Interpretability and Fuzzy Data Science

Bernadette Bouchon-Meunier and Christophe Marsala

Sorbonne Université, CNRS, Laboratoire d'Informatique de Paris 6,
LIP6, F-75005 Paris, France
Bernadette.Bouchon-Meunier@lip6.fr, Christophe.Marsala@lip6.fr

Extended abstract

Nowadays, data science is very popular due to the overwhelming mass and streams of data available in the digital world. Its main purpose is to extract information from data, in order to satisfy user's expectations, needs and requests. In this setting, interpretability is one of the key elements in the evaluation of the quality of this process or of its steps. In this paper, we consider fuzzy data science where fuzzy systems are a way to improve interpretability.

The concept of interpretability is complex and multifaceted [5]. For instance, *explainable AI* is very popular and it is a hot topic in the statistical machine learning community [2]. Moreover, interpretability is also highly dependent on the end-user, of his/her knowledge, and of the application domain.

We focus on two aspects of interpretability on which fuzzy systems prove their efficiency [3, 1]. First of all, interpretability of a learning model enables the user to understand the model on which the system is based to provide information, for instance to identify a class or a decision, or to present the answer to a query. Secondly, the interpretability of the information provided by a machine learning classifier can be regarded through the capacity of a human being, non expert in machine learning, to understand this result, for instance expressed in natural language-like form.

The *interpretability of a learning algorithm* is important and it has often been taken into account in data mining. For instance, according to end users, the understandability of the decision tree-based learning algorithm is considered as higher than the one of the SVM algorithm. Indeed, this last algorithm requests more specific mathematical notions (matrix, vector, hyper-plane, model optimisation,...) than those involved in the previous one.

On this topic, few works have been proposed in the fuzzy community. The interpretability of a machine learning algorithm involves the understandability of its validity and the proof of the appropriateness of the learning mechanisms it is based on. The validity is defined here as the way the algorithm fits the theoretical process of construction of the model that have been thought about. For instance, the decision tree algorithm validity is based on the use of information theory that provides strong and rigorous explanations about its building process. In a similar way, to provide a better understandability of the fuzzy decision tree construction algorithm, justifications have been proposed to explain the extension of the classical machine learning algorithm to an algorithm adapted to handle fuzzy or imperfect knowledge [4].

Indeed, this kind of interpretability and formal proofs of the validity of the fuzzy algorithm are crucial to better explain to end users how the algorithm proceeds and why she/he could be confident in it.

The *interpretability of the results of a classifier* is important when these results are proposed to an end-user. Such a result should be either understandable by itself, or explainable by means of additional information.

The output can be semantically explained, and in this case it should refer to classic expressions in the involved domain. Moreover, it should also refer to intuitive knowledge such as "membership to a class", easily understandable even by a user unaware of fuzzy modelling. In this kind of interpretability, even classic data mining algorithms can be concerned as it is easy to provide them with a post-processing step in which the output of the model is fuzzified to offer linguistic labels to the user.

On the other hand, explanations could be provided to help the user to understand the obtained results. In this case, in statistical machine learning, current works focus on providing mathematical explanations to interpret classification results [6]

Thus, in both of these aspects, fuzzy systems provide solutions because of their ability to interact with human beings. The importance of fuzzy solutions to various steps of data science methods has for instance been highlighted in [5].

References

1. Casillas, J., Cordón, O., Triguero, F.H., Magdalena, L.: Interpretability issues in fuzzy modeling, vol. 128. Springer (2013)
2. Gunning, D.: Explainable artificial intelligence (xai). Defense Advanced Research Projects Agency (DARPA), nd Web (2017)
3. Hüllermeier, E.: Fuzzy sets in machine learning and data mining. Applied Soft Computing 11, 1493–1505 (2011)
4. Marsala, C., Bouchon-Meunier, B.: Choice of a method for the construction of fuzzy decision trees. In: Proc. of the IEEE Int. Conf. on Fuzzy Systems, FUZZ-IEEE. pp. 584–589. St Louis, USA (May 2003)
5. Marsala, C., Bouchon-Meunier, B.: Fuzzy data mining and management of interpretable and subjective information. Fuzzy Sets and Systems 281, 252–259 (Dec 2015)
6. Ribeiro, M.T., Singh, S., Guestrin, C.: Why should I trust you?: Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1135–1144. ACM (2016)