Sélection de variables et auto-régression pour la prédiction de l'évolution de la Sclérose Latérale Amyotrophique

Thibault Anani[2], François Delbot[1,2], Jean-François Pradat-Peyre[1,2]

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

Mots-clés : *machine learning, sélection de variables, santé, optimisation, heuristiques.*

Introduction. La Sclérose Latérale Amyotrophique (SLA), ou plus communément appelée maladie de Charcot, est une maladie neurodégénérative pour laquelle il n'existe actuellement aucun traitement. L'espérance de vie médiane au moment de l'apparition des symptômes varie entre 3 et 5 ans. L'évolution des fonctions physiques d'un patient est calculée grâce à l'ALSFRS qui est une échelle largement adoptée par la communauté médicale variant de 40, quand le patient n'est pas affecté, à 0, pour indiquer une paralysie totale. Le décès du patient intervenant le plus souvent lors de la paralysie des muscles respiratoires. L'établissement d'un pronostic fiable est un enjeu majeur car il conditionne la prise en charge du patient et sa qualité de vie. Les méthodes basées sur du machine learning sur de grands jeux de données ont permis d'identifier des corrélations sous-jacentes dans les données des patients, comme dans le défi Pro-Act [1, 2]. Cependant, en raison d'une forte hétérogénéité des patients les modèles pronostics restent peu précis et peu fiables. De plus, 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 et par conséquent conduisent à des modèles peu exploitables pour prédire la progression de la maladie. 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 à une 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. Dans une étude précédente nous les avons comparées, et nous avons montré que dans la majorité des cas l'évolution différentielle était la métaheuristique la plus efficace permettant d'obtenir les modèles les plus performants et fiables [3].

Dans ce travail, nous effectuons une régression en utilisant un jeu de données composé de 2983 patients inclus dans des essais cliniques provenant des bases de données Pro-Act et Exhonit [1, 4]. Nous utilisons les données récupérées chez les patients entre le premier mois (T0) et le troisième mois (T3) de la maladie pour prévoir son évolution au cours de la première année (T6, T9 et T12). Nous proposons un modèle autorégressif permettant à chaque prévision de l'ALSFRS chez un patient de réinjecter cette valeur prédite dans le modèle afin de prévoir sa nouvelle valeur 3 mois plus tard. De plus, une nouvelle version améliorée de l'évolution différentielle est appliquée afin de déterminer le sous-ensemble de variables optimal.

Résultats. Le tableau 1 indique le RMSE [1] et le R^2-ajusté [2] obtenus avec et sans sélection de variables. En effectuant une sélection de variables nous parvenons à améliorer les performances de notre modèle peu importe la période. En effet, le RMSE indique que globalement, le modèle

1. La racine de l'erreur quadratique moyenne (RMSE) indique l'écart moyen entre les valeurs prédites et les valeurs observées. Plus le score est proche de 0, plus le modèle est adapté.

2. Le R^2-ajusté indique le pourcentage de variance de la variable cible expliqué par les variables explicatives. Plus le score est proche de 1, plus le modèle est adapté

est à une distance de 3.03 points en moyenne de la valeur réelle tandis que le R^2-ajusté indique que le modèle permet d'expliquer 84.37% de la variation de l'ALSFRS chez les patients.

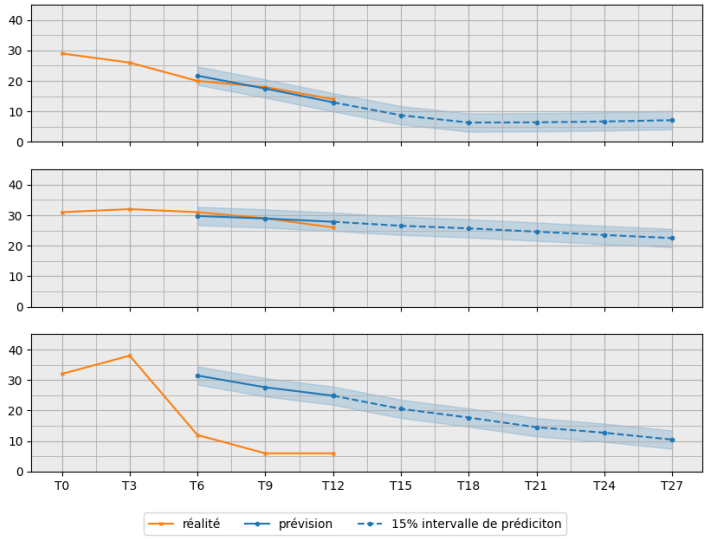

FIG. 1 – Comparaison de la progression de la maladie entre la prévision et la réalité sur 3 patients différents.

Nous obtenons également une meilleure prédiction que dans [5] sur la pente de l'ALSFRS de T3 à T12 avec un RMSE de 0.443.
Pour chacun des patients, une estimation de l'évolution de l'ALSFRS est tracée en fonction de ses caractéristiques. Le modèle fonctionne bien lorsqu'un patient présente une progression lente ou modérée de la maladie. En revanche, le modèle ne parvient pas à prévoir un rétablissement ou une très forte progression sur un court laps de temps.

Score	**Méthode**	**Variables**	**T6**	**T9**	**T12**	**Total**
RMSE	Sans Sélection	40	2.12	2.99	4.05	3.07
	Métaheuristique	29	2.10	2.95	3.98	3.03
R^2-ajusté	Sans Sélection	40	90.59	83.12	72.66	83.60
	Métaheuristique	29	91.17	84.57	75.26	84.37

TAB. 1 – Résultats des modèles avec et sans sélection de variables sur les données

Conclusion. La sélection de variables par une métaheuristique permet d'éliminer 11 variables sur les 40 présentes initialement, tout en améliorant la prédiction de la progression de la maladie. De plus, l'utilisation d'un modèle autorégressif permet de réaliser des prévisions quelle que soit la progression de la maladie chez le patient.

Références

[1] N. Atassi, J. Berry, A. Shui, N. Zach, A. Sherman, E. Sinani, J. Walker, I. Katsovskiy, D. Schoenfeld, M. Cudkowicz and M. Leitner. *The PRO-ACT database : design, initial analyses, and predictive features. Neurology*, 2014.

[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] T. Anani, J.F. Pradat-Peyre, F. Delbot. *Experimental Comparison of Metaheuristics for Feature Selection in Machine Learning in the Medical Context.* Artificial Intelligence Applications and Innovations, 2022.

[4] V. Meininger, B. Asselain, P. Guillet, P.N. Leigh, A. Ludolph, L. Lacomblez, W. Robberecht, *Pentoxifylline in als : a double-blind, randomized, multicenter, placebo-controlled trial.* Neurology, 2006.

[5] C. Pancotti, G. Birolo, C. Rollo, T. Sanavia, B. Di Camillo, U. Manera, A. Chiò and P. Fariselli. *Deep learning methods to predict amyotrophic lateral sclerosis disease progression.* Scientific Reports, 2022.