Constrained Interval Type-2 Fuzzy Classification Systems for Explainable AI (XAI)

Pasquale D'Alterio, Jonathan M. Garibaldi and Robert I. John
IMA and LUCID Research Groups
School of Computer Science, University of Nottingham
Nottingham, UK
{pasquale.d'alterio, jon.garibaldi, robert.john}@nottingham.ac.uk

***Abstract*—In recent year, there has been a growing need for intelligent systems that not only are able to provide reliable classifications but can also produce explanations for the decisions they make. The demand for increased explainability has led to the emergence of explainable artificial intelligence (XAI) as a specific research field. In this context, fuzzy logic systems represent a promising tool thanks to their inherently interpretable structure. The use of a rule-base and linguistic terms, in fact, have allowed researchers to create models that are able to produce explanations in natural language for each of the classifications they make. So far, however, designing systems that make use of interval type-2 (IT2) fuzzy logic and also give explanations for their outputs has been very challenging, partially due to the presence of the type-reduction step. In this paper, it will be shown how constrained interval type-2 (CIT2) fuzzy sets represent a valid alternative to conventional interval type-2 sets in order to address this issue. Through the analysis of two case studies from the medical domain, it is shown how explainable CIT2 classifiers are produced. These systems can explain which rules contributed to the creation of each of the endpoints of the output interval centroid, while showing (in these examples) the same level of accuracy as their IT2 counterpart.**

***Index Terms*—Constrained interval type-2, XAI, explainable type-2 fuzzy systems**

I. Introduction

Classification models have been widely adopted in recent years to tackle problems in a variety of fields, ranging from image classification to medical data analysis. Although state of the art classifiers are able to produce good predictions, many of them are unable to provide meaningful explanations for the classifications. Therefore, it can be very challenging to understand the decision process followed by the classifiers, especially in models such as neural networks that behave as *black boxes* [1]. However, in scenarios that significantly affect users, an interpretable model is required to ensure fair, non-discriminatory treatment, to validate the output of the system against experts' knowledge, and to detect any inconsistencies in the classification process [2], [3]. As a consequence of this, the field of explainable artificial intelligence (XAI) has risen in popularity in recent years [4]. Its ambitious goal is to build a new generation of intelligent models that not only are reliable in their predictions but can also be interpreted by their end-users.

Thanks to their rule-based structure and the use of linguistic labels [5], fuzzy logic systems (FLSs) inherently represent a promising tool to tackle this new challenge. Albeit their level of interpretability heavily depends on factors such as the number of rules and membership functions (MF) involved [2], [6]. In the literature, there are many successful applications of explainable type-1 (T1) FLSs. In many published articles, e.g. [7]–[9], it is possible to see how the rule-based structure, together with the use of linguistic labels, can be used to provide an explanation in natural language for each of the classifications produced by the systems.

Generating explanations for each of the outputs of an interval type-2 (IT2) [10] FLS, on the other hand, remains very challenging due to the different nature of the defuzzification process and the presence of the additional type-reduction step. When the endpoints of the interval centroid are computed using a procedure like the Karnik-Mendel (KM) algorithm [11] or one of its enhanced derivatives, it is not straightforward to create a direct relation between the embedded sets (ES) that generated the endpoints and the rule base of the system [12], [13]. Since IT2 FLSs have been shown to outperform T1 FLSs in many areas including classification (e.g. [14]), their use in explainable systems could lead to a similar improvement in performance.

Constrained interval type-2 (CIT2) fuzzy sets were first introduced by Garibaldi and Guadarrama [15] as a new way to model vague concepts starting from a T1 fuzzy set modeling the same concept, called a *generator set*. CIT2 modeling constrains the shape of the footprint of uncertainty (FOU) [16] that can be generated and considers as acceptable only the ES that have the same shape as the generator set — thereby introducing the concept of acceptable embedded sets (AESs). The goal is to use these additional constraints to keep a semantic connection between the CIT2 FS and the word it models while ensuring that only embedded sets with a meaningful shape are processed in operations such as the centroid defuzzification [13].

In this paper, the CIT2 defuzzification algorithm proposed by D'Alterio et al. [12] will be used to design CIT2 FLSs that provide explanations for each of their classifications. For both endpoints of the interval centroid, the AES, the rules and the input variables that contributed to their creation will be identified, adding valuable information for the understanding of the internal decision process of the system.

The rest of the paper is organized as follows: after a brief introduction on CIT2 fuzzy sets and the reasons why they were introduced, the creations of the explanations for CIT2

978-1-7281-6932-3/20/$31.00 ©2020 IEEE

FLSs will be discussed; this approach will then be applied to two case-studies in the medical domain, showing how the explanations can be obtained and the level of information they are able to provide while briefly discussing why the same level of explanation is harder to achieve with the standard IT2 representation.

II. Constrained Interval Type-2 Fuzzy Sets

Constrained interval type-2 (CIT2) fuzzy sets have been proposed to maintain a more meaningful relationship between an IT2 FS and the linguistic label it models, suitable for use in scenarios in which it is necessary to produce an FLS with a high degree of interpretability. In fact, in some contexts the standard IT2 representation can lead to FLS outputs that are hard to interpret semantically and for which it is hard to provide explanations [15], [17].

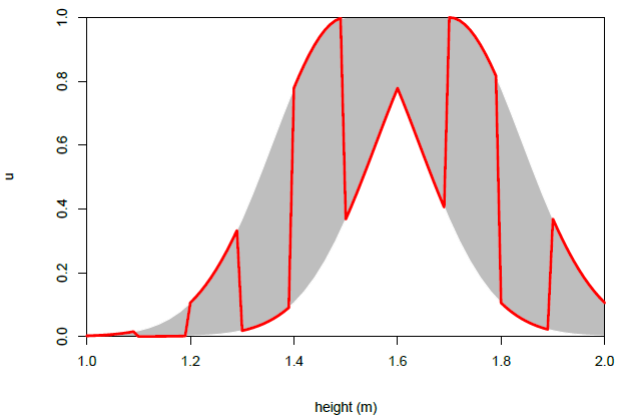


Fig. 1. IT2 FS modeling the word *medium*. Although mathematically acceptable, the ES in red could hardly represent the same word.

