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

William Raynaut, Chantal Soulé-Dupuy, Nathalie Vallès-Parlangeau

▶ **To cite this version:**

William Raynaut, Chantal Soulé-Dupuy, Nathalie Vallès-Parlangeau. Caractérisation d'instances d'apprentissage pour méta-mining évolutionnaire.. Journées Francophones Extraction et Gestion de Connaissances (EGC 2016), Jan 2016, Reims, France. pp.1-6. hal-03627164

HAL Id: hal-03627164
https://hal.science/hal-03627164

Submitted on 1 Apr 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Open Archive TOULOUSE Archive Ouverte (OATAO)

OATAO is an open access repository that collects the work of Toulouse researchers and makes it freely available over the web where possible.

This is an author-deposited version published in : http://oatao.univ-toulouse.fr/
Eprints ID : 17181

The contribution was presented at EGC 2016 :
http://egc2016.univ-reims.fr/index.php/Pr%C3%A9sentation

To cite this version : Raynaut, William and Soulé-Dupuy, Chantal and Vallés-Parlangeau, Nathalie *Caractérisation d'instances d'apprentissage pour méta-mining évolutionnaire.* (2016) In: Journées Francophones Extraction et Gestion de Connaissances (EGC 2016), 18 January 2016 - 22 January 2016 (Reims, France).

Any correspondence concerning this service should be sent to the repository administrator: staff-oatao@listes-diff.inp-toulouse.fr

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

Résumé. Les nombreuses techniques d'apprentissage et de fouille de données peuvent se révéler d'importants atouts dans divers domaines, mais choisir la technique la plus appropriée pour une application précise est une tâche très complexe pour un non-expert. Notre objectif est ainsi de produire un assistant de modélisation répondant à ce besoin, par une approche à la frontière du méta-apprentissage et des heuristiques évolutionnaires. Nous présentons ici le fonctionnement prévu de cet assistant, suivi d'une discussion de notre approche du problème de caractérisation des instances d'apprentissage, qui reste un verrou majeur du méta-apprentissage et méta-mining.

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'études catégorisant l'adéquation du biais d'algorithmes particuliers avec certains domaines ou situations, tels (Aha, 1992) et (Gama et Brazdil, 1995) qui y ont appliqué des techniques d'apprentissage de règles afin de décrire les conditions engendrant des différences de performance significatives entre algorithmes. Ces applications de l'apprentissage à l'étude de sa propre applicabilité ont posé les fondations du domaine du *méta-apprentissage*. Malgré bien d'autres applications fructueuses (Kalousis et Hilario, 2001b), et perspectives de recherches, telles la récursion infinie d'apprentissages auto-adaptatifs (Vilalta et Drissi, 2002), le problème du méta-apprentissage est toujours d'actualité, et les perspectives applicatives restent nombreuses. Parmi ces dernières, on considérera en particulier les récentes approches "*meta-mining*", qui s'affranchissent du cadre restreint de l'apprentissage au meta-niveau pour adresser une question plus large : *Considérant un besoin et un jeu de données, quelle suite de traitements donnera les meilleurs résultats ?*

Un algorithme est considéré capable d'apprentissage si sa performance s'améliore avec l'expérience (Mitchell, 1997). Dans le cas du méta-apprentissage, la plupart des approches impliquent la génération de méta-données, c'est-à-dire de données sur le problème de l'apprentissage, qui sont ensuite utilisées pour entrainer un algorithme d'apprentissage au méta-niveau. Notre perspective sur le sujet est que l'on peut définir un ensemble de méta-données comme une population de méta-instances d'apprentissage, chacune décrivant l'application de traitements précis à un jeu de données particulier, et qu'une bonne solution à un problème de méta-apprentissage peut être obtenue par l'emploi d'une heuristique évolutionnaire au sein de cette population. Le potentiel d'une telle approche n'a, à notre connaissance, pas encore été étudié, mais des approches comparables présentent des résultats encourageants, notamment pour des problèmes de satisfiabilité (SAT) (Xu et al., 2012) ou de sélection d'instance (Leyva et al., 2015).

