










FLocalX - Local to Global Fuzzy Explanations for Black Box Classifiers

Guillermo Fernandez¹ , Riccardo Guidotti² , Fosca Giannotti³ ,
Mattia Setzu² , Juan A. Alejo¹ , Jose A. Gámez¹ , and Jose M. Puerta¹ 

¹ Intelligent Systems and Data Mining Lab, Albacete, Spain
{Guillermo.Fernandez, JuanAngel.Aledo, Jose.Gamez, Jose.Puerta}@uclm.es

² University of Pisa, Pisa, Italy
{riccardo.guidotti, mattia.setzu}@unipi.it

³ Scuola Normale Superiore, Pisa, Italy
fosca.giannotti@sns.it

Abstract. The need for explanation for new, complex machine learning models has caused the rise and growth of the field of *eXplainable Artificial Intelligence*. Different explanation types arise, such as *local explanations* which focus on the classification for a particular instance, or *global explanations* which aim to show a global overview of the inner workings of the model. In this paper, we propose FLocalX, a framework that builds a fuzzy global explanation expressed in terms of fuzzy rules by using local explanations as a starting point and a metaheuristic optimization process to obtain the result. An initial experimentation has been carried out with a genetic algorithm as the optimization process. Across several datasets, black-box algorithms and local explanation methods, FLocalX has been tested in terms of both fidelity of the resulting global explanation, and complexity. The results show that FLocalX is successfully able to generate short and understandable global explanations that accurately imitate the classifier.

Keywords: XAI · Optimization · Metaheuristics · Fuzzy Rule-Based Systems · Local Explanations · Global Explanations

1 Introduction

In recent years, the increasing amount of data has allowed new, more complex models to be incorporated into a wide range of tasks [5, 6, 26]. However, the increasing complexity usually causes a decrease in model interpretability [2], which may not be advisable or suitable in certain critical fields, i.e., medicine, law, aviation, etc. Current European legislation also deals with this topic by means of the *right to explanation* included in the General Data Protection Regulation [18], which affects both humans and artificial intelligence techniques. *eXplainable Artificial Intelligence* (XAI) [3, 9] aims to push the usage of interpretability and explainability in order to gain an understanding of complex black box models used in sensitive contexts and critical areas.

Within the XAI taxonomy, one important distinction is whether a method generates *local* or *global* explanations. *Local explanations* are aimed at individual instances, and explain the decisions made by the model in a small neighborhood of the feature space around an instance, while *global explanations* aim to explain the entire behavior of the model. A common type of local explanation are factual and counterfactual explanations [7,8]. Factual explanations explain the reasoning behind a decision, while counterfactual explanations highlight the necessary changes to revert that decision. Focusing on decision rules as explanations, LORE (LOcal Rule-based Explainer) [8] is a well-known XAI algorithm that generates both factual and counterfactual local explanations by learning a proper neighborhood of the given instance, then inducing a *crisp* decision tree from which crisp rules are extracted. Further building on this idea, FLARE¹ instead leverages a *fuzzy* decision tree, extracting fuzzy, rather than crisp, rules. Due to their ease of extraction and high accuracy, local explanations have become a building block for global ones, blurring the line between the two. In [11] the authors turn local Shapley values into global explanations by means of functional decomposition. Other works merge local and global explanations through feature importance [15], concept relevance [19], saliency maps [20] and strategy summaries [13]. Most related to our application on rules as explanations, GLocalX [22], from which this paper takes inspiration, merges local crisp explanations to build a *global explanation theory*.

In this paper, we introduce *FLocalX*, a framework to create an agnostic global explanation theory for a black-box classifier in the form of a *fuzzy* rule-based system built using *local fuzzy explanations*. This global fuzzy explanation theory mimics the behavior of the underlying black-box classifier, and can be used to provide factual explanations for novel, previously non-explained instances whiel providing a general understanding of the model. This way, a user can better understand how the classifier works, and how it will behave upon new instances, rather than generate explanations *ex-novo*. Building a global theory with fuzzy, rather than crisp, rules leads to additional benefits, making the global explanation more understandable, flexible, and faithful to the black-box model. Fuzzy rules leverage linguistic labels, which improve their readability by associating high-level human-understandable concepts with their premises and have been widely used to design explainable systems [16,25] Fuzziness also allows us to infer several, rather than one, explanations per instance, effectively providing the user with alternative explanations. Performance-wise, fuzzy rule-based systems are particularly apt to leverage different types of local explanations [23].

The rest of the paper is structured as follows. Section 2 presents the problem and identifies the relevant elements. Section 3 illustrates the workflow of our proposal. Section 4 shows the experiments and behavior of FLocalX. Finally, Sect. 5 presents the conclusions and indicates some future research lines.