CIT2 FSs impose additional mathematical constraints on the definition of IT2 FS, limiting the shapes of both the FOU and the embedded sets (ESs) that are considered as acceptable (valid). As a consequence, they represent a special case of IT2 FSs. Since ESs are needed for some widely used fuzzy operators such as type-reduction and centroid defuzzification (implemented, for example, by the KM procedure), processing ESs that carry little underlying semantic meaning but are mathematically acceptable (e.g. Fig. 1) for the computation of the output, leads to results for which providing a meaningful explanation can be challenging.

Some key concepts of CIT2 sets are now reproduced (for readability). The CIT2 approach, generates an IT2 fuzzy set starting from a T1 set called a *generator set* (GS) modeling a linguistic label. The valid ESs of a CIT2 FS (i.e. the ones that can be processed by fuzzy operators), are obtained by translating the generator set along the x-axis in a given interval, called a displacement set (DS). Formally [13]:

Definition 1. *A displacement set (DS), denoted D, is a closed set of real numbers such that:*

$$D \subseteq \mathbb{R}, 0 \in D \tag{1}$$

Given a T1 generator set in conjunction with a displacement set, it is possible to define the T1 membership functions that will represent the acceptable embedded sets (AES) of the CIT2 FS.

Definition 2. *The collection of T1 acceptable embedded sets (CAES), is a set of T1 FSs obtained from the shifting of a T1 generator set G. Formally, each of the acceptable embedded sets (AES) S in a CAES can be expressed as:*

$$S = \{(x, \mu_S(x)) \,|\, x \in X\} \tag{2}$$

where

$$\mu_S : X \mapsto [0,1], \exists c \in D : \mu_S(x) = \mu_G(x - c), \forall x \in X \tag{3}$$

given a universe of discourse (UOD) X, a DS D, a T1 generator set G.

Therefore, all the AES have the same shape of the generator set. Given a CAES, we can generate a CIT2 FS:

Definition 3. *A constrained interval type-2 fuzzy set (CIT2 FS) $\breve{A}$, is defined as follows:*

$$\breve{A} = \{((x,u),\, 1) \,|\, x \in X,\, u \in \bigcup_{S \in CAES_{\breve{A}}} \mu_S(x)\} \tag{4}$$

with $CAES_{\breve{A}}$ being the CAES from which we obtain $\breve{A}$.

The upper and lower bound of the FOU can be expressed as follows:

Definition 4. *Given an CIT2 FS $\breve{A}$, we define its upper MF $\overline{\mu}_{\breve{A}}$ and lower MF $\underline{\mu}_{\breve{A}}$ as follows:*

$$\overline{\mu}_{\breve{A}}(x) = \max_{S \in CAES_{\breve{A}}} \mu_S(x) \tag{5}$$

$$\underline{\mu}_{\breve{A}}(x) = \min_{S \in CAES_{\breve{A}}} \mu_S(x) \tag{6}$$

Also the centroid can be computed using only the AES of a CIT2 FS:

Definition 5. *The constrained centroid of a CIT2 FS $\breve{A}$ is defined as the union of the centroids of its AES:*

$$C(\breve{A}) = \int_{A' \in CAES_{\breve{A}}} C(A') \tag{7}$$

Working with only these acceptable embedded sets provides increased ability for the operations that use the ESs to provide more interpretable results, thanks to the meaningful shapes of the AESs [13].

III. Explainable Constrained Interval Type-2 Fuzzy Systems

This subsection shows how the mathematical restrictions of CIT2 fuzzy sets, together with the inference and defuzzification approach described by D'Alterio et al. [12], can be used to design CIT2 FLSs that are able to provide explanations for each of the output centroids they produce.

A recent novel defuzzification algorithm for CIT2 sets [12] selects the two AES to determine the endpoints of the interval centroid of a CIT2 FLS. Each of the AESs is generated as the aggregation (by the use of the *or* operator) of all the MFs that appear as consequents in the rule-base. Each CIT2 consequent is replaced with one of its AESs (more on this below) and

