0.9523	**0.9584**	0.9639
ringnorm	7400	20	**0.8976**	0.8252	0.954	**0.9638**	0.9056	0.6902
satellite	6435	36	0.8578	**0.8784**	0.9097	0.9083	**0.9099**	0.9087
spambase	6301	57	**0.9129**	0.7735	0.9461	**0.947**	0.9295	0.7978
vehicle	846	18	**0.7166**	0.6067	0.7495	**0.7603**	0.7134	0.6495
waveform	5000	21	0.7565	**0.7899**	0.8218	0.8239	**0.8362**	0.8197
zoo	101	16	0.8773	**0.908**	0.8993	0.9037	**0.9263**	0.782

TAB. 1 – Moyenne des scores de précision sur les ensembles de test pour différentes méthodes :
— arbres : DT : arbre de décision (Breiman [2]), KMDT : arbre de décision k-means
— ensembles : RF : forêt d'arbres (Breiman [3]), ET : extra-arbres (Geurts [4]), KMRF : forêt d'arbres k-means,
— autre : KN : k-plus-proches-voisins
et différents ensembles de données avec n le nombre d'exemples et d la dimension.

plus-proches-voisins, tous avec optimisation des paramètres. Les résultats obtenus sur certains ensembles par les KMDT dépassent de plusieurs points ceux des arbres de décision classiques.

L'un des critères déterminants pour la qualité du modèle est sa taille. Quand la taille du KMDT dépasse considérablement (4-6 fois) celle d'un DT et est de l'ordre (en nombre de noeuds total) du nombre d'exemples de la base de données, le modèle tend à être moins efficace en prédiction (ringnorm, spambase, vehicle). C'est également sur ces ensembles que la méthode des k-plus-proches-voisins est la moins efficace, reflétant une structure des données moins décisive. Dans les autres ensembles de données testés, les KMDT sont entre 1.3 et 4 fois plus grands que les DT, ce qui est inévitable puisqu'ils s'arrêtent quand ils atteignent des feuilles pures sans optimiser ce critère pour la construction, contrairement aux DT. La formation d'ensembles permet d'améliorer les performances mais moins que dans le cas des DT.

4 Conclusions et perspectives

La méthode permet, lorsque les classes attendues sont adaptées aux k-means (i.e "rondes") d'obtenir de bons résultats en apprentissage et un critère de taille du modèle permet de prédire son efficacité. Les résultats encourageants de ce premier travail sur des modèles de classification structurelle pour l'apprentissage automatique ouvrent la voie pour l'utilisation de modèles plus généraux comme les systèmes de classes totalement équilibrés (Lehel [6]) et vers la généralisation des arbres de décision.

Références

[1] Alexander Böcker, Swetlana Derksen, Elena Schmidt, Andreas Teckentrup and Gisbert Schneider. A hierarchical clustering approach for large compound libraries *J. Chem. Inf. Model.*, 45 :807–815, Janvier 2005.

[2] Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. *Classification and Regression Trees.*. Wadsworth and Brooks, Monterey, CA, 1984.

[3] Leo Breiman. Random forests. *Mach. Learn.*, 45(1) :5–32, Octobre 2001.

[4] Pierre Geurts, Damien Ernst, Louis Wehenkel. Extremely Randomized Trees. *Mach. Learn.*, 63(1) :3–42, Avril 2006.

[5] Tin Kam Ho. The random subspace method for constructing decision forests. *IEEETrans. Pattern Anal. Mach. Intell.*, 20(8) :832–844,1998.

[6] Jenö Lehel. A characterization of totally balanced hypergraphs. *Discrete Math.*, 57(1–2), 59–65, Novembre 1985.