Nous envisageons donc l'application de cette méthodologie à un assistant de modélisation, dont le fonctionnement prévu est illustré par la figure 1 et est présenté de manière plus complète dans (Raynaut et al., 2015).

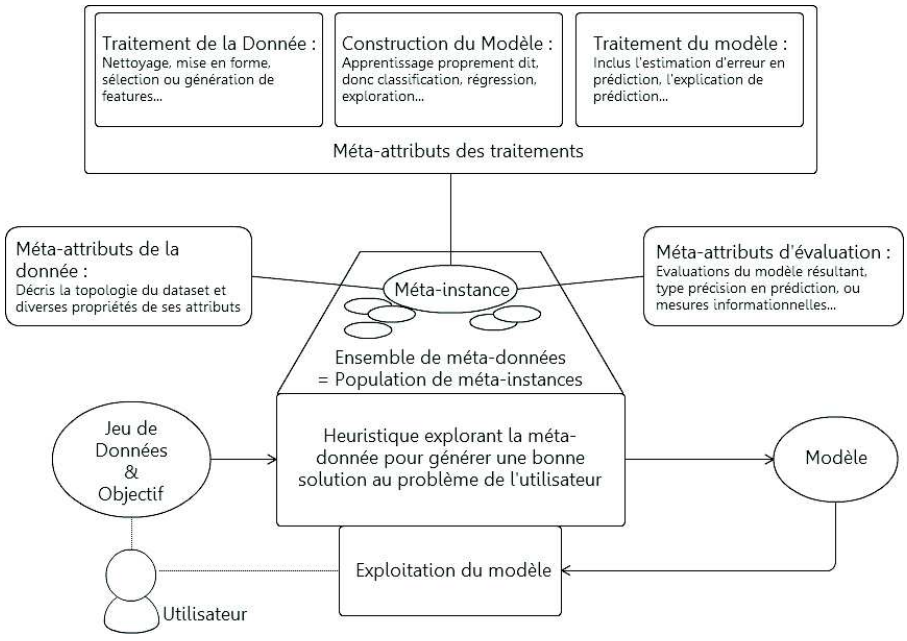

FIG. 1 – *Assistant de modélisation.*

Un élément crucial pour la réalisation de cette approche, ainsi que le sujet de ce papier, sera la caractérisation de ces méta-instances. Ce problème peut être considéré comme une version étendue du problème de caractérisation des jeux de données, étudié par la plupart des approches de méta-apprentissage. Ce dernier consiste en la définition d'un ensemble de propriétés (alors appelées méta-attributs) caractérisant finement des jeux de données, mais néanmoins soumis aux pré-requis imposés par les algorithmes d'apprentissage qui leur seront appliqués. Ces der-

niers 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, 2001b). Ce problème pourrait être résolu par l'emploi au méta-niveau d'une heuristique évolutionnaire dont la fonction de "*fitness*" reposerait sur une *dissimilarité* entre méta-instances. Cette dissimilarité pouvant être construite sur l'ensemble des informations disponibles, on s'affranchit des restrictions de représentation, se rapprochant en cela des approches "anti-essentialistes" telles que décrites par Duin (2015).

2 Caractérisation des instances d'apprentissage

Cette section concerne les attributs qui décriront nos méta-instances, ainsi appelés méta-attributs. L'ensemble de ces méta-attributs devrait être suffisamment important pour permettre une caractérisation fine de toute expérience de data-mining, mais un équilibre devra être trouvé pour éviter l'abondance de méta-attributs trop dépendants et limiter la complexité computationnelle. De plus, afin de pouvoir discriminer des méta-attributs ou méta-instances, la comparaison de méta-attributs au sein d'une méta-instance, tout comme la comparaison d'un même méta-attribut entre plusieurs méta-instances, devra être possible et sensée.

