LISIC
Laboratoire d'Informatique Signal & Image de la Côte d'Opale

ulco Université Littoral Côte d'Opale

Stage de Master 2 Recherche 2025

—

Algorithmes d'optimisation multi-objectifs pour la sélection d'attributs

Encadrement

Arnaud LIEFOOGHE `arnaud.liefooghe@univ-littoral.fr`
Sébastien VEREL `sebastien.verel@univ-littoral.fr`

Localisation

Équipe OSMOSE, laboratoire LISIC, ULCO, site de Calais

Période

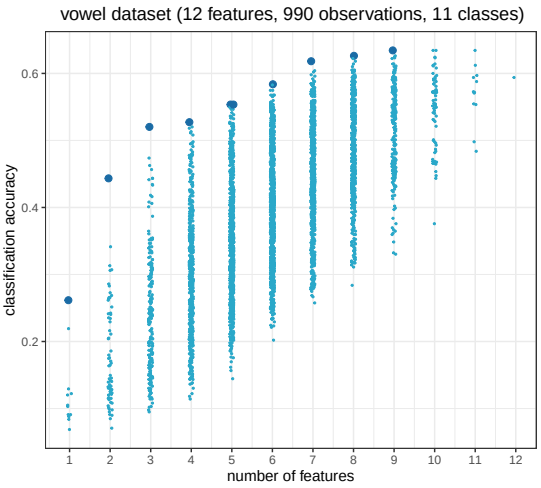
De mars à juillet 2025 environ (selon le calendrier de formation du/de la candidat·e)

Description du sujet

La **sélection d'attributs** constitue un élément clé de l'intelligence artificielle explicable. Elle consiste à choisir un sous-ensemble d'attributs caractérisant un jeu de données afin de construire un modèle d'inférence. Le but est double : (1) simplifier le modèle pour faciliter son interprétation en mettant en évidence les attributs essentiels aux bonnes prédictions ; (2) réduire le temps de calcul avec un entraînement et une validation plus rapides.
Les technologies avancées de sélection d'attributs formulent généralement ce problème comme un problème d'optimisation, qui s'inscrit dans une classe plus large de problèmes pseudo-booléen de sélection de sous-ensembles. Ce sujet s'inscrit donc pleinement dans la **thématique de l'intelligence artificielle et de l'optimisation**. Cependant, sa nature combinatoire, son coût de calcul élevé et son aspect "boîte noire" posent divers défis que les méthodes d'optimisation se doivent de surmonter. Pour un jeu de données ayant n attributs, il s'agit d'identifier le sous-ensemble minimal de $p \leqslant n$ attributs qui maximise la précision du modèle d'inférence. La qualité d'un sous-ensemble est évaluée en entraînant un modèle d'apprentissage supervisé avec les attributs sélectionnés comme prédicteurs. Ce problème est NP-difficile [AK98].

De surcroît, il s'avère intrinsèquement multi-objectifs. En effet, il est évident que chaque score d'induction engendre un problème distinct, avec ses propres caractéristiques et ses propres solutions. Cette complexité soulève des défis supplémentaires, puisque la sélection d'attributs implique ainsi différents objectifs [JN+24] combinant diverses fonctions de score tout en minimisant le nombre d'attributs sélectionnés. Par exemple, le score de classification est généralement défini comme le ratio entre le nombre de prédictions correctes et le nombre total d'observations. D'autres scores de classification comprennent la F-mesure, la précision et le rappel. Pour la régression, les mesures courantes incluent le coefficient de détermination, l'erreur quadratique ou encore l'erreur absolue.

En **optimisation multi-objectifs**, l'utilisateur ne recherche pas une solution unique, mais un ensemble de compromis optimaux entre les objectifs, appelé front Pareto. C'est au sein de cet ensemble qu'il pourra sélectionner la solution correspondant le mieux à ses préférences. Il est donc essentiel d'explorer les problèmes multi-objectifs de sélection d'attributs pour décrypter les interactions complexes entre les jeux de données, les modèles d'inférence et les fonctions de score. Cette compréhension approfondie des compromis et des interactions pourrait conduire à des algorithmes de sélection d'attributs plus efficaces, améliorant ainsi la performance et l'explicabilité des modèles tout en réduisant les coûts de calcul. Un récent tutoriel sur la sélection de sous-ensembles soulève la nécessité de clarifier l'impact de l'optimisation multi-objectifs sur le nombre d'optima locaux, et si cette approche en facilite véritablement la résolution [Qia24].



Depuis les années 1990, de nombreux algorithmes de recherche locale et d'évolution artificielle ont été développés pour l'optimisation multi-objectifs. Cependant, dans le domaine de la sélection d'attributs, les algorithmes actuels ne tirent pas pleinement parti de la nature combinatoire du problème. De plus, ils n'exploitent pas l'efficacité prouvée des approches de **recherche locale** pour d'autres problèmes multi-objectifs de sélection de sous-ensembles [JN+24]. En outre, en raison de son caractère chronophage, la sélection d'attributs s'appuie fréquemment sur des modèles d'inférence peu coûteux, tels que les k plus proches voisins (kNN) [DDK22]. Or, nos expériences préliminaires révèlent que le sous-ensemble optimal d'attributs peut varier considérablement selon le modèle d'inférence choisi [LTV24]. Ainsi, l'utilisation de kNN pourrait aboutir à des solutions différentes de celles obtenues avec le modèle d'inférence souhaité par l'utilisateur. Enfin, la question persiste quant à la similitude des défis d'optimisation inhérents aux divers modèles d'inférence. Ce sont précisément tous ces défis que nous nous proposons d'explorer dans le cadre de ce stage de Master 2 Recherche.