¹ https://dsi.uclm.es/descargas/technicalreports/DIAB-24-02-1/FLARE_Tech_Rep.pdf.

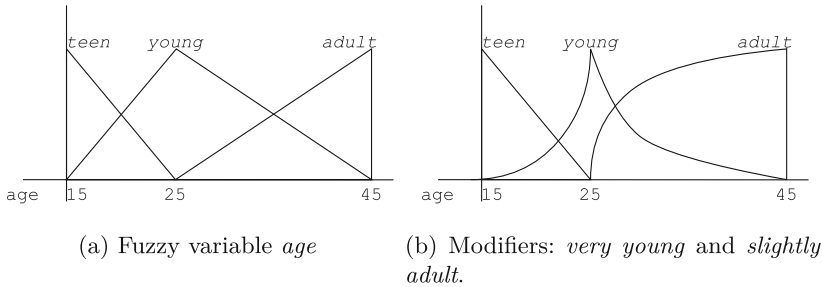


Fig. 1. Strong fuzzy partition for a fuzzy variable *age*

2 Setting the Stage

The Local to Global Explanation Problem [22] we aim to solve consists of finding a function g that, from a set of local explanations extracted from a black box classifier, yields a global explanation theory that describes its underlying logic.

First, let us revise some related concepts. In a classification problem, an instance $x = (x_1, \dots, x_n) \in \mathcal{X}_1 \times \dots \times \mathcal{X}_n$, where $\mathcal{X}_1, \dots, \mathcal{X}_n$ are n sets of *input variables*, is mapped to a decision $y \in \mathcal{Y} = \{y_1, \dots, y_n\}$ by a function (classifier) $f : \mathcal{X}_1 \times \dots \times \mathcal{X}_n \rightarrow \mathcal{Y}$. We write $f(x) = y$ to denote the classification y given to x . Let us denote by n_{cont} (resp. n_{disc}) the number of continuous (resp. discrete) variables in \mathcal{X} , s.t. $0 \leq n_{cont}, n_{disc} \leq n$, $n_{cont} + n_{disc} = n$. Let us assume that, associated with each continuous input variable \mathcal{X}_i , there is a fuzzy (linguistic) variable $\mathcal{F}_i = \{v_{i,1} \dots, v_{i,k_i}\}$ defined through a Ruspini partition [1] of k_i ordered fuzzy sets (see Fig. 1)². We use v_{i,z_i} to denote both the fuzzy set and its corresponding associated linguistic label, indistinctly. A triangular fuzzy set is defined by a triple of real-valued points: (start, peak, end), i.e. $teen = (15, 15, 25)$ and $young = (15, 25, 45)$ in Fig. 1a. . If we know the minimum and maximum values of $dom(\mathcal{X}_i)$, the partition becomes specified by $k_i - 2$ values. Given a value $\delta \in dom(\mathcal{X}_i)$, let $\mu_i(\delta) = (\mu_{i,1}(\delta), \dots, \mu_{i,k_i}(\delta))$ be the vector of membership degrees of δ to the k_i fuzzy sets of \mathcal{F}_i . In other words, $\mu_{i,z_i}(\delta)$ is the membership degree of δ to the set v_{i,z_i} . A linguistic hedge, or linguistic modifier, is a function that alters the membership function of a fuzzy set, which can modify the shape of the fuzzy set (see Fig. 1b). In this work, we use two of the most common linguistic hedges, “very” and “slightly”: $\mu_{i,z_i}^{very}(x_i) = (\mu_{i,z_i}(x_i))^2$ and $\mu_{i,z_i}^{slightly}(x_i) = \sqrt{\mu_{i,z_i}(x_i)}$. Finally, for discrete variables, we can interpret each value as a linguistic label whose associated fuzzy set has membership degree 1 in case the instance takes that value and 0 otherwise.

Let $b()$ be a classifier whose decision-making process needs to be explained, i.e., a black-box model, learned from a training dataset $TR =$

² Triangular membership functions are used in this article to illustrate the proposed method for simplicity/convenience. The framework allows other types of membership functions (Gaussian, trapezoidal, etc.) to represent the underlying fuzzy sets. However, the partitions must cover the complete domain for Eq. 1 to be valid.