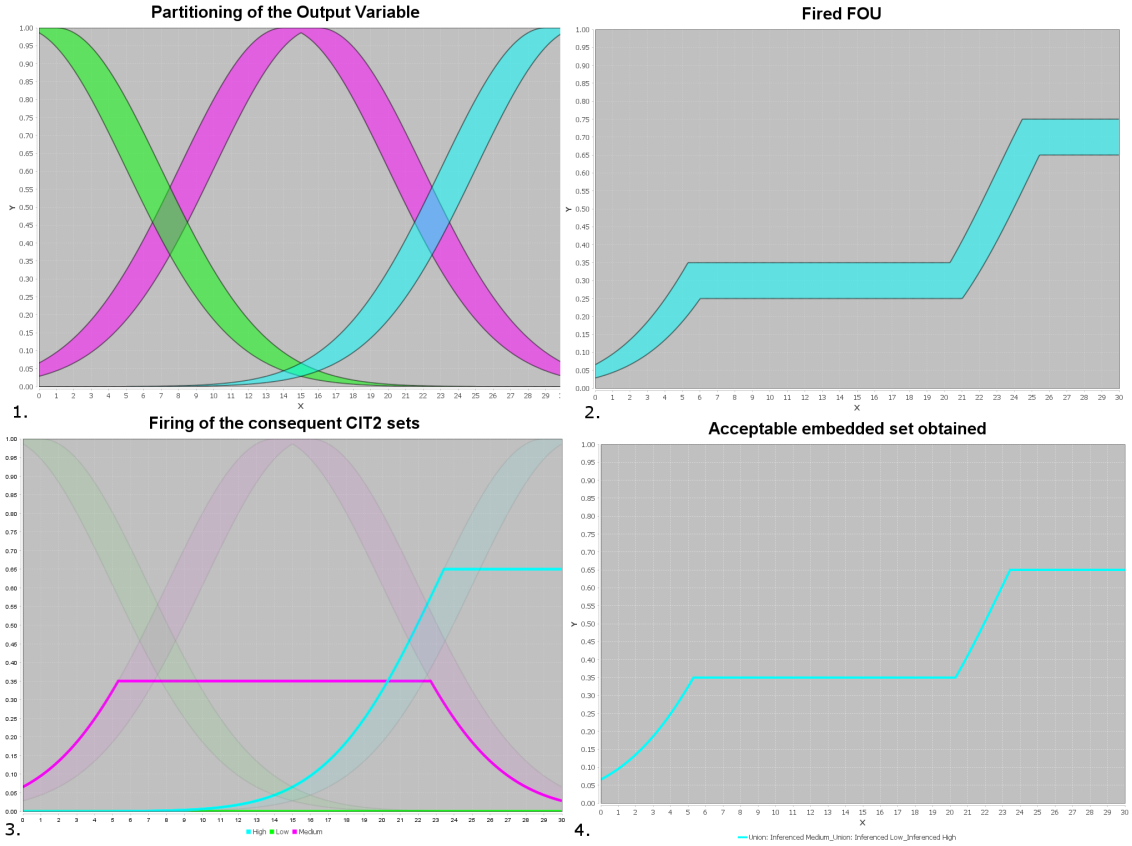


Fig. 2. Creation of the AES of the fired output (2.) that determines the left endpoint of the constrained centroid. First the partitioning of the output variable (1.) is shown, then for each consequent MF one AES is selected and inferenced (3.). Finally, the inferenced sets are aggregated to produce the final AES (4.).

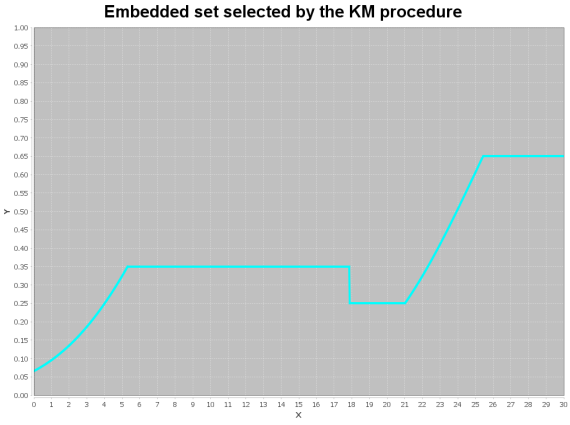


Fig. 3. The ES determining the left endpoint of the centroid of the same set as that shown in Fig. 2.2 using the KM procedure

then one of the endpoints of the firing interval of the rule they belong to is used to carry out the inference. The latter choice depends on the index value assigned to the consequent MF and on the *switch index* value that has been chosen by the algorithm [12]. By noting the rules and the firing value used for the inference on each consequent MF, it is possible to build an explanation for the final output.

The algorithm can be briefly summarized in the following steps [12]:

1) Give each CIT2 consequent MF an ordinal index by sorting them in ascending order of the minimum value of their support set.
2) For each CIT2 consequent set, compute its firing interval as the maximum lower and maximum upper values of the firing strengths of all the rules where it appears as a consequent.
3) If computing the right endpoint of the constrained centroid (i.e. to generate the AES with the maximum centroid value), replace each consequent MF with its rightmost AES; if computing the left endpoint, take the leftmost AES instead.
4) Test all the possible switch index values, between 0 and the maximum index given to the consequent MFs:
   i. If computing the left endpoint, use the upper value of the firing interval to utilise the MFs with an index smaller than the switch index and *switch* to the lower value afterwards; for the right endpoint instead, use the lower value of the firing interval before the switch index and the upper one after it.
   ii. Do the union of the AES resulting from the inference and defuzzify the set obtained.
5) Return, as the final constrained centroid, the lowest and highest centroid values obtained from the defuzzification at the previous step.

The process that leads to the creation of one of the acceptable embedded sets that determine the constrained centroid is

also shown in Fig. 2. It is straight-forward to see that the AES has been obtained as the union of two MFs (*medium* and *high*); additionally, the respective firing strengths of the rules that were used are also identifiable (i.e. the 'truncation heights' in Fig. 2.2), producing an easily interpretable AES. Once each consequent MF is replaced with one of its AESs (the leftmost or rightmost one) and for each one of them an inference value is chosen (i.e. one of the endpoints of the firing interval), all the operations are carried out using T1 mathematics. For this reason, as can also been seen in the example in Fig. 2, the AESs that determine the endpoints of the constrained centroid keep the same level of interpretability as any fuzzy output of a T1 FLS. In other words, while CIT2 FLSs allow for the modeling of uncertainty around the membership function (making use of the FOU) they also keep the same level of interpretability as T1 FLSs. On the other hand, the IT2 modeling struggles to achieve the same properties. The lower ES chosen by the KM procedure to defuzzify the same output set as that shown in Fig. 2.2 is shown in Fig. 3. Compared to the one selected by the constrained approach (Fig. 2.4), it is harder to identify how the consequent MFs contributed to its creation and it can be challenging to link it to the rules of the system and their firings [12]. This is because the KM procedure selects the two ES that solve a well-defined mathematical problem but that do not necessarily carry a semantic meaning.

Furthermore, as will be demonstrated in the next subsection and in the case studies in Sec. IV, these properties of CIT2 FLSs can be used to produce a human-readable explanation for each output of the system.

