Méthodes d'explication d'algorithmes de Machine Learning pour des séries temporelles

Laurence Rozé, IRISA -Lacodam, INSA Rennes, laurence.roze@irisa.fr
Véronique Masson, IRISA -Lacodam, Univ. Rennes 1, veronique.masson@irisa.fr

Lieu: Irisa, équipe Lacodam, Rennes

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

Contexte

Les algorithmes de Machine Learning sont aujourd'hui extrêmement répandus et utilisés. Ils sont présents dans une multitude de domaines comme le monde de la finance, la médecine, l'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 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, etc.).

Depuis quelques années, des méthodes pour expliquer les sorties d'algorithmes de Machine Learning ont été développées. Elles peuvent être de différents types :

- Agnostiques ou spécifiques. Une méthode agnostique peut être appliquée à n'importe quel type d'algorithme de Machine Learning tandis qu'une méthode spécifique sera dédiée à un type particulier d'algorithme.
- Locales ou globales. Une explication locale cherche à expliquer le résultat de l'algorithme de ML pour un exemple donné (quelque soit cet exemple) tandis qu'une méthode globale essaye d'expliquer l'algorithme pour l'ensemble des exemples.

Description du stage

Ce sujet de stage s'intéresse aux méthodes locales et agnostiques, telles que LIME [1] et SHAP [2], c'est-à-dire cherchant à expliquer le résultat d'un exemple donné indépendamment de l'algorithme de machine learning utilisé. Mais des limites à ces méthodes subsistent.

Certains domaines d'application, ou type de données, sont actuellement encore peu explorés, comme les séries temporelles, et c'est à ce type de données que nous nous intéressons ici. Une série temporelle est une suite de valeurs numériques représentant l'évolution d'une quantité spécifique au cours du temps. Par exemple, si l'on prend le relevé sur 24h d'un compteur électrique on aura une série temporelle contenant 24 valeurs, si le pas de temps d'obtention des valeurs est 1h, et 24x60 si le pas de temps est la minute. LEFTIST [3], développé dans Lacodam, est 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).

Un autre souci connu des algorithmes actuels de méthodes agnostiques et locales d'interprétabilité [1][2][3] est un problème de stabilité : lorsque l'on exécute plusieurs fois ces algorithmes sur un même exemple, le résultat retrouvé n'est pas toujours le même. Quelques premiers travaux [4][5] ont été réalisés pour essayer d'améliorer la stabilité du résultat. Notre sujet consiste à prendre en compte le problème de stabilité pour des algorithmes prenant en entrée des séries temporelles. 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.

Ce travail pourra être validé sur des séries temporelles provenant de consommation d'énergie.

Compétences

compétences en apprentissage automatique (Machine Learning) nécessaires
connaissances en statistique appréciées

Références bibliographiques

[1] « "Why Should I Trust You?": Explaining the Predictions of Any Classifier », *M. T. Ribeiro, S. Singh, C. Guestrin*, in KDD ‘16: The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining : 1135--1144, 2016

[2] « A Unified Approach to Interpreting Model Predictions », *Scott M. Lundberg, Su-In Lee ;* in Advances in Neural Information Processing Systems 30 : 4765--4774, 2017

[3] « Agnostic Local Explanation for Time Series Classification » *M. Guillemé, V. Masson, L. Rozé, A. Termier*, in IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI) : 432--439, 2019

[4] « Statistical stability indices for LIME: obtaining reliable explanations for Machine Learning models », *G. Visani, E. Bagli, F. Chesani, A. Poluzzi, D. Capuzzo*, arXiv, 2020

[5] « OptiLIME: Optimized LIME Explanations for Diagnostic Computer Algorithms », *G. Visani, E. Bagli, F. Chesani,* in AIMLAI, workshop on Advances in Interpretable Machine Learning and Artificial Intelligence, CIKM Conference, oct. 2020 (à paraître)