On peut intuitivement diviser les méta-attributs selon trois dimensions, comme illustré en figure 1. 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 modèles résultants 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), et employée dans une majorité d'études postérieures (Kalousis, 2002; Vilalta et Drissi, 2002; Leyva et al., 2015), 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. On note parfois l'emploi de techniques de sélection d'attributs à ce méta-niveau (Kalousis et Hilario, 2001a), mais les résultats expérimentaux ne permettent pas de conclure à la supériorité de quelconque mesure indépendamment du méta-apprentissage employé (Todorovski et al., 2000).
- Le second axe d'approche 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. Introduit comme "*landmarking*" par Pfahringer et al. (2000), cette approche emploie initialement le taux d'erreur d'un ensemble d'algorithmes basiques comme méta-donnée. Comme précédemment, les résultats suggèrent une forte dépendance de l'efficacité de cette approche avec le choix des algorithmes de base et du méta-niveau, ne révélant aucune combinaison uniformément supérieure. Des développements postérieurs ont introduit des mesures plus complexes, tel Peng et al. (2002) proposant comme méta-attributs des propriétés structurelles d'un arbre de décision construit sur la donnée.

Les expériences conduites par Fürnkranz et Petrak (2002) sur ces différentes approches tendent à conclure que toutes peuvent réaliser de bonnes performances dans diverses parties de l'ensemble des jeux de données, sans qu'aucune ne domine globalement.

La transition de paradigme entre notre 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*. Le coeur du problème réside en la comparaison de méta-attributs décrivant des attributs particuliers de la donnée, comme illustré dans l'exemple suivant.

Exemple *Considérons deux jeux de données, **A** et **B** illustrés en figure 2. **A** décrit 12 attributs de 100 individus, et **B** 10 attributs de 200 individus. On souhaite comparer les résultats de 5 mesures statistiques et informationnelles relevées sur les attributs individuels de ces jeux de données (comme illustré sur le second attribut de **A**). L'information complète que l'on souhaite comparer est donc un vecteur de 60 valeurs pour **A** et de 50 pour **B**.*

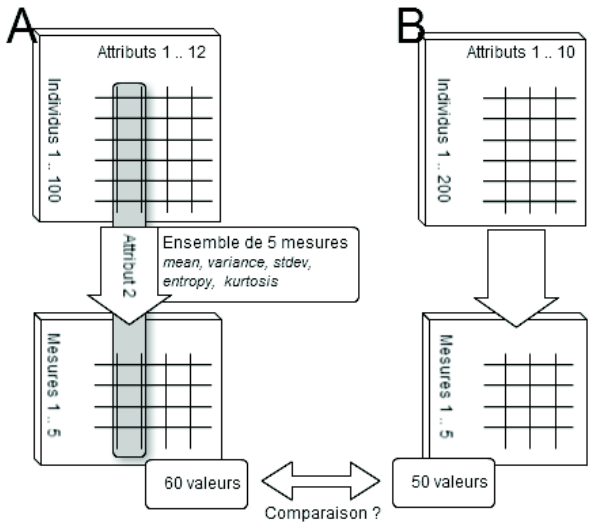

FIG. 2 – *Comparaison de mesures relevées sur des attributs individuels.*

*Notre approche est de comparer les attributs de **A** et **B** par paires les plus similaires, comparant les attributs en surnombre à d'hypothétiques attributs vides. L'hypothèse émise ici est qu'un attribut absent équivaut à un attribut dont aucune valeur n'est connue. Pour en revenir à l'exemple, la comparaison des 5 mesures s'effectuera donc entre l'attribut de **A** et l'attribut de **B** les plus similaires* selon ces mêmes mesures*, puis sur les second plus similaires et ainsi de suite, pour finir par comparer les mesures relevées sur les deux attributs surnuméraires de **A** avec leur valeur sur un hypothétique attribut vide. Cette comparaison par paire permet de s'affranchir de l'ordre de présentation des attributs, qui ne recèle aucune information, se concentrant sur la topologie réelle du jeu de données.*

Supposant qu'une comparaison très expressive résultera en un "*fitness*" plus juste, nous envisageons d'utiliser l'ensemble des mesures, statistiques comme issues de la théorie de l'information, présentées dans la littérature citée précédemment, ainsi que les différents méta-attributs construits sur les approches de "landmarking". Des recherches plus poussées seront cependant nécessaires pour limiter la duplication d'information entre les deux approches.