A. Generation of the explanation

In the examples provided in this paper, the explanations for the classification systems are divided into two parts: first the predicted class is presented, together with the interval centroid that generated it; then, for both endpoints of the centroid, the AESs, the rules and firing values that produced them are shown. Each rule has a different consequent MF, showing the firing strength for each of the possible classes.

The interpretable AESs provided give an intuitive idea of the firings of each class while the description with the rules that fired gives a more detailed and accurate description of the decision process followed by the FLS. The creation processes of the AESs themselves are illustrated: for each consequent MF in the rulebase one AES is chosen and inferenced using one of the endpoints of the firing interval; the union of all the inferenced set gives the AES of the fired FOU of the rulebase.

While similar explanations have already been produced for T1 FLS before (e.g. [8], [9]), they represent a novelty in the T2 field. In fact, producing explanations for IT2 and T2 FLS outputs has been very challenging since to compute the left and right endpoints of the interval centroid, all the embedded sets are processed regardless of their shape. As a consequence of that, the embedded sets that determine the endpoints of the interval centroid in the standard IT2 approach do not carry any particular meaning (making them harder to interpret), nor do they have a direct link with any of the rules of the rulebase (making the generation of an explanation less straightforward).

At this stage, there is no data gathered from users (e.g. with surveys) that determine the usefulness of the explanations of CIT2 FLS compared with IT2 ones. The superior explainability claimed in this paper is therefore based on the *ability* of CIT2 FLS to produce explanations for their classifications rather than on the users' feedback. Future work will focus on validating these claims by the use of surveys in which both approaches are compared in order to understand if the additional information provided by CIT2 FLSs is perceived as useful by domain experts.

IV. Case Studies

In this section, two case studies taken from the medical domain are analyzed. The goal is to demonstrate that the use of CIT2 FLS can be beneficial in situations in which it is important to understand the decision process behind the system classification to detect possible inconsistent decisions and/or to guarantee a fair treatment. At the same time it will be shown that, in these examples, both CIT2 and IT2 FLS achieve the same level of accuracy. The CIT2 FLSs presented below have been implemented using the novel Java library *Juzzy Constrained* [18] that supports CIT2 sets and systems.

A. Recommendation of post-operative chemotherapy for breast cancer

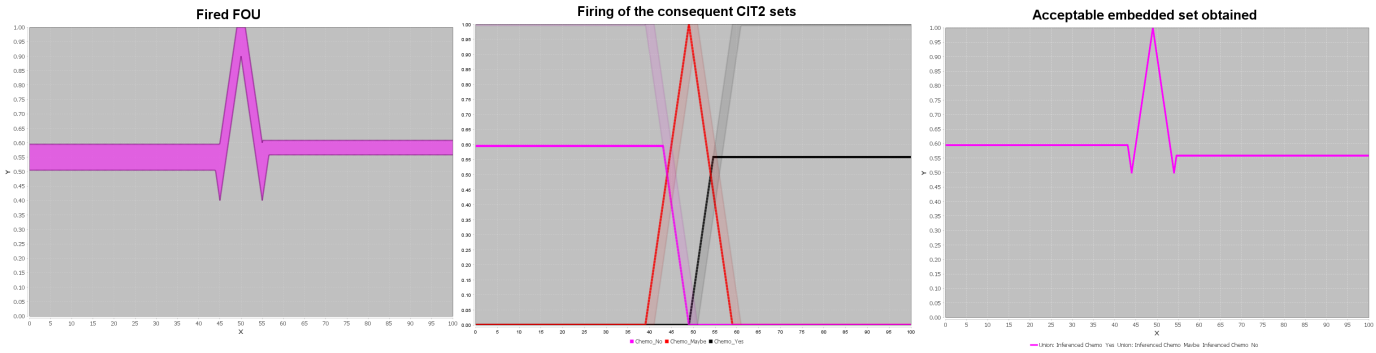
The first classification system presented here concerns the recommendation of post-operative chemotherapy for breast cancer. After the surgery to remove the tumor, a team of physicians makes a recommendation for the best additional therapy for the patient. In this case, the goal of the system is to replicate the decision process of the group of physicians with respect to the recommendation of chemotherapy. The three possible outcomes are *yes*, *no* and *maybe* with the first two cases denoting a decision in favor or against the use of chemotherapy and the latter represents the scenario in which a clear recommendation cannot be provided (e.g. because there is not an agreement among the physicians) and the post-operative therapy needs to be further discussed with the patient. The problem has already been analyzed by Garibaldi et al. [19], whereby different T1 and non-stationary [20] fuzzy systems have been designed and compared. The CIT2 FLS proposed in this paper, is based on the T1 FLS denoted as VI-F previously [19]. Its T1 MFs are used as generator sets for the corresponding CIT2 MFs; the displacement set $[-a, a]$ (i.e. the "shifting interval" used to obtain the FOU and the acceptable embedded sets, see Sec. II) has been experimentally chosen so that for each MF $|2a| = 2\%$ of the size of the universe of discourse.

The rule-base, as previously [19], is based on a written protocol provided by the Nottingham University Hospitals Trust, in order to assure a high level of interpretability. Additionally, each of the MFs used in the system models a word, such as *negative, positive, high, low* and *medium*. Fig. 4 shows an explanation provided for a case that has been classified as *maybe*, in which the output variable *chemo recommendation* is partitioned as shown in Fig. 6.

