AfIA
Association française
pour l'Intelligence Artificielle

Recherches en IA explicable dans l'équipe LFI du LIP6

Par **Christophe MARSALA**
Isabelle BLOCH
Marie-Jeanne LESOT
Sabrina TOLLARI
Jean-Noël VITTAUT
LIP6 / LFI
Sorbonne Université, CNRS
{prénom}.{nom}@lip6.fr
http://lfi.lip6.fr

L'équipe *Learning Fuzzy and Intelligent systems* (LFI) du laboratoire LIP6 de Sorbonne Université développe des recherches en interprétabilité des méthodes d'intelligence artificielle dans les domaines de l'aide à la décision, la science des données et l'apprentissage automatique. Les objectifs scientifiques et applicatifs sont de concevoir et de proposer des approches à la fois explicables durant leur construction et lors de leur utilisation.

Cet article présente brièvement ces contributions développées dans le cadre de multiples collaborations, en évoquant les tâches d'apprentissage automatique, dans des approches *by design* (dès la conception) ou *post-hoc*, la caractérisation de données par résumés linguistiques, l'interprétation d'images ou la formulation d'explications dans un cadre logique.

Variables linguistiques et logique floue

La logique floue a été proposée, par Zadeh en 1965, avec l'objectif de modéliser le raisonnement humain, fournissant dès sa conception même un outil pour faciliter l'interprétation des manipulations effectuées : la représentation de degrés de vérité au-delà d'une dichotomie manichéenne vrai/faux, de transitions progressives entre ces cas extrêmes et de limites imprécises donne une plus grande souplesse et une meilleure lisibilité des traitements réalisés.

En particulier, ce formalisme offre la possibilité de modéliser la sémantique vague de termes linguistiques, comme *proche*, défini sur l'univers des distances $\mathbb{R}^+$, *jeune*, défini sur l'univers des âges, ou plus généralement *faible, moyen, élevé* pour toute valeur numérique. De plus, l'interprétation de ces termes peut être précisée par l'utilisateur selon son expertise d'interprétation des variables. Ce formalisme permet donc une intégration aisée de connaissances propres à l'utilisateur, et par là la personnalisation des outils proposés. La représentation floue des variables numériques à l'aide de termes linguistiques offre ainsi une grande intelligibilité pour un utilisateur humain non spécialiste en IA [16]

De façon générale, la logique floue et la théorie des sous-ensembles flous constituent des outils naturels pour le domaine de l'IA explicable, mis en œuvre dans de nombreux cadres par l'équipe LFI et illustrés tour à tour ci-dessous.

Interprétabilité by design

De nombreux travaux de l'équipe LFI portent sur l'amélioration de l'interprétabilité des modèles d'apprentissage à la fois durant leur construction et dans leur utilisation pour

la classification.

La représentation floue des termes linguistiques permet de prendre en compte la gradualité des frontières, tout en limitant la complexité des modèles construits par apprentissage. Ces approches reposent sur l'extension des algorithmes d'apprentissage pour leur permettre de prendre en compte cette représentation floue, tout en respectant leur dimension numérique, ce qui les éloigne des approches strictement symboliques. Cela passe en premier lieu par une étude des mesures impliquées dans ces algorithmes et leur généralisation pour pouvoir les appliquer à des données floues [7], tout en conservant les propriétés intrinsèques de l'algorithme d'apprentissage. Le modèle d'étude de base dans ces travaux est, dans un cadre d'apprentissage supervisé, le modèle de construction de règles de décision, par exemple l'apprentissage d'arbres de décision flous [15, 17].

Interprétabilité post-hoc

Les approches d'interprétabilité *post-hoc* visent à proposer des explications intelligibles à tout système construit par apprentissage automatique, indépendamment de cette phase d'apprentissage. Elles peuvent être agnostiques, c'est-à-dire ne pas faire d'hypothèse sur le type de classifieur utilisé, ou au contraire exploiter des informations sur celui-ci.