$\{(x_1^t, \dots, x_n^t, y^t)\}_{t=1}^T$. Let $e = \{r_1, \dots, r_e\}$ be a multi-rule explanation formed by one (or more) fuzzy decision rules. Each rule $r = P(r) \rightarrow y(r)$ consists of a set of premises in conjunctive form $P(r) = p_{s_1} \wedge \dots \wedge p_{s_r}$ and an outcome $y(r) \in \mathcal{Y}$. Each premise $p_i = \langle \mathcal{F}_i, v_{i,z_i} \rangle$ is an attribute-value pair. For the continuous variables, \mathcal{F}_i is a fuzzy variable and v_{i,z_i} is one of its corresponding fuzzy sets. For the discrete variables, $\mathcal{F}_i = \mathcal{X}_i$ and v_{i,z_i} is a value from its domain. As an example, let us consider the following explanation for a loan request for a user $x = \{(age = 30), (job = Accountant), (amount = 20k)\}$:

$$e = \{(r_1 = age \text{ is young} \wedge job \text{ is Accountant} \rightarrow accept), \\ (r_2 = age \text{ is adult} \wedge amount \text{ is high} \rightarrow accept)\}$$

One property of multi-rule explanations is that, given an explanation e that explains the instance x , then $y(r) = b(x)$ for all $r \in e$. Fuzzy rules differ from crisp rules in that, while a crisp rule has a binary (0 or 1) match with an instance x , a fuzzy rule r has a *matching degree* with the instance, $md(r, x)$, defined as:

$$md(r, x) = \min_{i \in \{s_1, \dots, s_r\}} \{\mu_{i,z_i}(x_i)\} \in [0, 1]$$

An explanation theory $E = e_1 \cup \dots \cup e_q$ consists of a union of explanations which may have different outcomes.

Thus, the Local to Global Explanation Problem can be defined as follows: Given a black box $b()$, a set of instances $X = \{x^1, \dots, x^q\}$ and their local explanations $\{e_1, \dots, e_q\}$, the Local to Global Explanation Problem consists in deriving a global explanation theory $E_G = e'_1 \cup \dots \cup e'_q$ that aggregates the local explanations in order to summarize the logic of b .

3 Fuzzy Local to Global Explanation Framework

In this paper we propose FLocalX, a Fuzzy Local to Global Explanation framework that generates a global explanation theory which mimics a black box classifier given an initial set of local explanations. FLocalX takes the following elements as input a set of instances X and an explanation theory $E_L = e_1 \cup \dots \cup e_q$ formed by the union of the explanations of every instance in X , and generates the global explanation theory E_G by applying the following steps:

- First, it *transforms*, i.e., maps, the local fuzzy sets \mathcal{F}^j defined for each $e_j \in E_L$ to a common definition of fuzzy sets \mathcal{F}^C . This ensures that all local explanations in E_L share the same set of fuzzy variables. We name this explanation theory with common fuzzy sets E_C .
- Second, it *encodes* E_C into a simple, unique representation that will be the initial configuration C^{E_C} of the optimization process.
- Third, it *generates* the global explanation theory E_G from C^{E_C} through an optimization process.

This process results in a global explanation theory E_G that closely resembles the behavior of $b()$, and can provide a factual explanation for novel instances. Factual explanations can be extracted from E_G by obtaining, for instance, the *minimum robust factual explanation* defined in [7].

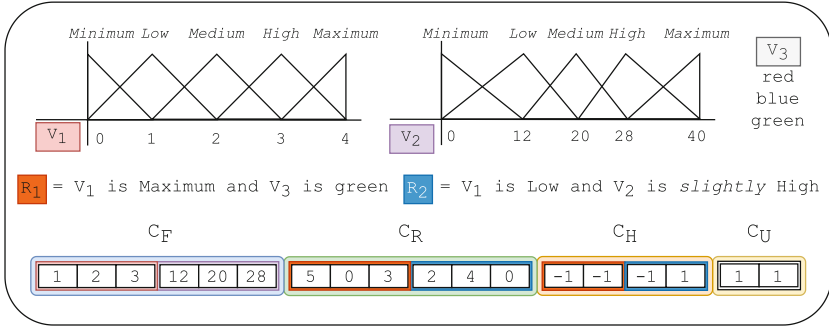


Fig. 2. Representation of the encoding of a FRBS

3.1 Local to Global Fuzzy Set Transformation

Depending on the method employed to extract the local explanations, they may not share the same fuzzy variable definitions, thus the same linguistic features may be defined by different fuzzy sets. For the sake of homogeneity, we uniform the fuzzy variable definitions $\mathcal{F}_i^1, \dots, \mathcal{F}_i^{|E_L|}$ of a given variable \mathcal{X}_i , and establish a global fuzzy variable definition \mathcal{F}^C by partitioning the domain of the numerical variables into equal-width sets, unless expert-provided sets are available.

