Comparaison expérimentale de métaheuristiques pour la sélection de variables dans le cadre de l'apprentissage automatique appliqué au domaine médical

Thibault Anani[2], F. Delbot[1,2], JF Pradat-Peyre[1,2]

[1] Université Paris Nanterre, Nanterre, France
[2] LIP6, Sorbonne Université, Paris, France
{thibault.anani,francois.delbot,Jean-Francois.Pradat-Peyre}@lip6.fr

Mots-clés : *apprentissage automatique, selection de variables, santé, optimisation.*

Introduction. Il est courant d'utiliser les méthodes d'apprentissage automatique pour développer des modèles de pronostics précis et fiables permettant d'établir une classification des patients atteints d'une certaine maladie [1]. La classification des patients, permet de regrouper les individus en fonction de leurs besoins et donc d'adapter le traitement du patient de manière pertinente. La disponibilité récente de données de patients atteints de la SLA (Sclérose Latérale Amyotrophique) a permis l'étude de différentes méthodes de pronostic et de classification. Certaines de ces méthodes d'apprentissage automatique ont réussi à exploiter les corrélations présentes dans les données pour mieux comprendre la progression de la maladie [2].Cependant, la quantité d'informations, c'est-à-dire le nombre de variables associées à un patient, peut perturber l'apprentissage car certaines variables ne sont pas pertinentes. Il est donc nécessaire de sélectionner un sous-ensemble des variables les plus appropriées de sorte à maximiser la qualité prédictive du modèle. La difficulté de cette stratégie est qu'elle est confrontée au problème de l'explosion combinatoire. En effet, le nombre de combinaisons possibles étant exponentiel, une énumération complète des sous-ensembles n'est pas réaliste. L'utilisation de méthodes statistiques et/ou de métaheuristiques permet d'approcher la solution optimale. De plus, un sous-ensemble de variables pourra mener à des performances différentes en fonction de la méthode d'apprentissage choisie (Régression logistique, Random forest, etc.). Le choix de la méthode d'apprentissage est généralement réalisé expérimentalement. La question est donc de déterminer, pour un jeu de données, le meilleur couple (métaheuristique, méthode d'apprentissage) permettant de maximiser la qualité prédictive du modèle obtenu.

Dans ce travail, nous effectuons une comparaison expérimentale de 7 métaheuristiques parmi les plus courantes (algorithmes génétiques, recuit simulé, essaim de particules, etc.) afin de déterminer le meilleur sous-ensemble de variables. Nous les associons avec 9 méthodes d'apprentissage parmi les plus courantes. Chaque couple (métaheuristique, méthode) est appliqué sur 13 jeux de données benchmarks ainsi que sur des jeux de données provenant du domaine médical, pour un total de 945 expériences.

Résultats. En effectuant une sélection de variables nous parvenons à améliorer le score sur tous les jeux de données sans aucune exception. Le tableau 1 indique, pour chaque jeu de données benchmark : 1) Le score obtenu avec la meilleur méthode d'apprentissage sans réaliser de sélection de variables. 2) Le score obtenu par les méthodes statistiques de l'état de l'art [3]. 3) Le score obtenu grâce au meilleur couple (métaheuristique, méthode d'apprentissage).

Nos expériences semblent indiquer que les métaheuristiques sont plus efficaces que les méthodes statistiques utilisées par [3] puisque nous arrivons à obtenir un meilleur score sur 11

des 13 jeux de données. L'amélioration de la qualité prédictive peut aller jusqu'à 8 points (par exemple pour le jeu de données *Tian*).

