Une nouvelle approche du *boosting* face aux données bruitées

Emna Bahri*
Mondher Maddouri**

* Laboratoire Eric, Université Lyon 2, 5 avenue Pierre Mendes France, 69676 Bron Cedex
Emna.Bahri@univ-lyon2.fr,
http://eric.univ-lyon2.fr
**INSAT, zone urbaine la charguia II Tunis, 1002 Tunisie
Mondher.Maddouri@fst.rnu.tn,
http://www.insatech.net

Résumé. La réduction de l'erreur en généralisation est l'une des principales motivations de la recherche en apprentissage automatique. De ce fait, un grand nombre de travaux ont été menés sur les méthodes d'agrégation de classifieurs afin d'améliorer, par des techniques de vote, les performances d'un classifieur unique. Parmi ces méthodes d'agrégation, le *boosting* est sans doute le plus performant grâce à la mise à jour adaptative de la distribution des exemples visant à augmenter de façon exponentielle le poids des exemples mal classés. Cependant, en cas de données fortement bruitées, cette méthode est sensible au surapprentissage et sa vitesse de convergence est affectée. Dans cet article, nous proposons une nouvelle approche basée sur des modifications de la mise à jour des exemples et du calcul de l'erreur apparente effectuées au sein de l'algorithme classique d'*AdaBoost*. Une étude expérimentale montre l'intérêt de cette nouvelle approche, appelée Approche Hybride, face à *AdaBoost* et à BrownBoost, une version d'*AdaBoost* adaptée aux données bruitées.

1 Introduction Générale

L'émergence des bases de données modernes qui présentent d'énormes capacités de stockage et de gestion, associée à l'évolution des systèmes de transmission et des techniques d'acquisition automatique des données contribuent à la construction d'une masse de données qui dépasse de loin les capacités humaines à les traiter. Ces données sont des sources d'informations pertinentes qui nécessitent des outils de synthèse et d'interprétation. Les recherches se sont orientées vers des systèmes d'intelligence artificielle puissants permettant l'extraction des informations utiles et aidant à la prise des décisions. Pour une meilleure synthèse et interprétation, la fouille de données ou *data mining* est née en puisant ses outils au sein de la statistique, de l'intelligence artificielle et des bases de données. La méthodologie du data mining offre la possibilité de construire un modèle de prédiction d'un phénomène à partir d'autres phénomènes plus facilement accessibles, qui lui sont liés, en se basant sur le processus d'extraction des connaissances à partir des données qui n'est qu'un processus de classification intelligente des données. Cependant, le modèle construit peut parfois engendrer des erreurs