Dans un cadre agnostique, les travaux menés dans l'équipe LFI se placent en particulier dans le domaine de la génération d'exemples contrefactuels [11], qui expliquent une prédiction en indiquant les modifications à apporter à la donnée considérée pour changer la classe prédite. Ces travaux, initiés en collaboration avec l'équipe R&D de Marcin DETYNIECKI du groupe AXA, se prolongent dans le cadre du *Trustworthy and Responsible AI Lab* (TRAIL), laboratoire de recherche commun à Sorbonne Université et AXA créé en décembre 2021.

Dans un cadre non-agnostique considérant l'apprentissage profond pour l'analyse d'images, il s'agit d'expliquer des résultats de classification ou segmentation. Les explications *post-hoc* exhibent les zones de l'image ou les caractéristiques qui contribuent le plus à la décision. Des travaux en imagerie biologique et médicale sont menés avec le LTCI / Télécom Paris, l'ISIR / Sorbonne Université, des équipes industrielles et plusieurs hôpitaux parisiens (par exemple [9, 18, 22]). Les recherches en cours portent sur la recherche d'explications non seulement locales, mais aussi structurelles, impliquant plusieurs objets dans les images et leur organisation spatiale.

Résumés linguistiques

La génération de textes à partir de données numériques ou catégorielles, tâche aussi appelée *data-to-text*, est une approche classique d'interprétabilité, les formulations linguistiques étant considérées comme faciles à comprendre par tout utilisateur. Les résumés linguistiques ajoutent l'objectif de fournir une vue synthétique, facilitant plus encore la compréhension du contenu des données. Les résumés par protoformes, introduits initialement par Yager, s'écrivent par exemple, dans leur forme basique, QRX sont P, où X représente les données à résumer, Q un quantificateur, comme la plupart ou quelques, et R et P des termes linguistiques correspondant à des propriétés d'intérêt. Q, P et R peuvent être représentés comme des variables linguistiques dans le formalisme des sous-ensembles flous, par exemple pour représenter un résumé tel que « la plupart des vols ayant un retard important sont des vols longs ».

Outre des questions d'efficacité calculatoire posées par l'explosion combinatoire de l'exploration des résumés possibles, étudiées en collaboration avec l'IRISA-Lannion [23], de nombreuses questions d'interprétabilité sont

néanmoins à soulever [13] : la formulation linguistique peut s'avérer être un piège en ce qu'elle semble intuitive, mais peut être ambiguë. En effet, il se peut que l'utilisateur n'ait pas la même compréhension de la phrase que le sens exprimé par la mesure de qualité implémentée, pour laquelle de multiples définitions peuvent être proposées [21]. Il est aussi possible que l'interprétation considérée des termes ne soit pas en adéquation avec la structure sous-jacente des données, ce qui peut conduire à des expressions linguistiques trompeuses [14].

Dans le cas d'expressions numériques approximatives, du type *environ* x où x est un nombre, une adéquation cognitive doit également être prise en compte, afin d'éviter le risque d'interprétation erronée. Des travaux menés en collaboration avec des psychologues cogniticiens de CHART-Paris VIII ont montré qu'elle doit tenir compte de la magnitude, du dernier chiffre significatif, de la granularité et de la complexité de x [12].

Au-delà des questions d'interprétation des phrases isolées, l'interprétabilité d'un résumé porte aussi sur les phrases dans leur ensemble, notamment pour tenir compte de relations de redondance, qui nuisent à l'intelligibilité, voire de contradictions à justifier [19, 20].

