Construction de noyaux pour l'apprentissage supervisé à partir d'arbres aléatoires

Vincent Pisetta*, Pierre-Emmanuel Jouve**
Djamel.A Zighed***

*RITHME, 59 bd Vivier Merle 69003 Lyon
vpisetta@rithme.eu,
**FENICS, 59 bd Vivier Merle 69003 Lyon
pjouve@fenics.com
***Laboratoire ERIC, 5 av. Pierre Mendès France 69500 Bron
abdelkader.zighed@univ-lyon2.fr

Résumé. Nous montrons qu'un ensemble d'arbres de décision avec une composante aléatoire permet de construire un noyau efficace destiné à l'apprentissage supervisé. Nous étudions théoriquement les propriétés d'un tel noyau et montrons que sous des conditions très souvent rencontrées en pratique, il existe une séparabilité linéaire entre exemples de classes distinctes dans l'espace induit par celui-ci. Parallèlement, nous observons également que le classique *vote à la majorité* d'un ensemble d'arbres est un hyperplan (sans garantie d'optimalité) dans l'espace induit par le noyau. Enfin, comme le montrent nos expérimentations, l'utilisation conjointe d'un ensemble d'arbres et d'un séparateur à vaste marge (SVM) aboutit à des résultats extrêmement encourageants.

1 Introduction

Parmi les techniques d'apprentissage statistique les plus performantes se trouvent les méthodes à base de noyaux dont le représentant le plus célèbre est très certainement le Séparateur à Vaste Marge (Vapnik, 1996). Son emploi, motivé initialement par les premiers résultats théoriques de l'apprentissage statistique s'est encore plus largement répandu suite aux nombreux succès empiriques apportés par cet apprenant (Pavlidis et al. (2004); Markowska-Kaczmar et Kubacki (2005); Polat et Günes (2007)). L'une des particularités du SVM est d'utiliser la fameuse *astuce du noyau* pour introduire de la non-linéarité dans la frontière de décision sans augmenter la complexité algorithmique de l'apprentissage. Ainsi, le choix du noyau est sans doute le paramètre le plus important de l'algorithme, différents noyaux pouvant aboutir à des résultats très différents.

Plusieurs indicateurs ont été proposés dans la littérature afin d'évaluer la qualité d'un noyau a priori, autrement dit avant même l'exécution de l'apprentissage. Le plus connu est probablement le *Kernel Target Alignment* (KTA) (Cristianini et al., 2002), bien que d'autres tels que FSM (Nguyen et Ho, 2008) ou la *polarisation* (Baram., 2005) aient également montré des résultats intéressants. L'intérêt de ces indicateurs est qu'ils permettent d'évaluer la pertinence