Institut de Recherche
en Informatique de Toulouse

Offre de thèse

Raisonnement formellement certifié en apprentissage automatique

L'intelligence artificielle est présente un peu partout dans l'industrie, et se fraie une place dans des applications où ses décisions peuvent avoir des conséquences dramatiques (véhicules autonomes, décisions financières ...). L'explicabilité des modèles issus de l'apprentissage automatique vise à aider les humains à comprendre les décisions prises par ces modèles complexes. Le besoin d'explicabilité est un des défis pour l'utilisation des techniques d'apprentissage automatique, dans les systèmes critiques mais également pour des problèmes d'éthiques (aide à la décision pour des banques ou la médecine par exemple).

Expliquer les décisions prises par un classifieur est important mais souvent difficile d'un point de vue de sa complexité. Il existe plusieurs formes d'explication (abductive, contrastive) et plusieurs requêtes possibles (par exemple savoir si un attribut protégé appartient à une explication ou énumérer toutes les explications). Pour certaines familles de classifieurs et pour certaines requêtes, la compléxité est polynomiale (par exemple les classifieurs monotones ou encore les arbres de décision). Pour avoir une garantie de la fiabilité des explications fournies par ces algorithmes, il faut vérifier formellement ces algorithmes.

Cette thèse portera sur la recherche de nouvelles familles de problèmes d'explication traitables en compléxité polynomiale, suivi de la construction de logiciels d'explication. L'utilisateur devra avoir confiance en ces logiciels, donc soit ils seront prouvés corrects, soit ils devront générer un certificat de correction des réponses. Le but final est de rendre public un logiciel efficace et fiable.

Références

[1] Clément Carbonnel, Martin C. Cooper, and João Marques-Silva. Tractable explaining of multivariate decision trees. In Pierre Marquis, Tran Cao Son, and Gabriele Kern-Isberner, editors, *Proceedings of the 20th International Conference on Principles of Knowledge Representation and Reasoning, KR 2023*, pages 127–135, 2023.

[2] Martin C. Cooper and João Marques-Silva. Tractability of explaining classifier decisions. *Artif. Intell.*, 316 :103841, 2023.

[3] Aurélie Hurault and João Marques-Silva. Certified logic-based explainable AI - the case of monotonic classifiers. In Virgile Prevosto and Cristina Seceleanu, editors, *Tests and Proofs - 17th International Conference, TAP 2023, Leicester, UK, July 18-19, 2023, Proceedings*, volume 14066 of *Lecture Notes in Computer Science*, pages 51–67. Springer, 2023.

Informations pratiques Il est attendu du candidat un bagage en informatique théorique nottament en compléxité et/ou en méthodes formelles.

La thèse sera financée par le projet ANR ForML (ANR-23-CE25-0009) à l'IRIT, en cotutelle entre Toulouse INP encadrée par Aurélie Hurault et Université Toulouse 3 encadré par Martin Cooper.

Un stage de Master2 est également disponible pour les étudiants intéressés par ce sujet de thèse.

Contacts : aurelie.hurault@enseeiht.fr et Martin.Cooper@irit.fr..

Institut de Recherche
en Informatique de Toulouse

Ph.D. thesis proposal

Formally Certified Reasoning in Machine Learning

Artificial intelligence is present everywhere in industry, and is making its way into applications where its decisions can have dramatic consequences (such as autonomous vehicles, finance and medicine). Explanation of models derived from machine learning is crucial to help humans understand the decisions made by these complex models. Explainability is one of the major challenges facing the use of machine learning techniques, especially in safety-critical systems but also concerning ethical issues (e.g. decision support for banks or medicine).

Explaining decisions made by a classifier is important but often difficult from the point of view of computational complexity. There are several forms of explanation (abductive, contrastive) and several possible queries (for example, knowing if a protected attribute belongs to an explanation or listing all explanations). For certain families of classifiers and for certain queries, complexity is polynomial (for example, monotone classifiers or decision trees). Furthermore, to have a guarantee of the reliability of the explanations provided by these algorithms, it is necessary to formally verify these explanation algorithms.

This thesis will focus on the search for new families of explanation problems that can be handled in polynomial complexity, followed by the construction of explanation software. The user must have confidence in these programs, so either they will be proven correct or they will generate a certificate to justify correctness. The ultimate goal is to make publicly available efficient and reliable software.

References

[1] Clément Carbonnel, Martin C. Cooper, and João Marques-Silva. Tractable explaining of multivariate decision trees. In Pierre Marquis, Tran Cao Son, and Gabriele Kern-Isberner, editors, *Proceedings of the 20th International Conference on Principles of Knowledge Representation and Reasoning, KR 2023*, pages 127–135, 2023.

[2] Martin C. Cooper and João Marques-Silva. Tractability of explaining classifier decisions. *Artif. Intell.*, 316:103841, 2023.

[3] Aurélie Hurault and João Marques-Silva. Certified logic-based explainable AI - the case of monotonic classifiers. In Virgile Prevosto and Cristina Seceleanu, editors, *Tests and Proofs - 17th International Conference, TAP 2023, Leicester, UK, July 18-19, 2023, Proceedings*, volume 14066 of *Lecture Notes in Computer Science*, pages 51–67. Springer, 2023.

Practical information The candidate is expected to have a background in theoretical computer science, particularly in complexity and/or formal methods.

The thesis will be funded by the ANR ForML project (ANR-23-CE25-0009) at IRIT, and co-supervised by Aurélie Hurault from Toulouse INP and Martin Cooper from the University of Toulouse 3.

A Master's degree internship is also available for candidates interested by this thesis proposal.

Contacts : aurelie.hurault@enseeiht.fr and Martin.Cooper@irit.fr