Given two fuzzy sets $v_{i,z_i} \in \mathcal{F}_i$ and $v'_{i,z'_i} \in \mathcal{F}'_i$, we compute their similarity as

$$S(v_{i,z_i}, v'_{i,z'_i}) = A(v_{i,z_i} \cap v'_{i,z'_i}) / A(v_{i,z_i} \cup v'_{i,z'_i})$$

where $A(v)$ is the area of the fuzzy set v . As usual in the literature, we use *min* as the intersection and *max* as the union. Then, given a variable \mathcal{X}_i , we define

$$M(v'_{i,z'_i}, \mathcal{F}_i) = \arg \max_{v_{i,z_i} \in \mathcal{F}_i} S(v_{i,z_i}, v'_{i,z'_i}), \tag{1}$$

which takes a fuzzy set $v'_{i,z'_i} \in \mathcal{F}'_i$ and returns the set $v_{i,z_i} \in \mathcal{F}_i$ with the greatest similarity. We get E_C by applying Eq. 1 to every premise of each $e_i \in E_L$.

3.2 Global Fuzzy Set Theory Encoding

In FLocalX, we frame the objective of building a global explanation theory as the process of optimizing the Fuzzy Rule-Based System (FRBS) formed by the set of fuzzy decision rules in E_C . To this aim we need an encoding of E_C , this is, a representation of a potential solution to the problem which will be used by the metaheuristic algorithm in the optimization process.

The objective of the optimization process used by FLocalX is twofold. First, maintaining the degree in which the FRBS mimics the black-box classifier as accurate as possible. Second, making the FRBS as compact as possible (in terms of number of rules), to favor interpretability [9]. Inspired by [4], we design a procedure to tune FRBS maintaining interpretability by using a genetic algorithm. To this aim, there are two elements of the FRBS that must be optimized:

- **Surface Structure.** It is a shallow description that defines the rule as the relation between the input and output variables. We optimize it by (i) using linguistic hedges, and by (ii) altering the premises of a rule. Optimization can modify the linguistic hedge applied to a particular premise $p = \langle \mathcal{F}_i, v_{i,z_i} \rangle$, the fuzzy set v_{i,z_i} associated with p , and whether or not \mathcal{F}_i appears in a rule.
- **Deep Structure.** It is a more specific description which expands the surface structure with the definitions of the membership functions. Optimization only affects the membership functions of the fuzzy sets. Using a metaheuristic algorithm we can reduce the explainability of the system in exchange for greater accuracy. We control this by preserving the shape of the fuzzy partitions, i.e., triangular Ruspini partitions as explained in Sect. 2.

Configuration. Each configuration C of the optimization process represents an explanation theory E , shown graphically in Fig. 2. For this purpose, we will use a four-part configuration ($C_F + C_R + C_H + C_U$) as follows:

- C_F is the encoding of the fuzzy variables. We assume that the minimum and maximum values of $dom(\mathcal{X}_i)$ are known. As an example, in Fig. 1a we know $k_i = 3$, $min = 15$ and $max = 40$, and so we only have a free value (25) to codify the three fuzzy sets: $\{(15, 15, 25); (15, 25, 40); (25, 40, 40)\}$. Just by changing the value 25 to, e.g. 20, we modify the fuzzy semantics of the variable, obtaining a new partition: $\{(15, 15, 20); (15, 20, 40); (20, 40, 40)\}$. Thus, C_F has a length of $(\sum_{i=1}^{n_{cont}} k_i - 2)$, all of them being real numbers.
- C_R is the encoding of the rules. It has a length of $n \cdot |E|$ elements, where $|E|$ is the number of rules in the explanation theory, i.e., in the FRBS. Each n consecutive elements codify a rule with an ordinal encoding from the set $\{0, \dots, k_i\}$, where 0 represents that the i -th variable does not appear in the rule and 1 to k_i identify each fuzzy set or value of the variable \mathcal{F}_i , depending on whether \mathcal{X}_i is numerical or categorical.
- C_H is the encoding of the linguistic hedges. It has a length of $n_{cont} \cdot |E|$ elements, where each element belongs to the set $\{-1, 0, 1\}$ representing no linguistic hedge (-1), *very* (0) or *slightly* (1), for that particular continuous (fuzzy) variable.
- C_U is the encoding of the used rules. $|E|$ elements-long, encodes whether a rule is used in the final FRBS (1) or not (0).

3.3 Global Explanation Theory Generation

In order to generate the global explanation theory E_G , we exploit the encoding illustrated in the previous section to create the chromosomes of a genetic optimization process. Since the optimization process aims to simultaneously (i) accurately mimic the black box $b()$, and to (ii) have a compact FRBS, we designed an objective function that takes into account both aspects. In particular, we measure the first goal as the Area Under the ROC curve (*AUC*) to correctly handle imbalanced datasets, while we measure the second goal as the number of rules used in the system. Specifically, the objective function $f(C)$ that we

maximize in our experimentation is defined as follows:

$$f(C) = \alpha \cdot \left(1 - \frac{\sum_{i=0}^{|C_U|-1} C_U[i]}{|C_U|}\right) + (1 - \alpha) \cdot AUC$$

with α used to balance the two values optimized. Note that this is an implementation of the objective function, but others may be used.

