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.

La principale limitation de ces approches tend aux pré-requis imposés par les algorithmes d'apprentissage employés au méta-niveau. Ces derniers imposent notamment l'utilisation de vecteurs de méta-attributs de taille fixe, ce qui implique des agrégations (par exemple la variance individuelle des différents attributs d'un jeu de données devient la variance moyenne sur ce jeu de données...) et donc une importante perte d'information (Kalousis et Hilario, 2001).

2 Proposition

Pour contourner cette limitation, on se propose de conserver toute l'information disponible dans nos méta-instances et d'adresser le problème de méta-apprentissage par l'emploi d'une heuristique évolutionnaire au sein de la population des méta-instances. En effet, on peut alors construire le *fitness* d'une telle heuristique comme dissimilarité entre une méta-instance et la caractérisation d'une solution idéale au problème soumis. Les méta-attributs constitueront alors le génome des méta-instances, dont l'évolution contrainte par divers mécanismes heuristiques devra permettre la découverte des traitements apportant une réponse satisfaisante au besoin de l'utilisateur.

La transition de paradigme entre cette approche et le méta-apprentissage traditionnel nous affranchit des pertes d'information causées par les agrégations évoquées plus tôt, mais nous place face à un nouveau défi : il est possible de caractériser librement les méta-instances, mais il faut pouvoir les *comparer de manière sensée*. La validité de l'approche repose donc intégralement sur la définition d'une telle dissimilarité.

Références

Kalousis, A. et M. Hilario (2001). Model selection via meta-learning : a comparative study. *International Journal on Artificial Intelligence Tools 10*(04), 525–554.

Michie, D., D. J. Spiegelhalter, et C. C. Taylor (1994). *Machine Learning, Neural and Statistical Classification*. Upper Saddle River, NJ, USA : Ellis Horwood.

Pfahringer, B., H. Bensusan, et C. Giraud-Carrier (2000). Tell me who can learn you and i can tell you who you are : Landmarking various learning algorithms. In *Proceedings of the 17th international conference on machine learning*, pp. 743–750.

Wolpert, D. H. (1996). The lack of a priori distinctions between learning algorithms. *Neural computation 8*(7), 1341–1390.

Summary

Machine learning has proven to be a powerful tool in diverse fields, and is getting more and more widely used by non-experts. One of the foremost difficulties they encounter lies in the choice and calibration of the machine learning algorithm to use. Our objective is thus to provide assistance in the matter, using a meta-learning approach based on an evolutionary heuristic. We introduce here this approach as a potential solution to the limitation of current data characterization.