Caractérisation d'instances d'apprentissage pour un méta-mining évolutionnaire

William Raynaut*, Chantal Soule-Dupuy*, Nathalie Valles-Parlangeau*,
Cedric Dray**, Philippe Valet**

*IRIT UMR 5505, UT1, UT3
Universite de Toulouse, France
prenom.nom@irit.fr
**INSERM, U1048
Universite de Toulouse, France
prenom.nom@inserm.fr

1 Motivation

L'apprentissage a été un secteur très prolifique ces dernières décennies, produisant nombre de techniques et algorithmes. Cependant, leurs performances sont sujettes à d'importantes variations d'un jeu de données à l'autre. On retrouve ainsi dans les *"No free lunch theorems"* (Wolpert, 1996) l'idée qu'il n'existe pas de solution meilleure en toute situation, d'apprentissage meilleur dans tous les domaines. Firent suite nombre d'applications de l'apprentissage à l'étude de sa propre applicabilité, posant les fondations du domaine du *méta-apprentissage*. Malgré bien d'autres applications fructueuses (Kalousis et Hilario, 2001), le problème du méta-apprentissage est toujours d'actualité, et les perspectives applicatives restent nombreuses.

L'un des principaux verrous actuels du méta-apprentissage, est la caractérisation des instances d'apprentissage, aussi appelées méta-instances. Cette caractérisation prend la forme d'un ensemble de méta-attributs, qui devra permettre une caractérisation fine de toute expérience d'apprentissage. On peut intuitivement diviser les méta-attributs selon trois dimensions, description du jeu de donnée, de traitements et algorithmes utilisés, et de la performance de ces traitements. Pour des raisons de volume, on ne s'intéressera ici qu'aux méta-attributs décrivant les jeux de données, ceux décrivant les traitements employés et l'évaluation des résultats seront privilégiés dans de futurs travaux.

Le problème de caractérisation d'un jeu de données a été étudié selon deux axes :

- Le premier consiste en l'emploi de mesures statistiques et information-théorétiques pour décrire le jeu de données. Cette approche, notamment mise en avant par le projet STATLOG (Michie et al., 1994), présente nombre de mesures très expressives, mais sa performance repose intégralement sur l'adéquation entre le biais de l'apprentissage effectué au méta-niveau et l'ensemble de mesures choisies.
- Le second axe d'approche, introduit comme "*landmarking*" par Pfahringer et al. (2000), considère quant à lui non pas des propriétés intrinsèques du jeu de données étudié, mais plutôt la performance d'algorithmes d'apprentissage simples exécutés dessus.