The predicted class is **MAYBE**, from the midpoint of the output [49.16, 52.2]
The leftmost centroid (49.16) is obtained from firing the following rules:

1. Chemo_no: 0.6, obtained because NPI IS High [1, 1] AND ER IS Not_Negative [1, 1] AND age IS Old [0.5, 0.6], using the upper membership degree of each input term
2. Chemo_maybe: 1, obtained because NPI IS High [1, 1] AND ER IS Not_Negative [1, 1], using the upper membership degree of each input term
3. Chemo_yes: 0.56, obtained because NPI IS High [1, 1] AND ER IS Weak [0.56, 0.61], using the lower membership degree of each input term

Aggregating these output terms produces the embedded set below, with the centroid 49.16:



The rightmost centroid (52.2) is obtained from firing the following rules:

1. Chemo_No: 0.5, obtained because NPI IS High [1, 1] AND ER IS Not_Negative [1, 1] AND age IS Old [0.5, 0.6], using the lower membership degree of each input term
2. Chemo_Maybe: 1, obtained because NPI IS High [1, 1] AND ER IS Not_Negative [1, 1], using the lower membership degree of each input term
3. Chemo_Yes term: 0.611, obtained because NPI IS High [1, 1] AND ER IS Weak [0.56, 0.611], using the upper membership degree of each input term

Aggregating these output terms produces the embedded set below, with the centroid 52.2:

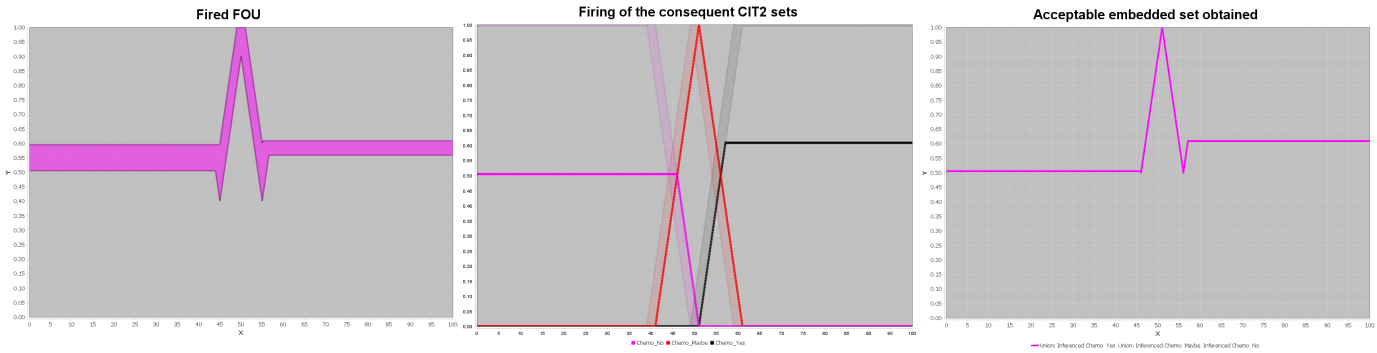


Fig. 4. Example of explanation of the output for the classification of the post-operative breast cancer treatment CIT2 FLS.

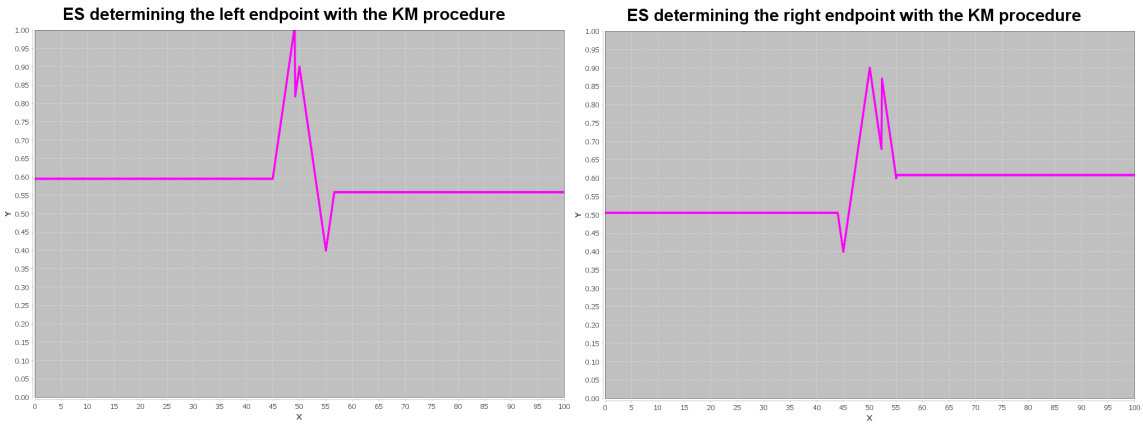


Fig. 5. Embedded sets selected by the KM procedure to defuzzify the fired FOU in Fig. 4

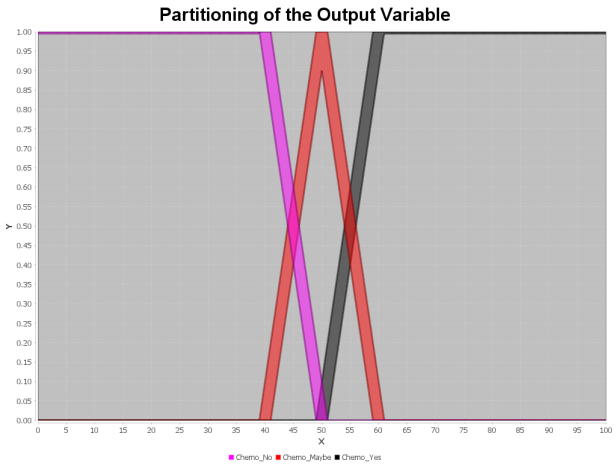


Fig. 6. Partitioning of the *chemo recommendation* variable. The FS, from left to right, model the words *no, maybe* and *yes*

Since CIT2 fuzzy sets are a subset of IT2 sets [13], the inferencing can also be carried out using the standard IT2 approach. Using the KM procedure to defuzzify the same FLS output, results in endpoints determined by the ESs shown in Fig. 5. When using the midpoint of the centroid to perform the classification, both the CIT2 and IT2 methodologies (using the KM defuzzification procedure) have an accuracy of 72.29% when tested on the same dataset used in [12].

