Optimisation multiobjectif pour le diagnostic de pathologies via biomarqueurs

Sara Tari, Laetitia Jourdan, Marie-Eléonore Kessaci, Julie Jacques, Lucien Mousin

Optimisation multiobjectif pour le diagnostic de pathologies via biomarqueurs

Sara Tari[1], Laetitia Jourdan[1], Marie-Éléonore Kessaci[1], Julie Jacques[2], Lucien Mousin[2]

[1] Université de Lille, CNRS, UMR 9189 CRIStAL, 59650 Villeneuve d'Ascq, France
{sara.tari,laetitia.jourdan,marie-eleonore.kessaci}@univ-lille.fr
[2] Université Catholique de Lille, Faculté d'économie, gestion et sciences, CNRS, UMR 9189 CRIStAL, France
{julie.jacques,lucien.mousin}@univ-catholille.fr

Mots-clés : *optimisation multiobjectif, classification, paramétrage automatique.*

1 Introduction

Les êtres vivants émettent une grande variété de composés organiques volatils (COVs) qui constituent une signature biologique individuelle. La composition et quantité des COVs émis pour un individu varient selon divers facteurs, dont les pathologies. Ils constituent donc des biomarqueurs d'intérêt pour le diagnostic de diverses pathologies, notamment les cancers ou les maladies telles que le diabète et l'insuffisance rénale. C'est dans ce contexte que s'inscrit le projet PATHACOV [1], visant à établir un diagnostic non invasif de maladies et particulièrement du cancer du poumon.

Ici, nous présentons notre approche pour extraire des règles permettant de formuler un diagnostic. Cela consiste à transformer le problème de classification supervisée en problème d'optimisation multiobjectif selon le modèle proposé dans [3]. Cette transformation présente plusieurs avantages par rapport aux méthodes de classification classiques : gestion de grands volumes de données, du déséquilibre des données et obtention d'un modèle interprétable.

Ce problème est traité pour la première fois au moyen d'une généralisation de recherches locales multiobjectif couplée à un configurateur automatique d'algorithmes pour optimiser la sélection des mécanismes et paramètres numériques menant vers de bonnes solutions.

2 Classification de patients par des méthodes d'optimisation

Dans un problème de classification, chaque individu est décrit par un ensemble d'attributs et une classe à prédire. Ici, les attributs correspondent aux COVs, leurs valeurs sont des réels correspondants à la quantité mesurée pour chaque COV dans chaque échantillon. La classe à prédire indique si l'individu est positif ou négatif à la pathologie investiguée. Les attributs des jeux de données sont discrétisés au moyen de méthodes classiques afin de rendre le problème combinatoire et ainsi de simplifier la notion de voisinage.

2.1 Modélisation sous forme de problème multiobjectif

Une solution du problème correspond à un modèle de classification supervisée décrit par un ensemble de règles, chacune étant une conjonction de plusieurs termes. Un terme est composé d'un attribut, d'un opérateur ($<$,$>$,$=$) et d'un intervalle. Les différentes règles d'une solution sont utilisées pour déterminer quels individus sont positifs pour une pathologie.

Plusieurs objectifs sont considérés simultanément pour optimiser la qualité des modèles de classification supervisée. Les objectifs à maximiser sont la confiance et la sensibilité, deux

1. Ce projet est financé par le programme Interreg France-Wallonie-Vlaanderen, avec le soutien du Fonds européen de développement régional.

objectifs souvent contradictoires. Un troisième objectif, à minimiser, correspond au nombre de termes afin de limiter les effets de *bloat* qui surviennent lorsqu'une solution se complexifie sans apport en termes de qualité.

Le voisinage d'une solution décrit tous les ensembles de règles dont un terme est ajouté, supprimé ou modifié. Modifier un terme consiste à changer son opérateur ou son intervalle.

2.2 Approche de résolution

Le problème est traité avec une métaheuristique multiobjective correspondant à l'unification des recherches locales multiobjectif présentée dans [2]. Pour cette stratégie, on distingue les paramètres catégoriques, correspondants aux stratégies possibles pour les différentes composantes de l'algorithme et les paramètres numériques, correspondant aux valeurs utilisées pour chaque composante.

MO-ParamILS [1], un configurateur automatique d'algorithmes, est utilisé pour sélectionner les paramètres catégoriques et numériques les plus adaptés à la résolution du problème selon une métrique donnée. La solution retournée correspond à celle du front Pareto ayant la meilleure f-mesure qui combine la valeur prédictive positive et la sensibilité.

3 Expérimentations

L'approche proposée est appliquée sur plusieurs versions discrétisées des jeux de données issus de prélèvements décrits brièvement dans le tableau 1, chacun correspondant à une instance. Nous utilisons ces instances pour nous comparer à des algorithmes classiques de classification supervisée implémentés dans `Weka` et `Scikit-learn`, sélectionnés et paramétrés avec un configurateur automatique en fonction d'une métrique donnée.

Nom	Diagnostique	#individus	#atteints	#attributs
T3	Dialyse	72	36	346
T4	Dialyse	74	37	341
P1	Cancer Prostate	103	59	137

TAB. 1 – Description des jeux de données réels issus de prélèvements de patients

Pour chaque triplet (instance, approche, métrique), 30 exécutions sont conduites pour une même durée maximale. Pour chacune d'entre elles, les modèles sont construits en utilisant des validations croisées 5-plis et 10-plis afin de limiter le surapprentissage.

Plusieurs indicateurs sont considérés pour optimiser le paramétrage de la classification et les trois approches sont comparées selon plusieurs métriques classiques en apprentissage artificiel (sensibilité, spécificité, aire sous courbe ROC, coefficient de corrélation de Matthews). Les résultats seront présentés lors de la conférence.

Références

[1] Aymeric Blot, Holger H Hoos, Laetitia Jourdan, Marie-Éléonore Kessaci-Marmion, and Heike Trautmann. Mo-paramils : A multi-objective automatic algorithm configuration framework. In *International Conference on Learning and Intelligent Optimization*, pages 32–47. Springer, 2016.

[2] Aymeric Blot, Marie-Éléonore Kessaci, and Laetitia Jourdan. Survey and unification of local search techniques in metaheuristics for multi-objective combinatorial optimisation. *Journal of Heuristics*, 24(6) :853–877, 2018.

[3] Julie Jacques, Julien Taillard, David Delerue, Laetitia Jourdan, and Clarisse Dhaenens. The benefits of using multi-objectivization for mining pittsburgh partial classification rules in imbalanced and discrete data. In *Proceedings of the 15th annual conference on Genetic and evolutionary computation*, pages 543–550. ACM, 2013.