Programme de travail et échéancier prévisionnel

Mois 1 — Étude bibliographique et inventaire des algorithmes existants pour la sélection d'attributs mono- et multi-objectifs. Sélection de jeux de données *benchmark* variés en nombre d'observations, nombre et type d'attributs, et type d'inférence (régression ou classification).

Mois 2-3 — Conception et développement d'un algorithme de résolution multi-objectifs pour la sélection d'attributs. Cette méthode prendra en entrée : un jeu de données, le type d'inférence (régression ou classification), le modèle d'apprentissage supervisé et les objectifs (score(s) et/ou nombre d'attributs sélectionnés). En sortie, l'algorithme fournira une approximation du front Pareto ainsi que diverses mesures de performance d'exécution (temps CPU, nombre d'entraînements effectués, etc.).

Mois 3-4 — Analyse comparative des performances de l'algorithme développé par rapport à l'état de l'art, en tenant compte des spécificités du problème de sélection d'attributs considéré. Évaluation de l'écart avec les meilleures solutions connues et interprétation des sous-ensembles obtenus, notamment la fréquence de sélection des attributs en fonction de leur importance relative.

Mois 5 — Raffinement de l'algorithme proposé en fonction des résultats observés, mise en œuvre de variantes algorithmiques et évaluation des améliorations en termes d'efficacité et de qualité des solutions obtenues. Mise en évidence des forces et des faiblesses des différentes approches étudiées.

Mois 6 — Rédaction d'un rapport sous la forme d'un article de recherche. Mise à disposition des données et des méthodes développées selon les principes de reproductibilité scientifique. Finalisation du rapport de stage et préparation de la soutenance.

Laboratoire d'accueil

Le stage se déroulera à Calais, au sein de l'équipe OSMOSE (optimisation, simulation et modélisation évolutionnaire) du laboratoire d'informatique, signal et image (LISIC) de l'Université du Littoral Côte d'Opale. L'étudiant·e évoluera dans un environnement scientifique dynamique, en interaction directe avec les membres de l'équipe et notre réseau de collaboration international. Ce stage sera en partie mené en collaboration avec le Japon. Un espace de travail équipé d'un accès internet et aux diverses installations du laboratoire sera mis à sa disposition. De plus, l'étudiant·e aura accès aux ressources de calcul haute performance de l'université pour mener ses expérimentations. Cette opportunité lui permettra de développer une expertise solide en optimisation, en apprentissage automatique et en intelligence artificielle. Elle lui offrira également une expérience enrichissante dans un contexte de recherche international.

La rémunération du stage est conforme à la législation en vigueur. Une opportunité de poursuite en thèse pourra être envisagée à l'issue du stage.

Profil recherché

Ce stage de recherche s'adresse aux étudiant·es en dernière année de Master ou d'école d'ingénieur en informatique. Les candidat·es devront posséder de solides connaissances en algorithmique, en optimisation et en apprentissage automatique. Une bonne maîtrise de la programmation en Python est également indispensable.

Modalités de candidature

Les candidatures doivent être envoyées à `arnaud.liefooghe@univ-littoral.fr` en incluant les documents suivants :

- CV détaillé
- Lettre de motivation
- Lettre(s) de recommandation
- Relevés de notes des trois dernières années

Bibliographie

[AK98] E. Amaldi, V. Kann: On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems. Theoretical Computer Science 209(1), 237–260 (1998)

[BD+16] M. Basseur, B. Derbel, A. Goëffon, A. Liefooghe: Experiments on greedy and local search heuristics for $d-$dimensional hypervolume subset selection. Genetic and Evolutionary Computation Conference (GECCO 2016), pp 541–548, Denver, USA (2016)

[DDK22] T. Dökeroglu, A. Deniz, H.E. Kiziloz: A comprehensive survey on recent metaheuristics for feature selection. Neurocomputing 494, 269–296 (2022)

[FP+94] F.J. Ferri, P. Pudil, M. Hatef, J. Kittler: Comparative study of techniques for large-scale feature selection. Machine Intelligence and Pattern Recognition 16, 403-413 (1994)

[JN+24] Jiao, B.H. Nguyen, B. Xue, M. Zhang: A survey on evolutionary multiobjective feature selection in classification: Approaches, applications, and challenges. IEEE Transactions on Evolutionary Computation 28(4), 1156–1176 (2024)

[LTV24] A. Liefooghe, R. Tanabe, S. Verel: Contrasting the landscapes of feature selection under different machine learning models. International Conference on Parallel Problem Solving from Nature (PPSN 2024), Lecture Notes in Computer Science, vol 15148, Hagenberg, Austria (2024)

[Qia24] C. Qian: Pareto optimization for subset selection: Theories and practical algorithms. International Conference on Parallel Problem Solving from Nature (PPSN 2024), Hagenberg, Austria (2024)

[XZ+16] B. Xue, M. Zhang, W.N. Browne, X. Yao: A survey on evolutionary computation approaches to feature selection. IEEE Transactions on Evolutionary Computation 20(4), 606–626 (2016)