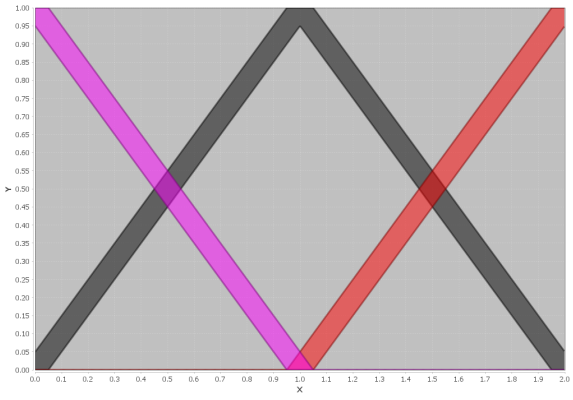


Fig. 7. Partitioning used for each of the variable in the thyroid CIT2 FLS

B. Thyroid disease diagnosis

In this case study the aim of the system is to predict whether a patient suffers from a thyroid disease (hypothyroidism or hyperthyroidism) on the basis of the analysis of some physiological data. For this system, there was no expert knowledge available from which it was possible to build the rule-base and the MFs. To build an interpretable FLS for this problem, each input variable has been partitioned with three MFs modeling the words *low, medium* and *high*, with the first and last one being implemented as triangular shoulders with their peaks being the endpoints of the universe of discourse, while the *medium* MF is as an isosceles triangle with its peak in the midpoint of the universe of discourse. The partitioning strategy described is shown in Fig. 7. The output variable is partitioned in the same way, with the 3 MFs representing respectively the terms *hypothyroidism, normal* and *hyperthyroidism*. The displacement set $[-a, a]$ (i.e. the "shifting interval" of the generator set to obtain the FOU and the acceptable embedded sets) has been experimentally chosen so that for each MF $|2a| = 5\%$ of the size of the universe of discourse.

For the rule-base, ten rules have been created using the same genetic approach described in [13] for the first stage of the optimization. Although this is one of many ways in which it is possible to generate a FLS from data, this method has been chosen with the only goal of generating a compact rulebase in which each MF identifies a meaningful linguistic label, to keep a high level of interpretability [6]. The dataset used for the learning phase is the "newthyroid" dataset available on the KEEL website [21]. The accuracy of the system produced on this dataset is 88.37% using the KM defuzzification method and 88.84% for the CIT2 version. Fig. 8 shows the explanation produced by the CIT2 FLS for one of the entries of the dataset. In comparison, the ESs that determine the endpoint of the centroid for the same FLS fuzzy output using the KM procedure are shown in Fig. 9.

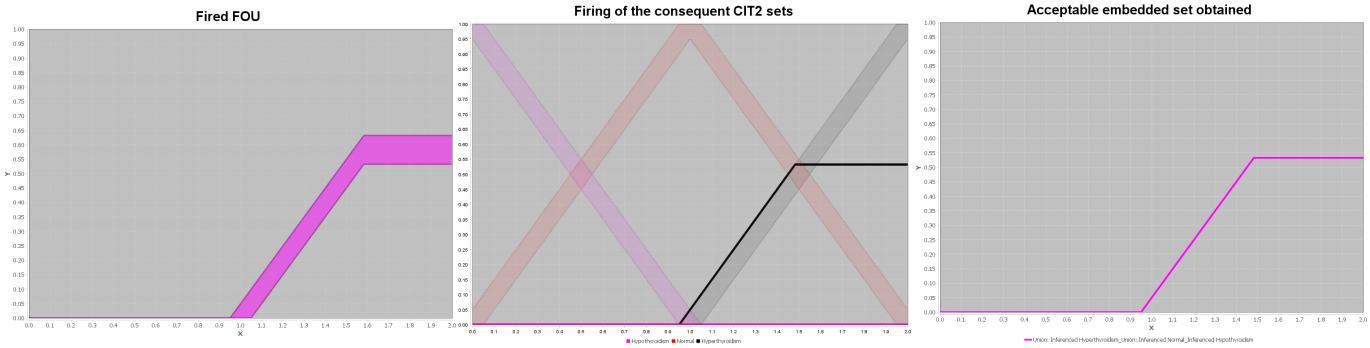
V. Discussion

In both the case studies provided, it has been shown how the previously proposed algorithm [12] can be used to produce explainable CIT2 FLSs (Figs. 4, and 8). Each of the outputs, in addition to the predicted class, also provides the interval centroid from which it was determined and an explanation for its generation. Each endpoint is then accompanied by the AES that determined it. For each of these AES an explanation for their creation is also provided, showing which rules contributed, their firing strength and the membership degree of the input values. We believe these explanations provide valuable information to understand the decision process followed by system for the following reasons:

- The presence of the interval centroid shows the effect of the uncertainty on the final output. Intuitively a 'wider' centroid represents a more uncertain result.
- As it is possible to see in Figs. 4 and 8, the AESs keep the same level of interpretability of T1 fuzzy outputs, i.e. it is possible to recognize the different terms involved (the consequent MFs) and the firing strengths of the rules they belong to (their 'truncation' heights). This provides an intuitive idea of how the constrained centroid has been obtained.
- Lastly, illustrating the rules that generated each of the AES and the membership degrees of the antecedent terms, provides a more technical and detailed explanation for the final output of the system.