FLocalX can employ any metaheuristic algorithm as optimizer in order to obtain the global explanation theory. For this work, we adopted a genetic algorithm [12]. Inspired by evolutionary adaptation, genetic algorithms encode solutions in a chromosome space, and sequentially evolve them, each generation selecting, merging, and improving on the previous one. In our case, merging is encoded by (i) a *crossover* operation, which generates new solutions by blending existing ones, and (ii) a *mutation* operation, which randomly alters a subset of the current solutions. In genetic fashion, a *selection* operation picks the best solutions (according to an objective function) which will be carried on to the next generation. Next, we detail these crucial aspects of the genetic algorithm:

- **Initial Population.** The initial population of size $\rho + 1$ for the genetic algorithm is generated in an informed manner, i.e., by altering a known configuration (C^{Ec}) rather than generating all elements at random. Given an initial configuration representing of a FRBS, each part C_F, C_R, C_H and C_U , generates $\lceil \rho/4 \rceil$ chromosomes by applying the mutation operator to that part (see below). The original configuration is also included in the initial population.
- **Crossover.** It selects pairs of chromosomes and crosses them with a probability p_{cross} . Due to the encoding adopted, the chromosome is divided in two, and different crossovers are applied:
 - First, a min-max-arithmetic crossover [10] is applied in the C_F part, generating four children.
 - Second, a six-point crossover is applied in the remaining chromosome, choosing two points for each part (i.e. two for C_R , two for C_H , and two for C_U). This generates two children.
 After recombining both parts, eight children are generated. The two best children are selected in order to keep the same population size.
- **Mutation.** It selects chromosomes and mutates them with a probability p_{mut} . The mutation over each part of the chromosome is performed applying an operation to a single bit $C[i]$ of each part of the chromosome as follows:
 - For C_F , the bit is randomly generated by sampling a real number from a uniform distribution in the range of the continuous variable.
 - For C_R , the bit is randomly chosen in the set $\{0, \dots, k_i\} \setminus C[i]$.
 - For C_H , the bit is randomly chosen in the set $\{-1, 0, 1\} \setminus C[i]$.
 - C_U is generated as $1 - C[i]$, i.e., altering the bit.
- **Selection.** A rank-based selection with respect to the fitness is used.
- **Replacement.** A replacement with elitism is performed, i.e., the best configuration from the previous population is kept.
- **Stop Criterion.** The genetic algorithm stops when the fitness of the best individual does not offer enough improvement beyond a threshold ϵ over a consecutive period of κ iterations.

These operations provide great flexibility and generality, and allow to directly *learn*, rather than *define*, the evolution of fuzzy rules. A challenging task such as the Local to Global one, which is not directly differentiable, requires flexible algorithms able to explore vast non-differentiable solution spaces, and adapt to a wide variety of users, and thus objectives. Optimizing global explanations to both be understandable by a user, as well as comprehensive enough to mimic a complex black-box classifier, is thus a perfect fit for our purpose.

4 Experiments

We evaluated FLocalX on three widely used multi-class datasets, i.e., *Iris*³, *Wine*⁴, and *Beer*⁵. The decision to use small datasets is driven by the main objective of developing and showcasing a framework to extract global explanations, rather than focusing on a specific metaheuristic (in this case, the genetic algorithm). As metaheuristic algorithms are resource-intensive and time-intensive processes, they often require specific optimizations made for each case and algorithm in order to tackle different problems. By employing simpler datasets, we can shift our focus towards illustrating the capability of the framework of working with different types of local explanations, as well as how it can seamlessly mimic a variety of black box algorithms. This is a first step in the line of work of a more complex experimentation where multiple metaheuristic algorithms are used and optimized with this framework in order to tackle much more complex problems. The implementation of FLocalX is available on Github⁶. **Experimental Setting.** We adopted the following metrics to evaluate the performance of FLocalX and the other classifiers used as baselines.

- *Accuracy.* It measures how close is the global explainer to the ground truth. We measure the accuracy of the black box (Acc-B), of the explanation theory formed by the union of the local explanations (Acc-U), and of the global explanation theory after applying FLocalX (Acc-F).
- *Fidelity.* It measures how well the global explainer mimics the black box classifier. We measure the fidelity of the explanation theory formed by the union of the local explanations (Fid-U), and the fidelity of the global explanation theory after applying FLocalX (Fid-F).
- *Number of Rules.* The total number of rules in the system. More rules indicate a more complex system and so a less interpretable system. We measure the number of rules before (#R) and after applying FLocalX (#R-F).
- *Number of Premises.* The number of premises in the antecedent of the rules. More premises are sometimes (falsely) perceived as being more helpful [14],

³ <https://archive.ics.uci.edu/dataset/53/iris>.

