Stage M2 – Classification multi-label à l'aide d'optimisation multi-objectif pour la détection du cancer du poumon

Equipe ORKAD - CRIStAL - Université de Lille

Présentation de la structure et contexte scientifique

ORKAD est une équipe de recherche du groupe thématique OPTIMA du laboratoire CRIStAL (Centre de Recherche en Informatique, Signal et Automatique de Lille) (UMR CNRS 9189) de l'Université de Lille. L'objectif principal de l'équipe ORKAD est d'exploiter simultanément l'optimisation combinatoire et l'extraction de connaissances pour résoudre des problèmes d'optimisation. Les métaheuristiques ont souvent été utilisées avec succès pour résoudre différentes tâches de machine learning [DhaenensJourdan2022]. En particulier, l'algorithme MOCA-I [Jacques2013-a], permet de classifier des données hétérogènes et mal réparties par méthode d'optimisation, sur des données médicales [Jacques2020]. L'équipe ORKAD a des partenariats avec le CHU de Lille ; notamment dans le cadre du projet européen PATHACOV pour la détection du cancer du poumon à partir de la concentration en composés organiques volatils dans l'air expiré [Hulo2023]. Dans ce stage, nous nous intéressons à l'extension de ces travaux aux données du projet ALCOVE, suite du projet PATHACOV, où l'objectif est de distinguer différentes classes de sujets: sain / malade (avec le stade : I, II, III, IV) ; opérable / non opérable.

Description du sujet

Dans le problème de classification multi-label, un enregistrement du jeu de données peut être associé à plusieurs labels : par exemple « cancer du poumon » et « opérable ». Des approches à base de métaheuristiques ont été proposées par le passé pour gérer ce problème, comme par exemple les colonies de fourmis [Otero2010]. La classification multi-label est souvent associée à une répartition déséquilibrée des différents labels à prédire [Tarekegn2021] et une des spécificités de l'algorithme MOCA-I est justement sa capacité à gérer ce déséquilibre [Jacques2013-a]. Dans MOCA-I, la modélisation est adaptée pour la classification binaire partielle (représentation de la solution, opérateurs d'initialisation et de voisinage,…). L'objectif de ce stage est de proposer une nouvelle représentation et opérateurs adaptés au problème de classification multi-label. Des méthodes de configuration automatique d'algorithmes comme irace [López-Ibáñez2016] seront utilisées pour identifier si les nouveaux opérateurs et stratégies proposés sont efficaces sur les benchmarks sélectionnés.

Compétences recherchées

Programmation Objet (Python ou C++) ; Connaissances en machine learning
Des connaissances en C++ et recherche opérationnelle, optimisation combinatoire seraient un plus.

Modalités du stage

Lieu : Laboratoire CRISTAL, Equipe ORKAD (Université de Lille, France)

Date de début : janvier 2025 (ou selon candidat)
Poursuite en thèse possible à l'issue du stage

Modalités de candidature

Joindre CV + lettre de motivation (si poursuite en thèse : + lettre(s) de recommandation + bulletins L3/M1/M2)

Julie Jacques – Maître de conférences, équipe ORKAD (CRIStAL / Université Lille) (julie.jacques@univ-lille.fr)
Marie-Éléonore Kessaci – Professeur, équipe ORKAD (CRIStAL / Université Lille) (mkessaci@univ-lille.fr)

Références

[DhaenensJourdan2022] C. Dhaenens, and L. Jourdan. Metaheuristics for Data Mining: Survey and Opportunities for Big Data . Annals of Operations Research 314, no 1 : 117-40. https://doi.org/10.1007/s10479-021-04496-0.
[Hulo2023] S. Hulo, J. Jacques, F. Sihrener, E. Wasielewski, Laetitia Jourdan, G. Poslednik, C. Poulet, A. Turlotte, T. Gey, Y. Douadi, L. Thiberville, M. Dewolf, J-M. Lecerf,I. Estevié, V. Ricard, J. Martin, A-C. Romain, N. Locoge, R. Matran, A. Scherpereel 160P Non-invasive analysis of VOCs in exhaled air can distinguish healthy controls from lung cancer patients and may improve the effectiveness of lung cancer screening Journal of Thoracic Oncology, 2023, 18 (4), pp.S128. ⟨10.1016/S1556-0864(23)00414-8⟩
[Jacques2013-a] J. Jacques, J. Taillard, D. Delerue, L. Jourdan, and C. Dhaenens : MOCA-I: discovering rules and guiding decision maker in the context of partial classification in large and imbalanced datasets, Learning and Intelligent OptimizatioN, Lecture Notes in Computer Science (LNCS), 2013
[Jacques2020] Julie Jacques, Helene Martin-Huyghe, Justine Lemtiri-Florek, Julien Taillard, Laetitia Jourdan, Clarisse Dhaenens, David Delerue, Arnaud Hansske, Valérie Leclercq The Detection of hospitalized patients at risk of testing positive to multi-drug resistant bacteria using MOCA-I, a rule-based “white-box” classification algorithm for medical data International Journal of Medical Informatics, 2020, October 2020, 142,
[López-Ibáñez2016] M. López-Ibáñez, J. Dubois-Lacoste, L. Pérez Cáceres, T. Stützle, and M. Birattari. The irace package: Iterated Racing for Automatic Algorithm Configuration. Operations Research Perspectives, 3:43–58, 2016.
[Otero2010] F. Otero, Alex A. Freitas, and Colin G. Johnson. « A Hierarchical Multi-Label Classification Ant Colony Algorithm for Protein Function Prediction ». Memetic Computing 2, no 3 (1 septembre 2010): 165-81. https://doi.org/10.1007/s12293-010-0045-4.
[Tarekegn2021] Tarekegn, Adane Nega, Mario Giacobini, et Krzysztof Michalak. « A review of methods for imbalanced multi-label classification ». Pattern Recognition 118 (1 octobre 2021): 107965. https://doi.org/10.1016/j.patcog.2021.107965.