The last 2 points described above represent a novelty in the IT2 field. In fact, modern algorithms like the KM [11] one and its enhanced versions are nowadays considered the standard for the defuzzification of IT2 FSs. They work by quickly identifying the embedded sets with the lowest and highest centroid value to compute the interval centroid of a set. However, although these embedded sets are mathematically acceptable and solutions to a well-defined optimization

The predicted class is **Hyperthyroidism**, from the midpoint of the output [1.59, 1.66]
The leftmost centroid (1.59) is obtained from firing the following rules:
    1. Hyperthyroidis: 0.6, obtained because T3resin IS Medium [0.53, 0.63] AND Thyroxin
        IS Medium [0.66, 0.76] AND Triiodinthyronine IS Medium [0.91, 1] AND TSH_value IS Low [0.92, 1]
        using the lower membership degree for each input term
Aggregating these output terms produces the embedded set below, with the centroid 1.59:



The rightmost centroid (1.66) is obtained from firing the following rules:
    1. Hyperthyroidism: 0.66, obtained because T3resin IS Medium [0.53, 0.63] AND Thyroxin
        IS Medium [0.66, 0.76] AND Triiodinthyronine IS Medium [0.91, 1] AND TSH_value IS Low [0.92, 1]
        using the upper membership degree for each input term
Aggregating these output terms produces the embedded set below, with the centroid 1.66:

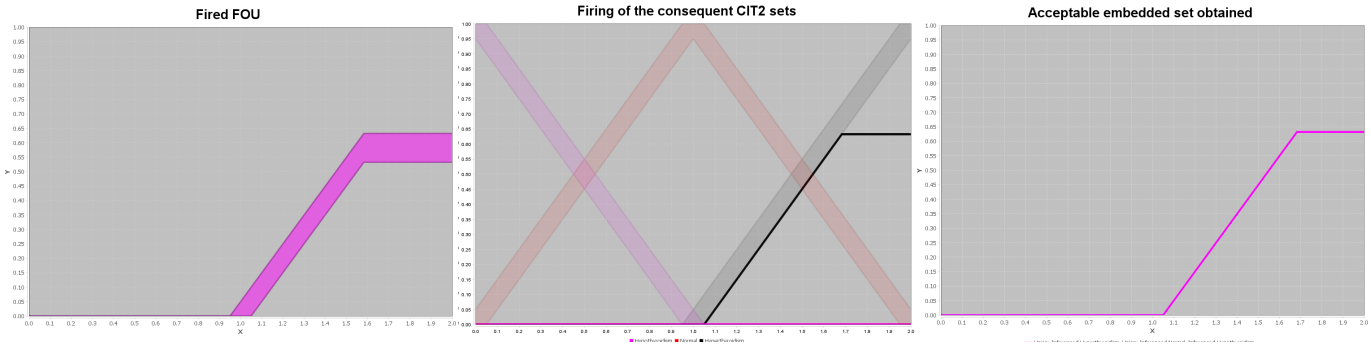


Fig. 8. Example of explanation of the output for the classification of thyroidal disease CIT2 FLS

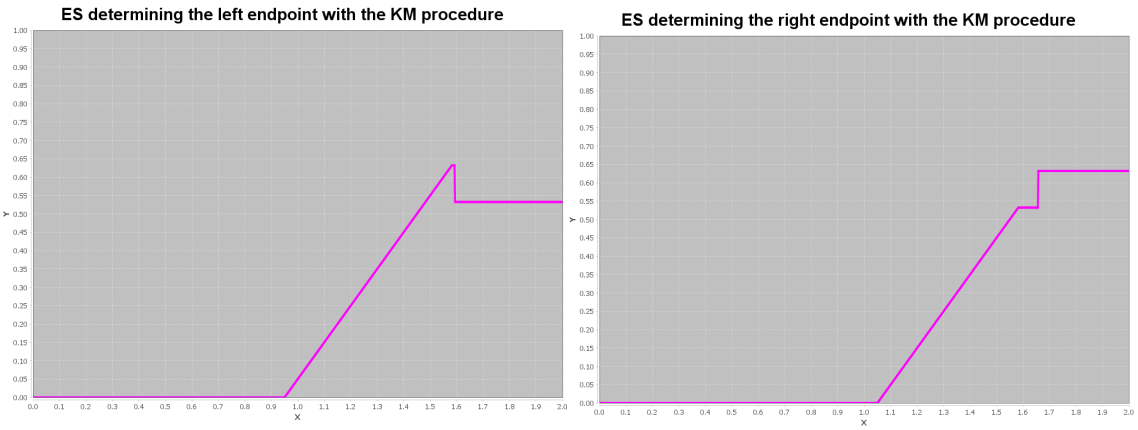


Fig. 9. Embedded sets selected by the KM to defuzzify the fired FOU shown in Fig. 8

problem, their shapes may not carry any particular meaning *in specific contexts*. That is because all the embedded sets are processed, regardless of their shape. Consequently, giving a semantic meaning to the embedded sets determined by the KM procedure may be challenging. Our claims are supported by the comparison between the embedded set chosen by the KM procedure in Figs. 5 and 9, and those produced by the constrained approach, in the explanations in Figs. 4 and 8, respectively. While the constrained embedded sets have the same level of interpretability of a T1 FLS output in which the different MFs and firing strengths are clearly identifiable, the same can not be said for the embedded sets of the KM approach. Particularly, due to the presence of the *switch point* (that is crucial for the identification of these embedded sets), the shape of the original MFs are partly lost and it is challenging to determine a direct relation between the rules of the FLS and the generation of such shapes. Therefore, building an explanation similar to the one offered by CIT2 FLS would not be straightforward.

We believe that the properties of CIT2 FLSs and the level of detailed shown in the explanations presented in our case studies, make CIT2 a valid and attractive *alternative* to IT2 FLS, in any context in which the interpretability of the system *and* a degree of explainability of the output is required.

VI. Conclusion

In this paper it has been described how the defuzzification algorithm recently proposed by D'Alterio et al. [12] can be used to design explainable CIT2 classification systems in which explanations can be provided for each of the classes predicted. In addition, it has been shown that the embedded sets processed by the CIT2 approach have a higher level of interpretability since they are built in a way that makes the identification of the linguistic terms and the firing strengths easier (see Sec. V).

