Une version optimisée de l'évolution différentielle pour la sélection de variables

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. L'application du machine learning à des données réelles, telles que celles issues d'essais cliniques, est une pratique courante. Ces jeux de données ont une disparité importante dans le nombre de variables, dont toutes ne sont pas nécessairement pertinentes pour le processus d'apprentissage en cours. La nature même de ces données implique des redondances, des omissions, des erreurs et un degré de subjectivité, tous ayant un impact négatif sur la qualité des modèles entraînés. La sélection de variables vise à optimiser la qualité des modèles de machine larning en identifiant un sous-ensemble de variables explicatives les plus adaptées à la prédiction de la variable cible, tout en éliminant les variables redondantes, bruyantes ou nocives. Il s'agit d'une étape cruciale dans la construction de modèles statistiques et de machine learning, car en éliminant les variables inutiles, il est possible de simplifier considérablement le modèle, améliorant ainsi son interprétabilité et sa robustesse tout en évitant des problèmes tels que le surajustement. Bien que diverses méthodes aient été proposées, de méthodes de filtre à des stratégies basées sur des heuristiques, la sélection de variables demeure un problème ouvert en raison du théorème du "No Free Lunch" [1], affirmant qu'aucune méthode universelle n'offre des performances supérieures pour tous les problèmes.

Dans le cadre de cette étude nous nous penchons en particulier sur la Sclérose Latérale Amyotrophique (SLA), également connue sous le nom de maladie de Charcot. Notre objectif est de prédire la survie à un an des patients en se basant sur les données disponibles après les trois premiers mois de leur prise en charge. La qualité des données à disposition peut avoir un impact significatif sur les performances des méthodes de sélection de variables et des méthodes d'apprentissage [2, 3]. Pour évaluer les performances des différentes méthodes de sélection de variables, nous utilisons plusieurs jeux de données, construits artificiellement à partir de 'Baseline', un jeu de données (lui-même artificiel) sans anomalie particulière : Données mal étiquetées (Class Noise), variables bruitées (Attribute Noise), variables redondantes (Redundant), déséquilibre entre les classes (Imbalanced), nombre conséquent de variables (Features 1), plus de variables que d'instances (Features 2) ou une combinaison de tout cela (All). De plus, le jeu de données 'Madelon' est ajouté au sein de l'expérience de manière à pouvoir comparer nos résultats avec ceux de l'état de l'art [4]. Nous présentons une amélioration de l'heuristique de l'évolution différentielle [5] utilisée pour la sélection de variables que nous nommons Tournament in Differential Evolution (TiDE). Notre heuristique est basé sur une stratégie de mutation ad-hoc utilisant un tournoi, une initialisation basé sur une méthode de filtre et une optimisation des paramètres de l'heuristique.

Résultats. Le tableau 1 indique l'exactitude pondérée[1] obtenu avec et sans sélection de variables pour chacun des jeux de données. Les résultats indiquent un impact significatif de

1. L'exactitude pondérée est la moyenne entre la sensibilité ($\frac{TP}{TP+FN}$) et la spécificité ($\frac{TN}{TN+FP}$). Représente le pourcentage moyen des scores de prédictions de chaque classe pondéré par la taille de chaque classe, permettant à un modèle de correctement identifier les individus de chaque classe, sans être influencé par le déséquilibre. Un score de 1 indique aucune erreur commise par le modèle.

la sélection de variables sur la qualité prédictive d'un modèle, en particulier quand le nombre de variables est élevé (Madelon, Features 1 et Features 2). Par exemple, la méthode 'TiDE (ReliefF)' sur Features 2 atteint un score de 97.97% comparé à 63.23% sans sélection, une amélioration de 34.67 points de pourcentage. L'évolution différentielle, et en particulier notre variante hybride, surpassent les autres méthodes dans l'ensemble des jeux de données dont la SLA.

