Construction de descripteurs pour classer à partir d'exemples bruités

Nazha Selmaoui*, Dominique Gay*, Jean-François Boulicaut**

*Université de la Nouvelle-Calédonie, ERIM EA3791, PPME EA3325
BP R4 98851 Nouméa, Nouvelle-Calédonie
{nazha.selmaoui, dominique.gay}@univ-nc.nc
**Université de Lyon, CNRS
INSA-Lyon, LIRIS UMR5205, 69621 Villeurbanne, France
jean-francois.boulicaut@insa-lyon.fr

Résumé. En classification supervisée, la présence de bruit sur les valeurs des descripteurs peut avoir des effets désastreux sur la performance des classifieurs et donc sur la pertinence des décisions prises au moyen de ces modèles. Traiter ce problème lorsque le bruit affecte un attribut classe a été très étudié. Il est plus rare de s'intéresser au bruit sur les autres attributs. C'est notre contexte de travail et nous proposons la construction de nouveaux descripteurs robustes lorsque ceux des exemples originaux sont bruités. Les résultats expérimentaux montrent la valeur ajoutée de cette construction par la comparaison des qualités obtenues (e.g., précision) lorsque l'on utilise les méthodes de classification à partir de différentes collections de descripteurs.

1 Introduction

Lorsqu'il s'agit de décrire un ensemble d'objets au moyen de descripteurs, les valeurs de ces derniers peuvent être collectées de façon plus ou moins fiable, par exemple lorsqu'elles sont le résultat d'un processus complexe d'acquisition de mesures. En classification supervisée, nous savons que la présence de bruit dans les exemples d'apprentissage peut avoir un impact négatif sur la performance des modèles construits et donc sur la pertinence des prises de décisions associées. Il existe deux types de problèmes de bruits. Le problème du *bruit de classe* (affectant uniquement l'attribut classe) a été très étudié ces dernières années. Plusieurs approches ont été proposées pour, par exemple, l'élimination, la correction du bruit (Zhu et Wu, 2004), ou encore la pondération des instances (Rebbapragada et Brodley, 2007). Le contexte du *bruit d'attributs* affectant uniquement les attributs non-classe ou descripteurs est moins traité. Nous trouvons des travaux sur la modélisation et l'identification du bruit (Kubica et Moore, 2003; Zhang et Wu, 2007) ainsi que des techniques de filtrage pour "nettoyer" les attributs bruités (Zhu et Wu, 2004; Yang et al., 2004).

Nous nous intéressons à ce problème de la classification en présence de descripteurs (attributs non classe) bruités. Plus précisément, nous voulons apporter une réponse à la question suivante : *comment construire des modèles prédictifs robustes à partir de données dont les attributs Booléens sont a priori bruités ?*