To support these claims, two case studies have been analyzed, both belonging to the medical domain: the selection of post-operative therapy for breast cancer and the thyroidal disease treatment problem. In both tasks the goal of the system was to analyze some physiological data belonging to the patient in order to make a therapy recommendation or a medical decision. The CIT2 approach has been compared to the standard IT2 one, showing that CIT2 FLSs are able to produce detailed explanations for the system outputs while having similar performances in terms of the accuracy of the classification. For each classification produced, the rules involved and the firing strengths used for each of the endpoints of the centroid have be shown, providing valuable information for the understanding of the decision process of the system.

In future work, we plan on gathering statistical data from surveys to explore whether the explanations provided by CIT2 FLS are perceived as more interpretable than the IT2 ones by end-users and experts. Additionally, the information in the explanations will be reorganized in order to generate a more coherent piece of text in natural language, similarly to what has been done for T1 FLSs in other work [8], [9]. Additional work is also needed to understand how interpretable the CIT2 explanations are for the end users compared to the ones produced by T1 systems.

References

[1] D. Castelvecchi, "Can we open the black box of ai?" *Nature News*, vol. 538, no. 7623, p. 20, 2016.

[2] J. M. Alonso, C. Castiello, and C. Mencar, *Interpretability of Fuzzy Systems: Current Research Trends and Prospects*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2015, pp. 219–237.

[3] B. Goodman and S. Flaxman, "European union regulations on algorithmic decision-making and a "right to explanation"," *AI Magazine*, vol. 38, no. 3, pp. 50–57, 2017.

[4] D. Gunning, "Explainable artificial intelligence (xai)," *Defense Advanced Research Projects Agency (DARPA), nd Web*, 2017.

[5] L. A. Zadeh, "Fuzzy logic= computing with words," in *Computing with Words in Information/Intelligent Systems 1*. Springer, 1999, pp. 3–23.

[6] A. Fernandez, F. Herrera, O. Cordon, M. Jose del Jesus, and F. Marcelloni, "Evolutionary fuzzy systems for explainable artificial intelligence: Why, when, what for, and where to?" *IEEE Computational Intelligence Magazine*, vol. 14, no. 1, pp. 69–81, Feb 2019.

[7] N. Potie, S. Giannoukakos, M. Hackenberg, and A. Fernandez, "On the need of interpretability for biomedical applications: Using fuzzy models for lung cancer prediction with liquid biopsy," in *International Conference on Fuzzy Systems (FUZZ-IEEE 2019)*, 2019.

[8] I. Baaj and J.-P. Poli, "Natural language generation of explanations of fuzzy inference decisions," in *International Conference on Fuzzy Systems (FUZZ-IEEE 2019)*, 2019.

[9] J. M. Alonso, A. Ramos-Soto, E. Reiter, and K. van Deemter, "An exploratory study on the benefits of using natural language for explaining fuzzy rule-based systems," in *2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, July 2017, pp. 1–6.

[10] J. M. Mendel, R. I. John, and F. Liu, "Interval type-2 fuzzy logic systems made simple," *IEEE Transactions on Fuzzy Systems*, vol. 14, no. 6, pp. 808–821, Dec 2006.

[11] N. N. Karnik and J. M. Mendel, "Centroid of a type-2 fuzzy set," *Information Sciences*, vol. 132, no. 1-4, pp. 195–220, 2001.

[12] P. D'Alterio, J. M. Garibaldi, and R. I. John, "A fast and explainable centroid defuzzification method for constrained interval type-2 fuzzy systems," *submitted to IEEE Transactions on Fuzzy Systems*, 2020.

[13] P. D'Alterio, J. M. Garibaldi, R. John, and A. Pourabdollah, "Constrained interval type-2 fuzzy sets," *IEEE Transactions on Fuzzy Systems*, pp. 1–1, 2020.

[14] A. H. M. Pimenta and H. A. Camargo, "Interval type-2 fuzzy classifier design using genetic algorithms," in *International Conference on Fuzzy Systems*, July 2010, pp. 1–7.

[15] J. M. Garibaldi and S. Guadarrama, "Constrained type-2 fuzzy sets," in *Advances in Type-2 Fuzzy Logic Systems (T2FUZZ), 2011 IEEE Symposium on*. IEEE, 2011, pp. 66–73.

[16] J. M. Mendel and R. John, "Footprint of uncertainty and its importance to type-2 fuzzy sets," in *Proceedings 6th IASTED Int'l. Conf. on Artificial Intelligence and Soft Computing (ASC 2002)*, July 2002, pp. 587 – 592.

[17] P. D'Alterio, J. M. Garibaldi, and R. John, "On the concept of meaningfulness in constrained type-2 fuzzy sets," in *International Conference on Fuzzy Systems (FUZZ-IEEE 2019)*, 2019.

[18] P. D'Alterio, J. M. Garibaldi, R. I. John, and C. Wagner, "Juzzy constrained: Software for constrained interval type-2 fuzzy sets and systems in Java," in *2020 IEEE World Congress on Computational Intelligence (WCCI 2020)*, July 2020.

[19] J. M. Garibaldi, S.-M. Zhou, X.-Y. Wang, R. I. John, and I. O. Ellis, "Incorporation of expert variability into breast cancer treatment recommendation in designing clinical protocol guided fuzzy rule system models," *Journal of Biomedical Informatics*, vol. 45, no. 3, pp. 447 – 459, 2012.

[20] J. M. Garibaldi, M. Jaroszewski, and S. Musikasuwan, "Nonstationary fuzzy sets," *IEEE Transactions on Fuzzy Systems*, vol. 16, no. 4, pp. 1072–1086, Aug 2008.

[21] J. Alcalá-Fdez, A. Fernández, J. Luengo, J. Derrac, S. García, L. Sánchez, and F. Herrera, "Keel data-mining software tool: data set repository, integration of algorithms and experimental analysis framework." *Journal of Multiple-Valued Logic & Soft Computing*, vol. 17, 2011.