TAB. 1 – Résultats des méthodes de sélection de variables pour chacun des jeux de données. Le score (exactitude pondérée) indiqué est le meilleur obtenu parmi 7 méthodes d'apprentissage ; la régression logistique, la régression Ridge, la machine à vecteurs de support, les K plus proches voisins, la forêt aléatoire, l'analyse discriminante linéaire et la classification naïve bayésienne.

Type	Méthode	SLA	Madelon	Baseline	Class Noise	Attribute Noise	Redundant	Imbalanced	Features 1	Features 2	All
—	Sans Sélection	71.89%	69.17%	73.23%	67.90%	75.07%	80.73%	74.21%	65.57%	63.23%	69.30%
Filtre	Corrélation (Spearman)	66.65%	87.17%	77.82%	72.16%	77.82%	76.82%	76.77%	77.83%	70.00%	69.94%
	Anova	65.16%	87.83%	77.82%	73.66%	78.49%	77.32%	77.06%	77.83%	71.33%	72.30%
	Information mutuelle	66.47%	76.67%	71.66%	64.82%	70.66%	77.32%	77.06%	77.83%	71.33%	72.30%
	MRMR	67.94%	72.00%	76.16%	69.49%	75.82%	76.66%	75.03%	75.83%	71.33%	64.29%
	ReliefF	66.98%	86.33%	88.82%	79.82%	88.82%	77.49%	81.80%	89.00%	72.00%	72.95%
Heuristique	Recherche Aléatoire	81.07%	78.05%	79.70%	74.38%	79.88%	82.99%	78.92%	72.07%	72.10%	71.41%
	Recherche Tabou	77.50%	88.22%	88.47%	80.83%	88.63%	87.86%	87.15%	81.70%	91.50%	80.40%
	Algorithme Génétique	81.37%	88.07%	88.49%	80.59%	88.74%	87.84%	86.96%	81.04%	89.43%	80.42%
	PBIL	81.87%	90.17%	90.86%	79.36%	91.04%	88.86%	89.45%	82.34%	95.89%	81.68%
	Evolution Différentielle	81.83%	95.13%	90.85%	81.08%	90.87%	88.90%	89.62%	85.75%	97.70%	86.90%
	TiDE (ReliefF)	81.96%	95.85%	91.53%	86.40%	91.60%	89.45%	90.11%	92.30%	97.97%	84.54%

Conclusion. Dans ce travail, nous avons comparé les performances expérimentales de différentes méthodes de sélection de variables dans le but d'optimiser la qualité prédictive des modèles entraînés. En particulier, nous avons examiné des données issues du domaine médical, qui présentent plusieurs défis (faible volume de données, données manquantes, données bruyantes et/ou inutiles). L'heuristique de l'évolution différentielle semble se distinguer. Nous avons donc proposé une nouvelle version optimisée (initialisation de la population à l'aide d'une méthode de filtre, nouvelle stratégie de mutation) qui obtient des résultats encore meilleurs. Cette bonne performance est intrigante étant donné son manque de popularité dans la littérature scientifique pour le problème spécifique de la sélection de variables. Nous recommandons donc l'utilisation de l'évolution différentielle plutôt que d'autres méthodes. À l'avenir, il serait intéressant d'expliquer cette supériorité apparente d'un point de vue théorique, ou d'identifier des jeux de données pour lesquels ses performances seraient plus mesurées.

Références

[1] D.H. Wolpert and W.G. Macready. *No free lunch theorems for optimization.* IEEE Transactions on Evolutionary Computation, 1997.

[2] 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.

[3] K. Hopf, S. Reifenrath. *Filter Methods for Feature Selection in Supervised Machine Learning Applications - Review and Benchmark.* arXiv, 2021.

[4] A. Bommert, X. Sun, B. Bischl, J. Rahnenführer, M. Lang. *Benchmark for filter methods for feature selection in high-dimensional classification dat.* Computational Statistics & Data Analysis, 2020.

[5] R. Storn and K. Price. *Differential Evolution – A Simple and Efficient Heuristic for global Optimization over Continuous Space.* Journal of Global Optimization, 1997.