Jeu de données	**Meilleur résultat toutes les variables**	**Meilleur résultat** [3] **sélection de variables**	**Meilleur résultat avec notre méthode**
Bioresponse	79.42% (ERT)	79.90%	80.35% (RDC + algorithme génétique)
Burczynski	94.43% (RRC)	98.00%	98.95% (RRC + évolution différentielle)
Chin	86.45% (LR, SVM, GNB)	89.90%	90.68% (RDC + algorithme génétique)
Chowdary	97.14% (LR)	98.00%	99.04% (LR + algorithme génétique)
Gina Prior	95.56% (ERT)	96.30%	96.34% (ERT + évolution différentielle)
Gina Agnostic	94.49% (ERT)	95.50%	95.53% (RDC + évolution différentielle)
Gravier	75.03% (LDA)	77.60%	85.12% (LDA + évolution différentielle)
Hiva Agnostic	96.74% (ERT)	97.00%	97.26% (RDC + évolution différentielle)
Internet	96.37% (ERT)	97.30%	97.47% (RDC + essaims de particules)
Madelon	75.54% (KNN)	87.50%	80.96% (RDC + essaims de particules)
Scene	98.59% (LDA)	98.90%	99.04% (RRC + évolution différentielle)
Tian	80.37% (LDA, GNB)	81.10%	89.02% (GNB + évolution différentielle)
Yeoh	97.16% (RRC)	100%	99.33% (LR + évolution différentielle)

TAB. 1 – Résultats des modèles constitués avec sélection de variables sur les benchmarks

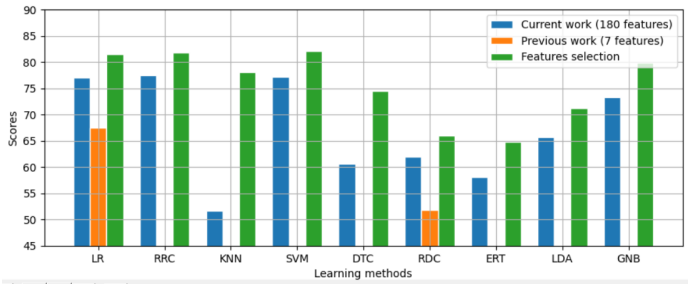

FIG. 1 – Comparaison de la qualité obtenue sur la classification de patients atteints de la SLA.

Nous avons appliqué notre méthodologie sur les données de patients atteints de la SLA. Nous obtenons une classification bien meilleure que dans [4], permettant ainsi une meilleure prise en charge des patients atteints de la SLA.

Conclusions. Pour chacun de nos jeux de données la métaheuristique qui nous permet d'obtenir le meilleur résultat est une métaheuristique à base de population. En particulier, l'évolution différentielle semble être la métaheuristique la plus efficace. Les métaheuristiques à base de parcours semblent au contraire inadaptées. Les performances des métaheuristiques à base de population semblent décorrélées du choix de la méthode d'apprentissage. A partir de ces expériences, nous recommandons d'utiliser l'évolution différentielle pour sélectionner un sous-ensemble de variables, quelle que soit la méthode d'apprentissage utilisée.

Références

[1] V. Grollemund, P.F. Pradat, G. Querin, F. Delbot, G. Le Chat, J.F. Pradat-Peyre and P. Bede. Machine Learning in Amyotrophic Lateral Sclerosis : Achievements, pitfalls, and future directions. *Frontiers in Neuroscience*, 2019.

[2] A.M. Antoniadi, M. Galvin, M. Heverin, O. Hardiman and C. Mooney *Prediction of caregiver burden in amyotrophic lateral sclerosis : a machine learning approach using random forests applied to a cohort study.* BMJ Open, 2020.

[3] A. Bommert, X. Sun, B. Bischl, J. Rahnenführer, M. Lang. *Benchmark for filter methods for feature selection in high-dimensional classification data.* CS & DA, 2020.

[4] V. Grollemund, P.F. Pradat, M.S. Secchi-Buhour, F. Delbot, G. Le Chat, J.F. Pradat-Peyre and P. Bede. Development and validation of a 1-year survival prognosis estimation model for Amyotrophic Lateral Sclerosis using manifold learning algorithm UMAP. *Scientific Reports*, 2020.