⁴ <https://archive.ics.uci.edu/dataset/109/wine>.

⁵ https://gitlab.citius.usc.es/ilia.stepin/fcfexpgen/-/tree/master/all_datasets/BEER_exp1.

⁶ GitHub: <https://github.com/Kaysera/flocalx>. FLocalX was programmed in Python 3.10, using libraries such as `numpy` and `scikit-learn` to properly manage the data structures and efficiently generate the explanations. To guarantee reproducibility, all the experiments are also published in a separate public Github repository <https://github.com/Kaysera/ida2024-experiments>.

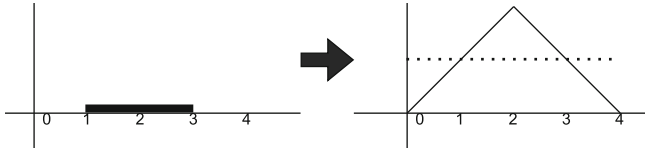


Fig. 3. LORE interval transformation to fuzzy set.

so shortening the rule together with a proper communication of attribute importance is a good practice. We measure the number of premises before ($\#P$) and after applying FLocalX ($\#P$ -F).

We used a train-validation-test (60%-30%-10%) split for the experimentation. The training split was used to train the black box classifiers using default hyperparameters. The validation partition was used to fit the hyperparameters of the local explanation methods, as well as to extract the local explanations (E_L). The test partition was used to measure the accuracy score for all algorithms. The genetic algorithm was repeated 20 times, altering the random seed and averaging the result between them. The parameters were chosen empirically⁷ as follows: population size (ρ) = 128, size pressure (α) = 0.1, # iterations (κ) = 20, threshold (ϵ) = 0.01, # fuzzy sets (k_i) = 5, p_{mut} = 0.15 and p_{cross} = 0.8. The fuzzy sets for Iris and Wine were obtained using equal-width partitions, while the fuzzy sets for Beer were obtained from [24].

We experiment with FLocalX with a set of different alternatives:

- *Black-Box Models*: We used *SVM*, *Neural Network (NN)* and *Random Forest (RF)* as baseline classifiers as implemented by *scikit-learn* [17].
- *Rule-Based Models*: Algorithms from which a ruleset that can be used for both prediction and explanation can be extracted. They are used as global explanation systems. The algorithms used are *Fuzzy Decision Tree (FDT)* [21], *LORE* [8] and *FLARE*.
- *Local to Global Approaches*: They set local explanations and merge them into a global explanation theory that is able to predict and explain instances of the dataset. We considered:
 - *FLocalX + LORE*: We used LORE to extract local explanations and then applied FLocalX. As FLocalX takes fuzzy rules, the intervals were expanded into fuzzy sets as if they were an α -cut of 0.5 of the corresponding fuzzy set. For example, the interval [1, 3] would become the fuzzy set (0, 2, 4) as shown in Fig. 3.
 - *FLocalX + FLARE*: We used FLARE to extract local explanations and then FLocalX was applied.

⁷ With these datasets, a large population size which is a power of 4 shows better results, and a small size pressure allows for faster convergence with a high accuracy. The rest of the parameters are standard for genetic tuning.

