Foreword

Bernadette Bouchon-Meunier

"Foreword", chapter in Paving the Way from Interpretable Fuzzy Systems to Explainable AI Systems, vol. 970, Studies in Computational Intelligence, pp. vii-ix, (Springer International Publishing), (ISBN: 978-3-030-71098-9) (2021)

Foreword

The explainability of decisions based on Artificial Intelligence systems goes beyond a scientific issue, as Artificial Intelligence is nowadays ubiquitous and many organizations have drawn our attention to the necessity to provide the decision-maker and the decision subject with an easy to understand explanation. After the DARPA's advocacy for Explainable AI[1], the European Commission published guidelines[2] to make AI systems and their decisions transparent and explained to all parties involved. Many efforts are currently made by AI researchers to provide solutions to this problem, generally approaching only one of the aspects of explainability, mainly interpretability, understandability, expressiveness or transparency. These aspects are clearly interrelated and lead to a better acceptability of decisions, as well as trustworthiness in the decision-making process.

Of all these components, interpretability and expressiveness are often considered to be of primary importance for the acceptability of systems based on artificial intelligence by users, considered from the point of view of natural language expression of decisions and their reasons. However, most knowledge in human brains is implicit and therefore not verbalizable, so that human beings themselves are used to decisions based on non-expressible reasoning. It should also be noted that words are not the only knowledge representation easily understandable by users, as graphs, charts, histograms or statistics can be well appreciated, depending on their clarity and the expertise of the user.

In addition to interpretability and expressiveness, Explainable Artificial Intelligence requires that the reasons for the decision based on artificial intelligence be presented to the user. The automatic generation of explanations has various forms and we can cite for instance the use of feature importance vectors that quantify the relative impact of each feature in the prediction[3], an

[1] https://www.darpa.mil/program/explainable-arti_cial-intelligence

[2] https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines- trustworthy-Ai

[3] D. Baehrens, T. Schroeter, S. Harmeling, K. Motoaki, K. Hansen, K.R. Muller: How to Explain Individual Classification Decisions. Journal of Machine Learning Research 11, (2010) 1803-1831.

example of which is the interpretability method LIME[4]. Visualization provides another form of explanation, in particular through partial dependence plots[5] which show the marginal effect of a feature over the model outcome. The comparison of decisions with particular ones is another solution to explain the decision, for instance using prototypes[6] or counterfactuals[7]. The detection of the most influential training instances influential for a given prediction gives another form of explanation[8]. A general overview of these approaches[9] provides more aspects of automatic explanations.

The most popular methods to address the problem of the explainability of models based on artificial intelligence are based on decision rules[10] and decision trees[11]. They are at the core of the very comprehensive book prepared by J. M. Alonso, C. Castiello, L. Magdalena and C. Mencar in the specific realm of a fuzzy knowledge representation.

A certain conception of interpretability and expressiveness can be dealt with immediately by fuzzy systems, since fuzzy sets were created by Lotfi Zadeh to treat classes or concepts in a manner similar to the human way of manipulating them. The paradigm of Computing With Words[12] has, in particular, been introduced to establish a bridge between computation and natural language.

---

[4] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should I trust you?": Explaining the predictions of any classifier. In Proc. of the 22nd ACM SIGKDD Int. Conf. On Knowledge Discovery and Data Mining (KDD'16), pages 1135–1144, 2016.

[5] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 29(5):1189–1232, 2001.

[6] Been Kim, Cynthia Rudin, and Julie A Shah. The Bayesian Case Model: A generative approach for case-based reasoning and prototype classification. In Advances in Neural Information Processing Systems, pages 1952–1960, 2014.

[7] David Martens and Foster Provost. Explaining Data-Driven Document Classifications. Mis Quarterly, 38(1):73–99, 2014.

[8] Mayank Kabra, Alice Robie, and Kristin Branson. Understanding classifier errors by examining influential neighbors. Proc. of the IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR'15), pages 3917–3925, 2015.

[9] Th. Laugel, Local Post-hoc Interpretability for Black-box Classifiers, PhD Thesis, Sorbonne Université, July 2020.

[10] Ryan Turner. A model explanation system. NIPSWorkshop on Black Box Learning and Inference, 2015.

[11] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Dino Pedreschi, Franco Turini, and Fosca Giannotti. Local rule-based explanations of black box decision systems. arXiv preprint 1805.10820, 2018a.

[12] L. A. Zadeh, Fuzzy logic = computing with words, in IEEE Transactions on Fuzzy Systems, vol. 4, no. 2, pp. 103-111, 1996

Lotfi Zadeh showed later that a fuzzy knowledge-based representation can be employed to represent the meaning by means of a constraint-based semantics of natural languages[13].

To approach the meaning of artificial intelligence-based decision supports is certainly one of the requests of Explainable Artificial Intelligence. In this sense, fuzzy classes or categories appear more acceptable to users than crisp classes, as they avoid the risk of an arbitrary boundary preventing the users to reach an expected decision category of which they are very close. Expressed in natural language, such categories look familiar to the user and easily understandable. Moreover, fuzzy rule-based systems and fuzzy decision trees add to the general capabilities of decision rules and decision trees to be directly approached the flexibility and familiarity of classes with gradual boundaries.

This remarkable book explains in details how fuzzy models can participate in the construction of decision systems that are interpretable, and moreover explainable. It goes far beyond the classic representation of fuzzy classes and fuzzy rule-based systems. It establishes a bridge with natural language processing and automatic text generation and it proposes solutions to deal with the well-known interpretability-accuracy trade-off. All aspects of interpretability of fuzzy rule-based systems are covered, from interpretability rating to directly exploitable software.

It paves the way for a greater involvement of fuzzy methods in Explainable Artificial Intelligence, participating in the general search for greater acceptability of artificial intelligence-based decision systems by users, more confidence in the resulting decisions and, hopefully, more fairness in decisions.

Paris, July 2020
Bernadette Bouchon-Meunier

[13] L.A. Zadeh. From computing with numbers to computing with words. From manipulation of measurements to manipulation of perceptions. *Ann N Y Acad Sci*. 2001;929:221-252.