Cadre d'Evaluation pour la Méta Analyse de Données

William Raynaut*, Chantal Soule-Dupuy*, Nathalie Valles-Parlangeau*

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

Disclaimer - Ce texte est un aperçu de *"Meta-Mining Evaluation Framework : A large scale proof of concept on Meta-Learning"*, accepté pour publication à AI 2016, *"29th Australasian Joint Conference on Artificial Intelligence"* (Raynaut et al., 2016).

La méta analyse de données désigne la recherche d'une méthode efficace ou optimale permettant d'adresser un problème d'analyse de données. Cela recouvre une grande variété de tâches, dont certaines ont d'ores et déjà été abondamment étudiées. Par exemple, pour le problème de satisfiabilité booléenne (SAT), différentes approches de type *portfolio* ont été développées (Xu et al., 2012), reposant sur la sélection d'un algorithme approprié à la résolution d'une instance particulière du problème. La sélection d'algorithmes a également été employé pour des problèmes d'apprentissage, donnant lieu à diverses approches de méta-apprentissage. Ces problèmes particuliers ont été étudiés isolément, mais le prochain défi de la méta analyse de données réside en leur unification. En particulier, la recommandation de chaine de traitement d'analyse de données a reçu un intérêt croissant ces dernière années (Zakova et al., 2011; Serban et al., 2013). Ce problème consiste en la construction de chaînes de traitement permettant de résoudre différents problèmes d'analyse de données.

L'émergence de ces nouvelles approches amène la question de leur évaluation et comparaison. En effet, les critères employés dans l'évaluation des méthodes dédiées à des sous-problèmes spécifiques diffèrent souvent. Comparer la performance d'une recherche de motif et d'une régression n'est pas trivial. Afin de pouvoir évaluer et comparer les méthodes existantes et futures de méta analyse de données, nous nous attachons à construire un cadre général basé sur un critère unifié.

D'autre part, l'analyse de données reposant toujours principalement sur l'expertise humaine, la connaissance du méta-domaine est partielle et souvent implicite. Un moyen de construire explicitement cette connaissance pourrait consolider notre compréhension du domaine et aider à orienter les recherches à venir. Nous accorderons donc une grande importance à la compréhensibilité des résultats et à la qualification de leur validité.

Une expérience à grande échelle a été réalisée pour démontrer la praticabilité du cadre d'évaluation, et les tests statistique employés pour l'exploration des résultats valident les connaissances produites. Ils permettent par ailleurs d'étudier par une visualisation intuitive diverses questions que l'on peut se poser sur l'analyse de données, comme illustré en Figure 1 : *Quelle sélection d'attributs employer au méta-niveau ?* On peut y remarquer certains groupes

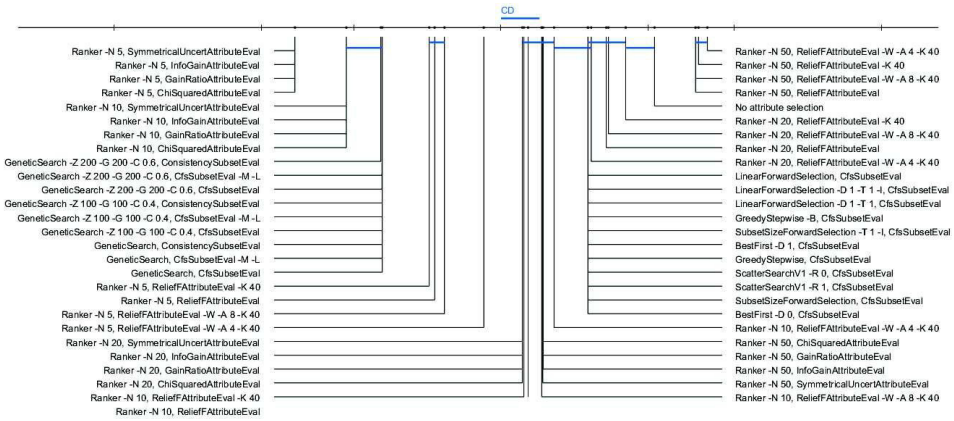

FIG. 1 – *Resultats du test de Nemenyi sur les méthodes de sélections d'attributs employées au méta-niveau. Les méthodes sont classées par performance croissante et les groupes connectés ne sont pas jugés significativement différents.*

de méthodes virtuellement équivalentes, ainsi qu'y visualiser l'impact des paramètres de certaines méthodes, mais le résultat le plus frappant est que la majorité des méthodes sont significativement *moins* performantes que l'absence de sélection d'attributs au méta-niveau, suggérant l'importance de grands ensembles de méta-attributs pour la performance du méta-apprentissage.

Au delà de son intérêt naturel dans l'évaluation de nouvelles méthodes de méta analyse de données, notre approche permet d'étudier divers aspects du méta-niveau encore mal connus. En particulier, des expériences conçues selon ce cadre d'évaluation sont en cours pour étudier l'impact de la caractérisation des jeux de données sur la performance du méta-apprentissage.

Références

Raynaut, W., C. Soule-Dupuy, et N. Valles (2016). Meta-mining evaluation framework : A large scale proof of concept on meta-learning. In *29th Australasian Joint Conference on Artificial Intelligence*.

Serban, F., J. Vanschoren, J.-U. Kietz, et A. Bernstein (2013). A survey of intelligent assistants for data analysis. *ACM Computing Surveys (CSUR) 45*(3), 31.

Xu, L., F. Hutter, J. Shen, H. H. Hoos, et K. Leyton-Brown (2012). Satzilla2012 : improved algorithm selection based on cost-sensitive classification models. *Balint et al.*, 57–58.

Zakova, M., P. Kremen, F. Zelezny, et N. Lavrac (2011). Automating knowledge discovery workflow composition through ontology-based planning. *Automation Science and Engineering, IEEE Transactions on 8*(2), 253–264.