Table 1. Performance and Explainability of Different Models

	Method	Black Box	Fid-U	Fid-F	Acc-B	Acc-U	Acc-F	#R	#R-F	#P	#P-F
Iris	FDT	–	–	–	–	1.00	–	12.00	–	1.42	–
	FLARE	NN	0.97	0.93	1.00	0.95	0.91	32.00	5.05	1.31	1.27
		RF	0.94	0.91	0.93	0.94	0.91	26.00	4.16	1.81	1.70
		SVM	0.93	0.95	1.00	0.95	0.92	19.00	4.47	1.26	1.32
	LORE	NN	0.93	0.94	1.00	0.94	0.93	45.00	4.79	1.96	1.55
		RF	0.97	0.93	0.93	0.97	0.93	45.00	4.16	2.07	1.85
SVM		0.92	0.95	1.00	0.94	0.92	45.00	4.37	1.58	1.36	
Wine	FDT	–	–	–	–	0.94	–	36.00	–	3.00	–
	FLARE	NN	0.86	0.76	0.89	0.82	0.77	48.00	8.68	1.42	1.54
		RF	0.61	0.61	1.00	0.61	0.61	41.00	2.89	1.39	1.23
		SVM	0.99	0.73	0.67	0.71	0.76	17.00	4.95	1.00	1.32
	LORE	NN	0.77	0.78	0.89	0.76	0.76	54.00	4.79	2.52	1.94
		RF	0.90	0.88	1.00	0.90	0.88	54.00	6.21	3.19	3.02
SVM		0.93	0.76	0.67	0.68	0.71	52.00	5.05	1.02	1.56	
Beer	FDT	–	–	–	–	1.00	–	69.00	–	2.42	–
	FLARE	NN	0.69	0.71	0.80	0.67	0.79	128.00	20.42	1.85	1.78
		RF	0.87	0.88	1.00	0.87	0.88	129.00	26.68	2.34	2.18
		SVM	0.86	0.77	0.85	0.85	0.82	99.00	15.21	1.96	1.93
	LORE	NN	0.74	0.76	0.80	0.70	0.80	119.00	13.42	2.01	1.98
		RF	0.92	<i>0.89</i>	1.00	0.92	<i>0.88</i>	119.00	14.63	2.54	2.60
SVM		0.78	0.82	0.85	0.67	0.86	119.00	15.58	2.02	2.07	

Results. We compare the results of FLocalX for two different local explanation methods, using the union of the local explanations as a global explainer and studying how much improvement our framework provides. We also use a rule-based white box method (i.e., FDT) as baseline. Table 1 reports both the performance of the global explainers, as well as its level of complexity.

As one objective of the optimization process is to minimize the size of the rule-based system, testing the impact on performance is necessary. We can observe that problems where Acc-U is really high (i.e. >0.9), Acc-F is lower than Acc-U, likely because most rules are necessary to achieve that degree of accuracy. However, that decrease in accuracy is not so much as to lose trust in the explainer. On the other hand, in more complex problems where the starting point is not as good (the Beer dataset with FLARE and NN, or LORE and SVM are examples of this), the optimization process can even improve the starting point’s accuracy. This suggests that a metaheuristic approach, while time-consuming, benefits hard-to-solve problems. Finally, it is worth mentioning that LORE rules tend to be a better starting point for FLocalX than FLARE rules. This finding might suggest that either crisp rules are better than fuzzy rules as a starting point, or that more premises provide a better starting point. More experiments will be done to explore the cause.

Turning to explanation complexity, the most relevant part is in the reduction of rules from the union of local explanations ($\#R$) to after FLocalX is applied ($\#R-F$). We can observe that we need around 10% – 15% of the number of rules from which FLocalX starts. Beer shows the largest explanation theories (at around 20 rules for FLARE and 14 for LORE), which are still readable for humans. Moreover, there is a great reduction from the baseline white-box classifiers, needing around 40% of the rules in simpler datasets and around 20% – 30% of the rules in more complex problems. The number of rules generated by the FDT increases with the complexity of the problems, which makes it unfit as a global explainer for difficult problems, i.e., valid for Iris and Wine but unreasonably long at 70 rules for Beer. On the other hand, looking at the number of premises, most rules have around 1–3 premises, also manageable for a human reader. $\#P-F$ is only marginally smaller than $\#P$ because $f(C)$ does not consider the length of the rule. Finally, we can see that LORE global explanations usually have fewer rules than FLARE, with some more premises per rule.

The results of this preliminary experimentation, with a single optimization algorithm (i.e., a genetic algorithm) and smaller datasets, showcase the flexibility of the framework, which can generate compact and performant global explanation theories that can be useful to a human reader.

5 Conclusions and Future Work

This work introduces FLocalX, a model agnostic local to global explanation framework based on fuzzy logic that leverages the power of evolutionary computing to obtain a global explanation of a black-box model. FLocalX uses local explanations formed as fuzzy rules as the starting point from which it builds a global fuzzy explanation that summarizes the model underneath. Using a genetic algorithm as the optimization method, the experimentation carried out in this paper shows that FLocalX is able to generate a short and accurate global explanation theory, improving upon the trivial union of local explanations, as well as upon the used baseline white box model. As future research directions, we intend to perform a comprehensive study on the different hyperparameters, as well as different operators and objective functions for the genetic tuning process of FLocalX. Moreover, we would like to study the usage of a different metaheuristic algorithm to replace the genetic procedure. Finally, the difference in performance shown between using FLARE to generate the local explanation theory and using LORE motivates the need to experiment with other local explainers.