Approches hybrides

Le cadre de l'IA hybride vise à modéliser et utiliser à la fois des connaissances et des données en combinant plusieurs pans de l'IA. Des approches logiques permettent de représenter et raisonner sur des connaissances. Celles-ci sont ensuite transcrites dans des modèles computationnels structurels (graphes, hypergraphes, ontologies, treillis de concepts). Ces structures sont enrichies de modèles des imprécisions inhérentes aux descriptions linguistiques des connaissances dans la théorie des ensembles flous [4, 5] (par exemple « la structure A est à droite de la structure B »). Le problème du fossé sémantique entre les concepts abstraits et les domaines concrets des données est résolu là encore dans le cadre des ensembles flous à l'aide de la notion de variable linguistique présentée au début de l'article. Ces approches, appliquées à l'interprétation d'images, permettent de conserver le lien entre les données et les connaissances, et les connaissances utilisées pour une tâche de décision (reconnaissance d'objets par exemple) fournissent des explications des décisions (par exemple, les relations spatiales utilisées pour reconnaître des objets, confrontées aux connaissances a priori sur l'organisation spatiale des objets dans la scène observée). Ces résultats peuvent alors être exprimés sous forme de descriptions linguistiques du contenu des images.

Une réflexion en cours, en particulier avec des radiologues et des philosophes des sciences [10], porte sur les questions éthiques pour lesquelles les approches hybrides de l'explicabilité pourraient apporter des éclairages.

Interfaces explicatives

Au-delà de la génération d'explications, la construction d'interfaces permettant de les visualiser et de les manipuler s'avère une composante essentielle de leur adoption par les utilisateurs : de nombreux outils d'explications s'adressent à des spécialistes d'apprentissage automatique et d'intelligence artificielle, et non à des utilisateurs sans cette expertise. Des travaux menés en collaboration avec AXA et CHART-Paris VIII sur des interfaces de présentation d'explications de type instances contrefactuelles et vecteurs d'importance locale montent l'intérêt de la contextualisation et de l'interaction, à la fois pour la compréhension objective et la satisfaction subjective [8].

Approches symboliques

Des approches purement symboliques, en particulier logiques, sont également dévelop-

pées dans l'équipe, selon deux directions. Une première approche, en collaboration avec le MICS / CentraleSupelec, le LAMSADE / Université Paris Dauphine, le CRIL / Université d'Artois et l'ULA (Merida, Vénézuela), porte sur des méthodes d'abduction, où une observation est expliquée par une formule logique en fonction d'une base de connaissances. Des exemples concrets d'opérateurs d'abduction ont été proposés dans le cadre de la morphologie mathématique, d'abord en logique propositionnelle, puis dans un cadre plus général englobant, entre autres, la logique floue [1, 2, 3].

Une deuxième direction, inspirée des travaux de Halpern et Miller, repose sur des arbres causaux. Des travaux en cours portent sur les liens entre les modèles structurels causaux et les systèmes d'argumentation abstraits (avec l'ISIR / Sorbonne Université), ainsi que sur une formalisation floue des explications par contraste dans le cadre des modèles structurels causaux [6].

Références

[1] M. Aiguier, J. Atif, I. Bloch, and R. Pino Pérez. Explanatory relations in arbitrary logics based on satisfaction systems, cutting and retraction. *International Journal of Approximate Reasoning*, 102 :1–20, 2018.

[2] M. Aiguier and I. Bloch. Logical dual concepts based on mathematical morphology in stratified institutions. *Journal of Applied Non-Classical Logics*, 29(4) :392–429, 2019.

[3] J. Atif, C. Hudelot, and I. Bloch. Explanatory reasoning for image understanding using formal concept analysis and description logics. *IEEE Transactions on Systems, Man and Cybernetics : Systems*, 44(5) :552–570, May 2014.

[4] I. Bloch. Fuzzy sets for image processing and understanding. *Fuzzy Sets and Systems*, 281 :280–291, 2015.

[5] I. Bloch. Mathematical morphology and spatial reasoning : Fuzzy and bipolar setting. *TWMS Journal of Pure and Applied Mathematics – Special Issue on Fuzzy Sets in Dealing with Imprecision and Uncertainty : Past and Future Dedicated to the memory of Lotfi A. Zadeh*, 12(1) :104–125, 2021.

[6] I. Bloch and M.-J. Lesot. Vers une formulation floue des explications par contraste. In *Rencontres Francophones sur la Logique Floue et ses Applications*, pages 191–198, 2021.

