Amélioration de l'interprétabilité des explications de *SHAP* grâce à la découverte de sous-groupes

Maëlle Moranges*, Thomas Guyet*

*AIstroSight, Inria, Hospices Civils de Lyon, UCBL
maelle.moranges@inria.fr

Si les modèles prédictifs sont désormais largement employés en médecine, la question de fournir des explications compréhensibles et cliniquement pertinentes reste un défi. *SHAP* (Lundberg et Lee, 2017), aujourd'hui largement utilisé, produit des explications locales et globales mais présente plusieurs limites : 1) il indique l'importance moyenne des variables sans préciser leurs effets concrets, 2) il ne capture pas les interactions entre variables et 3) il peut masquer des comportements propres à des sous-populations. Les règles décisionnelles *SI–ALORS* constituent une alternative intéressante car elles sont proches du raisonnement clinique et permettent de décrire des profils spécifiques. Toutefois, les approches existantes génèrent des règles uniquement locales (Ribeiro et al., 2018; Guidotti et al., 2019) ou uniquement globales (Yuan et al., 2022). Pour répondre à ces limitations, nous proposons une méthode agnostique au modèle combinant *SHAP* et la découverte de sous-groupes (Wrobel, 1997), afin de produire des règles cohérentes, multidimensionnelles, et dont les prémisses sont exploitables comme explications locales et globales.

La découverte de sous-groupes identifie des règles R sous forme de conjonctions de conditions décrivant un sous-groupe associé à une classe cible c. Par exemple, "$\mathsf{age} \in [80, 95[\wedge \mathsf{gender} = \mathsf{female}$" forme une règle caractérisant un sous-groupe à risque de *crise cardiaque*. La qualité d'un sous-groupe peut être évaluée par la *WRAcc* (Lavrač et al., 2004).

Nous proposons de faire une extraction de règles similaires en adaptant la mesure qualité de la règle de sorte que la prémisse d'une règle n'implique que des termes qui correspondent à des variables ayant des valeurs de Shapley élevées, i.e. qui sont fortement pris en compte par le modèle pour forger la décision de la règle. On souhaite ainsi privilégier les sous-groupes cohérents avec les contributions explicatives du modèle.

Nous proposons ainsi la *WRAcc* pondérée d'une règle R par les valeurs SHAP :

$$\mathrm{WRAcc}_{\phi}(R) \;=\; \frac{W\left(\mathrm{cov}(R)\right)}{W\left(D\right)} \cdot \left(\frac{W\left(\mathrm{cov}_{(c)}(R)\right)}{W\left(\mathrm{cov}(R)\right)} - \frac{W\left(D_{(c)}\right)}{W\left(D\right)} \right)$$

où $W(X)$ somme les contributions des valeurs de SHAP de toutes les variables de la règle R pour tous les exemples d'un ensemble X, D est l'ensemble des exemples et $\mathrm{cov}(R)$ représente les exemples qui satisfont la prémisse de la règle R. L'indice $^{(c)}$ précise que les exemples de ces ensembles doivent en plus être prédits dans la classe c par le modèle à expliquer.

Les règles maximisant WRAcc_{ϕ} sont recherchées via une *beam search*. Les deux paramètres d'entrée sont : la profondeur maximale et le nombre de règles par classe.

$\text{WRAcc}_\phi(R)$	R	c
0.084	prevalentHyp=0	$\to 0$
0.076	male=0 $\wedge$ prevalentHyp=0	$\to 0$
0.076	male=0	$\to 0$
0.073	age < 0.26	$\to 0$
0.071	diabetes=0 $\wedge$ prevalentHyp=0	$\to 0$
0.071	BPMeds=0 $\wedge$ prevalentHyp=0	$\to 0$
0.069	prevalentHyp=0 $\wedge$ prevalentStroke=0	$\to 0$
0.063	BPMeds=0 $\wedge$ male=0	$\to 0$
0.084	prevalentHyp=1	$\to 1$
0.082	age ≥ 0.74	$\to 1$
0.077	sysBP ≥ 0.32	$\to 1$
0.076	male=1	$\to 1$
0.069	prevalentHyp=1 $\wedge$ prevalentStroke=0	$\to 1$
0.068	age ≥ 0.74 $\wedge$ prevalentStroke=0	$\to 1$
0.066	age ≥ 0.74 $\wedge$ diabetes=0	$\to 1$
0.065	prevalentHyp=1 $\wedge$ sysBP ≥ 0.32	$\to 1$

$\tilde{\phi}_x(R)$	R	c
0.2174	prevalentHyp=1 $\wedge$ sysBP ≥ 0.32	$\to 1$
0.1473	sysBP ≥ 0.32	$\to 1$
0.0828	male=1	$\to 1$
0.0701	prevalentHyp=1	$\to 1$

FIG. 1 – *Exemples d'explications globales (à gauche) et locales (à droite) pour le jeu framingham.*

Les règles extraites offrent des explications globales, mais il est également possible d'en dériver une explication locale. Pour une instance, nous retenons les règles globales qui la couvrent et dont toutes les variables présentent une contribution SHAP positive. Elles sont classées selon leur contribution moyenne et constituent l'explication locale.

Nous avons évalué l'approche sur quatre jeux de données médicaux (*Framingham*, *Heart-attack*, *Covid19*, *Obesity*), chacun associé à un modèle performant. Les valeurs SHAP sont obtenues via `SHAP` et les sous-groupes via `pysubgroup`. Pour évaluer les règles locales, nous mesurons : (1) la *fidélité* (accord entre l'explication et le modèle), (2) la *précision* (accord avec la vérité terrain), (3) la *complétude* (proportion d'instances expliquées), (4) la *cohérence* (accord entre règles locales). Pour les règles globales, nous rapportons la *WRAcc* et le *lift*. Les règles extraites présentent une fidélité élevée (> 0.9) sur les jeux de données binaires et une bonne complétude (> 0.8). Elles ont des *WRAcc* toujours positives et des *lifts* supérieurs à 1, confirmant leur pertinence pour caractériser les classes cibles.

Références

Guidotti, R., A. Monreale, F. Giannotti, D. Pedreschi, S. Ruggieri, et F. Turini (2019). Factual and counterfactual explanations for black box decision making. *IEEE Intelligent Systems 34*(6), 14–23.

Lavrač, N., B. Kavšek, P. Flach, et L. Todorovski (2004). Subgroup discovery with CN2-SD. *Journal of Machine Learning Research 5*, 153–188.

Lundberg, S. M. et S.-I. Lee (2017). A unified approach to interpreting model predictions.

Ribeiro, M. T., S. Singh, et C. Guestrin (2018). Anchors : High-precision model-agnostic explanations. In *Proceedings of the AAAI conference*, pp. 1527–1535.

Wrobel, S. (1997). An algorithm for multi-relational discovery of subgroups. In *European symposium on principles of data mining and knowledge discovery*, pp. 78–87. Springer.

Yuan, J., B. Barr, K. Overton, et E. Bertini (2022). Visual exploration of machine learning model behavior with hierarchical surrogate rule sets. *IEEE Transactions on Visualization and Computer Graphics 30*(2), 1470–1488.