UNIVERSITÉ DE RENNES 1

Direction de la Recherche et de l'Innovation

Appel à projets - Année 2022

Contrats doctoraux «Ordinaires»

ECOLE DOCTORALE : MathSTIC **POLE DE RECHERCHE :**

Unité de recherche d'accueil

Unité de Recherche : IRISA

Equipe : LACODAM

Encadrement

Directeur de thèse : ALEXANDRE TERMIER **Unité de recherche** : IRISA

Employeur/Établissement d'origine : UNIVERSITE DE RENNES 1

HDR : X oui ☐ non **Nb de thèses en cours** : 4

Mail : alexandre.termier@univ-rennes1.fr **Tél**. : 02 99 84 71 13

Co-directeur de thèse : **Unité de recherche** :

Employeur/Établissement d'origine :

HDR : ☐ oui ☐ non **Nb de thèses en cours** :

Mail : **Tél.** :

Sujet

Titre : Dépasser la divination : stabiliser l'interprétabilité des algorithmes de machine learning

Mots clés : apprentissage automatique (Machine Learning), interprétabilité (eXplainable Artificial Intelligence), séries temporelles

Cotutelle internationale (oui/non) :non **Si oui, préciser l'établissement et le pays de rattachement** :

Descriptif (15-30 lignes) :

Les algorithmes d'apprentissage automatique (Machine Learning) sont aujourd'hui extrêmement répandus et utilisés. Ils sont présents dans une multitude de domaines (finance, médecine, aide à la décision, etc.). Malheureusement, les méthodes de Machine Learning ne sont pas parfaites et dans des domaines comme la médecine ou des vies peuvent être en jeu, il est indispensable de pouvoir justifier le résultat retourné. C'est aussi un enjeu pour que ces systèmes soient réellement utilisés, il faut que les utilisateurs puissent avoir confiance en eux. C'est pourquoi un nouveau domaine de recherche a vu le jour : XAI (eXplainable Artificial Intelligence). Ce domaine est particulièrement utile pour les systèmes de Machine Learning dont le fonctionnement interne ne permet pas de comprendre facilement comment le résultat a été obtenu (réseaux de neurones, méthodes ensemblistes).

Depuis quelques années, des méthodes pour expliquer les sorties d'algorithmes de Machine Learning ont été développées [1]. Les approches de référence, comme LIME [2] et SHAP [3], sont dites agnostiques (elles s'appliquent en sortie de n'importe quel type d'algorithme de Machine Learning) et locale (elles donnent une explication de la prédiction pour un exemple particulier). Bien que ces méthodes puissent donner de bonnes explications, leur principal souci est la stabilité des explications : lorsque l'on exécute plusieurs fois ces algorithmes sur un même exemple, l'explication trouvée n'est pas toujours la même. Cela rend actuellement difficile leur déploiement dans des scénarios réels, où l'on attend des réponses constantes sur les mêmes problèmes.

Quelques premiers travaux [3][4] ont été réalisés pour essayer d'améliorer la stabilité du résultat des algorithmes d'explication. Dans ce sujet, nous nous attaquons au sujet de la stabilité des méthodes d'interprétabilité (ou explicabilité) dans le cas particulier où les données d'entrée de l'algorithme de

Machine Learning sont des séries temporelles (séquences ordonnées dans le temps de valeurs numériques, en général issues de capteurs comme des relevés de température). Ce type de données est encore assez peu considéré dans la communauté XAI ([6],[7]), malgré sa grande importance applicative (de plus en plus de données issues de capteurs).

Notre équipe a proposé LEFTIST [5], un premier algorithme permettant de retourner une explication pour la classification de séries temporelles. Dans cet algorithme une explication est un ensemble de sous-séries temporelles extraites de la série temporelle à expliquer (les parties de la série temporelle ayant été jugées responsables de la classification).

Notre sujet consiste à prendre en compte le problème de stabilité pour des algorithmes prenant en entrée des séries temporelles (comme LEFTIST). La notion de stabilité est ici plus complexe que pour des données de type attribut/valeur. En effet deux attributs sont, soit égaux, soit différents, tandis que deux sous séries temporelles peuvent se chevaucher et être seulement partiellement identiques.

Les notions de stabilité des explications et de fidélité de l'algorithme d'explication au système d'apprentissage initial sont à définir pour les séries temporelles et à affiner pour tous les types de données. L'objectif est ici de proposer un type d'algorithme d'explication locale et agnostique qui s'appuie sur ces définitions de stabilité et de fidélité pour mettre en œuvre un bon compromis de ces propriétés.

1. Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, Dino Pedreschi, "A Survey of Methods for Explainng Black Box Models", ACM computing surveys, volume 51, Issue 5, 2019, p1-p42
2. Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin, "Why Should I Trust You?", KDD'16, p1135-1144
3. Lundberg, Scott M. and Su-In Lee. A Unified Approach to Interpreting Model Predictions. NIPS 2017
4. R. Guidotti and S. Ruggieri, "On The Stability of Interpretable Models", International Joint Conference on Neural Networks (IJCNN), 2019
5. G. Visani, E. Bagli, F. Chesani, A. Poluzzi, D. Capuzzo, "Stability indices for LIME : obtaining reliable explanations for Machine Learning models", 2020 arXiv
6. Maël Guillemé, Véronique Masson, Laurence Rozé, Alexandre Termier, « Agnostic local explanation for time series classification ». ICTAI, 2019, p432-439
7. Udo Schlegel, Duy Lam Vo, Daniel A. Keim, Daniel Seebacher, TS-MULE : Local Interpretable model-agnostic explanations for time series forecast models, AIMLAI 2021