[7] B. Bouchon-Meunier and C. Marsala. Entropy and monotonicity in artificial intelligence. *International Journal of Approximate Reasoning*, 124 :111–122, 2020.

[8] C. Bove, J. Aigrain, M.-J. Lesot, C. Tijus, and M. Detyniecki. Contextualization and exploration of local feature importance as explanations to improve understanding and satisfaction of non-expert users. In *International Conference on Intelligent User Interfaces*, 2022.

[9] V. Couteaux, S. Si-Mohamed, O. Nempont, T. Lefevre, A. Popoff, G. Pizaine, N. Villain, I. Bloch, A. Cotten, and L. Boussel. Automatic knee meniscus tear detection and orientation classification with Mask-RCNN. *Diagnostic and Interventional Imaging*, 100 :235–242, 2019.

[10] V. Israël-Jost et al. L'éthique en radiologie : quand, comment ? premiers éléments. *Journal d'Imagerie Diagnostique et Interventionnelle*, 4 :238–240, 2021.

[11] T. Laugel, M.-J. Lesot, C. Marsala, X. Renard, and M. Detyniecki. The dangers of post-hoc interpretability : Unjusti-

fied counterfactual explanations. In *28th International Joint Conference on Artificial Intelligence*, pages 2801–2807, 2019.

[12] S. Lefort, M.-J. Lesot, E. Zibetti, C. Tijus, and M. Detyniecki. Interpretation of approximate numerical expressions : Computational model and empirical study. *International Journal on Approximate Reasoning*, 82 :193–209, 2017.

[13] M.-J. Lesot, G. Moyse, and B. Bouchon-Meunier. Interpretability of fuzzy linguistic summaries. *Fuzzy Sets and Systems*, 292(1) :307–317, 2016.

[14] M.-J. Lesot, G. Smits, and O. Pivert. Adequacy of a user-defined vocabulary to the data structure. In *International Conference on Fuzzy Systems*, 2013.

[15] C. Marsala. Fuzzy decision trees for dynamic data. In *IEEE Symposium on Evolving and Adaptive Intelligent Systems*, pages 17–24, 2013.

[16] C. Marsala and B. Bouchon-Meunier. Fuzzy data mining and management of interpretable and subjective information. *Fuzzy Sets and Systems*, 281 :252–259, 2015.

[17] C. Marsala and D. Petturiti. Rank discrimination measures for enforcing monotonicity in decision tree induction. *Information Sciences*, 291 :143–171, 2015.

[18] G. Martin, S. El-Madafri, A. Becq, J. Szewczyk, and I. Bloch. Instruments Segmentation in X-ray Fluoroscopic Images for Endoscopic Retrograde Cholangio Pancreatography. In *Medical Informatics Europe*, 2022.

[19] G. Moyse, M.-J. Lesot, and B. Bouchon-Meunier. Oppositions in fuzzy linguistic summaries. In *International Conference on Fuzzy Systems*, 2015.

[20] A. Oudni, M.-J. Lesot, and M. Rifqi. Processing contradiction in gradual itemset extraction. In *International Conference on Fuzzy Systems*, 2013.

[21] A. Oudni, M.-J. Lesot, and M. Rifqi. Accelerating effect of attribute variations : accelerated gradual itemsets extraction. In *International Conference on Information Processing and Management of Uncertainty*, pages 395–404. Springer, 2014.

[22] A. Pirovano, L. G. Almeida, S. Ladjal, I. Bloch, and S. Berlemont. Computer-aided diagnosis tool for cervical cancer screening with weakly supervised localization and detection of abnormalities using adaptable and explainable classifier. *Medical Image Analysis*, 73 :102167, 2021.

[23] G. Smits, P. Nerzic, O. Pivert, and M.-J. Lesot. Frels : Fast and reliable estimated linguistic summaries. In *International Conference on Fuzzy Systems*, 2021.