Acknowledgements. This work has been funded by the following projects: SBPLY/21/180225/000062 (Government of Castilla-La Mancha and ERDF funds); PID2019-106758GB-C33, and FPU19/02930 (MCIN/AEI/10.13039/501100011033 and ERDF Next Generation EU); and 2022-GRIN-34437 (Universidad de Castilla-La Mancha and ERDF funds), EU NextGenerationEU programme PNRR-PE-AI FAIR (Future Artificial Intelligence Research), PNRR-SoBigData.it - Prot. IR0000013, H2020-INFRAIA-2019-1: Res. Infr. G.A. 871042 *SoBigData++*, ERC-2018-ADG G.A. 834756 *XAI*, and CHIST-ERA-19-XAI-010 SAI.

References

1. Alonso, J.M., et al.: Explainable fuzzy systems: paving the way from interpretable fuzzy systems to explainable AI systems. In: SCI (2021)
2. Angelov, P.P., et al.: Explainable artificial intelligence: an analytical review. WIREs Data Mining Knowl. Discov. **11**(5), e1424 (2021)
3. Arrieta, A.B., et al.: Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf. Fusion **58**, 82–115 (2020)
4. Casillas, J., et al.: Genetic tuning of fuzzy rule deep structures preserving interpretability and its interaction with fuzzy rule set. IEEE TFS **13**(1), 13–29 (2005)
5. Chen, T., et al.: Xgboost: extreme gradient boosting. R package **1**(4), 1–4 (2015)
6. Dai, Z., et al.: Coatnet: marrying convolution and attention for all data sizes. Adv. Neural. Inf. Process. Syst. **34**, 3965–3977 (2021)
7. Fernández, G., et al.: Factual and counterfactual explanations in fuzzy classification trees. IEEE Trans. Fuzzy Syst. **30**(12), 5484–5495 (2022)
8. Guidotti, R., et al.: Factual and counterfactual explanations for black box decision making. IEEE Intell. Syst. **34**(6), 14–23 (2019)
9. Guidotti, R., et al.: A survey of methods for explaining black box models. ACM Comput. Surv. **51**(5), 93:1–93:42 (2019)
10. Herrera, F., et al.: Fuzzy connectives based crossover operators to model genetic algorithms population diversity. Fuzzy Sets Syst. **92**(1), 21–30 (1997)
11. Hiabu, M., et al.: Unifying local and global model explanations by functional decomposition. In: AISTATS, vol. 206, pp. 7040–7060. PMLR (2023)
12. Holland, J.H.: Adaptation in Natural and Artificial Systems: an Introductory Analysis with Applications to Biology, Control, and AI. MIT Press, Cambridge (1992)
13. Huber, T., et al.: Local and global explanations of agent behavior: integrating strategy summaries with saliency maps. Artif. Intell. **301**, 103571 (2021)
14. Kliegr, T., et al.: A review of possible effects of cognitive biases on interpretation of rule-based machine learning models. Artif. Intell. **295**, 103458 (2021)
15. Lundberg, S.M., et al.: From local explanations to global understanding with explainable AI for trees. Nat. Mach. Intell. **2**(1), 56–67 (2020)
16. Maria, A.J., et al.: Explainable fuzzy systems: Paving the way from interpretable fuzzy systems to explainable AI systems. SCI **970** (2021)
17. Pedregosa, F., et al.: Scikit-learn: ML in python. JMLR **12**, 2825–2830 (2011)
18. Regulation, G.D.P.: General data protection regulation (GDPR). Intersoft Consulting, Accessed Oct 24 **1** (2018)
19. Schrouff, J., et al.: Best of both worlds: local and global explanations with human-understandable concepts. CoRR (2021)
20. Schrouff, J., et al.: Best of both worlds: local and global explanations with human-understandable concepts. CoRR **abs/2106.08641** (2021)
21. Segatori, A., et al.: On distributed fuzzy decision trees for big data. IEEE Trans. Fuzzy Syst. **26**(1), 174–192 (2017)
22. Setzu, M., et al.: Glocalx-from local to global explanations of black box AI models. Artif. Intell. **294**, 103457 (2021)
23. Stepin, I., et al.: Generation and evaluation of explanations for decision trees and fuzzy rule-based classifiers. In: FUZZ, pp. 1–8. IEEE (2020)
24. Stepin, I., Catala, A., Pereira-Fariña, M., Alonso, J.M.: Factual and counterfactual explanation of fuzzy information granules. In: Pedrycz, W., Chen, S.-M. (eds.) Interpretable Artificial Intelligence: A Perspective of Granular Computing. SCI, vol. 937, pp. 153–185. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-64949-4_6

25. Varshney, A.K., et al.: Literature review of the recent trends and applications in various fuzzy rule-based systems. In: IJFS, pp. 1–24 (2023)
26. Zhang, S., et al.: The diversified ensemble neural network. *Adv. Neural. Inf. Process. Syst.* **33**, 16001–16011 (2020)