3 Conclusion et perspectives

Les méta-attributs présentés ici donnent un aperçu de la structure des méta-instances qui seront manipulées par l'heuristique évolutionnaire. En d'autres termes, ces méta-attributs constitueront 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.

Cependant, afin de compléter et d'évaluer notre approche, plusieurs tâches importantes restent à compléter. Tout d'abord, une représentation du besoin de l'utilisateur permettant son élicitation automatique ou semi-automatique sera nécessaire. En effet, l'assistant s'adressant à des non-experts, il devra aider à la définition formelle du besoin qui indiquera la cible de l'évolution. De plus, la sémantique d'une telle représentation pourra permettre de réduire la complexité computationnelle de l'heuristique en éliminant les méta-instance ne pouvant pas répondre au besoin de l'utilisateur.

La dissimilarité entre méta-instances devra ensuite être formalisée afin de construire la fonction fitness de l'heuristique, qui devra à son tour être définie et calibrée. La création de mécanismes "prédateurs" limitant la prolifération de méta-instances "inutiles" devra également être considérée. On s'intéressera en particulier aux approches génétiques (Bacardit et Llorà, 2013) et mémétiques (Smith, 2007) présentant des propriétés désirables, telles l'absence de contrainte quant à la structure du génome, et un parallélisme natif très souhaitable pour une approche computationnellement très complexe.

Références

Aha, D. W. (1992). Generalizing from case studies : A case study. In *Proc. of the 9th International Conference on Machine Learning*, pp. 1–10.

Bacardit, J. et X. Llorà (2013). Large-scale data mining using genetics-based machine learning. *Wiley Interdisciplinary Reviews : Data Mining and Knowledge Discovery 3*(1), 37–61.

Duin, R. P. (2015). The dissimilarity representation for finding universals from particulars by an anti-essentialist approach. *Pattern Recognition Letters 64*, 37 – 43.

Fürnkranz, J. et J. Petrak (2002). Extended data characteristics. Technical report, METAL consortium. Accessed 12/11/15 at `citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.97.302`.

Gama, J. et P. Brazdil (1995). Characterization of classification algorithms. In *Progress in Artificial Intelligence*, pp. 189–200. Springer.

Kalousis, A. (2002). *Algorithm selection via meta-learning*. Ph. D. thesis, Universite de Geneve.

Kalousis, A. et M. Hilario (2001a). Feature selection for meta-learning. In *Proceedings of the 5th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, PAKDD '01, London, UK, UK, pp. 222–233. Springer-Verlag.

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

Leyva, E., A. Gonzalez, et R. Perez (2015). A set of complexity measures designed for applying meta-learning to instance selection. *Knowledge and Data Engineering, IEEE Transactions on 27*(2), 354–367.

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

Mitchell, T. M. (1997). Machine learning. *Burr Ridge, IL : McGraw Hill 45*.

Peng, Y., P. A. Flach, P. Brazdil, et C. Soares (2002). Decision tree-based data characterization for meta-learning. Accessed 12/11/2015 at `citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.101.4890`.

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.

Raynaut, W., C. Soule-Dupuy, et N. Valles-Parlangeau (2015). Addressing the meta-learning problem with metaheuristics (extended abstract). In *Metaheuristics International Conference*.

Smith, J. E. (2007). Coevolving memetic algorithms : a review and progress report. *Systems, Man, and Cybernetics, Part B : Cybernetics, IEEE Transactions on 37*(1), 6–17.

Todorovski, L., P. Brazdil, et C. Soares (2000). Report on the experiments with feature selection in meta-level learning. In *Proceedings of the PKDD-00 workshop on data mining, decision support, meta-learning and ILP : forum for practical problem presentation and prospective solutions*, pp. 27–39. Citeseer.

Vilalta, R. et Y. Drissi (2002). A perspective view and survey of meta-learning. *Artif. Intell. Rev. 18*(2), 77–95.

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

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.

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 expand here previous work presenting the intended workflow of a modeling assistant by describing and discussing the characterization of learning instances we intend to use.